

Semi-parametryczna estymacja punktu zmiany parametrów w szeregach niegaussowskich metodą maksymalizacji wielomianu

Streszczenie. Zastosowano metodę maksymalizacji wielomianu do syntezy adaptacyjnych algorytmów dla estymacji punktu zmiany wartości średniej oraz wariancji szeregu losowego w trybie a posteriori. Analiza i wyniki modelowania statystycznego wykazały, że uwzględniając parametry nie-gaussowskich danych statystycznych w oszacowaniu wielomianowym uzyskuje się istotny wzrost dokładności.

Abstract. An application of the maximization technique in the synthesis of polynomial adaptive algorithms for a posterior (retrospective) estimation of the change-point of the mean value or variance of random sequences is presented. Statistical simulation shows a significant increase in the accuracy of polynomial estimates, which is achieved by taking into account the non-Gaussian character of statistical data. (Semi-parametric estimation of the change-point of parameters of the non-Gaussian sequences by polynomial maximization method)

Słowa kluczowe: punkt zmiany parametrów, estymacja a posteriori, wielomian stochastyczny, wartość średnia, wariancja, kumulanty.
Keywords: change-point, a posteriori estimation, stochastic polynomial, mean value, variance, cumulant coefficients.

Wstęp

Technikę pomiarową w coraz większym stopniu stosuje się w nauce, przemyśle, medycynie i innych zastosowaniach do badania właściwości sygnałów o składowych modelowanych procesami stochastycznymi, a przy równomiernym ich próbkowaniu - szeregami losowymi. Zadania takie rozwiązuje się metodami statystycznymi zarówno w czasie rzeczywistym jak i a posteriori, czyli metodami retrospektywnymi. Podejście a posteriori wymaga dłuższego okresu na uzyskanie podstaw do podjęcia decyzji o dalszych działaniach, ale można wówczas wykorzystać całą informację o badanym obiekcie zawartą w próbie sygnału. Zapewnia ono bardziej dokładne i wiarygodne oszacowanie badanych parametrów procesu.

Jednym z takich zadań jest wykrywanie momentu występowania "zaburzenia" w obserwowanym procesie, np. dla celów kontrolnych lub diagnostycznych, czyli pomiar położenia punktu zmiany jego parametrów statystycznych. Punkt taki nazywa się w literaturze statystycznej krótko punktem zmiany [1]. Oszacowanie a posteriori punktu zmiany parametrów procesów stochastycznych jest wykorzystywane w praktyce w wielu dziedzinach, takich jak wykrywanie wahań klimatu [2], analiza genetyczna szeregów czasowych [3], wykrywanie włamań w sieciach komputerowych [4] i telekomunikacyjnych [5], segmentacja sygnałów mowy i wiadomości z portali społecznościowych [6], wykrywanie pojawienia się sygnałów emisji akustycznej [7] i detekcja sygnałów wideo [8]. Można też je zastosować w badaniu przebiegu ciągłych procesów przemysłowych

Tak szeroki zakres zadań aplikacyjnych wymaga opracowania wielu różnych modeli matematycznych i algorytmów do przetwarzania danych pozyskiwanych w pomiarach procesów losowych. Badania dotyczące metod wykrywania punktu zmiany obejmują głównie klasę procesów losowych modelowanych rozkładem gęstości prawdopodobieństwa (ang. akronim PDF) w postaci funkcji Gaussa. Istnieje też wiele procesów rzeczywistych, które bardziej dokładnie modelują inne rozkłady PDF. Metody oparte na rozkładach PDF nazywa się parametrycznymi.

Podstawowe trudności podejścia parametrycznego (wg Bayesa i metody największej wiarygodności) to: wymóg posiadania a priori informacji o postaci rozkładu prawdopodobieństwa, potencjalnie duża złożoność analizy właściwości statystycznych i trudność zastosowania jej w praktyce. Brakuje też rekomendacji ujednolaczącej szacowanie niepewności takich pomiarów. Szereg ostatnich badań teoretycznych ukierunkowanych jest na stosowanie

alternatywnych metod statystycznych, które umożliwiały by zmniejszenie lub minimalizację niezbędnej a priori ilości informacji. Wykorzystuje się tu odporne (ang. *robust*) procedury przetwarzania statystycznego, które są nieczułe na niedokładności modeli probabilistycznych, lub też stosuje testy nieparametryczne, nie związane z danym typem rozkładu. Ceną za tę "niezależność" metod alternatywnych są nieco gorsze ich właściwości jakościowe niż optymalnych metod parametrycznych [9].

Jednym z alternatywnych sposobów rozwiązywania zadań z przetwarzania danych nie-gaussowskich procesów losowych jest użycie statystyk (momentów i kumulantów) wyższego rzędu [10]. Taki sposób modelowania stosuje się w różnych dziedzinach do wykrywania zaburzenia w procesie mierzonym, czyli do estymacji położenia punktu zmiany jego parametrów.

W niniejszej pracy, do wykrywania w trybie a posteriori punktu zmiany parametrów sygnału, zastosowano metodą maksymalizacji wielomianu (ang. *polynomial maximization method* - akronim PMM). Yu. Kunchenko zaproponował wykorzystanie do tego celu nie-gaussowskich wielomianów stochastycznych [10]. Stosuje się je w połączeniu z opisem danych próbki za pomocą momentów i kumulantów. Upraszcza to znacznie syntezę adaptacyjnych algorytmów statystycznych, które opierają się na probabilistycznych właściwościach pozyskanych danych i umożliwiają istotną poprawę dokładności i wiarygodności oszacowania (mniej błędnych decyzji przy mniejszych wariancjach ocen). Opis danych losowych poprzez momenty i kumulanty jest przybliżony. Oparte na nim metody statystyczne, wraz ze wzrostem rzędu użytej statystyki, umożliwiają uzyskiwanie rezultatów zbliżających się asymptotycznie do optymalnych. Metody te w polskiej literaturze statystycznej nazywa się semi-parametrycznymi w ślad za ang. *semiparametric* (choć dosłowne tłumaczenie - to pół-parametryczne).

Celem niniejszej pracy jest omówienie algorytmów do szacowania w trybie a posteriori punktu zmiany wartości średniej lub wariancji w próbkach sygnału opisywanego nie-gaussowskim szeregiem losowym. Stosuje się tu metodę opartą na maksymalizacji wielomianu statystycznego. Skuteczność proponowanej metody zbada się za pomocą modelowania statystycznego metodą Monte Carlo.

Matematyczne sformułowanie zagadnienia

Załóżmy, że w wyniku pomiarów procesu losowego w czasie dyskretyzowanym, np. przez regularne próbkowanie, otrzymano próbkę $\Gamma \hat{x} = \{x_1, x_2, \dots, x_n\}$. Przy czym elementy

tej próbki można traktować jako zbiór n statystycznie niezależnych zmiennych losowych, tj. nieskorelowanych ze sobą (czyli próbkowanie było nie za częste na tyle, że dla procesów o ograniczonym widmie można nie uwzględniać autokorelacji danych). Probabilistyczne właściwości próbek określa się np. przez jej wartość oczekiwaną θ , wariancję σ^2 i wartości współczynników kumulantów γ_l aż do rzędu $l = \overrightarrow{2s}$ (s stopień wielomianu).

Rozpatrzmy kilka przypadków, gdy do pewnej nieznanej a priori chwili τ czasu dyskretnego występuje w procesie wartość średnia θ_0 i wariancja σ_0^2 , a w chwili $\tau+1$ jeden z tych parametrów zmienia się istotnie (np. wartość średnia z θ_0 na θ_1 , lub wariancja z σ_0^2 na σ_1^2). Na podstawie analizy całej próbki \bar{x} należy oszacować położenie τ punktu zmiany. Warianty tego zadania mogą różnić się zdobytą a priori informacją o wartości zmieniającego się parametru (przed lub po wystąpieniu punktu zmiany) oraz wartościami innych parametrów modelu szeregu losowego.

Oszacowanie a posteriori punktu zmiany średniej metodą maksymalnego prawdopodobieństwa

Jednym z głównych sposobów badania a posteriori punktu zmiany stało się podejście oparte na maksymalizacji prawdopodobieństwa, opracowane przez Hinckley'a [11]. Zaproponował on ogólną asymptotyczną metodę otrzymywania *post factum* rozkładów oszacowań punktu zmiany metodą maksymalnego prawdopodobieństwa (MMP). Zastosowanie tej metody wymaga posiadania a priori informacji o rozkładzie prawdopodobieństwa danych losowych przed i po zmianie ich parametru.

Dla danych szeregu losowego o rozkładzie Gaussa ocena średniej metodą MMP pokrywa się z metodą momentów (MM), tj.

$$(1) \quad \hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

Oszacowanie (1) jest zgodne (ang. *consistent*) i nieobciążone. Ocena $\hat{\theta}$ takiej próbki przy $n \rightarrow \infty$ nie jest przesunięta względem wartości średniej θ całej populacji, a pokrywa się z nią. Metodę tę jako oszacowanie nieparametryczne można wykorzystywać do szacowania wartości średniej zmiennych losowych o dowolnym rozkładzie prawdopodobieństwa. Jednakże ocena ta jest efektywna tylko dla modelu Gaussa. Dla tego modelu przy znanej wariancji σ^2 logarytm funkcji maksymalnego prawdopodobieństwa przekształca się w statystykę [1]:

$$(2) \quad T_r(\theta_0, \theta_1) = r \sum_{i=1}^r (x_i - \theta_0)^2 + (n-r) \sum_{i=r+1}^n (x_i - \theta_1)^2$$

Ma ona maksimum w okolicy rzeczywistej wartości τ położenia punktu zmiany. Tak więc, wynik $\hat{\tau}$ oceny tego punktu znajduje się według algorytmu:

$$(2a) \quad \hat{\tau} = \arg \max_{1 \leq r \leq n-1} T_r(\theta_0, \theta_1)$$

Hinckley [11] rozważył dla rozkładu Gaussa również przypadek nieznanymi parametrami θ_1 i θ_2 . Wówczas oszacowanie metodą MMP punktu zmiany średniej

$$(3) \quad \hat{\tau} = \arg \max_{1 \leq r \leq n-1} \left[r(\hat{\theta}_{0,r} - \hat{\theta})^2 + (n-r)(\hat{\theta}_{1,r} - \hat{\theta})^2 \right]$$

gdzie

$$(4) \quad \hat{\theta}_{0,r} = \frac{1}{r} \sum_{i=1}^r x_i, \quad \hat{\theta}_{1,r} = \frac{1}{n-r} \sum_{i=r+1}^n x_i$$

Statystyki (2) i (3) nie zależą od innych parametrów probabilistycznych. Można wykorzystywać je do nieparametrycznego szacowania punktu zmian średniej szeregu losowego o dowolnym rozkładzie. Jednak wówczas, tak jak i przy ocenie średniej na podstawie (1), algorytm nieparametryczny traci optymalność. Dlatego też pada się nowe algorytmy estymacji nieliniowej oparte na metodzie maksymalizacji wielomianu (MMW). Pozwalają one w sposób dosyć prosty uwzględnić stopień niegaussowości danych statystycznych.

Oszacowanie a posteriori punktu zmiany wariancji metodą maksymalnego prawdopodobieństwa

Dla rozkładu Gaussa oszacowanie wariancji metodą MMP przy znanej wartości oczekiwanej θ ma postać:

$$(5) \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \theta)^2$$

Oszacowanie (5) jest niesprzeczne (zgodne) i nieobciążone (asymptotycznie nieprzesunięte).

Metoda momentów (ang. akronim MM) może być wykorzystywana do oceny nieparametrycznej szacowania wariancji zmiennych losowych o dowolnym rozkładzie. Dla tego modelu probabilistycznego logarytm funkcji maksymalnego prawdopodobieństwa ma postać [1]:

$$(6) \quad \ln L(\sigma_0^2, \sigma_1^2) = -\frac{n}{2} \ln(2\pi) - \frac{r}{2} \ln(\sigma_0^2) + \\ -\frac{n-r}{2} \ln(\sigma_1^2) - \frac{\sum_{i=1}^r (x_i - \theta)^2}{2\sigma_0^2} - \frac{\sum_{i=r+1}^n (x_i - \theta)^2}{2\sigma_1^2}$$

Jeśli wartość θ jest z góry znana i niezmienna, to próbkę \bar{x} można scentrować, tj. przyjąć za równą zero jej wartość średnią. Wówczas można utworzyć statystykę o postaci:

$$(7) \quad T_r(\sigma_0^2, \sigma_1^2) = -n \ln(2\pi) - r \ln(\sigma_0^2) + \\ - (n-r) \ln(\sigma_1^2) - \frac{\sum_{i=1}^r x_i^2}{\sigma_0^2} - \frac{\sum_{i=r+1}^n x_i^2}{\sigma_1^2}$$

Ma ona maksimum w okolicy rzeczywistej wartości τ punktu zmiany.

Dla nieznanymi a priori wartości wariancji szeregu losowego przed i po zmianie, w algorytmie można zastosować ich chwilowe (a posteriori) szacunki

$$(8) \quad \hat{\sigma}_{0,r}^2 = \frac{1}{r} \sum_{i=1}^r x_i^2, \quad \hat{\sigma}_{1,r}^2 = \frac{1}{n-r} \sum_{i=r+1}^n x_i^2$$

Wówczas algorytm estymacji punktu zmiany ma postać [1]:

$$(9) \quad \hat{\tau} = \arg \max_{1 \leq r \leq n-1} \left[n \ln(\hat{\sigma}^2) - r \ln(\hat{\sigma}_{0,r}^2) - (n-r) \ln(\hat{\sigma}_{1,r}^2) \right]$$

w którym szacowaną wartość $\hat{\sigma}^2$ oblicza się dla całej próbki ze wzoru (5).

Ponieważ statystyki (7) i (9), nie zależą od innych parametrów, to otrzymane wyniki można stosować do nieparametrycznego szacowania punktu zmiany wariancji

sekwencji losowej o dowolnym rozkładzie. Jednakże, podobnie jak przy ocenie średniej, nieparametryczne algorytmy tracą swoją optymalność. Oznacza to, że uzyska się dokładność znacznie mniejszą niż dla oszacowania optymalnego. Dlatego też poniżej rozpatrzy się nowe wielomianowe algorytmy do retrospektywnej oceny punktu zmian w nie-gaussowskim szeregu losowym.

Ogólny algorytm wielomianowej metody estymacji punktu zmiany parametru

Metodą MMW [10], dla rozmieszczonych równomiernie w wyniku próbkowania elementów próbki \bar{x} można znaleźć oszacowanie dowolnego parametru ϑ rozwiązując równanie stochastyczne o postaci

$$(10) \quad \sum_{j=1}^s h_j(\vartheta) \left[\frac{1}{n} \sum_{i=1}^n x_i^j - \alpha_j(\vartheta) \right] \Big|_{\vartheta=\hat{\vartheta}} = 0$$

w którym: s - stopień wielomianu użytego do szacowania parametrów; $h_j(\vartheta)$ - współczynniki, $\alpha_j(\vartheta)$ - teoretyczne momenty początkowe i -tego rzędu.

Współczynniki $h_j(\vartheta)$ dla $j = \overline{1, \dots, s}$ znajduje się rozwiązując układ równań algebraicznych liniowych (dla danego stopnia wielomianu s) otrzymany z warunku minimum wariancji dla poszukiwanego oszacowania parametru ϑ , np.

$$(11) \quad \sum_{j=1}^s h_j(\vartheta) F_{i,j}(\vartheta) = \frac{d}{d\vartheta} \alpha_i(\vartheta)$$

gdzie: $F_{i,j}(\vartheta) = \alpha_{i+j}(\vartheta) - \alpha_i(\vartheta)\alpha_j(\vartheta)$ - scentrowane korelanty o wymiarze (i, j) .

Układ równań (10) rozwiązuje się metodą Kramera:

$$h_j(\vartheta) = \frac{\Delta_{js}}{\Delta_s} \quad \text{dla } j = \overline{1, s}$$

gdzie: $\Delta_s = \det \|F_{i,j}\|$, $i, j = \overline{1, s}$ - wielomian stochastyczny o wymiarze s ; Δ_{js} - wyznaczniki otrzymane po zamianie w Δ_s kolumny j na kolumnę wyrazów wolnych układu (10).

Poniżej proponuje się zastosowaniu metody MMW do oszacowania w trybie a posteriori punktu zmiany parametru sygnału losowego. W tym celu wykorzystuje się właściwości wielomianów stochastycznych o postaci:

$$(12) \quad l_{sn}(\bar{x}/\vartheta) = nk_0(\vartheta) + \sum_{j=1}^s k_j(\vartheta) \sum_{i=1}^n x_i^j$$

gdzie

$$(12a) \quad k_0(\vartheta) = \int_a^{\vartheta} \sum_{j=1}^s [h_j(\vartheta) \alpha_j(\vartheta)] d\vartheta,$$

$$(12b) \quad k_j(\vartheta) = \int_a^{\vartheta} h_j(\vartheta) d\vartheta \quad \text{dla } j = \overline{1, \dots, s}$$

Wartość oczekiwana $E\{l_{sn}(\bar{x}/\vartheta)\}$ funkcji od ϑ osiąga maksimum w punkcie rzeczywistej wartości tego parametru.

Jeśli więc rzeczywista wartość parametru ϑ z pewnego przedziału (a, b) będzie maksymalizować wielomian stochastyczny o postaci (12), to aby znaleźć wartość parametru, przy której występuje zaburzenie (nagła zmiana

wartości z ϑ_0 na ϑ_1), tworzy się wielomian stochastyczny

$$(14) \quad P_r^{(s)}(\vartheta_0, \vartheta_1) = rk_0(\vartheta_0) + \sum_{j=1}^s k_j(\vartheta_0) \sum_{i=1}^r x_i^j + \\ + (n-r)k_0(\vartheta_1) + \sum_{j=1}^s k_j(\vartheta_1) \sum_{i=r+1}^n x_i^j$$

Ma on maksimum w pobliżu prawdziwej wartości τ punktu zmiany. Tak więc, ogólny algorytm dla poszukiwania oszacowania punktu zmiany przy zastosowaniu metody MMW ma postać:

$$(15) \quad \hat{\tau} = \arg \max_{1 \leq r \leq n-1} P_r^{(s)}(\vartheta_0, \vartheta_1)$$

Wielomianowa estymacja punktu zmiany wartości średniej

Oszacowanie średniej metodą maksymalnego prawdopodobieństwa (MMP) z zastosowaniem wielomianu stopnia $s=1$ pokrywa się z liniowym oszacowaniem MM wg wzoru (1) [10]. Stąd synteza wielomianowych algorytmów do oszacowania punktu zmian tego parametru ma znaczenie jedynie w przypadku stopni $s \geq 2$.

Dla stopnia wielomianu $s=2$ oszacowanie średniej znajduje się poprzez rozwiązywanie równania kwadratowego o postaci:

$$(16) \quad \gamma_3 \theta^2 - \left[2\gamma_3 \frac{1}{n} \sum_{i=1}^n x_i - \sigma(2 + \gamma_4) \right] \theta + \\ - \sigma(2 + \gamma_4) \frac{1}{n} \sum_{i=1}^n x_i + \gamma_3 \left[\frac{1}{n} \sum_{i=1}^n (x_i)^2 - \sigma^2 \right] \Big|_{\theta=\hat{\theta}} = 0.$$

gdzie: $\gamma_3 = \mu_3 / \sigma^3$ - współczynnik asymetrii, $\gamma_4 = \mu_4 / \sigma^4$ - współczynnik kurtozy, μ_l - moment centralny rzędu l .

Z (16) wynika, że na ocenę $\hat{\theta}_{s=2}$ dodatkowo wpływają współczynniki: asymetrii γ_3 i kurtozy γ_4 . Dla obu tych współczynników równych zero (jak dla rozkładu Gaussa) oszacowanie wg wielomianu (16) degeneruje się do oszacowania klasycznego wg (1).

W [12] wykazano, że przy zastosowaniu równania (16) uzyskuje się większą dokładność oszacowania

$$(17) \quad g_2(\gamma_3, \gamma_4) = 1 - \frac{\gamma_3^2}{2 + \gamma_4}$$

gdzie: g_2 - współczynnik zmniejszenia wariancji w stosunku do oszacowania wg (1) dla granicy $n \rightarrow \infty$.

Z (10) i (12) wynika, że wielomian stochastyczny (12) o stopniu $s=2$ maksymalizują w pobliżu prawdziwej wartości parametru θ następujące współczynniki:

$$(18) \quad k_0(\theta) = \frac{\sigma^3}{6\Delta_2} [2\gamma_3 \theta^3 + 3(2 + \gamma_4) \sigma \theta^2 - 6\gamma_3 \sigma^2 \theta], \\ k_1(\theta) = \frac{\sigma^3}{\Delta_2} [\gamma_3 \theta^2 + (2 + \gamma_4) \sigma \theta], \quad k_2(\theta) = -\frac{\sigma^3}{\Delta_2} \gamma_3 \theta.$$

gdzie $\Delta_2 = \sigma^6 (2 + \gamma_4 - \gamma_3^2)$

W [12] wykazano, że dla informacji a priori o wartości średniej θ_0 przed punktem zmiany i θ_1 po tym punkcie

oraz przy warunku $\theta_1 > \theta_0$, statystyka (14) dla stopnia wielomianu $s = 2$ ma postać

$$(19) \quad P_r^{(2)}(\theta_0, \theta_1) = (n-r) \left[\frac{1}{3} \gamma_3 (\theta_1^3 - \theta_0^3) + \frac{1}{2} \sigma (2 + \gamma_4) (\theta_1^2 - \theta_0^2) - \sigma^2 \gamma_3 (\theta_1 - \theta_0) \right] + [\gamma_3 (\theta_1^2 - \theta_0^2) + \sigma (2 + \gamma_4) (\theta_1 - \theta_0)] \sum_{i=r+1}^n x_i - \gamma_3 (\theta_1 - \theta_0) \sum_{i=r+1}^n x_i^2$$

Przy braku informacji a priori o wartościach średnich, w wielomianowej ocenie punktu zmiany (jak i w klasycznym przypadku) można nieznanne wartości tych parametrów zastąpić oszacowaniami o postaci (4). Tworzy się je dla każdego potencjalnego punktu zmiany. Wówczas algorytm adaptacyjny dla oszacowania punktu zmiany, oparty na metodzie MMW dla stopnia $s = 2$, można przedstawić jako:

$$(20) \quad \hat{t} = \arg \max_{1 \leq r \leq n-1} \left\{ r \left[\frac{4}{3} \gamma_3 \hat{\theta}_{0,r}^3 + \frac{3}{2} \sigma (2 + \gamma_4) \hat{\theta}_{0,r}^2 + \left[\sigma^2 + \sum_{i=1}^r (x_i)^2 \right] \gamma_3 \hat{\theta}_{0,r} \right] + (n-r) \left[\frac{4}{3} \gamma_3 \hat{\theta}_{1,r}^3 + \frac{3}{2} \sigma (2 + \gamma_4) \hat{\theta}_{1,r}^2 + \left[\sigma^2 + \sum_{i=r+1}^n (x_i)^2 \right] \gamma_3 \hat{\theta}_{1,r} \right] \right\}$$

Analiza struktury wielomianów statystyk (19) i (20) po raz kolejny potwierdza, że wykorzystanie metody MMP dla stopnia wielomianu $s = 2$ jest celowe tylko przy asymetrii ($\gamma_3 \neq 0$) funkcji rozkładu danych.

Wielomianowa estymacja punktu zmiany wariancji

W [12] wykazano, że oszacowanie wariancji σ^2 metodą maksymalizacji wielomianu MMW uzyskuje się tylko dla stopnia wielomianu $s \geq 2$. Dla $s = 2$ wielomianowe oszacowanie tego parametru (przy $\theta = 0$) ma postać

$$(21) \quad \hat{\sigma}_{s=2}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 + \frac{1}{2} \gamma_3^2 \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 + \gamma_3 \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \left\{ \frac{1}{4} \gamma_3^2 \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 + \frac{1}{n} \sum_{i=1}^n x_i^2 \right\}^{\frac{1}{2}}$$

Z (21) wynika, że dodatkowym parametrem, który ma wpływ na ocenę $\hat{\sigma}_{s=2}^2$ jest jedynie współczynnik asymetrii γ_3 . Dla jego wartości równej zero (symetria PDF) oszacowanie wielomianowe (21) degeneruje się do klasycznego (5).

W [10] wykazano, że w przypadku asymptotycznym $n \rightarrow \infty$ wykorzystanie (21) daje wzrost dokładności (zmniejszenie wariancji estymacji). Uzyskuje się wartość porównywalną jak dla oceny (5), opisaną w (17) jako g_2 .

Z (9) i (12a,b) wynika, że w pobliżu prawdziwej wartości σ^2 stochastyczny wielomian dla stopnia $s = 2$ o postaci (12) maksymalizują współczynniki:

$$(22) \quad k_0(\sigma) = \frac{-\ln(\sigma^2)}{(2 + \gamma_4 - \gamma_3^2)}, \quad k_1(\sigma) = \frac{2\gamma_3}{\sigma(2 + \gamma_4 - \gamma_3^2)},$$

$$k_2(\sigma) = \frac{-1}{\sigma^2(2 + \gamma_4 - \gamma_3^2)}.$$

Niezależne od σ^2 wyrażenie $(2 + \gamma_4 - \gamma_3^2)$ wystąpi w mianowniku każdego składnika (14) i przy tworzeniu potrzebnej statystyki można jego nie uwzględniać. Tak więc, przy istnieniu informacji a priori o wartości wariancji σ_0^2 przed zmianą i σ_1^2 po zmianie, statystykę wielomianową (14) rzędu $s = 2$ ma postać:

$$(23) \quad P_r^{(2)}(\sigma_0^2, \sigma_1^2) = -r \ln(\sigma_0^2) + 2\gamma_3 \frac{\sum_{i=1}^r x_i}{\sigma_0} - \frac{\sum_{i=1}^r x_i^2}{\sigma_0^2} + -(n-r) \ln(\sigma_1^2) + 2\gamma_3 \frac{\sum_{i=r+1}^n x_i}{\sigma_1} - \frac{\sum_{i=r+1}^n x_i^2}{\sigma_1^2}$$

Wykazano też [12], że dla znanych a priori wartości wariancji szeregu losowego σ_0^2 i σ_1^2 , postaci (23) odpowiada inna wygodniejsza w obliczeniach statystyka:

$$(24) \quad P_r^{(2)}(\sigma_0^2, \sigma_1^2) = -(n-r) \left[\ln(\sigma_1^2) - \ln(\sigma_0^2) \right] + 2\gamma_3 \left[\frac{1}{\sigma_1} - \frac{1}{\sigma_0} \right] \sum_{i=r+1}^n x_i - \left[\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right] \sum_{i=r+1}^n x_i^2$$

Przy nieznanach a priori wartościach wariancji szeregu losowego, w wielomianowym oszacowaniu punktu zmiany zastępuje się te wariancje oszacowaniami uzyskanymi a posteriori o postaci (5). Tworzy się to dla każdego potencjalnie możliwego punktu zmiany.

Jeśli na przykład nieznaną a priori jest tylko wartość wariancji $\hat{\sigma}_{1,r}^2$ po zmianie, to algorytm oszacowania punktu zmiany \hat{t} ma postać:

$$(25) \quad \hat{t}_1 = \arg \max_{1 \leq r \leq n-1} \left\{ (n-r) \left(\ln(\hat{\sigma}_{1,r}^2) - \ln(\sigma_0^2) + 1 - \frac{\hat{\sigma}_{1,r}^2}{\sigma_0^2} \right) + -2\gamma_3 \left(\frac{1}{\hat{\sigma}_{1,r}} - \frac{1}{\sigma_0} \right) \sum_{i=r+1}^n x_i \right\}$$

Gdy nie są znane a priori wartości obu parametrów $\hat{\sigma}_{0,r}^2$ przed zmianą i po $\hat{\sigma}_{1,r}^2$, to algorytm dla \hat{t} ma postać:

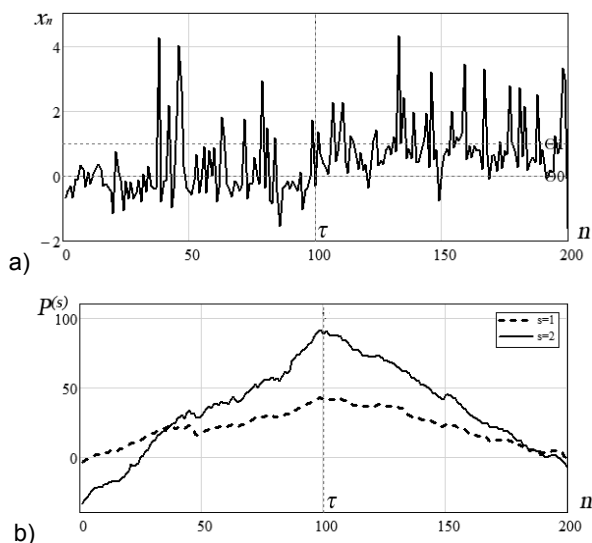
$$(26) \quad \hat{t}_2 = \arg \max_{1 \leq r \leq n-1} \left\{ 2\gamma_3 \left(\frac{\sum_{i=1}^r x_i}{\hat{\sigma}_{0,r}} + \frac{\sum_{i=r+1}^n x_i}{\hat{\sigma}_{1,r}} \right) + -r \ln(\hat{\sigma}_{0,r}^2) - (n-r) \ln(\hat{\sigma}_{1,r}^2) - n \right\}$$

Analiza struktury wielomianowych statystyk (23) - (26) po raz kolejny potwierdza, że wykorzystywanie metody MMW dla wielomianu stopnia $s = 2$ ma sens, gdy współczynnik asymetrii rozkładu danych statystycznych $\gamma_3 \neq 0$.

Statystyczne modelowanie oszacowania a posteriori punktu zmiany

Na podstawie wyników powyższej analizy opracowano w środowisku programowym MATLAB pakiet oprogramowania do modelowania statystycznego proponowanych semi-parametrycznych procedur estymacji punktu zmiany wartości średniej lub wariancji nie-gaussowskich szeregów losowych. Pakiet ten umożliwia modelowanie statystycznych eksperymentów zarówno pojedynczych, jak i wielokrotnych według metody Monte Carlo, aby można było porównać obliczeniowo dokładność algorytmów klasycznych i proponowanych algorytmów wielomianowych.

Na rysunku 1b przedstawiono wyniki jednego z modelowań procedury estymacji punktu zmiany średniej ($\theta_0 = 0$, $\theta_1 = 1$) w nie-gaussowskim szeregu losowym o parametrach $\sigma = 1$, $\gamma_3 = 2$ i $\gamma_4 = 5$ (rys.1a).

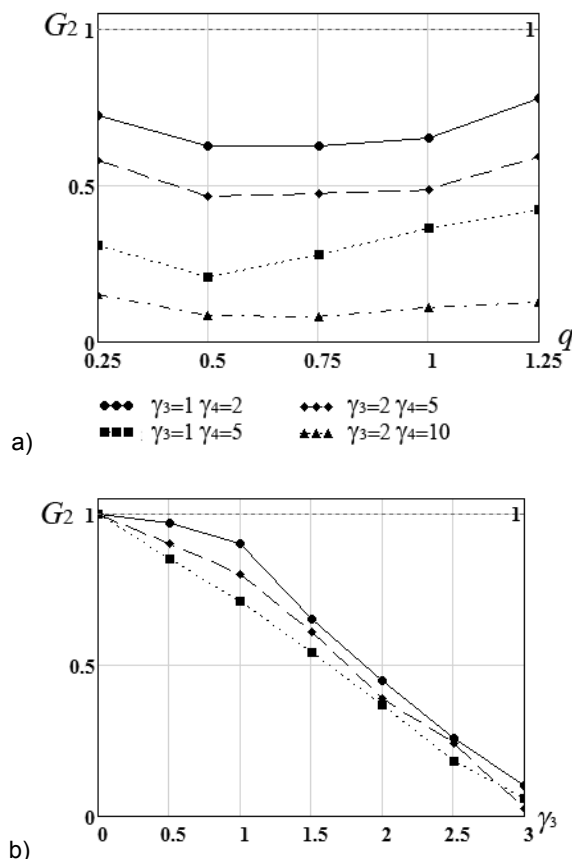


Rys 1 - Przykład szacowania a posteriori punktu zmiany wartości średniej.

W procedurze tej zastosowano klasyczną wersję algorytmu (2) do oszacowania a posteriori metodą MMP (odpowiadającą MMW dla $s=1$) i algorytm wielomianowy (12) MMW przy $s=2$ (rys.2b). Podane wyniki wykazują wizualnie, że dla statystyki wielomianowej stopnia $s=2$ uzyskuje się potencjalnie wyższą dokładność, ponieważ ma ona ostrzejszy wierzchołek w porównaniu do "rozmytego" maksimum statystyki dla $s=1$ (linia przerywana na rys 1b).

Wyniki pojedynczego eksperymentu nie wystarczają jednak do w pełni wiarygodnego porównywania dokładności algorytmów oceny statystycznej. W tym przypadku jako kryterium do porównywania skuteczności może być stosunek wariancji oszacowań punktu zmiany zaburzeń, otrzymywanych przy doświadczeniach powtarzanych wielokrotnie dla tych samych wartości początkowych parametrów modelu. Teoretycznie na wyniki modelowania statystycznego algorytmów do estymacji a posteriori punktu zmiany parametru może wpływać wiele czynników, w tym: względna wartość zmiany parametru, właściwości probabilistyczne (współczynniki kumulantów wyższych rzędów) próbki z nie-gaussowskiego szeregu losowego, zakres dostępnych a priori informacji o wartościach zmian parametrów. Ponadto, na dokładność szacowania punktu zaburzenia ma również wpływ liczność próbki n , a na dokładność wyznaczenia wariancji ocen - liczba symulacji m wykonanych metodą Monte Carlo dla jednakowych warunków początkowych.

Jako przykład, na rysunkach 2 i 3 podano wartości G_2 otrzymane z modelowania statystycznego szeregu przy $m=2000$ powtórzeniach próbek o $n=200$ danych. Charakteryzują one uzyskane metodą MMW zmniejszenie ocen punktu zmiany wariancji wielomianu stopnia $s=2$ w stosunku do otrzymywanych metodą klasyczną.



Rys. 2 - Wartości doświadczalne współczynników redukcji wariancji

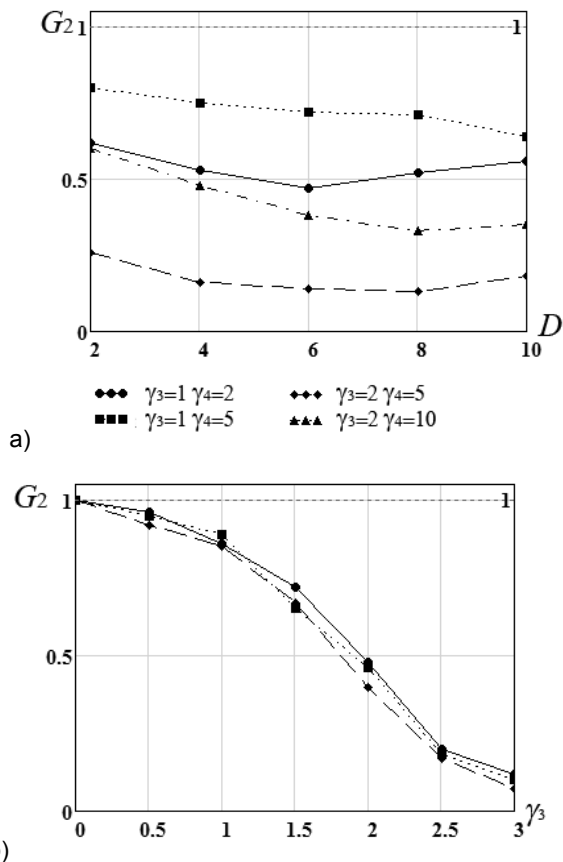
G_2 dla ocen punktu zmiany średniej:
 ●-●-● znane wartości średniej,
 ■-■-■ znane wartości średniej przed zmianą,
 ◆-◆-◆ nieznanne wartości średniej.

Rysunek 2a przedstawia zależności G_2 od względnej wartości zmiany $q = (\theta_1 - \theta_0)/\sigma$, uzyskane dla różnych wartości współczynników asymetrii γ_3 i kurtozy γ_4 , a rysunek 2b - zależności G_2 od γ_3 przy $\gamma_4 = 10$ i $q = 0,5$ dla różnej wielkości informacji a priori o wartościach średnich szeregów losowych przed i po punkcie zmiany.

Rysunek 3a (na następnej stronie) przedstawia zależności wartości G_2 od wartości względnej zmiany wariancji $D = \sigma_1^2/\sigma_0^2$ uzyskane dla różnych wartości współczynników asymetrii γ_3 i kurtozy γ_4 . Zaś na rysunku 3b podano zależności G_2 od γ_3 (dla $\gamma_4 = 10$ i $D = 4$), otrzymane dla różnej znanej a priori informacji i kilku wartości zmiennych parametrów rozkładu.

Analiza tych i innych wyników doświadczalnych potwierdza założenie teoretyczne o skuteczności wielomianowych oszacowań punktu zmiany parametru w szeregu losowym. W tym przypadku względny wzrost dokładności dla różnych wariantów zadania w przybliżeniu jest podobny i wynika z istnienia, lub braku informacji a priori o wartościach parametru podlegającego zmianie.

Wzrost dokładności ocen nie zależy znacząco od względnej wielkości zmiany, ale przede wszystkim od stopnia nie-gaussowskości rozkładu danych, wyrażonego liczbowo przez wartości bezwzględne współczynników kumulantów wyższego rzędu.



Rys. 3 - Doświadczalne wartości współczynników redukcji wariancji ocen punktu zmiany wariancji:

- $\bullet-\bullet-\bullet$ znane wartości średniej,
- $\blacksquare-\blacksquare-\blacksquare$ znane wartości średniej przed zmianą,
- $\blacktriangle-\blacktriangle-\blacktriangle$ nieznanne wartości średniej.

Wnioski

Wyniki wykonanych badań prowadzą łącznie do ogólnego wniosku o potencjalnie dużej skuteczności zastosowania metody maksymalizacji wielomianu (MMW) do syntezy prostych w realizacji adaptacyjnych algorytmów do szacowania punktów zmiany parametrów w procesach stochastycznych o statystykach nie-gaussowskich.

Nowością o charakterze naukowym jest opracowanie oryginalnego podejścia do tworzenia semi-parametrycznych algorytmów dla estymacji w trybie aposteriori punktu zmiany parametrów szeregu losowego z wykorzystaniem wielomianów stochastycznych.

Z wielu możliwych kierunków rozwoju tej metody powinno się uwzględnić:

- zwiększenie stopnia wielomianu stochastycznego dla uzyskania bardziej skutecznych rozwiązań zagadnienia, zwłaszcza dla nie-gaussowskich szeregów losowych o rozkładach symetrycznych;
- analizę wpływu dokładności wyznaczania parametrów modelu nie-gaussowskiego (statystyki wyższego rzędu) na stabilność algorytmów wielomianowych dla szacowania a posteriori punktu zmiany;

- syntezę wielomianowych algorytmów do szacowania punktu zmiany innych parametrów (współczynników regresji i korelacji) oraz położenia punktu równoczesnej zmiany kilku parametrów procesu (np. średniej i wariancji).

Przewiduje się, że zarówno opisaną tu metodę wykrywania aposteriori punktu τ zmiany parametru procesu stochastycznego jak i metody w czasie rzeczywistym, w tym polskie badania teoretyczne [13], będzie można wykorzystać w oprogramowaniu systemów pomiarowych do bieżących badań parametrów takich procesów [14].

LITERATURA

- Chen J., Gupta A. K. Parametric statistical change point analysis. Birkhaeuser, (2012), p.273
- Reeves J., Chen, X. L. Wang, R. Lund, and Q. Lu. A review and comparison of change-point detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6): 900–915, (2007).
- Wang Y., C. Wu, Z. Ji, B. Wang, and Y. Liang. Non-parametric change-point method for differential gene expression detection. *PLoS ONE*, 6(5):e20060, (2011).
- Yamanishi K., Takeuchi J., Williams G., and Milne P., On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2000) p. 320–324.
- Constantinos S. Hilaris, Ioannis T. Rekanos, and Paris Ast. Mastorocostas, "Change Point Detection in Time Series Using Higher-Order Statistics: A Heuristic Approach," *Mathematical Problems in Engineering*, vol. 2013, Article ID 317613, 10 pages, (2013).
- Liu S., Yamada, M., Collier, N., & Sugiyama, M. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks* vol.43, (2013) p.72-83.
- Lokajicek T., and Klima K., "A First Arrival Identification System of Acoustic Emission (AE) Signals by Means of a Higher-Order Statistics Approach". *Measurement Science and Technology*, Vol. 17 (2006), p.2461-66.
- Yih-Ru Wang, The signal change-point detection using the high-order statistics of log-likelihood difference functions, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP* (2008), vol., no., p. 4381- 4384.
- Brodsky B., Darkhovsky B. Nonparametric Methods in Change-Point Problems. Kluwer Academic Publishers, Dordrecht, the Netherlands, (1993).
- Kunchenko Y., Polynomial Parameter Estimations of Close to Gaussian Random variables. Shaker Verlag Aachen, Germany, (2002).
- Hinkley D. Inference about the change-point in a sequence of random variables *Biometrika*. v. 57 (1970) no 1, p. 1–17.
- Kunchenko Y. P., Zabolotnyii S. V., Polinomialnyje ocenki parametrov sluchainykh vielichyn. Chast II. Ocenka parametrov bliskih k gaussowskim sluchainykh vielichyn. Monography of Cherkasy State University of Technology ChITI, Cherkasy (2001) s.s.251 (in Russian)
- Nosek K., Metody wykrywania punktu zmiany przy ograniczeniach na kształt alternatyw. Praca doktorska. AGH 2012
- Korczyński M. J., Warsza Z. L.: *Properties of virtual instrument to on-line measurement of nonstationary processes. Proceedings (Abstracts) of 3rd International Conference Manufacturing 2010*, Poznań University of Technology, s.198

Autorzy: Doc. dr inż. Serhii W. Zabolotnij, Cherkaskij Nacional'nij Tehnicheskij Universitet, Cherkasy Ukraina e-mail: zabolotni@ukr.net

Doc.(em.) dr inż. Zygmunt L. Warsza, Przemysłowy Instytut Automatyki i Pomiarów PIAP, 02 486 Warszawa, Al. Jerozolimskie 202, e-mail: zlw@op.pl