

УДК 519.2

MSC 62G08

A NON-PARAMETRIC STATISTICAL TECHNIQUE FOR CHANGEPOINT DETECTION IN CYBER-PHYSICAL SYSTEMS

A. V. URASOVSKYI

Faculty of Computer Science and Cybernetics, Taras Shevchenko Kiev National University,
Kiev, Ukraine, E-mail: urazovskya@gmail.com, ORCID: 0000-0002-7918-2876

НЕПАРАМЕТРИЧНИЙ СТАТИСТИЧНИЙ МЕТОД ВИЯВЛЕННЯ ТОЧОК ПЕРЕХОДУ В КІБЕРФІЗИЧНИХ СИСТЕМАХ

А. В. УРАЗОВСЬКИЙ

Факультет комп'ютерних наук та кібернетики, Київський національний університет
імені Тараса Шевченка, Київ, Україна, E-mail: urazovskya@gmail.com,
ORCID: 0000-0002-7918-2876

АБСТРАКТ. Cyber-physical systems generate multidimensional time series describing the state of the system. When the state of the system changes, it is necessary to detect the transition point in the time series. The article describes a new nonparametric method for detecting the transition point in multidimensional time series generated by components of cyber-system components, using the principal component analysis (PCA) as a dimensionality reduction method, and this dimensionality reduction is accompanied by the application of Petunin statistics to one-dimensional data sets. Numerical and quasi-real experiments demonstrate the high accuracy and stability of the proposed algorithm over a wide range of distributions and hypothetical examples of cyber-physical systems. The accuracy is measured by the number of steps after the transition point when it was detected. There is also a comparison with the already known methods — the Wilcoxon test and the Kolmogorov-Smirnov consistency test. Accuracy up to 20 steps from the transition point was achieved, and in most cases even less — no more than 10 steps. This method provides a clear and human-understandable interpretation of algorithms and their results.

KEYWORDS: real-time series, changepoint, nonparametric statistics, cyber-physical systems, dimensionality reduction.

АНОТАЦІЯ. Кіберфізичні системи генерують багатовимірні часові ряди, що описують стан системи. Коли стан системи змінюється,

необхідно виявити точку переходу в часовому ряді. У статті описано новий непараметричний метод виявлення точки переходу у багатовимірних часових рядах, що генеруються компонентами кіберсистемних компонентів, з використанням методу головних компонент (РСА) як методу зменшення розмірності, і це зменшення розмірності супроводжується застосуванням статистики Петуніна до одновимірних наборів даних. Чисельні та квазіреальні експерименти демонструють високу точність та стабільність запропонованого алгоритму в широкому діапазоні розподілів та гіпотетичних прикладів кіберфізичних систем. Точність вимірюється кількістю кроків після точки переходу, коли вона була виявлена. Також є порівняння з вже відомими методами — критерієм Вілкоксона і критерієм узгодженості Колмогорова-Смірнова. Було досягнуто точності до 20 кроків від точки переходу, а в більшості випадків навіть менше — не більше 10 кроків. Цей метод забезпечує чітку та зрозумілу людиною інтерпретацію алгоритмів та їх результатів.

КЛЮЧОВІ СЛОВА: часові ряди, точка переходу, непараметрична статистика, кіберфізичні системи, зменшення розмірності.

1. Вступ

Автоматизовані системи та штучний інтелект започаткували нову еру можливостей для виявлення точок переходу у багатовимірних часових рядах. Цей технологічний прогрес має особливе значення в таких галузях, як медицина, інженерія, економіка та кібербезпека, де може бракувати вузькоспеціалізованих спеціалістів. Комп'ютерне моделювання та інтелектуальні системи відіграють ключову роль в оптимізації розподілу людських ресурсів, спрямовуючи їх на розв'язання критичних задач, що безпосередньо впливають на добробут людей.

Ще одна важлива сфера, де комп'ютерне моделювання та інтелектуальні системи мають вирішальне значення — це атомна енергетика. З огляду на багаторічне використання таких джерел енергії, забезпечення цілісності та працездатності ключових компонентів стає критично важливим для запобігання або мінімізації наслідків потенційних аварій. Відповідно, системи атомних електростанцій постійно вдосконалюються, охоплюючи резервні та диверсифіковані механізми безпеки, що забезпечують надійність і зміцнюють довіру до захисту працівників і населення навіть в умовах непередбачуваних ситуацій.

Щоб метод вважався ефективним у такому контексті, він має відповідати таким ключовим критеріям:

1. **Висока точність:** мінімізує кількість хибнопозитивних і хибнонегативних спрацьовувань, підвищуючи надійність виявлення точок переходу.
2. **Стабільність:** гарантує, що одиничні викиди або аномалії не спричиняють суттєвих викривлень у сигналі та не провокують хибне виявлення переходу.

3. **Незалежність від основного розподілу:** універсальність методу для різних ситуацій, процесів та об'єктів є бажаною властивістю, що робить його гнучким і широко застосовуваним.
4. **Низька обчислювальна вартість:** можливість працювати онлайн без істотного навантаження на сервери є практичною необхідністю.
5. **Збалансована чутливість:** важливо знайти правильний баланс — не надто чутливий (щоб уникати реагування на незначні зміни) і не надто нечутливий (щоб не пропустити критичні події на кшталт вибуху реактора).

У цій статті розглядається новий непараметричний метод виявлення точок переходу у багатовимірних часових рядах, що базується на критерії, запропонованому Ключиним і Петуніним [1]. Цей підхід демонструє явні переваги над традиційними статистичними критеріями, такими як статистики Колмогорова-Смірнова та Вілксона, що також відзначено у попередніх дослідженнях. Окрім того, обговорюється потенціал запропонованого методу у сфері медицини.

2. ОГЛЯД ЛІТЕРАТУРИ

В останнє десятиліття завдання своєчасного виявлення структурних зсувів у багатовимірних часових рядах стало ключовим для кіберфізичних систем: від енергетики й автономного транспорту до медичного моніторингу. Поява різкої зміни середнього, дисперсії або кореляційної структури сигналу часто сигналізує про критичні відхилення, отже методи детекції мають працювати онлайн, бути стійкими до викидів і не залежати від гіпотез щодо базового розподілу.

У літературі простежується дві парадигми. Пакетні (batch) алгоритми аналізують увесь масив даних постфактум, тоді як потокові (streaming) опрацьовують спостереження по мірі надходження, що критично, коли рішення потрібно приймати в реальному часі. Розгорнутий огляд офлайн-процедур наведено у [2], де систематизували класичні CUSUM-тести, сегментаційні схеми й байєсівські підходи до вже сформованих вибірок.

Однак із зростанням розмірності даних ефективність традиційних критеріїв падає: явище «прокляття розмірності» призводить до втрати статистичної потужності й вибухового зростання обчислювальних витрат. Це докладно продемонстровано у [3], де досліджували логарифмічну правдоподібність та розходження Кульбака-Лейблера для багатовимірних потоків. Проблему частково пом'якшують контроль-карти нового покоління, здатні фіксувати розріджені (sparse) зсуви середнього у високих вимірностях, що показали у [4].

Поряд із параметричними методами активно розвиваються непараметричні детектори. Прикладом є E-detectors — універсальна рамка онлайн-тестів, яка мінімізує припущення про розподіли й показує конкурентну швидкість на симульованих потоках. Близьку за ідеологією, Байєсівську стратегію пропонують у [5]; їхній алгоритм із лінійною складністю дозволяє

налаштовувати компроміс між точністю та швидкістю у випадку множинних змін.

Коли потрібна детальна локалізація сегментів, дослідники звертаються до методів мережевого кластерування станів [6]. Цей підхід демонструє високу роздільну здатність, але платить за це значними обчислювальними витратами, що обмежує його застосування в режимі реального часу. Високорозмірні задачі підштовхнули до розробки процедур, які обрізають пошуковий простір. Наприклад, у [7] запропонували швидку багатозмінну схему, проте вона покладається на припущення про незалежність спостережень, що не завжди реалізується у практиці.

Суттєвий виклик — наявність викидів; класичні CUSUM-тести можуть «сплутати» одиничний аномальний сплеск із тривалим трендом. Модифіковану байєсівську процедуру, стійку до аномалій, розробили у [8], продемонструвавши, що коректне обмеження апостеріорної ймовірності значно знижує кількість хибних спрацьовувань. Додатково у [9] показали, що функціональне «обрізання» CUSUM-статистики дає змогу отримати асимптотично оптимальні пороги при збереженні лінійної складності, що особливо актуально для хмарних серверів із жорсткими ресурсними квотами.

Окремий клас робіт присвячено адаптації байєсівської онлайн-детекції до бюджетних обчислень: у [10] описали схему, де розмір буфера й кількість перевірок динамічно коригуються, що дозволяє тримати затримку сповіщення в межах кількох десятків спостережень навіть у потоках з тисячами параметрів.

Методи машинного навчання широко використовуються для вирішення цих проблем, але вони спираються на навчальні дані, і отримані дані можуть мати зовсім різні розподіли [11]. Аналіз існуючих методів пошуку точок переходу та аномалій не показує універсальних тестів для вирішення цих проблем [12].

Статтю присвячено опису нового універсального статистичного методу виявлення точок змін та аномалій у багатовимірних часових рядах, що генеруються компонентами кіберфізичних систем. Методи машинного навчання широко використовуються для виявлення точок змін у багатовимірних часових рядах, що генеруються кіберфізичними системами, надаючи цінні можливості в різних галузях, таких як медицина, інженерія, економіка та кібербезпека [13].

Знаходження точки переходу можна здійснити різними способами, наприклад, простим скануванням часового ряду на наявність точки переходу або шляхом визначення конкретного місця розташування потрібної точки. Існування точки переходу визнається в [14], але локалізація та точність координат цільової точки є недостатніми.

Запропонований метод необхідний для аналізу безперервного потоку даних, оскільки він може виявляти викиди, не базується на жодних конкретних припущеннях та не залежить від інформації про розподіл даних. Нижче наведено роботи авторів, які дотримуються такої ж позиції з цього питання. Розширений байєсівський метод онлайн-виявлення точок зміни був створений у [15]. У [16] представлено стратегію дослідження географічних

даних. Автори [17] врахували алгоритми, які були запропоновані лише для експоненціальних моделей. Дані про тип розподілу були використані в [18] для зменшення обчислювальної складності. Подібний недолік можна знайти в статті [19].

Знаходження точок зміни моделі вимагає апріорних припущень щодо даних. Часто потрібна попередня обробка даних, щоб точно визначити точки зміни в багатовимірному часовому ряді [8]. Автори оцінили байєсівський метод розщеплення багатовимірних часових рядів за допомогою вибірки Гіббса та методу МСМС. Автори показали, що точки зміни надійно виявляються, а їхні координати локалізуються шляхом неявно аналізу структури залежності. Подібні пропозиції щодо розпізнавання жестів були зроблені в статті [20]. Через значні зрушення середніх значень оцінка точки зміни за допомогою оцінки моментів Юла-Вокера [21] є нестабільною.

Попри помітний прогрес, огляд виявляє кілька прогалин.

По-перше, відсутній єдиний репрезентативний бенчмарк, який би поєднував реалістичні викиди, сезонність і корельовані канали.

По-друге, більшість статей звітує про результати на синтетичних даних, тоді як публікацій із відкритими промисловими наборами (наприклад, телеметрія ядерних реакторів чи флоти БПЛА) обмаль.

Нарешті, залишається відкритим питання автоматичного налаштування ширини ковзного вікна й адаптивних порогів. Узагальнюючи, сучасна наукова спільнота просувається від статичних, розподіл-залежних моделей до гібридних, непараметричних і ресурсо-ефективних схем. Наступним кроком бачиться всебічна валідація цих підходів на реальних потоках даних та розробка інтегрованих систем моніторингу, де алгоритм детекції зміни є основою, а не ізольованим модулем.

3. ТЕОРЕТИЧНА ЧАСТИНА

1. Статистика Петуніна. В роботах [22, 23, 24] запропонована та досліджена p -статистика, що базується на теоремі Хілла [25], яка дозволяє оцінити ступінь близькості між двома вибірками. Перевірка гіпотези про однаковість функцій розподілу вибірок виконується саме за допомогою цієї статистики.

Розглядаються функції розподілу F_G та $F_{G'}$ для двох генеральних сукупностей G і G' . Нехай маємо дві вибірки:

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \in G, \quad \mathbf{x}' = (x'_1, x'_2, \dots, x'_m) \in G',$$

де $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ і $x'_{(1)} \leq x'_{(2)} \leq \dots \leq x'_{(m)}$ — відповідні порядкові статистики. Припустимо, що $F_G(u) = F_{G'}(u)$, тоді ймовірність того, що деяке x'_k належить інтервалу $(x_{(i)}, x_{(j)})$, дорівнює [25]

$$P\left(A_{ij}^{(k)}\right) = P\left(x'_k \in (x_{(i)}, x_{(j)})\right) = p_{ij}^{(n)} = \frac{j-i}{n+1}. \quad (1)$$

Якщо $x'_k \in \{x'_{(1)}, x'_{(2)}, \dots, x'_{(m)}\}$, то можна знайти частоту $h_{ij}^{(n)}$ події A_{ij} та побудувати довірчі інтервали $(\Delta_{ij}^{(1)}, \Delta_{ij}^{(2)})$ для p_{ij} на рівні значущості β :

тобто

$$B = \left\{ p_{ij} \in \left(\Delta_{ij}^{(1)}, \Delta_{ij}^{(2)} \right) \right\}, \quad P(B) = 1 - \beta. \quad (2)$$

Відповідно до [26], маємо:

$$\Delta_{ij}^{(1)} = \frac{h_{ij}^{(n)} n + \frac{g^2}{2} - g \sqrt{h_{ij}^{(n)} (1 - h_{ij}^{(n)}) n + \frac{g^2}{4}}}{n + g^2}, \quad (3)$$

$$\Delta_{ij}^{(2)} = \frac{h_{ij}^{(n)} n + \frac{g^2}{2} + g \sqrt{h_{ij}^{(n)} (1 - h_{ij}^{(n)}) n + \frac{g^2}{4}}}{n + g^2}. \quad (4)$$

Можна визначити рівень значущості інтервалу довіри

$$I_{ij}^{(n,m)} = (\Delta_{ij}^{(1)}, \Delta_{ij}^{(2)})$$

через значення g , припускаючи:

$$\varphi(g) = 1 - \frac{\beta}{2}, \quad (5)$$

де $\varphi(g)$ — функція щільності нормального розподілу. Згідно з правилом 3σ [27], значення $g = 3$ відповідає рівню значущості не більше ніж 0.05.

Загальна кількість довірчих інтервалів

$$N = \frac{n(n-1)}{2}, \quad (6)$$

а кількість таких інтервалів I_{ij} , що містять ймовірність $p_{ij}^{(n)}$, позначимо L . Таким чином, статистика Петуніна

$$h^{(n)} = \frac{L}{N}, \quad (7)$$

характеризує ступінь близькості вибірок \mathbf{x} і \mathbf{x}' . Перевірка гіпотези H на рівні значущості 0.05 проводиться шляхом підстановки значення h у формулу для довірчого інтервалу і перевіркою, чи потрапляє оцінене значення в межі $I = (\Delta^{(1)}, \Delta^{(2)})$ [1].

2. Доведення консистентності. Поведінка h під нульовою гіпотезою $H_0: F_G = F_{G'}$.

Для кожної пари (i, j) справджується:

$$P\{p_{ij} \in I_{ij}\} = 1 - \beta.$$

Введемо функцію-індикатор:

$$B_{ij} = \mathbf{1}\{p_{ij} \in I_{ij}\}.$$

Тоді математичне сподівання індикатора:

$$\mathbb{E}(B_{ij}) = 1 - \beta \quad \Rightarrow \quad \mathbb{E}L = (1 - \beta)N \quad \Rightarrow \quad \mathbb{E}h = (1 - \beta).$$

Оскільки $B_{ij} \in \{0, 1\}$ і таких індикаторів N штук, можна застосувати нерівність Хеффдінга для сум незалежних бернуллівських випадкових величин.

Лема 1 (Нерівність Хеффдінга для змінних Бернуллі). *Нехай X_1, X_2, \dots, X_n — незалежні випадкові величини, де $X_i \in [0, 1]$ та $\mathbb{E}X_i = \mu_i$. Тоді для будь-якого $t > 0$ виконується:*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mu_i\right| \geq t\right) \leq 2 \exp(-2nt^2).$$

У нашому випадку це застосовується до $X_i = B_{ij}$, а отже до $h = \frac{1}{N} \sum B_{ij}$, і отримуємо:

$$P(|h - (1 - \beta)| > \varepsilon) \leq 2e^{-2\varepsilon^2 N}.$$

При $n \rightarrow \infty$, відповідно $N = \binom{n}{2} \rightarrow \infty$, тому для будь-якого фіксованого $\varepsilon > 0$ маємо:

$$2e^{-2\varepsilon^2 N} \rightarrow 0,$$

а отже

$$P(|h - (1 - \beta)| > \varepsilon) \rightarrow 0.$$

Це означає, що $h \xrightarrow{P} (1 - \beta)$, тобто h є ймовірнісно консистентною оцінкою.

Поведінка h під альтернативною гіпотезою $H_1 : F_G \neq F_{G'}$. «Сильною» називають альтернативу, за якої F_G і $F_{G'}$ відрізняються у будь-якій точці, тобто $\sup_x |F_G(x) - F_{G'}(x)| = \delta > 0$.

Якщо ж F_G і $F_{G'}$ збігаються на якійсь підмножині (наприклад, різняться лише стрибком у точці x_0), то h буде збігатись не до 0, а до деякого числа $\gamma \in (0, 1 - \beta)$ — це нормально, критерій усе одно консистентний (ймовірність помилки II роду $\rightarrow 0$).

Ми хочемо показати, що при будь-якому «справжньому» відхиленні розподілів F_G і $F_{G'}$, значення h прямує до 0, а не до $1 - \beta$, як під H_0 . Тоді критерій, який відхиляє H_0 , коли h менше за поріг, буде консистентним.

Введемо функцію різниці $D(x) = F_G(x) - F_{G'}(x)$. Позначимо максимальну абсолютну різницю між функціями як $\delta = \sup_x |D(x)| > 0$.

Оберемо будь-яку точку x^* , де $|D(x^*)| = \delta$.

У впорядкованій вибірці X , $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ позначимо за k найбільший індекс, для якого $x_{(k)} \leq x^*$. Тобто ми затиснули $x_{(k)} \leq x^* < x_{(k+1)}$.

Як добре відомо, для великих n позиція k добре наближує квантиль $F(x^*)$ за формулою $k \approx n \cdot F(x^*)$.

За наших визначень, інтервал $(x_{(i)}, x_{(j)})$ перетинає x^* , якщо $i \leq k < j$, тобто точка x^* лежить всередині такого інтервалу. Порахуємо їхню кількість. Кількість значень $i - k$, кількість значень $j - n - k$, тому кількість таких пар (i, j) дорівнює $k(n - k)$, що можна наблизити як:

$$k(n - k) \approx F(x^*) \cdot (1 - F(x^*)) \cdot n^2.$$

Для будь-якої такої пари (i, j) , для якої інтервал $(x_{(i)}, x_{(j)})$ перетинає x^* , виконується, що теоретична ймовірність під H_0 це $p_{ij} = \frac{j-i}{n+1}$, а справжня ймовірність за $F_{G'}$ це $\lambda_{ij} = F_{G'}(x_j) - F_{G'}(x_i)$.

Тому:

$$|\lambda_{ij} - p_{ij}| = |F_{G'}(x_j) - F_{G'}(x_i) - F_G(x_j) + F_G(x_i)| = |D(x_j) - D(x_i)|.$$

За припущенням, існує точка x^* , де $|D(x^*)| = \delta > 0$. Без обмеження загальності розглянемо випадок, де $D(x^*) = \delta$ (якщо $-\delta$, усі знаки нижче міняються на протилежні).

Через неперервність D існує $\varepsilon > 0$ таке, що $D(x) \geq \frac{\delta}{2}$ для всіх $x \in (x^* - \varepsilon, x^* + \varepsilon) = (a, b)$.

Візьмемо будь-який індекс i такий, що $x_{(i)} \leq a$. Для нього $D(x_{(i)}) \leq D(a) \leq \frac{\delta}{2}$.

Візьмемо будь-який індекс j такий, що $x_{(j)} \in (a, b)$, щоб для нього виконувалось $D(x_{(j)}) \geq \frac{3\delta}{4}$.

Тоді для обраних індексів:

$$|D(x_j) - D(x_i)| \geq \frac{3\delta}{4} - \frac{\delta}{2} = \frac{\delta}{4} = \Delta.$$

Число таких індексів i приблизно дорівнює $n \cdot F(a) = c_1 n$, а число таких індексів j приблизно дорівнює $n \cdot p(a, b) = c_2 n$. Отже всіх комбінацій «поганих» пар, для яких індикатор дорівнює нулю дорівнює $M = c_1 n \cdot c_2 n = cn^2$, де $c = c_1 c_2 > 0$ не залежить від n .

Для пари (i, j) у другій вибірці спостерігаємо частоту $h_{ij} = \frac{L_{ij}}{m}$.

При рівні значущості $1 - \beta$ формула Вілсона дає:

$$\Delta_{ij}^{(1)} = \frac{h_{ij}^{(n)} m + \frac{g^2}{2} - g \sqrt{h_{ij}^{(n)} (1 - h_{ij}^{(n)}) m + \frac{g^2}{4}}}{m + g^2},$$

$$\Delta_{ij}^{(2)} = \frac{h_{ij}^{(n)} m + \frac{g^2}{2} + g \sqrt{h_{ij}^{(n)} (1 - h_{ij}^{(n)}) m + \frac{g^2}{4}}}{m + g^2}.$$

Напівширина цього інтервалу обчислюється як:

$$w_{ij} = \frac{g \sqrt{h_{ij}^{(n)} (1 - h_{ij}^{(n)}) m + \frac{g^2}{4}}}{m + g^2}.$$

Оскільки $0 \leq h_{ij} \leq 1 \Rightarrow h_{ij}(1 - h_{ij}) \leq \frac{1}{4}$, використаємо це для оцінки w_{ij} :

$$w_{ij} \leq \frac{g \sqrt{\frac{1}{4} m + \frac{g^2}{4}}}{m + g^2} = \frac{g}{2\sqrt{m + g^2}} \leq \frac{g}{2\sqrt{m}}.$$

Отже якщо ми хочемо, щоб $w_{ij} \leq \frac{\Delta}{2}$ для всіх інтервалів, достатньо взяти $m \geq \left(\frac{g}{\Delta}\right)^2$.

Пари, які здатні породити $|\lambda_{ij} - p_{ij}| < \delta$ це всі впорядковані пари (i, j) із N_δ точок, які лежать у $B_\delta = \{x : |D(x)| < \delta\}$.

Для великого n за законом великих чисел, $N_\delta = \#\{i : x_i \in B_\delta\} = np_\delta + o_p(n)$, тобто лінійно багато (в середньому np_δ).

Оцінимо кількість таких пар. Їх не більше, ніж $\binom{N_\delta}{2} = \frac{1}{2} p_\delta^2 n^2$. Обираючи δ достатньо малим, робимо $(p_\delta)^2$ достатньо малим, тобто щоб «гарних» було істотно менше, ніж «поганих». Маємо

$$\mathbb{E}L \leq \frac{1}{2} (1 - \beta) (p_\delta)^2 n^2 \Rightarrow \mathbb{E}h = \frac{\mathbb{E}L}{N} \leq (1 - \beta) (p_\delta)^2.$$

Оскільки $\delta \rightarrow 0$, то і $p_\delta \rightarrow 0$, отже $\mathbb{E}h \rightarrow 0$.

Далі робимо так само, як і пункт з H_0 . Використовуємо нерівність Хеф-фдінга:

$$\mathbb{P}(|h - \mathbb{E}h| \geq \varepsilon) \leq 2 \cdot \exp(-2\varepsilon^2 N).$$

Права частина прямує до 0 при будь-якому фіксованому ε . Разом із тим, що $\mathbb{E}h \rightarrow 0$, маємо, що $h \rightarrow 0$ за ймовірністю, коли $n, m \rightarrow \infty$.

Разом із результатом першого кроку ($h \rightarrow 1 - \beta$ під H_0) це доводить консистентність критерію: вибір порогу між $1 - \beta$ та 0 гарантує, що ймовірності помилки I та II роду прямують до 0.

3. Доведення потужності p -статистики Петуніна. Припущення про форму щільності. У більшості реальних задач функція щільності f є унімодальна (зростає до моди m , далі спадає). Це покриває нормальний розподіл, Лапласа, GEV ($k > -1$), логнормальний тощо.

Важливий наслідок. Функція щільності $f(x)$ є зростаючою на $(-\infty; m]$ і спадною на $[m; +\infty)$.

Теорема 1 (Теорема про потужність p -статистики Петуніна у випадку гіпотези про зсув). *Нехай F – неперервна, строго зростаюча функція розподілу з унімодальною щільністю $f > 0$ на \mathbb{R} . Є вибірки $X_{1:n} \stackrel{iid}{\sim} F$, $Y_{1:m} \stackrel{iid}{\sim} G = F(x - \mu)$, $\mu \neq 0$. (без обмеження загальності $\mu > 0$ – позитивний зсув). Тоді p -статистика Петуніна $h_{n,m} \xrightarrow{P} 0$ при $n, m \rightarrow \infty$, тобто критерій має асимптотичну потужність 1 при будь-якому ненульовому зсуві μ .*

Доведення. Введемо $D(x) = F(x) - G(x) = F(x) - F(x - \mu)$. Оскільки F строго зростаюча і неперервна, $\lim_{x \rightarrow \pm\infty} D(x) = 0$. Також $D'(x) = f(x) - f(x - \mu)$. Із припущення про унімодальність щільності f випливає:

- для $x < m$ та $\mu > 0$: $x - \mu < x < m \Rightarrow f(x) > f(x - \mu) \Rightarrow D'(x) > 0$,
- для $x > m + \mu$: $x > x - \mu > m \Rightarrow f(x) < f(x - \mu) \Rightarrow D'(x) < 0$.

Отже, існує $x^* \in (m, m + \mu)$ таке, що $D'(x^*) = 0$, і x^* – локальний максимум D . Позначимо $\delta = D(x^*) > 0$.

Візьмемо $\varepsilon \in (0, \frac{\delta}{2})$ і області:

$$A = \left\{ x : D(x) \geq \frac{\delta}{2} \right\},$$

$$B = \{ x : D(x) \leq \varepsilon \}.$$

Ймовірність потрапляння в A : $p_A = F(x^* + \eta) - F(x^* - \eta) > 0$, аналогічно для $p_B > 0$.

Нехай $N_A = \#\{i: X_i \in A\}$, $N_B = \#\{i: X_i \in B\}$. За законом великих чисел

$$\frac{N_A}{n} \xrightarrow{P} p_A, \quad \frac{N_B}{n} \xrightarrow{P} p_B$$

З ймовірністю $\rightarrow 1$ при великих n

$$N_A \geq \frac{1}{2} p_A n, \quad N_B \geq \frac{1}{2} p_B n$$

Беремо будь-який індекс $i \in B$, $j \in A$. Для них

$$|D(X_j) - D(X_i)| \geq \frac{\delta}{2} - \varepsilon = \Delta > 0$$

Кількість таких пар $M_n \geq \frac{1}{4}p_A p_B n^2 = cn^2$ — це були «погані» пари. Тепер порахуємо «гарні» пари.

Для будь-яких $i < j$

$$p_{ij} = \frac{j-i}{n+1} = F(X_j) - F(X_i)$$

У зсунутому розподілі $G(x) = F(x - \mu)$

$$\lambda_{ij} = G(X_j) - G(X_i) = F(X_j - \mu) - F(X_i - \mu)$$

Тоді отримаємо, що $0 \leq \lambda_{ij} \leq p_{ij}$, звідки

$$|\lambda_{ij} - p_{ij}| = p_{ij} - \lambda_{ij} \geq p_{ij} - p_{ij-1} = \frac{1}{n+1}$$

Тоді щоб $|\lambda_{ij} - p_{ij}| \leq \varepsilon$ необхідно мати $p_{ij} \leq \varepsilon + \frac{1}{n+1} \Rightarrow j-i \leq \lfloor (n+1)\varepsilon \rfloor$. Тому кількість таких пар індексів $|S_\varepsilon| \leq n \lfloor (n+1)\varepsilon \rfloor = O(\varepsilon n^2)$

Тепер можемо оцінити очікуване значення

$$\mathbb{E}h_{n,m} = \frac{\mathbb{E}L_{n,m}}{N} \leq (1-\beta) \frac{S_\varepsilon}{N} \leq 2(1-\beta)\varepsilon$$

При $\varepsilon \rightarrow 0$ отримаємо, що $\mathbb{E}h_{n,m} \rightarrow 0$. Тоді за нерівністю Хеффдінга отримаємо, що $h_{n,m} \xrightarrow{P} 0$. \square

Теорема 2 (Теорема про потужність p -статистики Петуніна у випадку гіпотези про масштаб). *Нехай F — неперервна, строго зростаюча функція розподілу з унімодальною щільністю $f > 0$ на \mathbb{R} . Є вибірки $X_{1:n} \stackrel{iid}{\sim} F$, $Y_{1:m} \stackrel{iid}{\sim} G = F\left(\frac{x}{\sigma}\right)$, $\sigma \neq 1$ (без обмеження загальності $\sigma > 1$ — чисте розтягування; ширший розподіл). Тоді p -статистика Петуніна $h_{n,m} \xrightarrow{P} 0$ при $n, m \rightarrow \infty$, тобто критерій має асимптотичну потужність 1 при будь-якому неединичному розтягуванні σ .*

Доведення. Введемо різницеву функцію $D(x) = F(x) - G(x) = F(x) - F\left(\frac{x}{\sigma}\right)$.

- При $x > 0$: $\frac{x}{\sigma} < x \Rightarrow F\left(\frac{x}{\sigma}\right) < F(x) \Rightarrow D(x) > 0$.
- При $x < 0$: $\frac{x}{\sigma} > x \Rightarrow F\left(\frac{x}{\sigma}\right) > F(x) \Rightarrow D(x) < 0$.
- При $x = 0$: $D(0) = 0$.

Також зрозуміло, що $D'(x) = f(x) - \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$.

На правій піввісі $x > t$, починаючи з деякого x_0 буде виконуватись те, що $D'(x) < 0$.

Для околу моди, тобто для x трохи більших за t , $f(x) > f\left(\frac{x}{\sigma}\right)$, бо обидва аргументи близькі до піку, але x ближче, тому $D'(x) > 0$. Тож існує точка x_+^* , для якої $D'(x_+^*) = 0$, причому x_+^* — максимум.

По аналогії на лівій піввісі ми знайдемо точку мінімуму x_-^* .

Позначимо $D(x_+^*) = \delta_+$, $D(x_-^*) = -\delta_-$. Візьмемо число $0 < \varepsilon < \frac{\delta}{2}$ і дві області:

$$A := \left\{ x : D(x) \geq \frac{\delta_+}{2} \right\} \supset (x_+^* - \eta, x_+^* + \eta),$$

$$B := \left\{ x : D(x) \leq -\frac{\delta_-}{2} \right\} \supset (x_-^* - \eta, x_-^* + \eta)$$

для деякого малого $\eta > 0$.

Ймовірність потрапити в цей інтервал:

$$p_A = F(x_+^* + \eta) - F(x_+^* - \eta) > 0, \quad p_B = F(x_-^* + \eta) - F(x_-^* - \eta) > 0$$

Нехай $N_A = \#\{i : X_i \in A\}$, $N_B = \#\{i : X_i \in B\}$. За законом великих чисел $\frac{N_A}{n} \xrightarrow{P} p_A$, $\frac{N_B}{n} \xrightarrow{P} p_B$

З ймовірністю $\rightarrow 1$ при великих n $N_A \geq \frac{1}{2}p_An$, $N_B \geq \frac{1}{2}p_Bn$.

Беремо будь-який індекс $i \in B, j \in A$. Для них

$$|D(X_j) - D(X_i)| \geq \left| \frac{\delta_+}{2} - \left(-\frac{\delta_-}{2} \right) \right| = \delta.$$

Кількість таких пар $M_n \geq \frac{1}{4}p_Ap_Bn^2 = cn^2$ — це були «погані» пари.

Тепер порахуємо «гарні» пари.

Для будь-яких $i < j$ $p_{ij} = \frac{j-i}{n+1} = F(X_j) - F(X_i)$. У зсунутому розподілі $G(x) = F\left(\frac{x}{\sigma}\right)$

$$\lambda_{ij} = G(X_j) - G(X_i) = F\left(\frac{X_j}{\sigma}\right) - F\left(\frac{X_i}{\sigma}\right)$$

Тоді отримаємо, що $0 \leq \lambda_{ij} \leq p_{ij}$, звідки

$$|\lambda_{ij} - p_{ij}| = p_{ij} - \lambda_{ij} \geq p_{ij} - p_{ij-1} = \frac{1}{n+1}$$

Тоді щоб $|\lambda_{ij} - p_{ij}| \leq \varepsilon$ необхідно мати $p_{ij} \leq \varepsilon + \frac{1}{n+1} \Rightarrow j - i \leq \lfloor (n+1)\varepsilon \rfloor$.

Тому кількість таких пар індексів $|S_\varepsilon| \leq n \lfloor (n+1)\varepsilon \rfloor = O(\varepsilon n^2)$

Тепер можемо оцінити очікуване значення

$$\mathbb{E}h_{n,m} = \frac{\mathbb{E}L_{n,m}}{N} \leq (1 - \beta) \frac{S_\varepsilon}{N} \leq 2(1 - \beta)\varepsilon.$$

При $\varepsilon \rightarrow 0$ отримаємо, що $\mathbb{E}h_{n,m} \rightarrow 0$. Тоді за нерівністю Хефдінга отримаємо, що і $h_{n,m} \xrightarrow{P} 0$. □

4. Метод головних компонент (Principal Component Analysis).

РСА був запропонований Карлом Пірсоном ще у 1901 році [28] як аналог теореми головних осей у механіці.

Головні компоненти визначаються як ортогональне лінійне перетворення, що трансформує дані у нову систему координат таким чином, що найбільша дисперсія (за деякою скалярною проекцією даних) припадає на першу координату (першу головну компоненту), друга за величиною дисперсія — на другу координату і так далі [29].

Розглянемо матрицю даних X розміру $n \times p$, центровану по кожному стовпцю (тобто емпіричне середнє кожного стовпця дорівнює нулю), де

кожен рядок $x_{(i)}$ — це одна реалізація експерименту, а кожен стовпець — окрема ознака (наприклад, вимір певного сенсора).

Математично перетворення задається як набір з l векторів-коефіцієнтів розмірності p :

$$\mathbf{w}_{(k)} = (w_1, \dots, w_p)_{(k)}$$

які проєктують кожен рядок $x_{(i)}$ матриці X у новий вектор головних компонент:

$$t_{(i)} = (t_1, \dots, t_l)_{(i)}, \quad \text{де } t_{k(i)} = x_{(i)} \cdot w_{(k)}, \quad i = 1, \dots, n; \quad k = 1, \dots, l$$

Це робиться так, що нові змінні t_1, \dots, t_l (які розглядаються по всій вибірці) послідовно наслідують максимальну можливу дисперсію від X . При цьому кожен вектор коефіцієнтів w нормований і є одиничним (тобто $\|w\| = 1$). Зазвичай кількість головних компонент l вибирається меншою за p , щоб досягнути зменшення розмірності.

4. ПРАКТИЧНА ЧАСТИНА

1. Опис алгоритму. Метою експериментів є перевірка ефективності алгоритму, що призначений для аналізу стаціонарних часових рядів, зосередженого на виявленні першої точки переходу та перевірці гіпотези однорідності. Алгоритм працює за наступною схемою:

1. **Підготовка даних.** Спочатку задається параметр ширини вікна w (width) та виділяється початкова вибірка x_1, x_2, \dots, x_w . Далі застосовується метод ковзного вікна.
2. **Основні підрахунки.** Для кожного вікна $(x_{i+1}, x_{i+2}, \dots, x_{i+w})$ виконується наступне:
 - **Аналіз головних компонент.** Проводиться проєкція вибірок методом головних компонент (Principal Component Analysis).
 - **Зменшення розмірності.** Отримуються проєкції для обох підвбірок: (p_1, \dots, p_w) та $(p_{i+1}, \dots, p_{i+w})$.
 - **Розрахунок статистики Петуніна.** Обчислюється p -статистика Петуніна ($pstat$) для кожної з підвбірок.
 - **Перевірка однорідності.** Якщо значення $pstat \geq 0,95$, то вважаємо, що розподіли співпадають. Якщо $pstat < 0,95$ — розподіли відмінні.
3. **Зсув вікна.** Вікно $(x_{i+1}, x_{i+2}, \dots, x_{i+w})$ зсувається на один елемент вправо. Алгоритм повторюється з кроку 2.
4. **Завершення.** Процедура повторюється доти, доки не буде проаналізовано всю послідовність. Якщо підвбірка після деякого елемента x_n демонструє неоднорідність, точку x_{n+1} розглядаємо як точку переходу.

Для демонстрації роботи алгоритму використовується ширина вікна $w = 40$, часовий ряд довжини $N = 200$, розділений на дві рівні частини з різними розподілами, де кожен елемент є 100-вимірним вектором. Процедуру повторюємо 100 разів та обчислюємо середнє значення p -статистики Петуніна.

На завершення результати візуалізуються: значення $pstat \geq 0,95$ позначаються синім кольором (однакові розподіли), а значення $pstat < 0,95$ — червоним (відмінні розподіли).

Крім p -статистики Петуніна, ми додатково використовуємо критерій Вілкоксона та Колмогорова–Смірнова для незалежних вибірок. За допомогою кожного з цих критеріїв також визначається позиція першої точки переходу, і результати порівнюються між собою за точністю локалізації.

2. Стрибок по середньому для рівномірного розподілу. У цьому прикладі розглядається часова послідовність довжини $N = 200$, яка складається з двох інтервалів з різними розподілами, де $U(a, b)$ — рівномірний розподіл на інтервалі $[a; b]$.

Часовий інтервал	T1	T2
0–99	$U(0; 1)$	—
100–199	—	$U(1; 2)$

ТАБЛ. 1. Опис розподілу для координат у часовому ряді.

Для кожного кроку застосовується РСА та обчислюються три типи статистик.

Результати візуалізуються у вигляді таких графіків: значення p -статистики з використанням порогового значення 0,95, KS-статистики Колмогорова–Смірнова та W -статистики Вілкоксона з використанням порогового значення 0,05.

Значення вище порогу (однорідні вибірки) позначаються синім кольором, нижче (неоднорідні вибірки) — червоним.

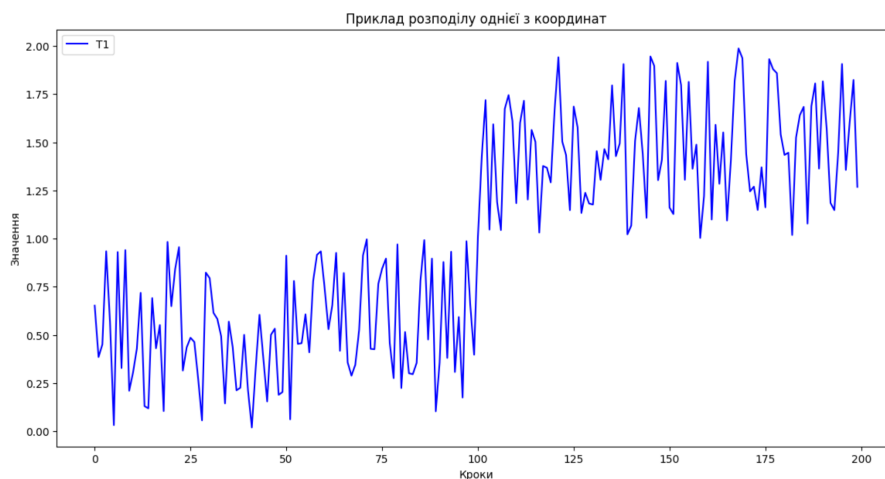


Рис. 1. Приклад розподілу однієї з координат.

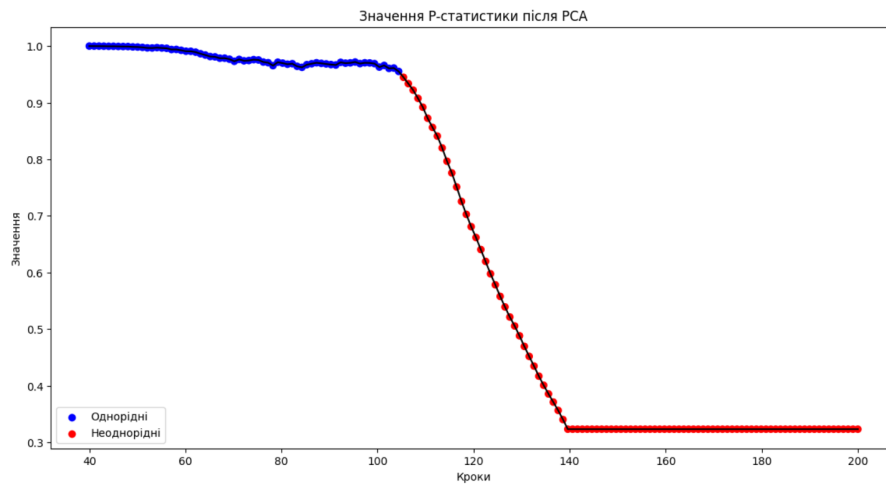


Рис. 2. Значення P-статистики після PCA. Перша точка переходу — 105.

Як видно з Таблиці 1, шукана перша точка переходу дорівнює 100, а потім відбулося відхилення від початкового розподілу.

З Рисунку 2 видно, що перша знайдена точка зміни дорівнює 105, тобто через 5 кроків після початку стрибка.

З Рисунку 3 видно, що перша знайдена точка зміни дорівнює 110, тобто через 10 кроків після початку стрибка.

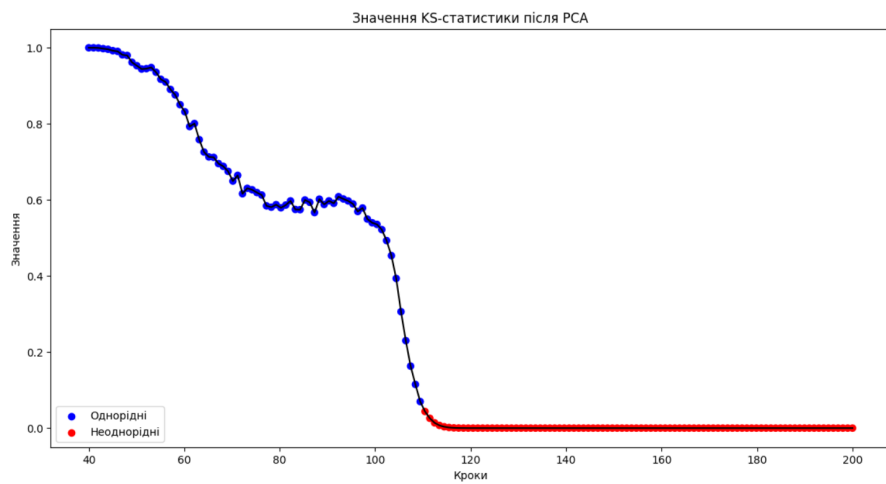


Рис. 3. Значення KS-статистики після PCA. Перша точка переходу — 110.

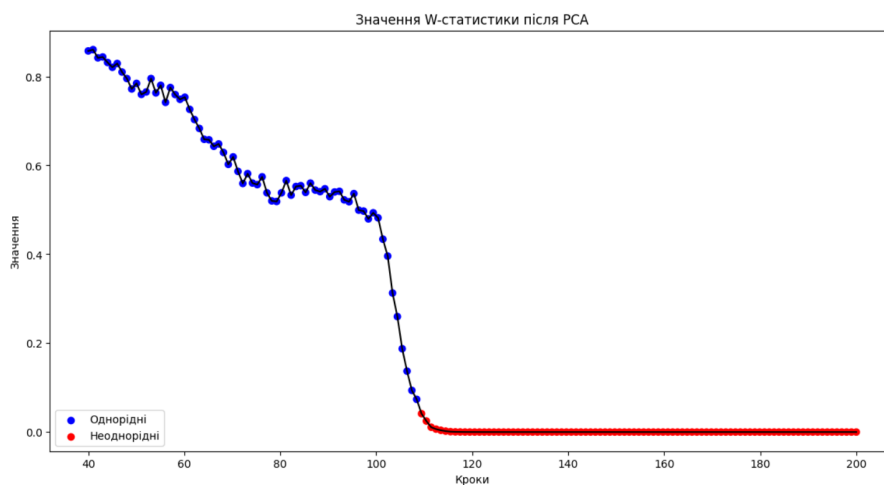


Рис. 4. Значення W-статистики після PCA. Перша точка переходу — 109.

З Рисунку 4 видно, що перша знайдена точка зміни дорівнює 109, тобто через 9 кроків після початку стрибка.

Як видно, кожна зі статистик доволі швидко розпізнала точку переходу, причому p -статистика виявилась не гіршою за інших.

3. Стрибок по дисперсії для нормального розподілу. У цьому прикладі розглядається часова послідовність довжини

$$N = 200,$$

яка складається з двох інтервалів з різними розподілами, де $N(\mu, \sigma)$ — нормальний розподіл з параметрами μ і σ .

Часовий інтервал	T1	T2
0–99	$N(0; 1)$	—
100–199	—	$N(0; 2)$

Табл. 2. Опис розподілу для координат у часовому ряді.

Для кожного кроку застосовується PCA та обчислюються три типи статистик.

Результати візуалізуються у вигляді таких графіків: значення p -статистики з використанням порогового значення 0,95, KS-статистики Колмогорова-Смірнова та W-статистики Вілксона з використанням порогового значення 0,05.

Значення вище порогу (однорідні вибірки) позначаються синім кольором, нижче (неоднорідні вибірки) — червоним.

Як видно з Таблиці 2, шукана перша точка переходу дорівнює 100, а потім відбулося відхилення від початкового розподілу.

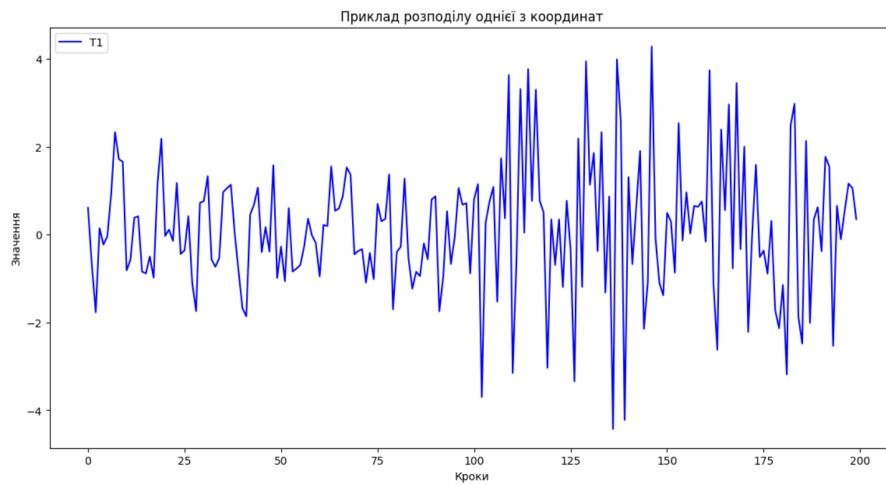


Рис. 5. Приклад розподілу однієї з координат.

З Рисунку 6 видно, що перша знайдена точка зміни дорівнює 108, тобто через 8 кроків після початку стрибка.

З Рисунку 7 видно, що перша знайдена точка зміни дорівнює 132, тобто через 32 кроки після початку стрибка.

З Рисунку 8 видно, статистика Вілкоксона не знайшла жодної точки переходу.

В цьому експерименті вже видно, що в гіпотезі про змінення масштабу p -статистика показує кращі результати за інші статистики.

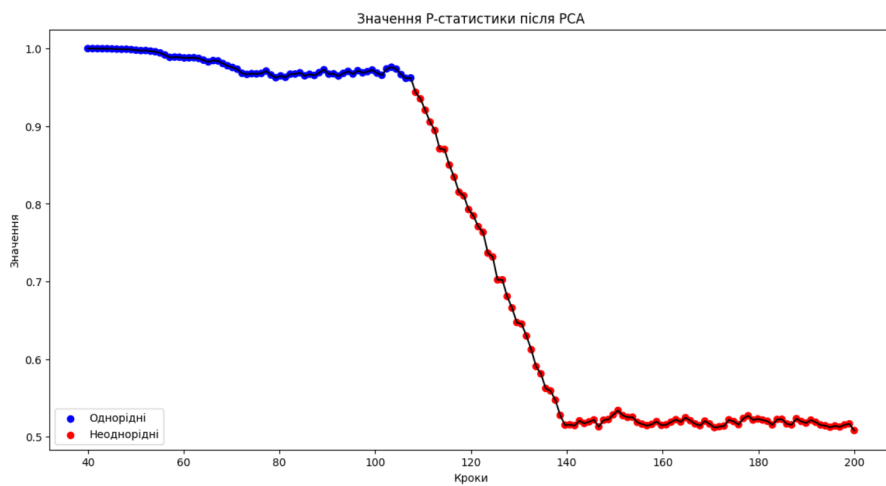


Рис. 6. Значення P -статистики після PCA. Перша точка переходу — 108.

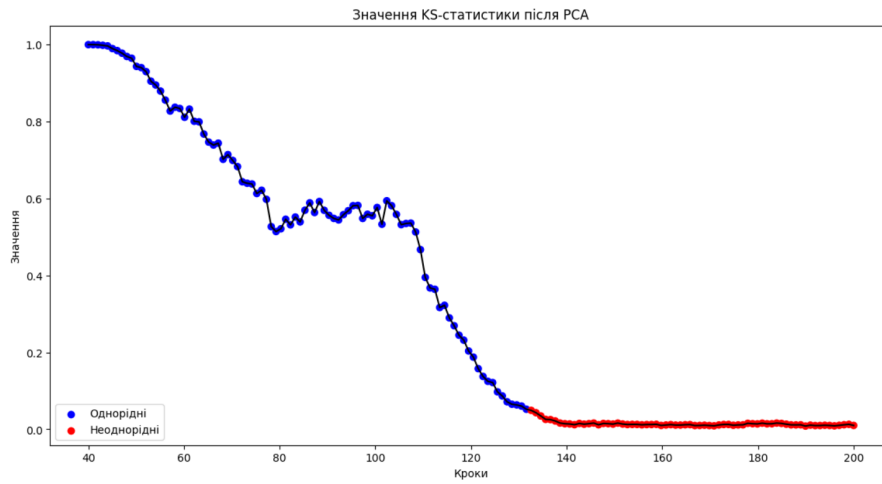


Рис. 7. Значення KS-статистики після PCA. Перша точка переходу — 132.

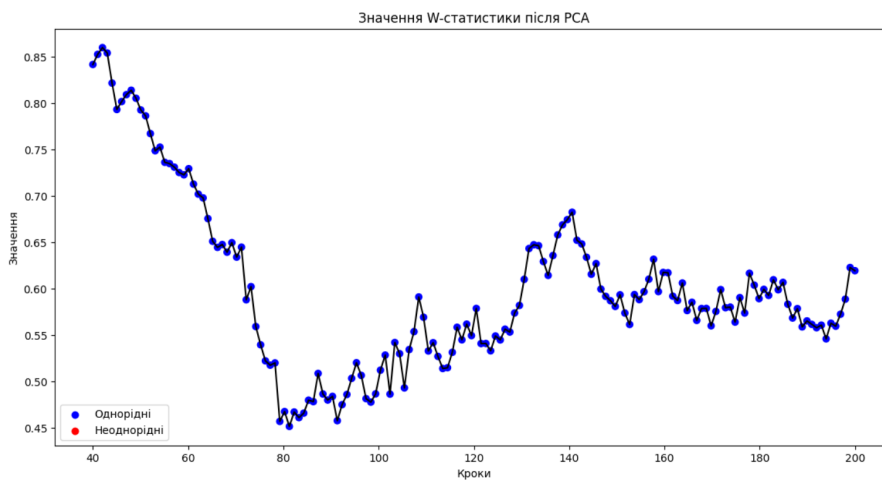


Рис. 8. Значення W-статистики після PCA. Не виявлено жодної точки переходу.

4. Стрибок по дисперсії для логнормального розподілу. У цьому прикладі розглядається часова послідовність довжини $N = 200$, яка складається з двох інтервалів з різними розподілами, де $\text{Log}N(\mu, \sigma)$ — логнормальний розподіл з параметрами μ і σ .

Часовий інтервал	T1	T2
0–99	$\text{Log}N(0; 1)$	—
100–199	—	$\text{Log}N(0; 0.5)$

ТАБЛ. 3. Опис розподілу для координат у часовому ряді.

Для кожного кроку застосовується PCA та обчислюються три типи статистик.

Результати візуалізуються у вигляді таких графіків: значення p -статистики з використанням порогового значення 0,95, KS-статистики Колмогорова-Смірнова та W -статистики Вілкоксона з використанням порогового значення 0,05. Значення вище порогу (однорідні вибірки) позначаються синім кольором, нижче (неоднорідні вибірки) — червоним.

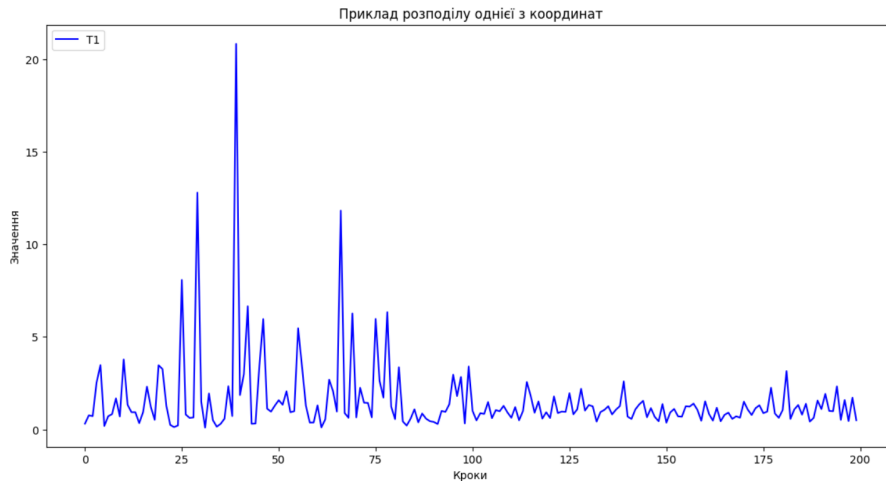


Рис. 9. Приклад розподілу однієї з координат.

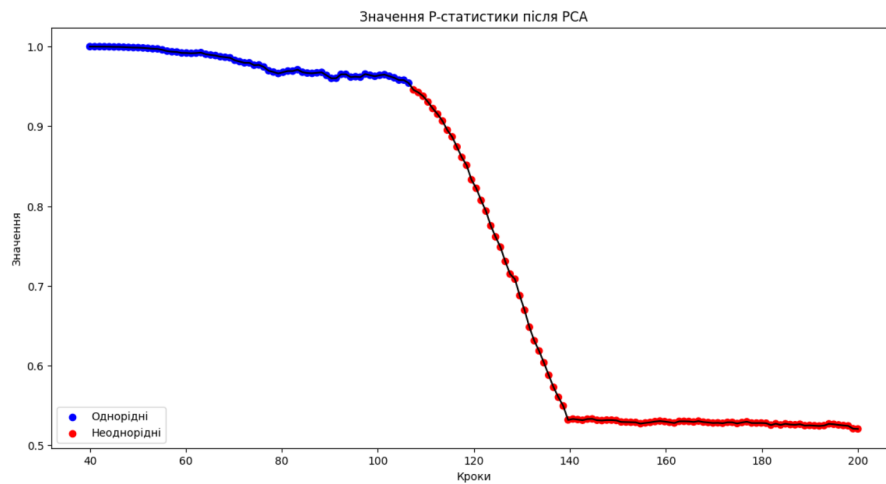


Рис. 10. Значення P -статистики після PCA. Перша точка переходу — 107.

Як видно з Таблиці 3, шукана перша точка переходу дорівнює 100, а потім відбулося відхилення від початкового розподілу.

З Рисунку 10 видно, що перша знайдена точка зміни дорівнює 107, тобто через 7 кроків після початку стрибка.

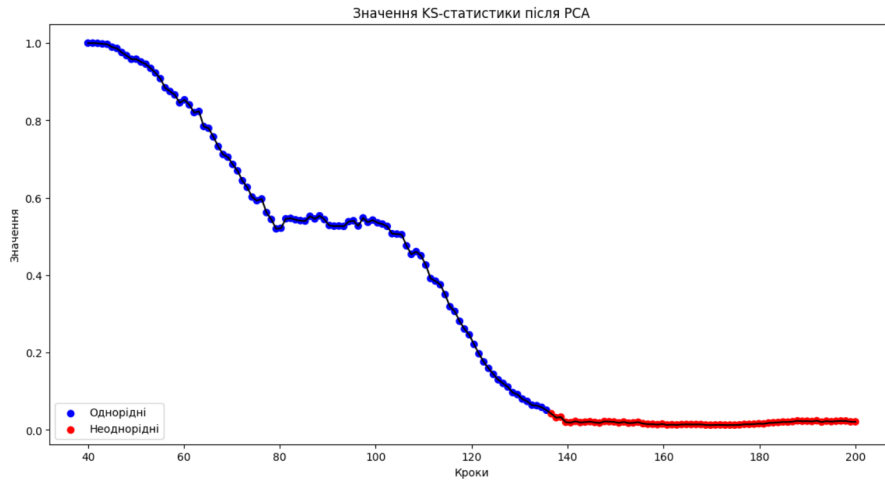


Рис. 11. Значення KS-статистики після PCA. Перша точка переходу — 136.

З Рисунку 11 видно, що перша знайдена точка зміни дорівнює 136, тобто через 36 кроків після початку стрибка.

З Рисунку 12 видно, статистика Вілкоксона не знайшла жодної точки переходу.

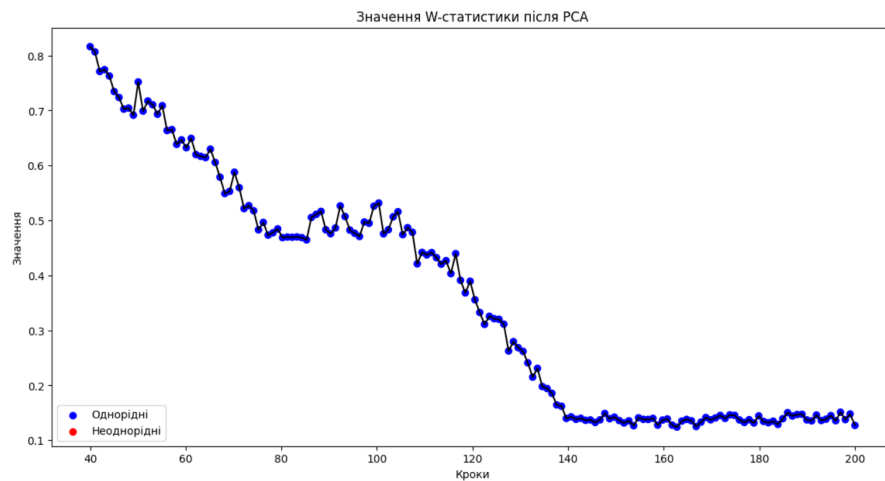


Рис. 12. Значення W-статистики після PCA. Не виявлено жодної точки переходу.

Так само як і в попередньому експерименті, p -статистика показує кращі результати за інші статистики.

5. Висновки

У цій статті представлено новий непараметричний підхід до виявлення точок переходу у багатовимірних часових рядах, згенерованих кіберфізичними системами. Запропонований алгоритм не потребує припущень щодо типу розподілу, поєднує зменшення розмірності (через PCA) із використанням P -статистики, та забезпечує ефективне і точне виявлення змін стану системи.

Експериментальні дослідження із застосуванням таких трьох статистик: P -статистики, KS -статистики (Колмогорова–Смірнова) та W -статистики (Манна–Уїтні–Вілкоксона) — продемонстрували такі результати.

При перевірці гіпотези про зсув середнього значення (рівномірний розподіл $U(0; 1) \rightarrow U(1; 2)$), усі три критерії показали подібну ефективність, своєчасно виявляючи першу точку переходу. У цьому випадку P -статистика не поступалася альтернативним методам.

Однак при перевірці гіпотези про зміну масштабу (нормальний розподіл $N(0, 1) \rightarrow N(0, 2)$), переваги запропонованого підходу є очевидними. Лише P -статистика та KS -статистика демонстрували здатність до виявлення точки переходу. Причому KS -статистика виявилась надмірно чутливою до викидів, оскільки ґрунтується на оцінці супремуму різниці емпіричних функцій розподілу.

У свою чергу, P -статистика виявила точку переходу раніше, стабільніше та точніше, ніж усі інші. Це свідчить про її придатність для задач виявлення аномалій, пов'язаних із варіативністю або розсіюванням даних. W -статистика в умовах зміни масштабу взагалі не забезпечувала задовільного результату.

Таким чином, саме запропонована P -статистика виявляється найбільш придатною для задач детектування аномалій, пов'язаних із варіативністю чи дисперсією даних. Її використання, у поєднанні з техніками зменшення розмірності (зокрема, PCA), забезпечує нову якість у вирішенні задач виявлення точок переходу у складних динамічних системах і ідентифікації зміни режимів роботи складних систем.

Автор заявляє про відсутність конфлікту інтересів щодо публікації цієї статті.

ЛІТЕРАТУРА

1. Klyushin D.A., Petunin Y.I., Nonparametric population equivalence test based on measure of closeness between samples. *Ukrainian Mathematical Journal*. 2003. Vol. 55. P. 181–198.
2. Truong C., Oudre L., Vayatis N. Selective review of offline change point detection methods. *Signal Processing*. 2020. Vol. 167. doi:10.1016/j.sigpro.2011.2.1072912.
3. Alippi C., Boracchi G., Carrera D., Roveri M. Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss. Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2016. P. 1368–1374. doi:10.48550/arXiv.1510.04850.
4. Wang Z., Zwetsloot I.M. A Change-Point Based Control Chart for Detecting Sparse Changes in High-Dimensional Heteroscedastic Data. 2021. arXiv:2101.09424v1. doi:10.48550/arXiv.2101.09424.

5. Sorba O., Geissler C. Online Bayesian inference for multiple change points and risk assessment. 2021. *arXiv preprint arXiv:2106.05834v1*. doi:10.48550/arXiv.2106.05834.
6. Navarro M., Allen G.I., Weylandt M. Network Clustering for Latent State and Change point Detection. 2021. *arXiv preprint arXiv:2111.01273v1*. doi:10.48550/arXiv.2111.01273.
7. Tickle S.O., Eckley I.A., Fearnadh P. A computationally efficient, high-dimensional multiple change point procedure with application to global terrorism incidence. 2020. *arXiv:2011.03599v2*. doi:10.1111/rssc.12695.
8. Fearnadh P., Rigaiil G. Change point Detection in the Presence of Outliers. *Journal of the American Statistical Association*. 2018. Vol. 114. P. 169–183. doi:10.1080/016214512.2017.1385466.
9. Romano G., Eckley I., Fearnadh P., Rigaiil G. Fast Online Change point Detection via Functional Pruning CUSUM statistics. 2021. *arXiv:2110.08205v2*. doi:10.48550/arXiv.2110.08205.
10. Wang Z., Lin X., Mishra A., Sriharsha R. Online Change point Detection on a Budget. International Conference on Data Mining Workshops (ICDMW), 2022. P. 414–420. doi:10.1109/ICDMW53433.2021.00057.
11. Windmann A., Steude H., Niggemann O. Robustness and Generalization Performance of Deep Learning Models on Cyber-Physical Systems: A Comparative Study. 2023. doi:10.48550/arXiv.2306.07737.
12. Steude H., Windmann A., Niggemann O. Learning Physical Concepts in Cyber-Physical Systems: A Case Study. 2021. doi:10.48550/arXiv.2111.14151.
13. Aminikhanghahi S., Cook D.J. A survey of methods for time series change point detection. *Knowledge and Information Systems*. 2017. Vol. 51. P. 339–367.
14. Jaehyeok S., Ramdas A., Rinaldo A. E-detectors: a nonparametric framework for online change point detection. 2022. *arXiv:2203.03532v1*. doi:10.48550/arXiv.2203.03532.
15. Wendelberger L., Gray J., Reich B., Wilson A. Monitoring Deforestation Using Multivariate Bayesian Online Change point Detection with Outliers. 2021. *arXiv preprint arXiv:2112.12899v2*. doi:10.48550/arXiv.2112.12899.
16. Adams P., Mackay D. Bayesian Online Change point Detection. 2007. *arXiv preprint arXiv:0710.3742v1*. doi:10.48550/arXiv.0710.3742.
17. Cooney P., White A. Change-point Detection for Piecewise Exponential Models. 2021. *arXiv preprint arXiv:2112.03962v1*. doi:10.48550/arXiv.2112.03962.
18. Hallgren K.L., Heard N.A., Turcotte M.J.M. Change point detection on a graph of time series. 2021. *arXiv:2102.04112v1*. doi:10.48550/arXiv.2102.04112.
19. Fotoohinasab A., Hocking T., Afghah F. A Greedy Graph Search Algorithm Based on Change point Analysis for Automatic ORs Complex Detection. *Computers in Biology and Medicine*. 2021. Vol. 130. P. 104208.
20. Renz K., Stache N., Fox G., Varol G., Albanie S. Sign Segmentation with Change point-Modulated Pseudo-Labeling. 2021. *arXiv:2104.13817v1*. doi:10.48550/arXiv.2104.13817.
21. Gallagher C., Killick R., Lund R., Shi X. Autocovariance Estimation in the Presence of Change points. 2021. *arXiv preprint arXiv:2102.10609v2*. doi:10.48550/arXiv.2102.10609.
22. Madreimov I., Petunin Yu.I. Characterization of a uniform distribution using order statistics. *Teor Ver Mat Statist*. 1982. 27. P. 96–102.

-
23. Matveichuk S.A., Petunin Yu.I. Generalization of Bernoulli schemes that arise in order statistics, I, *Ukrainian. Math. J.* 1990. Vol. 42. No. 4. P. 459–466.
 24. Matveichuk S.A., Petunin Yu.I. Generalization of Bernoulli schemes that arise in order statistics. II, *Ukrainian Math. J.* 1991. Vol. 43. No. 6. P. 728–734.
 25. Hill B.M. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *J Am Stat Assoc.* 1968. Vol. 63. P. 677–691.
 26. van der Waerden B.L. *Mathematische Statistik.* Springer-Verlag, Berlin, 1957; English transl. of 2nd ed., Springer-Verlag, Berlin and New York, 1969.
 27. Petunin Y.I., Klyushin D.A., Ganina K.P., Borodai N.V., Andrushkiv R.I. Computer diagnosis of breast cancer. *Bulletin of Kyiv University, Ser. Cybernetics.* 2001. Vol. 2. P. 58–68.
 28. Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11): 559–572, 1901. <https://doi.org/10.1080/14786440109462720>.
 29. Jolliffe I.T. *Principal Component Analysis.* Springer Series in Statistics, Springer-Verlag, New York, 2002. <https://doi.org/10.1007/b98835>.

Надійшла: 15.01.2025 / Прийнята: 21.02.2025