

Київський національний університет імені Тараса Шевченка

Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

УРАЗОВСЬКИЙ АНДРІЙ ВЛАДИСЛАВОВИЧ

УДК 519.6

ДИСЕРТАЦІЯ

**НЕПАРАМЕТРИЧНІ МЕТОДИ ПОШУКУ ТОЧОК ПЕРЕХОДУ В
БАГАТОВИМІРНИХ ЧАСОВИХ РЯДАХ**

113 Прикладна математика

11 Математика та статистика

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ А.В. Уразовський

Науковий керівник: Ключин Дмитро Анатолійович,
доктор фізико-математичних наук, професор

Київ – 2025

АНОТАЦІЯ

Уразовський А.В. Непараметричні методи пошуку точок переходу в багатовимірних часових рядах. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 113 «Прикладна математика». – Київський національний університет імені Тараса Шевченка, Київ, 2025.

Розробка ефективних методів виявлення структурних змін у багатовимірних часових рядах є надзвичайно актуальною та міждисциплінарною задачею, що має безпосереднє значення для широкого спектра кіберфізичних систем. До таких систем належать як складні інженерні комплекси в енергетиці, зокрема атомній та відновлюваній, так і сучасні медичні діагностичні та моніторингові пристрої, транспортні мережі з елементами автономного управління, а також системи забезпечення кібербезпеки критичної інфраструктури. У кожній із цих галузей зміни у внутрішньому стані об'єкта або процесу часто відбуваються раптово, а їхні наслідки можуть мати масштабний і навіть незворотний характер.

Такі зміни найчастіше проявляються як різкі зсуви у статистичних характеристиках сигналу — середньому значенні, дисперсії, кореляційних залежностях — або як комплексні багатовимірні відхилення, що важко виявити простими методами аналізу. У свою чергу, своєчасне та надійне виявлення подібних точок переходу дозволяє оперативно реагувати на потенційно небезпечні ситуації, запобігати аваріям та збоєм у роботі, оптимізувати процеси управління та підвищувати загальну ефективність функціонування системи. Особливо це важливо для систем реального часу, де навіть незначна затримка у виявленні критичної зміни може призвести до значних матеріальних збитків або загроз безпеці людей.

Метою дисертаційної роботи є:

1. Розробка універсального алгоритму непараметричної статистичної детекції точок переходу у багатовимірних часових рядах, який є стійким до викидів та не залежить від апріорних припущень щодо розподілу даних.
2. Дослідження розробленого алгоритму на предмет консистентності та статистичної потужності при збільшенні обсягів і розмірності даних.
3. Проведення порівняльного аналізу ефективності запропонованого методу з відомими підходами на широкому класі розподілів, виявлення його переваг та обмежень.

Поставлена мета передбачає побудову математичної моделі, яка дозволяє одночасно зменшити обчислювальну складність розв'язання задачі та зберегти високу чутливість до змін у структурі даних, а також забезпечує можливість обґрунтованого порівняння з традиційними методами статистичного аналізу.

Об'єктом дослідження є багатовимірні часові ряди, що описують стан кіберфізичних систем у динаміці їх роботи. **Предметом** дослідження є алгоритмічні та статистичні методи обробки даних для виявлення моментів зміни у процесах з невизначеними або складними законами розподілу.

Методика дослідження поєднує апарат непараметричної статистики, методи зменшення розмірності (зокрема, метод головних компонент), теоретичний аналіз стійкості алгоритмів до викидів та аномалій, а також чисельні експерименти на синтетичних і квазіреальних наборах даних. Запропонований підхід передбачає початкове зниження розмірності до одного виміру за допомогою РСА, після чого здійснюється оцінка статистичної близькості підпоследовностей даних із використанням статистики Петуніна.

Запропонований у дисертаційній роботі алгоритм призначений для виявлення першої точки переходу у багатовимірних часових рядах і базується на комбінації методів зменшення розмірності та непараметричної статистики. Основна ідея полягає у використанні ковзного вікна фіксованої ширини, в межах якого проводиться поетапний аналіз структурної однорідності даних.

На кожному кроці вікно ділиться на дві підвибірки рівної довжини, які представляють попередній і поточний стан системи. Для кожної підвибірки здійснюється зменшення розмірності за допомогою методу головних компонент (Principal Component Analysis), що дозволяє проєктувати багатовимірні дані на головну вісь з максимальним розсіюванням. Це спрощує подальший статистичний аналіз, зменшуючи обчислювальну складність і підвищуючи чутливість алгоритму до зміни розподілу.

Після проєкції обох підвбірок на першу головну компоненту обчислюється значення p -статистики Петуніна, яке відображає ступінь подібності або відмінності між їхніми розподілами. Якщо p -значення перевищує встановлений поріг (у даному дослідженні — 0,95), вважається, що розподіли статистично не відрізняються; у протилежному випадку робиться висновок про наявність змін у структурі даних.

Процедура аналізу повторюється після зсуву вікна на один елемент, що забезпечує поступове сканування всього часового ряду. Момент, після якого вперше виявляється статистично значуща відмінність, інтерпретується як перша точка переходу. Такий підхід дозволяє локалізувати момент зміни з високою точністю та мінімальною затримкою відносно його реального настання.

Додатково у дослідженні передбачена можливість використання для порівняння класичних критеріїв однорідності, зокрема критерію Колмогорова–Смірнова та критерію Вілкоксона для незалежних вибірок. Це

дає змогу провести об'єктивну оцінку ефективності запропонованого методу на тлі вже відомих статистичних інструментів та визначити його переваги і можливі обмеження у різних умовах.

Результати проведених досліджень показали, що розроблений метод забезпечує високу точність локалізації моментів структурних змін (у межах 10–20 кроків після їх настання) та стійкість до випадкових аномальних сплесків, не потребуючи попередніх припущень про розподіл даних. Порівняння з критеріями Колмогорова–Смірнова та Вілкоксона підтвердило переваги запропонованого алгоритму як за точністю, так і за обчислювальною ефективністю, що робить його придатним для інтеграції у системи моніторингу в режимі реального часу.

Наукова новизна полягає у:

1. Розробці нового універсального непараметричного алгоритму виявлення точок переходу у багатовимірних часових рядах на основі комбінації РСА та статистики Петуніна.
2. Обґрунтуванні та експериментальному підтвердженні стійкості методу до викидів та незалежності від форми базового розподілу.
3. Запропонуванні критерію точності та стабільності для оцінки ефективності алгоритмів в умовах реального часу.
4. Проведенні комплексних обчислювальних експериментів на даних різної природи, включаючи моделювання сценаріїв для атомної енергетики та інших технічних систем.

Практичне значення дисертаційної роботи визначається широкими можливостями її застосування в автоматизованих системах моніторингу та управління, що працюють в умовах обмежених обчислювальних ресурсів і потребують оперативного реагування на критичні зміни у стані об'єкта. Отримані результати можуть бути адаптовані для потреб промислових,

енергетичних, медичних та транспортних комплексів, а також для задач кібербезпеки, де безпомилкове та швидке виявлення змін є запорукою надійності та безпеки.

Ключові слова: часові ряди, точка переходу, непараметрична статистика, кіберфізичні системи, зменшення розмірності.

ABSTRACT

Urasovskyi A.V. Nonparametric methods for finding changepoints in multidimensional time series. – Qualification scientific work in the form of a manuscript.

Dissertation for the degree of Doctor of Philosophy in Physics and Mathematics in specialty 113 “Applied Mathematics”. – Taras Shevchenko National University of Kyiv, Kyiv, 2025.

Development of effective methods for detecting structural changes in multidimensional time series is an extremely relevant and interdisciplinary task that is of direct importance for a wide range of cyber-physical systems. Such systems include both complex engineering complexes in the energy sector, in particular nuclear and renewable, and modern medical diagnostic and monitoring devices, transport networks with elements of autonomous control, as well as systems for ensuring cybersecurity of critical infrastructure. In each of these industries, changes in the internal state of an object or process often occur suddenly, and their consequences can be large-scale and even irreversible. Such changes most often manifest themselves as sharp shifts in the statistical characteristics of the signal - the average value, variance, correlation dependencies - or as complex multidimensional deviations, which are difficult to detect using simple analysis methods. In turn, timely and reliable detection of such changepoints allows you to quickly respond to potentially dangerous situations, prevent accidents and malfunctions, optimize management processes and increase the overall efficiency of the system. This is especially important for real-time systems, where even a slight delay in detecting a

critical change can lead to significant material damage or threats to human safety.

The purpose of the dissertation is:

1. Development of a universal algorithm for non-parametric statistical detection of changepoints in multidimensional time series, which is robust to outliers and does not depend on a priori assumptions about the data distribution.
2. Study of the developed algorithm for consistency and statistical power with increasing data volumes and dimensions.
3. Conducting a comparative analysis of the effectiveness of the proposed method with known approaches on a wide class of distributions, identifying its advantages and limitations.

The purpose involves constructing a mathematical model that allows simultaneously reducing the computational complexity of solving the problem and maintaining high sensitivity to changes in the data structure, and also provides the possibility of a reasonable comparison with traditional methods of statistical analysis.

The **object** of the study is multidimensional time series that describe the state of cyber-physical systems in the dynamics of their operation. The **subject** of the research is algorithmic and statistical methods of data processing for detecting moments of change in processes with uncertain or complex distribution laws.

The research **methodology** combines the apparatus of nonparametric statistics, dimensionality reduction methods (in particular, the principal component method), theoretical analysis of the stability of algorithms to outliers and anomalies, as well as numerical experiments on synthetic and quasi-real data sets. The proposed approach involves an initial reduction of the dimensionality to one dimension using PCA, after which the statistical proximity of data subsequences is assessed using Petunin statistics. The algorithm proposed in the dissertation is designed to detect the first changepoint in multidimensional time series and is based on a combination of dimensionality reduction methods and nonparametric statistics. The main idea is

to use a fixed-width sliding window, within which a step-by-step analysis of the structural homogeneity of the data is performed. At each step, the window is divided into two subsamples of equal length, representing the previous and current state of the system. For each subsample, dimensionality reduction is performed using the Principal Component Analysis method, which allows projecting multidimensional data onto the principal axis with maximum dispersion. This simplifies further statistical analysis, reducing computational complexity and increasing the sensitivity of the algorithm to changes in the distribution. After projecting both subsamples onto the first principal component, the Petunin p-statistic is calculated, which reflects the degree of similarity or difference between their distributions. If the p-value exceeds the set threshold (in this study, 0.95), the distributions are considered to be statistically indistinguishable; otherwise, it is concluded that there are changes in the data structure. The analysis procedure is repeated after shifting the window by one element, which provides a gradual scan of the entire time series. The moment after which a statistically significant difference is first detected is interpreted as the first changepoint. This approach allows to localize the moment of change with high accuracy and minimal delay relative to its actual occurrence. Additionally, the study provides for the possibility of using classical homogeneity criteria for comparison, in particular the Kolmogorov–Smirnov criterion and Wilcoxon's criterion for independent samples. This allows us to objectively assess the effectiveness of the proposed method against the background of already known statistical tools and to determine its advantages and possible limitations in different conditions. The results of the conducted studies showed that the developed method provides high accuracy of localization of moments of structural changes (within 10–20 steps after their occurrence) and resistance to random anomalous spikes, without requiring prior assumptions about the data distribution. Comparison with the Kolmogorov–Smirnov and Wilcoxon criteria confirmed the advantages of the proposed algorithm both in terms of accuracy and computational efficiency, which makes it suitable for integration into real-time monitoring systems.

The scientific **novelty** consists in: 1. Development of a new universal non-parametric algorithm for detecting changepoints in multidimensional time series based on a combination of PCA and Petunin statistics. 2. Substantiation and experimental confirmation of the method's robustness to outliers and independence from the form of the underlying distribution. 3. Proposing an accuracy and stability criterion for assessing the effectiveness of algorithms in real-time conditions. 4. Conducting complex computational experiments on data of various natures, including modeling scenarios for nuclear power and other technical systems. The practical significance of the dissertation work is determined by the wide possibilities of its application in automated monitoring and control systems operating in conditions of limited computing resources and requiring prompt response to critical changes in the state of the object. The results obtained can be adapted for the needs of industrial, energy, medical and transport complexes, as well as for cybersecurity tasks, where accurate and rapid detection of changes is the key to reliability and safety.

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

Список публікацій здобувача:

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Klyushin D., Urazovskyi A. Nonparametric Test for Change-Point Detection of IoT Time-Series Data // *Chapter in: Kumar P., Obaid A., Cengiz K., Balas A. (Eds.) A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems*, Intelligent Systems Reference Library, vol. 210, Springer, 2022, pp. 99-122.
2. Klyushin D., Urazovskyi A. Non-Parametric Methods of Finding Changepoints in Multidimensional Time Series // *In: Proceedings of the 3rd International Workshop on Computational & Information Technologies for*

- Risk-Informed Systems* (CITRisk, Neubiberg, Germany, January 12, 2023) Pickl S. et al (Eds.), pp. 106–118.
3. Klyushin D., Urazovskyi A. Locating changepoints in multidimensional time series using non-parametric methods // *In: Proceedings of the The Sixth International Workshop on Computer Modeling and Intelligent Systems*, (CMIS, Zaporizhzhia, Ukraine, May 3, 2023)
 4. Klyushin D., Urazovskyi A. Statistical Algorithm for Change Point Detection in Multivariate Time Series of Medicine Data Based on Principles of Explainable Artificial Intelligence, 2023, *Medical Data Analysis and Processing using Explainable Artificial Intelligence*, pp. 58-90.
 5. Klyushin D., Urazovskyi A. Locating Changepoints in Multidimensional Time Series Using Non-Parametric Methods // *In: Proceedings of the Modern Machine Learning Technologies and Data Science Workshop*, (MoML&T&DS 2023), Emmerich M., Vysotska V., Lytvynenko V. (Eds.) Lviv, Ukraine, June 3, 2023, pp. 12–24.
 6. Klyushin D., Urazovskyi A. Comparison of Dimensionality Reduction Methods for Finding Changepoints in Multidimensional Real-time Series // *In: Proceedings of the 4rd International Workshop on Computational & Information Technologies for Risk-Informed Systems* (CITRisk, Neubiberg, Germany, January 26, 2024) Pickl S. et al (Eds.).
 7. Klyushin D., Urazovskyi A. A non-parametric statistical technique for anomaly detection in CPS. *Cyber Physical System 2.0: Communication and Computational Technologies. Volume 2*, Routledge, 2025.
 8. Urazovskyi, A. A non-parametric statistical technique for changepoint detection in cyber-physical systems. *Journal of Numerical and Applied Mathematics*, 2025, pp. 101-122.

Зміст

Вступ.....	13
Розділ 1. Сучасний стан досліджень	22
1.1. Загальні підходи до задачі пошуку точок переходу.....	22
1.2. Новітні методи та їх застосування в різних предметних галузях.....	25
1.3. Теоретичні результати та оцінки меж виявлення.....	36
1.4. Узагальнення та висновки.....	39
Розділ 2. Теоретична частина.....	41
2.1. Статистика Петуніна.....	42
2.2. Доведення консистентності.....	43
2.3. Доведення потужності p -статистики Петуніна.....	48
2.4. Метод головних компонент (Principal Component Analysis).....	53
2.5. Статистика Колмогорова-Смірнова.....	54
2.6. Статистика Вілкоксона.....	55
2.7. Еліпсоїди Петуніна.....	56
2.8. Лінійний дискримінант Фішера.....	59
Розділ 3. Практична частина.....	62
3.1. Одновимірний випадок.....	63
3.2. Використання еліпсоїдів Петуніна.....	85
3.3. Дослідження поведінки p -статистики Петуніна у тривимірному випадку.....	102
3.4. Порівняння методів FLD і PCA.....	110
3.5. Порівняння статистик у багатовимірних часових рядах.....	120
3.6. Симуляційні експерименти на квазіреальних даних.....	131
Висновки дисертації.....	149
Список використаних джерел.....	151

Список умовних позначень:

Розподіли: $N(\mu, \sigma^2)$ – нормальний розподіл з математичним сподіванням μ і дисперсією σ^2 ; $LogN(\mu, \sigma^2)$ – логнормальний розподіл з логарифмічним масштабом μ і формою σ^2 ; $Exp(\lambda)$ – експоненціальний розподіл з інтенсивністю λ ; $Gamma(k, \theta)$ або $\Gamma(k, \theta)$ – гамма розподіл з параметрами k і θ ; $U(a, b)$ – неперервний рівномірний розподіл на відрізку $[a, b]$; $G(\mu, \beta)$ – розподіл Гумбеля з параметром розміщення μ і параметром масштабу β ; $R(\sigma)$ – розподіл Релея з масштабом σ ; $L(\mu, \beta)$ – розподіл Лапласа з параметром розміщення μ і параметром масштабу β ; $Logis(\mu, \beta)$ – логістичний розподіл з параметром розміщення μ і параметром масштабу β .

Медичні показники: HR – частота серцевих скорочень; SpO2 – насичення крові киснем; T – температура.

$[x]$ – ціла частина числа.

ВСТУП

Актуальність теми

Сучасні кіберфізичні системи — від промислових виробничих комплексів і енергетичних установок до медичних моніторингових пристроїв, автономних транспортних засобів та інтернету речей — генерують великі обсяги багатовимірних часових рядів, що описують стан системи в реальному часі. У цих системах виникнення структурних змін у сигналах (так званих точок переходу або change-points) може бути наслідком різних явищ: технічних відмов, атак на керуючі підсистеми, фізичних аварій, зовнішніх впливів або різких змін у навантаженні. Своєчасне та надійне виявлення таких подій має критичне практичне значення — воно дозволяє запобігти аваріям, зменшити матеріальні збитки, врятувати життя і підвищити стійкість та надійність систем управління.

Разом з тим, класичні підходи до детекції змін часто виявляються недостатніми у сучасних умовах через кілька взаємопов'язаних проблем. По-перше, багатовимірність даних і «прокляття розмірності» призводять до різкого зниження статистичної потужності традиційних критеріїв і до зростання обчислювальних витрат при обробці потоків даних у реальному часі. По-друге, у багатьох прикладних задачах відсутні або ненадійні апріорні відомості про розподіл спостережень, отже параметричні методи або методи, що вимагають навчання на репрезентативних масивах, можуть бути непридатними або небезпечними для застосування. По-третє, реальні потоки часто містять поодинокі викиди та аномалії, що здатні «збити» алгоритм або спричинити хвилю хибних сповіщень; це вимагає від методів підвищеної робастності. І, нарешті, для багатьох операційних середовищ (медичні монітори, системи безпеки АЕС, керування БПЛА) потрібні онлайн-алгоритми з низькою обчислювальною вартістю, які можуть працювати з обмеженими ресурсами та з мінімальною затримкою сповіщення.

Через перелічені фактори сучасні дослідження рухаються в бік комбінованих рішень: поєднання зменшення розмірності (щоб пом'якшити «прокляття») із

непараметричними критеріями (щоб знизити залежність від припущень щодо розподілів) та з усуненням чутливості до викидів. Саме така комбінація є ядром запропонованого в роботі підходу: проектування ефективного потоку обробки, у якому багатовимірні вектори проєктуються на одну координату (за допомогою методу головних компонент або лінійного дискримінанта Фішера), а потім для одновимірних проєкцій застосовується непараметрична р-статистика Петуніна для перевірки однорідності підвбірок. Такий підхід дозволяє поєднати низьку обчислювальну вартість (за рахунок роботи з одновимірними проєкціями) зі стійкістю до невідомих розподілів і до випадкових викидів.

Крім очевидних практичних потреб, тема має значну наукову вагу. Питання побудови універсальних та обчислювально ефективних статистичних критеріїв для потоків багатовимірних даних містить низку фундаментальних математичних проблем: коректна формалізація двовибіркового тесту в умовах залежностей і високої розмірності, оцінка асимптотичної поведінки непараметричних статистик у потоковому режимі, обґрунтування стійкості до викидів та методів адаптивного вибору вікна. Розв'язання цих задач здатне не лише поліпшити практичні системи моніторингу, а й розширити теоретичні знання у галузі непараметричної статистики, теорії детекції аномалій та багатовимірної обробки сигналів.

На тлі сучасних тенденцій — широкого впровадження мереж сенсорів, розвитку інтернету речей, поширення автономних систем і зростання загроз кібербезпеки — підвищення точності та надійності алгоритмів детекції точок переходу залишається однією з ключових задач. Розробка підходів, здатних поєднати розпізнавання змін у реальному часі, незалежність від класу розподілів, робастність до викидів і прийнятні обчислювальні витрати, має важливе прикладне та наукове значення й відкриває поле для подальших досліджень і впроваджень у критичних інфраструктурах та прикладних системах.

Мета і завдання дослідження

Метою дисертаційної роботи є розробка універсального, обчислювально ефективного та робастного до викидів алгоритму непараметричної статистичної детекції точок переходу у багатовимірних часових рядах, який не залежить від апіорних припущень щодо розподілу даних та забезпечує високу точність виявлення змін у структурі сигналів у режимі, наближеному до реального часу.

Для досягнення поставленої мети у роботі сформульовано та вирішено такі основні завдання:

Аналіз сучасних методів детекції точок переходу у часових рядах, з особливим акцентом на багатовимірні непараметричні підходи, їх переваги та обмеження у контексті реальних кіберфізичних систем.

Розробка математичної моделі процесу виявлення точок переходу у багатовимірних часових рядах, яка враховує високу розмірність, можливу присутність викидів, невідомий розподіл даних та вимоги до обчислювальної ефективності.

Створення алгоритму, що поєднує методи зменшення розмірності (РСА або FLD) з непараметричними критеріями (p-статистикою Петуніна) для забезпечення стійкості до викидів і незалежності від типу розподілу.

Дослідження властивостей алгоритму — зокрема його консистентності, статистичної потужності та чутливості до різних типів змін у структурі даних при варіації розмірності та обсягу вибірок.

Проведення порівняльного аналізу запропонованого підходу з класичними методами (Колмогорова–Смірнова, Вілкоксона та іншими) на широкому спектрі розподілів і прикладних сценаріїв, визначення його переваг та обмежень.

Верифікація алгоритму на модельних та прикладних даних з галузей енергетики, медицини, інженерії та кібербезпеки з оцінкою часу реакції та точності виявлення.

Методи дослідження

У процесі виконання дисертаційної роботи використано комплекс взаємопов'язаних теоретичних, математичних, алгоритмічних та експериментальних методів, спрямованих на досягнення поставленої мети та вирішення визначених завдань.

Теоретичну основу дослідження становлять положення математичної статистики, непараметричних методів перевірки гіпотез, теорії випадкових процесів і часових рядів, а також методи багатовимірного аналізу даних. Особливу увагу приділено непараметричним двовибірковим критеріям, зокрема p -статистиці Петуніна, які не потребують апріорних знань про вид розподілу випадкових величин та забезпечують роботу в умовах присутності викидів.

Алгоритмічний апарат дослідження включає методи зменшення розмірності даних, такі як метод головних компонент (Principal Component Analysis, PCA) та лінійний дискримінант Фішера (Fisher's Linear Discriminant, FLD), що дають змогу проєктувати багатовимірні часові ряди на одну координату з максимальною збереженою інформативністю. Це дозволяє зменшити вплив «прокляття розмірності» та підвищити обчислювальну ефективність подальшої статистичної обробки.

Моделювання та чисельні експерименти проводилися для різних типів змін у структурі даних — зсуву середнього значення, зміни дисперсії, комплексних змін розподілу — на широкому класі ймовірнісних моделей (нормальний, рівномірний, експоненціальний, гамма-, Гумбелевий, Релєсвий розподіли тощо). При цьому враховувалась присутність поодиноких викидів та зміна розмірності вхідних даних.

Програмна реалізація алгоритмів виконана мовою Python із використанням бібліотек NumPy, SciPy, scikit-learn та інших пакетів для статистичного аналізу та обробки сигналів. Реалізовані програмні модулі дозволяють виконувати покрокову обробку даних у режимі ковзного вікна, проводити порівняльний аналіз різних методів і будувати графічні інтерпретації результатів.

Методи оцінювання ефективності включали обчислення показників затримки виявлення, відсотка хибних спрацювань, статистичної потужності тестів та середньої точності локалізації точки переходу. Для порівняння використовувалися результати роботи класичних критеріїв Колмогорова–Смірнова та Вілкоксона для незалежних вибірок.

Верифікація результатів здійснювалась як на синтетичних даних, так і на прикладних наборах з галузей енергетики, медицини, інженерії та кібербезпеки, що дозволило підтвердити універсальність та практичну цінність запропонованого алгоритму.

Наукова новизна отриманих результатів

У дисертаційній роботі отримано такі наукові результати, що характеризуються новизною та виносяться на захист:

1. Запропоновано новий універсальний алгоритм виявлення точок переходу у багатовимірних часових рядах, який поєднує методи зменшення розмірності з непараметричним критерієм p -статистики Петуніна, що дозволяє знизити обчислювальну складність та підвищити стійкість до викидів і невідомих розподілів даних.
2. Розроблено модифіковану процедуру ковзного вікна для багатовимірних потоків даних, яка забезпечує поетапне тестування структурної однорідності підвибірок та дозволяє локалізувати момент зміни з високою точністю у режимі, наближеному до реального часу.
3. Вперше проведено систематичний порівняльний аналіз роботи p -статистики Петуніна у поєднанні з методами зменшення розмірності на широкому класі розподілів (нормальний, рівномірний, експоненціальний, гамма-, Гумбелевий, Релеєвий тощо) та у різних сценаріях зміни структури даних.
4. Отримано нові оцінки затримки виявлення та статистичної потужності алгоритму для різних умов — зміни середнього, дисперсії, комплексних змін розподілу — і показано, що у багатьох випадках затримка не перевищує 10–20 кроків навіть при високій розмірності.

5. Розроблено програмну реалізацію запропонованого алгоритму з можливістю інтеграції у системи моніторингу кіберфізичних об'єктів, яка забезпечує онлайн-аналіз даних та візуалізацію результатів для користувача.
6. Показано універсальність та практичну цінність методу шляхом його верифікації на прикладних даних із галузей енергетики, медицини, інженерії та кібербезпеки, що підтверджує придатність алгоритму до впровадження у різних сферах.

Особистий внесок здобувача

Науковому керівнику належить постановка загальної задачі, визначення напрямку досліджень та рекомендації щодо вибору методів розв'язання проміжних підзадач.

Претендентом опубліковано 8 наукових робіт [1]–[8]: 4 наукові статті та 4 тези конференцій (у складі збірників праць міжнародних наукових заходів). Творчий внесок у опубліковані наукові роботи є наступним:

- Статті [1]–[7] виконано у співавторстві з науковим керівником. Науковому керівнику належать постановка задачі, вибір загальної концепції алгоритмічного підходу та формулювання окремих математичних тверджень. Автором самостійно реалізовано програмні модулі алгоритму, проведено серію чисельних експериментів, виконано аналіз результатів та оформлено текстові матеріали.
- Стаття [8] написана без співавторів. Постановка задачі та загальні рекомендації щодо методів розв'язання отримані від наукового керівника – Д.А. Ключина. Робота присвячена формулюванню математичної моделі непараметричної детекції точок переходу у багатовимірних часових рядах, розробці алгоритму на основі р-статистики Петуніна з використанням методів зменшення розмірності та

експериментальній перевірці його властивостей на синтетичних і прикладних даних.

Зокрема:

- стаття [1] описує порівняння з класичними критеріями в одновимірному випадку;
- стаття [2] містить результати дослідження в тривимірному випадку;
- стаття [3] містить результати застосування запропонованого методу до медичних даних у контексті пояснюваного штучного інтелекту.
- стаття [4] містить результати дослідження з використанням еліпсів Петуніна;
- стаття [5] описує модифікацію алгоритму з урахуванням різних сценаріїв структурних змін та викидів у даних;
- статті [6] – [8] присвячено розробці та тестуванню непараметричного підходу для багатовимірних часових рядів з використанням методу головних компонент та р-статистики Петуніна, а також порівняння з класичними критеріями;
- Постановка задач та ідеї досліджень, викладені у тезах конференцій, сформульовано за підтримки наукового керівника; виступи на усіх конференціях автор здійснив самостійно.

Апробація наукових результатів відбулась на чотирьох конференціях:

1. CITRisk'2022: 3rd International Workshop on Computational & Information Technologies for Risk-Informed Systems, December 21-22, 2022, Munich, Germany.
2. CMIS-2023: The Sixth International Workshop on Computer Modeling and Intelligent Systems , May 3, 2023, Zaporizhzhia, Ukraine

3. MoMLLeT&DS 2023: 5th International Workshop on Modern Machine Learning Technologies and Data Science, June 3, 2023, Lviv
4. CITRisk'2023: 4th International Workshop on Computational & Information Technologies for Risk-Informed Systems, January 26, 2024, Munich, Germany

Структура та обсяг дисертації

Дисертаційна робота викладена на 160 сторінках та складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків.

У першому розділі «Сучасний стан досліджень» проведено аналіз існуючих підходів до виявлення точок переходу у часових рядах, розглянуто класичні параметричні та непараметричні методи, особливості їх застосування у багатовимірних задачах та в умовах наявності викидів. Окрему увагу приділено практичним обмеженням існуючих алгоритмів та актуальним науковим викликам.

Другий розділ «Теоретична частина» присвячено формулюванню математичної постановки задачі непараметричної детекції точок переходу у багатовимірних часових рядах, обґрунтуванню вибору статистичних методів, розробці алгоритмічного підходу на основі p -статистики Петуніна та методів зменшення розмірності, а також доведенню властивостей розробленого алгоритму.

Третій розділ «Практична частина» містить опис програмної реалізації запропонованого методу, експериментальні результати на синтетичних та прикладних даних з різних галузей (енергетика, медицина, інженерія, кібербезпека), а також порівняльний аналіз із класичними алгоритмами за критеріями точності, затримки виявлення та стійкості до викидів.

У висновках наведено основні результати роботи, сформульовано практичні рекомендації та окреслено перспективи подальших досліджень у даному напрямі.

Зв'язок роботи з науковими програмами

Результати, отримані у дисертаційній роботі, мають відношення до проекту "Обчислювальні алгоритми і оптимізація для штучного інтелекту, медицини та оборони" (ДР 0122U002026), підтриманого Міністерством освіти та науки України.

Тематика дисертаційного дослідження узгоджується з метою та завданнями вказаної наукової програми і спрямована на розвиток методів аналізу та обробки багатовимірних часових рядів для підвищення ефективності моніторингу стану кіберфізичних систем. Отримані в роботі результати становлять частину комплексних наукових досліджень кафедри та використовуються у наукових і прикладних проєктах, що виконуються за вказаною темою.

Розділ 1 «Сучасний стан досліджень»

Даний розділ містить систематизований огляд сучасних досліджень, присвячених проблемі виявлення структурних змін у багатовимірних часових рядах, що генеруються кіберфізичними системами. Розглянуто математичні моделі, статистичні та алгоритмічні підходи, а також практичні реалізації як у класичних параметричних, так і в непараметричних постановках. Особливу увагу приділено методам, стійким до викидів та здатним працювати в умовах високої розмірності даних і обмежених обчислювальних ресурсів. Проведено аналіз сильних і слабких сторін існуючих рішень, визначено основні виклики та прогалини, що залишаються відкритими у даній галузі.

1.1. Загальні підходи до задачі пошуку точок переходу

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 [9], 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 [10], де досліджували логарифмічну правдоподібність та розходження Кульбека–Лейблера для багатовимірних потоків. Проблему

частково пом'якшують контроль-карти нового покоління, здатні фіксувати розріджені (sparse) зсуви середнього у високих вимірностях, що показали у [11].

Поряд із параметричними методами активно розвиваються непараметричні детектори. Прикладом є E-detectors — універсальна рамка онлайн-тестів, яка мінімізує припущення про розподіли й показує конкурентну швидкодію на симульованих потоках. Близьку за ідеологією, Байєсівську стратегію пропонують у [12]; 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 [13]. 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 [14] запропонували швидко 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 [15], продемонструвавши, що коректне обмеження апостеріорної ймовірності значно знижує кількість хибних спрацьовувань. Додатково у [16] 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 [17] 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. [18]. Аналіз існуючих методів пошуку точок переходу та аномалій не показує універсальних тестів для вирішення цих проблем [19]. Розділ 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 [20]. 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 [21], 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 [22]. У [23] представлено стратегію дослідження географічних даних. Автори [24] врахували алгоритми, які були запропоновані лише для експоненціальних моделей. Дані про тип розподілу були використані в [25] для зменшення обчислювальної складності. Подібний недолік можна знайти в статті [26].

Знаходження точок зміни моделі вимагає апріорних припущень щодо даних. Часто потрібна попередня обробка даних, щоб точно визначити точки зміни в багатовимірному часовому ряді [15]. 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 [27]. Через значні зрушення середніх значень оцінка точки зміни за допомогою оцінки моментів Юла-Вокера [28] 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.

1.2. Новітні методи та їх застосування в різних предметних галузях

Автори роботи [29] розробили нейромережевий підхід до виявлення точок зміни у новинних потоках, адаптувавши схему learning-by-confusion. Ідея полягає у тренуванні класифікаторів для розрізнення статей з різних часових відрізків, а точність класифікації використовується для оцінки статистичної відстані повної варіації між розподілами контенту. Значні піки цієї метрики трактуються як точки зміни. Перевагою підходу є здатність фіксувати не лише зміну тематики, а й семантичні зсуви тону та акцентів. Серед обмежень — ігнорування кореляцій між матеріалами в межах сегментів та складність валідації на реальних даних через відсутність точної розмітки. Автори вбачають потенціал застосування методу й для аудіо-, відео- та зображень.

В статті [30] розробили новий підхід до онлайн-навчання прихованих станів роботів, що працюють у складних та невпорядкованих середовищах.

Запропонована методика ґрунтується на розширеній моделі Generalized Hidden Parameter Markov Decision Process та дозволяє в реальному часі формувати виразне й інтерпретоване подання латентних факторів, які впливають як на динаміку переходів між станами, так і на функцію винагороди. Отримані результати свідчать, що поєднання структурованих латентних подань з онлайн-адаптацією відкриває перспективи для створення більш стійких, контекстно-обізнаних роботів, здатних до надійної роботи в умовах невизначеності та змін.

У роботі [31] запропонували розширену версію Bayesian Model Mixing (BMM) для побудови рівняння стану (EOS) матерії нейтронних зірок, яка поєднує обмеження з різних густинних діапазонів у єдину й узгоджену модель. Це дозволило локально зменшити невизначеність EOS та скорегувати її форму відповідно до нових фізичних обмежень. Методологія BMM з changepoint detection є гнучкою, розширюваною на багатовимірні параметри (густина та температура), та здатна інтегрувати нові дані в режимі «онлайн». Це робить її перспективним інструментом для майбутніх симуляцій злиттів нейтронних зірок і багатоканальних астрофізичних досліджень.

Автори [32] запропонували новий підхід до вибору параметра штрафу в алгоритмах Optimal Partitioning для задач changepoint detection. Результати показали, що рекурентні мережі загалом перевершують традиційні моделі (лінійну регресію з L1 та MLP) у задачах changepoint detection, особливо на великих наборах даних. Найкращу продуктивність продемонстрував GRU. Перевага підходу полягає у відсутності потреби в експертному налаштуванні ознак, що робить метод придатним для різних типів послідовностей.

У статті [33] запропонували метод WATCH (Weighted Adaptive Testing for Changepoint Hypotheses) для адаптивного моніторингу роботи ШІ/ML-систем у реальному часі. Ключовою новацією є використання зважених конформних мартингалів (WCTM), які узагальнюють стандартні конформні тестові

мартингали та дозволяють перевіряти ширший клас гіпотез, ніж лише ті, що базуються на припущенні незалежності та однакового розподілу даних.

В роботі [34] запропонували алгоритм CONCH (CONformal CHangepoint localization) для точної локалізації єдиної точки переходу в послідовності даних за мінімальних статистичних припущень. Результати експериментів на синтетичних і реальних наборах показали, що CONCH формує вузькі довірчі інтервали та зберігає гарантоване покриття у скінченних вибірках. Це робить алгоритм універсальним інструментом для офлайн-детекції та локалізації точок переходу в умовах відсутності жорстких статистичних припущень.

Далі у [35] представили робастний та адаптивний підхід для виявлення точок переходу у високовимірних даних, зокрема при наявності важкохвостих розподілів. Експерименти показали, що метод зберігає високу потужність як при важкохвостих розподілах, так і при нормальних даних, перевершуючи існуючі підходи у задачах детекції структурних змін у високовимірних середовищах.

Автор [36] запропонував універсальну методологію онлайн-детекції точок переходу, яка дозволяє використовувати офлайн-тести в реальному часі при надходженні даних послідовно. Симуляційні експерименти показали перевагу методу над сучасними аналогами в задачі виявлення зміни середнього.

У статті [37] представили метод аугментації нестационарних часових рядів, який поєднує офлайн пошук точки переходу з модифікованим iAAFT. Для ЕЕГ зміни виявляються на основі набору статистичних ознак, після чого кожен сегмент окремо синтезується та об'єднується. Для ЕКГ зберігаються піки R, що забезпечує морфологічну достовірність сигналу. Порівняння з традиційними методами показало покращення точності та зменшення хибних спрацювань, особливо у виявленні епілептичних нападів.

Автори [38] запропонували метод Bottom-up Iterative Anomalous Diffusion Detector (BI-ADD) для виявлення точок переходу у коротких молекулярних

траєкторіях, що описуються дробовим броунівським рухом. Метод ефективно працює, коли зміни у дифузійних параметрах значні, і показав високу точність на змаганні AnDi2 у завданні відеотраєкторій (1-е місце).

В роботі [39] розробили модель Multimodal-ChangePoint Detection (MCPD) для раннього виявлення учнів середньої школи, які перебувають у групі ризику. Експерименти на даних з 1248 учнів (Гонконг, 2019–2022) показали точність понад 75% і гнучкість при зміні визначення «учня групи ризику».

У статті [40] запропонували універсальний підхід до онлайн-моніторингу якості прогнозних моделей шляхом послідовного виявлення точок переходу у похибках прогнозу. Ідея полягає в тому, що зміни у вихідному процесі (зсув середнього, зміна дисперсії тощо), які призводять до погіршення прогнозів, швидше проявляються саме в похибках, ніж у вихідних даних, особливо для складних моделей. Експерименти на симульованих і реальних даних підтвердили ефективність методу у своєчасному виявленні структурних змін, що дозволяє швидше адаптувати прогнози та управлінські рішення.

Автори [41] розробили загальний метод постдетекційного локалізування точки переходу у послідовному аналізі. Симуляції підтвердили, що методи забезпечують гарантоване (неасимптотичне) покриття та працюють у ширшому спектрі задач, ніж класичні підходи на основі CUSUM.

У статті [42] пропонують новий статистично обґрунтований підхід до аналізу «C-dates-as-data» для виявлення змін у частоті подій у часі, зокрема шляхом ідентифікації changepoints. На відміну від поширених SPDs, підхід не вимагає жорстко заданих форм узагальнення, забезпечує коректні й гнучкі оцінки та надає інструменти для інтерпретації причин демографічних і екологічних зрушень.

Автори [43] представили алгоритм SPACETIME для одночасного виявлення причинно-наслідкових зв'язків, точок переходу та групування часових

інтервалів і наборів даних за спільними причинними механізмами. SPACETIME успішно працює на синтетичних і реальних даних, зокрема для аналізу річкового стоку та біосферно-атмосферних взаємодій, виявляючи зміни у структурі причинних мереж та періоди з подібними механізмами.

В роботі [44] запропонували ART — універсальний, дистрибутивно-незалежний та модельно-агностичний метод виявлення точок переходу з гарантіями для скінченних вибірок. ART підтримує багатомасштабний аналіз, оцінку кількох точок переходу і постдетекційний аналіз із контролем похибок, що робить його гнучким і надійним інструментом, особливо для високовимірних і машинно-навчених даних.

У статті [45] дослідили вплив частоти кадрів та сегментаційних алгоритмів на аналіз траєкторій мікрочастинок, що рухаються молекулярними моторами з чергуванням руху та пауз. Робота підкреслює, що результати changepoint-аналізу істотно залежать від частоти кадрів і параметрів сегментації, виявляючи компроміс між зміщенням і дисперсією оцінок, та закликає до пошуку статистично оптимальних режимів спостереження.

Автори [46] досліджують tipping points у музичних творах — моменти різкої зміни, що сприймаються слухачами як ключові поворотні точки. На прикладі записів мазурок Шопена вони порівнюють суб'єктивно визначені tipping points із точками переходу у трьох параметрах тональної напруги: дисонансі, віддаленості від тональності та гармонічному русі, а також із відхиленнями у темпі. Виявлено, що майже всі популярні tipping points збігалися зі статистично значущими точками переходу хоча б в одному з параметрів, що підкреслює роль changepoint-аналізу у виявленні структурних зламів у музичному виконанні.

В роботі [47] пропонують BONE — єдиний ймовірнісний фреймворк для онлайн-навчання в нестабільних середовищах, який прямо враховує точки

переходу. Така модульна структура дозволяє як відтворювати відомі методи, так і створювати нові алгоритми, здатні адаптуватися до як різких, так і поступових змін у даних, що робить BONE гнучким інструментом для задач прогнозування та виявлення структурних зрушень у потокових даних.

У статті [48] вказують, що у складних моделях стандартне мінімізування in-sample втрат для пошуку точок переходу часто призводить до зміщень і хибних оцінок через оверфітинг. Вони пропонують cross-fitting — використання незалежних вибірок для оцінки втрат out-of-sample, що забезпечує узгоджене визначення точок переходу, орієнтоване на реальну прогностичну точність у майже однорідних сегментах.

Автори [49] пропонують метод AJDN (Asynchronous Jump Detection under Nonstationary noise) для виявлення стрибків у високовимірних часових рядах, де порушуються типові припущення про стаціонарність шуму та синхронність змін між вимірами. На відміну від класичних підходів, AJDN може ідентифікувати точки переходу, що виникають у різний час у різних вимірах, та працює за наявності складної часової залежності й несталих автокореляцій.

В роботі [50] автор представив пакет changepointGA для R, що використовує генетичний алгоритм для швидкого пошуку точок переходу без перебору всіх конфігурацій. Він одночасно визначає порядок моделі та положення точок, підтримує власні функції пристосованості й оператори, працює паралельно та підходить для застосувань у науці й промисловості.

У статті [51] запропонували метод Lean Bonferroni Changepoint Detection для швидкого виявлення та локалізації точок зміни з гарантіями на скінчених вибірках. Метод базується на локальних тестах для «триплетів» моментів часу та їх комбінації через зважену поправку Бонферроні, що дозволяє досягати оптимального порогу виявлення без налаштувальних параметрів. LBD працює для нормальних, експоненціальних та непараметричних моделей, забезпечує

довірчі інтервали для позицій точок зміни та нижні межі їх кількості. Обчислювальна складність майже лінійна, що робить метод придатним для великих даних.

Автори [52] представили модель *dynamic multiplex random dot product graph* (DMRDPG) для багат шарових динамічних графів та метод оцінювання *doubly unfolded adjacency spectral embedding* (DUASE). Метод можна застосовувати для кластеризації, виявлення трендів і глобальних точок зміни, зокрема у поєднанні з алгоритмом *iso-mirror*, як показано на прикладах геополітичних та фінансових мереж.

Автор [53] проаналізував часову складність *binary segmentation* для пошуку точок переходу. Запропоновано методи обчислення меж кількості кандидатів на розбиття та показано, що на реальних даних алгоритм працює майже з оптимальною швидкістю. Недолік — квадратична складність підрахунку нижньої межі, планується її зменшення.

У роботі [54] запропонували ефективний *log-linear* алгоритм підбору оптимального кроку градієнтного спуску при мінімізації AUM і максимізації AUC. Метод швидший за *grid search*, зберігаючи таку ж точність у задачах класифікації та пошуку точок переходу, і формує повне подання AUM/AUC для вибору кроку на кожній ітерації.

В статті [55] ввели модель *piecewise stationary linear bandit* (PSLB) та алгоритм *PS ϵ BAI+*, який поєднує зміну виявлення (LCD) і вирівнювання контекстів (LCA) для швидкої ідентифікації ϵ -оптимального важеля з імовірністю $\geq 1-\delta$. Метод досягає майже мінімаксної вибіркової складності (оптимальної з точністю до логарифмічного множника) та показує значне зниження числа вибірок порівняно з базовими алгоритмами, зберігаючи конкурентність навіть без знання контекстів і точок зміни.

Автори [56] запропонували метод changepoint-аналізу біомедичних часових рядів на основі узагальнених змішаних моделей, що враховують їхню структуру. Метод легко розширюється для різних структур коваріацій, але потребує налаштування параметрів, що може обмежувати виявлення змін у високовимірних даних.

В статті [57] розробили ентропійний метод відбору ознак, що поєднує cross-fuzzy ентропію та changepoint-аналіз для виявлення впливів зовнішніх факторів. На моделі виверження Пінатубо метод зафіксував охолодження півночі та простежив причинний зв'язок від аерозолів у тропіках до зміни сонячного потоку.

Автори [58] запропонували TUNE — алгоритм-незалежний метод післядетекційного статистичного аналізу в changepoint-задачах. Він встановлює універсальний поріг для тестових статистик, напряду контролюючи FWER без селективних р-значень, працює з різними моделями та алгоритмами, зберігаючи надійність і гнучкість навіть у високовимірних даних.

У роботі [59] розробили байєсівський метод просторово-змінного виявлення та оцінки точок переходу для просторово-часових даних, щоб відстежити регіональний вплив виверження вулкана Пінатубо. Застосування до даних AOD і температури показало просторові патерни впливу та його прогресію після виверження.

В статті [60] вперше дослідили виявлення точок переходу для середнього напряду біваріатних кутових даних (тороїдальних і сферичних). Використовуючи геометрію тору, вони ввели "криву матрицю дисперсії" та побудували два непараметричні тести на основі аналога відстані Махаланобіса. Методи застосовано для виявлення змін у напрямках вітру-хвиль та траєкторії циклону Virogojo.

Автор [61] запропонував метод виявлення точок переходу у психометричних дослідженнях, що базується на статистиці псевдо-Score з відомими розподілами під нульовою та альтернативною гіпотезами. Підхід простий у розрахунках, не потребує ресемплінгу, перевершує стандартні TFSP для нормальних і бінарних даних та дозволяє виконувати аналіз потужності й планування вибірки. Метод реалізовано в пакеті R *segmented* і Shiny-додатку.

В роботі [62] запропонували підхід до підвищення точності виявлення точок переходу у динамічному програмуванні через прогнозування штрафного параметра λ за допомогою багат шарового перцептрона з ReLU. На тестах із геномними даними метод показав вищу точність і F1-метрику, хоча потребує ручного відбору ознак, що обмежує його масштабованість.

Автори [63] запропонували state-space модель з випадковими партиціями, що еволюціонують у часі, для виявлення точок переходу як змін у кластерній структурі багатовимірних часових рядів. Реалізація через Gibbs sampling показала ефективність на симуляціях і даних жестів людини, точно виявляючи моменти активності та зміни кластерів.

У статті [64] представили *tidychangepoint* — R-пакет, що уніфікує аналіз виявлення changepoints в одноваріантних часових рядах. Підтримує як детерміновані методи (PELT), так і випадкові генетичні алгоритми, спрощуючи крос-аналіз підходів.

В роботі [65] запропонували метод виявлення змін у спільнотній структурі високорозмірних атрибутованих динамічних графів, використовуючи GNN для оцінки модульності з урахуванням як топології, так і багатовимірних атрибутів вузлів. Значні зсуви модульності сигналізують про точку переходу, що підтверджено симуляціями та аналізом мережі Twitter з текстовими ознаками вузлів, де метод виявив події реального світу.

Автори [66] запропонували онлайн-байєсівський метод виявлення змін у прихованій спільнотній структурі блок-гомогенних мережевих процесів Пуассона, використовуючи фактор забування у варіаційному наближенні для послідовного оновлення постеріору. Підхід масштабується на великі мережі, виявляє зміни у швидкостях ребер та належності вузлів у реальному часі й продемонстрований на симуляціях та даних мережі велопрокату Santander Cycles.

У статті [67] застосували багатовимірний аналіз змін до статистики MLB з кінця 1800-х до 2021 року, виявляючи зміни не лише у середньому, а й у дисперсії показників. Метод дозволив емпірично відтворити історичні бейсбольні ери та визначити періоди зародження й занепаду командних «династій», демонструючи користь багатовимірного підходу для спортивної аналітики.

В роботі [68] розробили байєсівську модель часового бікластерування, що виявляє точки переходу у вимірюваннях, які змінюються в часі, всередині сталих груп суб'єктів.

Автори [69] запропонували метод пост-селекційного висновування для точок переходу у дисперсії, що дає коректні р-значення після виявлення змін. Ключ у тому, щоб обчислювати нульовий розподіл тестової статистики з урахуванням факту виявлення точки переходу.

В статті [70] запропонували онлайн-метод CPDMD для виявлення точок переходу в потокових даних із трендами, періодичністю та сезонністю. Метод непараметричний, не потребує попереднього навчання й показує високу точність для змін середнього, дисперсії, періодичності та структури на синтетичних і реальних наборах даних.

У роботі [71] ввели поняття first-order changepoint для часових рядів мереж, коли змінюється не стільки розподіл мережі, скільки швидкість її еволюції.

Вони доводять, що цей підхід дозволяє локалізувати точки переходу навіть при плавній зміні структури, демонструючи ефективність на симуляціях і даних мозкових органодів.

У статті [72] пропонують байєсівську функціональну графову модель, яка одночасно оцінює часово-сталі та змінні у часі залежності між багатовимірними функціональними даними та виявляє точки переходу у графовій структурі. Симуляції та аналіз температури поверхні моря показали перевагу підходу у точності виявлення та інтерпретації змін залежностей.

Автори [73] розробили дві сім'ї процедур послідовного моніторингу для виявлення точок переходу в параметрах GARCH(1,1), зокрема змін між стаціонарними та вибуховими режимами волатильності. Оцінено граничні розподіли затримки виявлення, а ефективність підтверджено симуляціями та аналізом фондових даних.

У статті [74] представили пакет SNSeg для R, що реалізує методику SNCP — виявлення точок переходу на основі самонормалізації. Підхід є непараметричним, стійким до часової залежності та дозволяє в єдиній схемі оцінювати зміни середнього, дисперсії, кореляції, квантилів та їх комбінацій у одно- та багатовимірних рядах. Розширення SNHD орієнтоване на багатовимірні дані. Пакет надає функції для аналізу, візуалізації та оцінювання параметрів сегментів, спрощуючи практичне застосування пошуку точок переходу.

Автори [75] розробили онлайн-метод виявлення точок переходу для розподілених сенсорних мереж, орієнтований на мінімізацію обміну даними між датчиками та хмарою. Метод забезпечує близьку до ідеального сценарію (без обмежень на комунікацію) потужність виявлення, суттєво скорочуючи витрати на передачу, і демонструє баланс між розміром вікна, локальним порогом та обмеженнями на ресурси.

1.3. Теоретичні результати та оцінки меж виявлення

У роботі [76] запропонували нову теоретичну основу для багатомасштабних статистичних тестів, що усуває залежність від вибору параметра згладжування та зберігає оптимальну асимптотичну здатність виявлення. Ключова ідея полягає у заміні адитивного штрафу на мультиплікативне зважування, що дозволило поширити методику на нестационарні нелінійні часові ряди з використанням спеціалізованого бутстреп-підходу. Запропонований підхід придатний для виявлення сигналів, перевірки адекватності регресійних моделей та задач багатоточкового виявлення змін.

В роботі [77] довели поріг виявлення точки переходу у моделях преференційного приєднання, коли доступний лише фінальний знімок мережі. У такій задачі параметр $\delta(t)$, що описує силу приєднання до вершин з високим ступенем, може змінюватися в момент τ_n через зовнішні події або інтервенції. Раніше було відомо, що простий тест за кількістю вершин мінімального ступеня виявляє пізню зміну, якщо вона трапилася не пізніше ніж за $O(\sqrt{n})$ кроків до кінця, і було висунуто гіпотезу, що пізніші зміни виявити неможливо.

Автори підтвердили цю гіпотезу, показавши, що жоден метод не може надійно виявити точку переходу, якщо вона сталася пізніше ніж $n - o(\sqrt{n})$, та що локалізувати її точніше, ніж з похибкою $o(\sqrt{n})$, теж неможливо. Доведення спирається на оцінки повної варіаційної відстані, використання нерівності Ефрона–Стайна та технік сполучення.

Робота визначає фундаментальну межу можливостей для виявлення структурних змін у випадкових графах і відкриває перспективи для дослідження складніших правил приєднання, множинних або тимчасових змін, а також альтернативних динамічних моделей мереж.

В роботі [78] роблять огляд застосувань теорії випадкових матриць (RMT) у багатовимірній статистиці, зокрема для задач changepoint detection. Автори підкреслюють, що аналіз спектральних властивостей великих випадкових матриць, особливо коваріаційних, дозволяє виявляти структурні зміни у даних, використовуючи як поведінку основного спектра, так і крайніх власних значень. Такий підхід є корисним для виявлення моментів різкої зміни у високовимірних часових рядах, де класичні методи малоефективні, і забезпечує нову теоретичну основу для сучасних методів статистичного аналізу.

Автор [79] досліджує мінімаксні швидкості в задачах тестування точок переходу у дисперсії та коваріації. Для одновимірних даних автор повністю характеризує складність виявлення зміни дисперсії. Для багатовимірних випадків отримано нижні межі швидкості за операторною нормою з урахуванням s -розрідженості, а також побудовано оптимальні та обчислювально ефективні тести, здатні виявляти зміни у матриці коваріацій.

В роботі [80] порівняли вісім алгоритмів виявлення точок переходу на штучних та реальних кардіосигналах, аналізуючи вплив шуму, толерантності в часі та аритмій. Найвищу точність на штучних даних показав mBOCD, а стійкість до екстрасистол — PELT2. Для задачі класифікації пацієнтів із RBD від здорових осіб найкращі результати дали ознаки, отримані з алгоритму RMDM. Автори підкреслюють, що вибір і налаштування методу CPD слід робити з урахуванням конкретної прикладної задачі.

Суттєвий внесок у розвиток теорії непараметричних методів виявлення точок переходу належить Б. Бродському та Б. Дарховському, які розробили низку фундаментальних підходів до аналізу стохастичних моделей із перемиканням режимів. Вони розробили систематичний набір непараметричних і послідовних / ретроспективних методів для виявлення моментів зміни (зміни середнього, дисперсії, розподілу тощо), обґрунтували математичні властивості

цих методів і застосували підхід до практичних задач контролю якості, фінансів, обробки сигналів.

У своїй праці [81] вони зібрали і розвинули нелінійні та непараметричні підходи до задач виявлення і оцінювання моменту зміни — як для офлайн-аналізу (*retrospective*), так і для послідовних процедур (*sequential*). Ця праця стала однією з опорних робіт у цій галузі.

У подальших працях і застосуваннях часто згадується так званий *Brodsky–Darkhovsky statistic* — статистика, яка має сильні властивості при виявленні структурних зламів (її максимум дає оцінку точки зміни). Ця статистика й підхід добре працюють при виявленні різких зламів у середньому/варіації та часто використовуються у прикладних дослідженнях (фінансові ряди, контроль якості тощо). [82]

На відміну від класичних параметричних тестів, вони акцентували на методах, які не потребують точного знання структури розподілів, використовуючи дивергенції, рангові/розподільні статистики тощо. Це робить їхні методи стійкими до відхилень від нормальності та до «тяжких хвостів».

Вони також працювали над мінімаксними послідовними тестами для багатогіпотезних задач та змістовно розвивали теорію оптимальних/мінімаксних процедур у задачах детекції та оцінювання точок переходу. Ці результати особливо важливі для задач, де потрібно швидко і надійно виявляти зломи при потоковому надходженні даних. [83]

Методи Бродського й Дарховського широко цитуються в літературі з детекції кількох змін, обробки фінансових часових рядів, статистичного контролю якості, систем раннього виявлення збоїв тощо; їхні підходи часто поєднують із сучасними непараметричними й ядровими методами. [84] [85]

1.4. Узагальнення та висновки

У сучасній літературі представлено значну кількість підходів до виявлення точок переходу у часових рядах, які охоплюють як офлайн, так і онлайн-методи, працюють з різними типами даних (одно- та багатовимірні) та орієнтовані на різні прикладні сфери — від фінансових ринків і промислового моніторингу до біомедицини та мережі IoT.

Позитивними рисами існуючих рішень є:

- Висока гнучкість — розроблені методи здатні працювати з різними типами змін (середнє, дисперсія, автоковаріація, квантилі) та різною природою даних (стаціонарні, нестаціонарні, з автокореляцією).
- Математична обґрунтованість — більшість підходів супроводжуються строгими теоретичними результатами (консистентність, асимптотичні розподіли, оцінки затримок).
- Прикладна спрямованість — алгоритми адаптовані до конкретних завдань: фінансовий моніторинг “бульбашок”, медична діагностика за фізіологічними сигналами, аналіз мережевих даних тощо.
- Програмна доступність — розроблені пакети (наприклад, SNSeg) полегшують використання методів практиками.

Водночас виявлені і суттєві обмеження:

- Спеціалізація на окремих типах змін — багато методів налаштовані лише на середнє чи дисперсію, що вимагає додаткових модифікацій для інших параметрів.
- Чутливість до параметрів налаштування — вибір вікна, вагових коефіцієнтів чи порогів часто критично впливає на якість, а оптимальний вибір не завжди очевидний.

- Залежність від статистичних припущень — порушення умов (наприклад, висока автокореляція, невідома післязмінна модель) може суттєво знизити ефективність виявлення.
- Комунікаційні та обчислювальні обмеження — особливо в онлайн-налаштуваннях з розподіленими сенсорними мережами, де передача даних є дорогою.

Напрями для покращення охоплюють:

- Розробку універсальних процедур, здатних виявляти різні типи змін без зміни базової методології.
- Автоматичне налаштування параметрів та адаптивний вибір вікон/порогів у процесі роботи.
- Підвищення стійкості до залежностей та шуму в даних.
- Інтеграцію методів з оптимізаційними або керуючими модулями, що дозволить не лише виявляти, а й оперативно реагувати на зміни.
- Розвиток комунікаційно-ефективних онлайн-алгоритмів, придатних для обмежених за ресурсами систем.

Підсумовуючи, хоча галузь виявлення точок переходу розвинута та багата на підходи, універсального методу, який би одночасно поєднував гнучкість, стійкість, теоретичну обґрунтованість та ефективність у реальних умовах, наразі не існує. Це залишає простір для подальших досліджень, зокрема у напрямі поєднання сильних сторін різних класів алгоритмів та розробки адаптивних гібридних рішень.

Розділ 2 «Теоретична частина»

У цьому розділі представлено комплекс теоретичних відомостей, необхідних для побудови та обґрунтування методів, що застосовуються в подальших розділах роботи. Розгляд розпочинається зі статистики Петуніна, яка використовується для виявлення відмінностей між вибірками та ідентифікації точок переходу. Детально подано її математичне визначення, розглянуто властивості, зокрема доведення консистентності та оцінювання потужності p -статистики Петуніна.

Далі описано метод головних компонент (Principal Component Analysis), який є ефективним інструментом для зменшення розмірності та виділення найінформативніших напрямків варіації даних. Цей метод особливо корисний при роботі з багатовимірними часовими рядами та великими масивами даних, де необхідне попереднє узагальнення інформації без істотної втрати змістовності.

Окрему увагу приділено класичним непараметричним статистикам, які широко застосовуються для порівняння вибірок без строгих припущень щодо закону розподілу даних. Зокрема, розглянуто статистику Колмогорова–Смірнова, що дозволяє оцінити відмінності між емпіричними розподілами, та статистику Вілкоксона, яка ефективно використовується для перевірки гіпотез щодо зсуву медіани або узагальнених позиційних характеристик.

Подальший виклад присвячено геометричним підходам до аналізу даних, зокрема еліпсоїдам Петуніна, які застосовуються для побудови областей допустимих значень у багатовимірному просторі ознак. Такі області можуть бути корисними як для виявлення аномалій, так і для визначення меж класів у задачах класифікації.

Завершує розділ розгляд лінійного дискримінанту Фішера, що є одним із базових методів лінійної класифікації. Він ґрунтується на пошуку проєкції даних, яка максимізує роздільну здатність між класами, та має тісний зв'язок з аналізом дисперсії.

Об'єднуючим фактором для всіх розглянутих у розділі методів є їхня спільна мета — підвищення точності та надійності виявлення змін у структурі даних, що безпосередньо сприяє ефективному пошуку точок переходу та побудові коректних алгоритмів моніторингу й аналізу.

2.1. Статистика Петуніна

В роботах [86][87][88] запропонована та досліджена р-статистика, що базується на теоремі Хілла [89], 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:

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

де $x_{(1)} \leq x_{(2)} \leq x_{(3)} \dots \leq x_{(n)}$ і $x'_{(1)} \leq x'_{(2)} \leq x'_{(3)} \dots \leq x'_{(m)}$ - 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 [89]

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

Якщо $x'_k \in \{x'_{(1)}, x'_{(2)}, x'_{(3)}, \dots, x'_{(m)}\}$, то можна знайти частоту $h_{ij}^{(n)}$ випадкової події A_{ij} та побудувати довірчі інтервали $(\Delta_{ij}^{(1)}, \Delta_{ij}^{(2)})$ для ймовірності p_{ij} на рівні значущості β , тобто

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

Відповідно до [90]

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

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

Можна визначити рівень значущості довірчого інтервалу $I_{ij}^{(n,m)} = (\Delta_{ij}^{(1)}, \Delta_{ij}^{(2)})$ через значення g , припускаючи $\phi(g) = 1 - \frac{\beta}{2}$, де $\phi(g)$ – функція щільності нормального розподілу. Згідно з правилом 3σ [91], значення $g = 3$ відповідає рівню значущості не більше ніж 0.05. Загальна кількість довірчих інтервалів $I_{ij} = (\Delta_{ij}^{(1)}, \Delta_{ij}^{(2)})$ позначимо як $N = \frac{n(n-1)}{2}$. Кількість таких інтервалів I_{ij} , що містять ймовірність $p_{ij}^{(n)}$ позначимо L . Таким чином, статистика Петуніна,

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

характеризує ступінь близькості вибірок x і x' . Перевірка гіпотези H на рівні значущості 0.05 проводиться шляхом підстановки значення $h^{(n)}$ у формулу для довірчого інтервалу і перевіркою, чи потрапляє оцінене значення в межі $I = (\Delta^{(1)}, \Delta^{(2)})$ [92].

2.2. Доведення консистентності

В подальшому замість $h^{(n)}$ будемо писати просто h .

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

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

Тоді математичне сподівання індикатора:

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

Оскільки $B_{ij} \in \{0,1\}$ і таких індикаторів N штук, то можемо застосувати нерівність Хеффіна для сум незалежних Бернулівських випадкових величин.

Лема 1 (Нерівність Хеффіна для змінних Бернуллі). Нехай $X_1, X_2, \dots, X_n$ – незалежні випадкові величини, де $X_i \in [0,1]$ та $EX_i = \mu_i$. 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 отримаємо:

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

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 тому ймовірність $P(|h - (1 - \beta)| > \varepsilon) \rightarrow 0$. 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 $X \quad x_{(1)} \leq x_{(2)} \leq x_{(3)} \dots \leq x_{(n)}$ 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 $k \cdot (n - k)$, що можна наблизити як

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

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} .

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

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».

Отже

$$\mathbb{E}L \leq \frac{1}{2}(1 - \beta)(p_\delta)^2 n^2$$

Отримуємо

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

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

Далі робимо так само, як і пункті з H_0 . Використуємо нерівність Хеффіна:

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

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.

2.3. Доведення потужності р-статистики Петуніна

Припущення про форму щільності. У більшості реальних задач функція щільності f є унімодальною (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 (про потужність р-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 $\mu > 0$ – додатний зсув).

Тоді р-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 різницеву функцію $D(x) = F(x) - G(x) = F(x) - F(x - \mu)$.

Оскільки F строго зростаюча і неперервна на $\mathbb{R}$, то $\lim_{x \rightarrow \pm\infty} D(x) = 0$.

Також зрозуміло, що $D'(x) = f(x) - f(x - \mu)$

Із припущення про унімодальність щільності випливає:

1) Для $x < m$ і $\mu > 0$: $x - \mu < x < m$, тобто $f(x) > f(x - \mu) \Rightarrow D'(x) > 0$.

2) Для $x > m + \mu$: $x > x - \mu > m$, тобто $f(x) < f(x - \mu) \Rightarrow D'(x) < 0$.

Так як похідна змінює знак, то існує точка $x^* \in (t, t + \mu)$ така, що $D'(x^*) = 0$, причому x^* - локальний максимум D .

Нехай $\delta = D(x^*) = F(x^*) - F(x^* - \mu) > 0$

Примітка. Якщо F симетрична довкола 0 (нормальний розподіл, Лапласа), то $t = 0$ і рівність $f(x) = f(x - \mu)$ дає розв'язок $x^* = \frac{\mu}{2}$.

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

Область поблизу точки максимуму x^* :

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

Область у хвості, де $D(x) \rightarrow 0$:

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

Так як D – неперервна, то існує інтервал $(x^* - \eta, x^* + \eta)$ із $D(x) \geq \frac{\delta}{2}$.

Ймовірність потрапити в цей інтервал це

$$p_A = F(x^* + \eta) - F(x^* - \eta) > 0$$

Так як $\lim_{x \rightarrow -\infty} D(x) = 0$, то знайдеться $b < x^* - \mu$ таке, що на $(-\infty, b]$ виконується $D(x) \leq \varepsilon$. Ймовірність потрапити в цей інтервал це

$$p_B = F(b) - F(-\infty) = F(b) > 0$$

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

За законом великих чисел

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

З ймовірністю $\rightarrow 1$ при великих n

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

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 $M_n \geq \frac{1}{4} p_A p_B n^2 = c n^2$ – 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 отримуємо, 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 Хеффінга отримаємо, що $h_{n,m} \xrightarrow{P} 0$. ■

Теорема 2 (про потужність r -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 r -статистика Петуніна $h_{n,m} \xrightarrow{P} 0$ при $n, m \rightarrow \infty$, тобто критерій має асимптотичну потужність 1 при будь-якому неодиначному 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 $x > 0$, починаючи з деякого x_0 буде виконуватись те, що $D'(x) < 0$.

Для околу моди, тобто для x трохи більших за m $f(x) > f\left(\frac{x}{\sigma}\right)$, бо обидва аргументу 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 це

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

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

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

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

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

Кількість таких пар

$$M_n \geq \frac{1}{4} p_A p_B n^2 = c n^2$$

– 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 Хеффінда
отримаємо, що і $h_{n,m} \xrightarrow{P} 0$. ■

2.4. Метод головних компонент (Principal Component Analysis)

РСА був запропонований Карлом Пірсоном ще у 1901 році [89] 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 проєкцією даних) припадає на першу координату (першу головну компоненту), друга за величиною дисперсія — на другу координату і так далі [90].

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 :

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

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

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

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.

2.5. Статистика Колмогорова-Смірнова

Нехай емпірична функція розподілу F_n , яка побудована за вибіркою

$X = (X_1, \dots, X_n)$ має вигляд:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

Де $I_{X_i \leq x}$ вказує, чи попало спостереження X_i в область $(-\infty, x]$:

$$I_{X_i \leq x} = \begin{cases} 1, & X_i \leq x; \\ 0, & X_i > x. \end{cases}$$

Виконується перевірка того, чи є вибірка, породжена випадковою величиною ξ з функцією розподілу $F(x)$. Статистика критерію для емпіричної функції розподілу $F(x)$ визначається наступним чином:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

Нехай задано дві незалежні вибірки випадкової величини ξ : $x = (x_1, x_2, \dots, x_n)$ та $x' = (x_1', x_2', \dots, x_m')$ і за ними побудовані емпіричні функції розподілу $F_{1,n}(x), F_{2,m}(x)$. Статистика критерію у даному випадку визначається наступним чином:

$$D_{n,m} = \sup_{x \in \mathbb{R}} |F_{1,n}(x) - F_{2,m}(x)|$$

Прийняття рішення відбувається за критерієм Смірнова:

Якщо статистика $\sqrt{\frac{nm}{n+m}} D_{n,m}$ перевищує квантиль розподілу Колмогорова K_α для заданого рівня значущості α , то нульова гіпотеза H_0 (про однорідність вибірок) відкидається. Інакше вона приймається на рівні α . [95]

2.6. Статистика Вілкоксона

Нехай є дві вибірки $x_1 = (x_{1,1}, \dots, x_{1,N})$ і $x_2 = (x_{2,1}, \dots, x_{2,N})$ одного розміру N .

1. Для кожної пари треба порахувати $|x_{2,i} - x_{1,i}|$ і $\text{sgn}(x_{2,i} - x_{1,i})$.
2. Пари, де $|x_{2,i} - x_{1,i}| = 0$ слід виключити.
3. Впорядкувати пари, що залишились, від найменшої до найбільшої абсолютної різниці $|x_{2,i} - x_{1,i}|$.
4. Порахувати ранги R_i пар вже у новому ряді.
5. Обчислити статистику W

$$W = \sum_{i=1}^N (\text{sgn}(x_{2,i} - x_{1,i}) \cdot R_i)$$

6. Для невеликих $n \leq 20$ критичне значення статистики можна подивитись у відповідних таблицях. Для великих n ця статистика має наближений

нормальний розподіл. Щоб обчислити критичне значення треба спочатку обчислити значення

$$z_{obs} = \frac{W - \mu}{\sigma}, \text{ де } \mu = \frac{n(n+1)}{4}, \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Перевірити, чи справджується, що $2P(z \geq z_{obs}) \leq 0.05$ вже для випадкової величини z , що має нормальний розподіл $z \sim N(0,1)$. Якщо це виконується, то нульова гіпотеза відхиляється. [96]

2.7. Еліпсоїди Петуніна

Нехай є набір випадкових точок $X = \{x_1, \dots, x_n\}$, де $x_i \in \mathbb{R}^d$. Задача полягає в пошуку оптимального еліпсоїда, що містить всі дані точки і має найменший об'єм серед усіх інших з тією ж властивістю. Щоб описати алгоритм побудови, розглянемо випадок $d = 2$.

Спочатку будуємо опуклу оболонку множини точок $X = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Знаходимо діаметр опуклої оболонки і її кінці (x_k, y_k) та (x_l, y_l) .

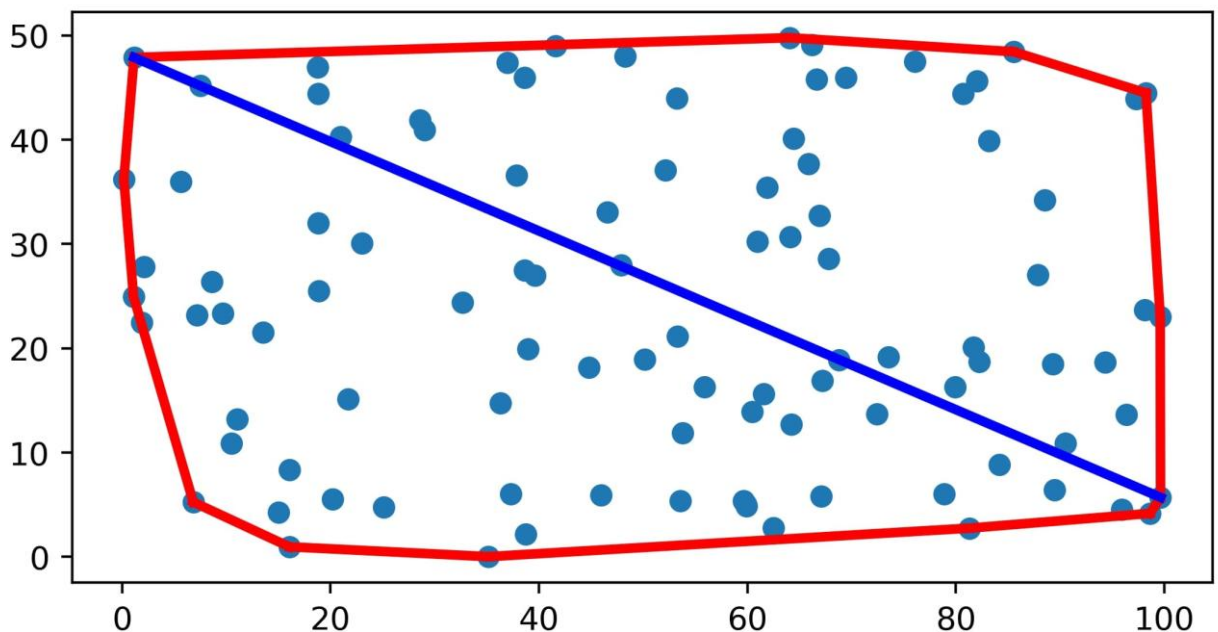


Рис. 2.7.1. Діаметр опуклої оболонки

Формуємо відрізок L , що з'єднує обидва кінця і шукаємо точки, що лежать на найбільшій відстані від L : (x_r, y_r) та (x_q, y_q) . Будуємо відрізки L_1, L_2 , що паралельні L , проходять через (x_r, y_r) та (x_q, y_q) і мають довжину 1.

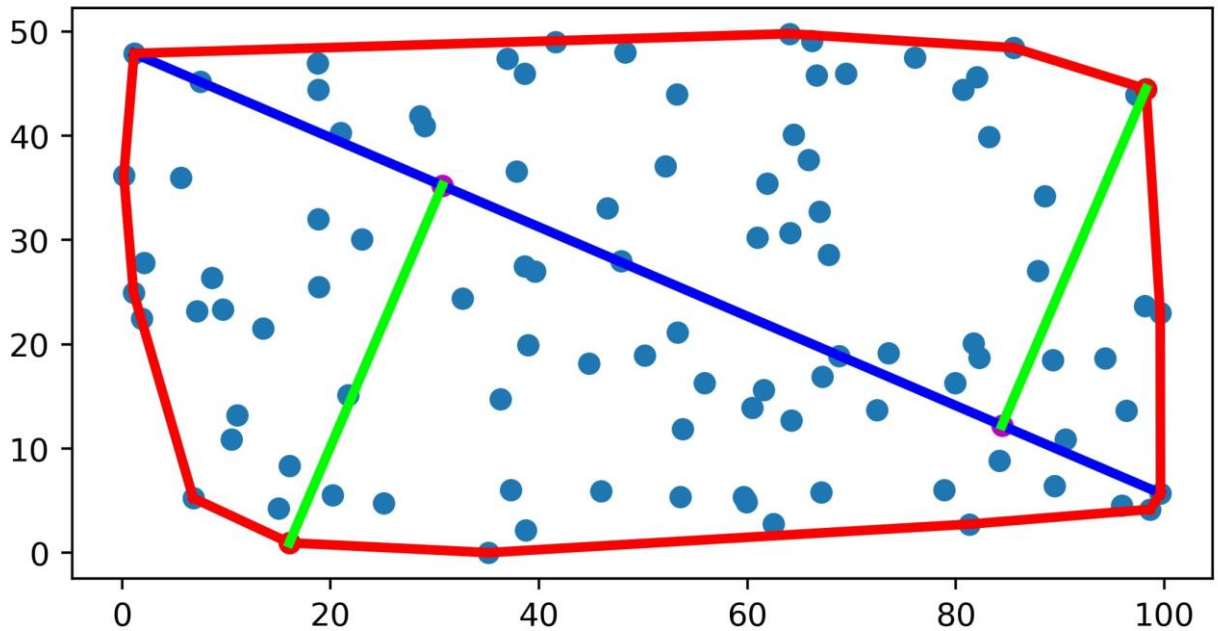


Рис. 2.7.2. Дві найбільш віддалені точки від діаметру

Також створюємо відрізки L_3 та L_4 , що перпендикулярні до L і проходять через (x_k, y_k) та (x_l, y_l) . Відрізки L_1, L_2, L_3 і L_4 утворили прямокутник Π . Позначимо меншу сторону за a , а більшу за b . Зробимо паралельний перенос, поворот і розтягнення з коефіцієнтом $\alpha = \frac{a}{b}$, щоб Π перетворився у квадрат Π' з центром (x'_0, y'_0) . Всі інші точки перетворились на точки $(x'_1, y'_1), \dots, (x'_n, y'_n) \in \Pi'$.

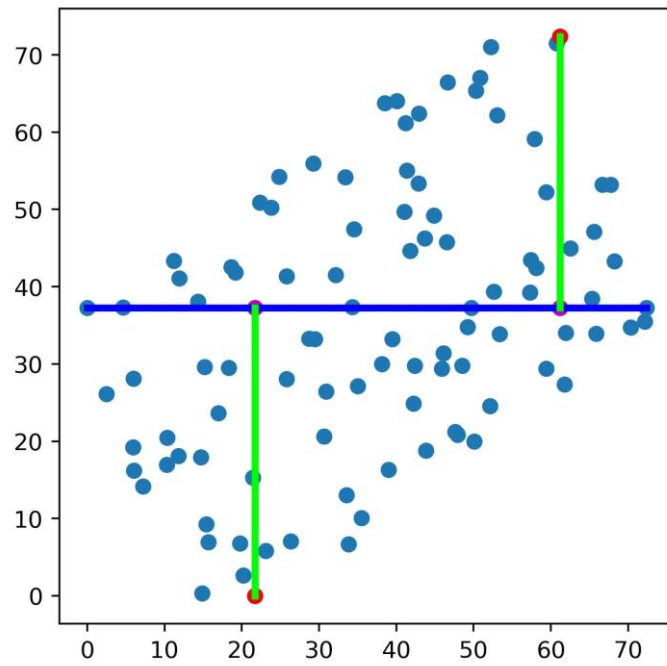


Рис. 2.7.3. Паралельний перенос, поворот і розтягнення

Позначимо за r_i – відстані від (x'_i, y'_i) до центра квадрата (x'_0, y'_0) . Нехай $R = \max(r_1, \dots, r_n)$. Побудуємо коло C з центром в точці (x'_0, y'_0) і радіусом R , а також всі інші концентричні кола з тим же центром і радіусами r_i .

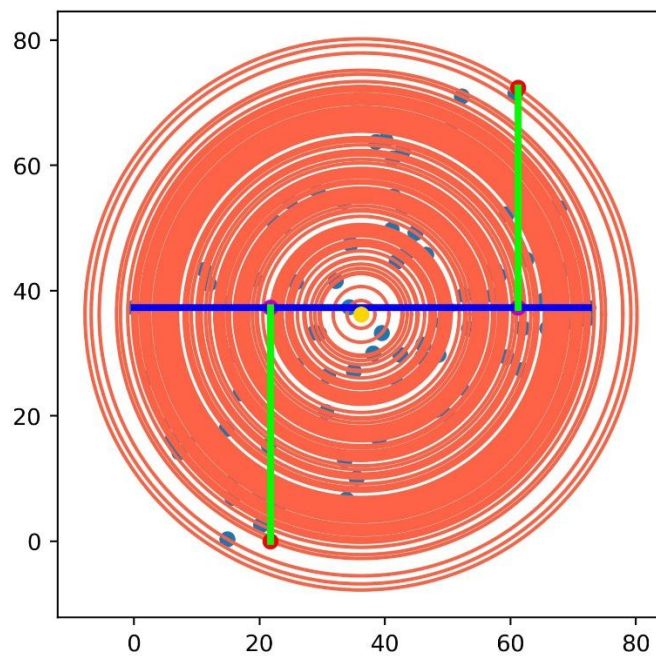


Рис. 2.7.4. Побудова концентричних кіл

Виконуючи обернені перетворення з цим колом, ми отримуємо шуканий еліпс E .

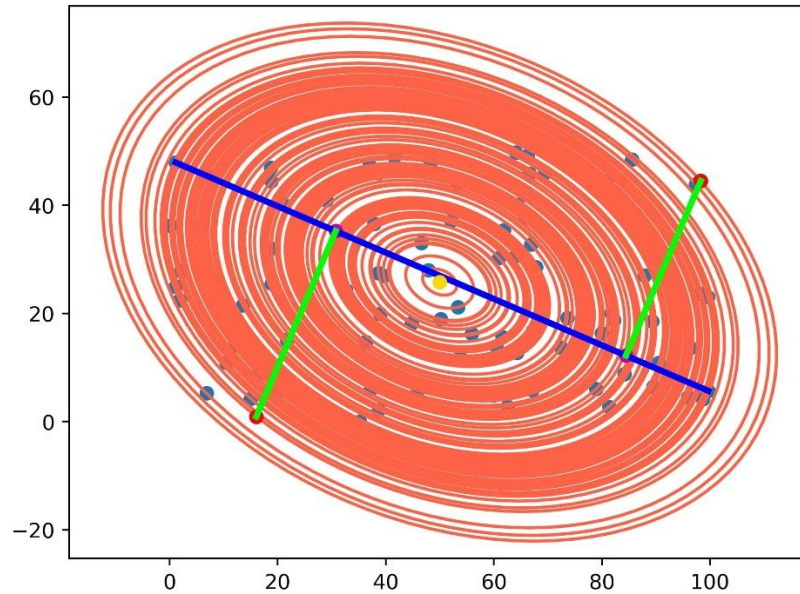


Рис. 2.7.5. Концентричні еліпси

Випадок $d > 2$ розглядається аналогічно. [97]

Еліпсоїд Петуніна має чудові властивості. По-перше, він унікально розташовує точки відповідно до їх статистичної глибини, оскільки за побудовою на поверхні з еліпсоїда Петуніна лежить лише одна точка. По-друге, ймовірність того, що випадкові точки з того самого розподілу лежать в межах еліпса Петуніна, дорівнює $\frac{n-1}{n+1}$. Таким чином, ми маємо не тільки еліпс, що містить заданий набір з n випадкових точки з точною ймовірністю, але також можемо знайти викиди.

2.8. Лінійний дискримінант Фішера

Хоча в оригінальній статті Фішера [98] описується дискримінант, який дещо відрізняється від LDA, терміни лінійний дискримінант Фішера та LDA (лінійний дискримінантний аналіз) часто використовуються як взаємозамінні. Деякі припущення LDA, такі як те, що класи мають нормальний розподіл або що коваріації класів рівні, не базуються на методі Фішера.

Розглянемо два класи спостережень, які мають коваріації Σ_0, Σ_1 та середні значення $\vec{\mu}_0, \vec{\mu}_1$. Якщо ми використаємо лінійну комбінацію ознак $\vec{w} \cdot \vec{x}$, то отриманий розподіл матиме середні значення $\vec{w} \cdot \vec{\mu}_i$, і дисперсії $\vec{w}^T \Sigma_i \vec{w}$ для $i = 0, 1$. Співвідношення дисперсії між класами до дисперсії всередині класів визначає, як Фішер оцінив відокремленість цих двох розподілів.

$$S = \frac{\sigma_{between}^2}{\sigma_{within}^2} = \frac{(\vec{w} \cdot \vec{\mu}_1 - \vec{w} \cdot \vec{\mu}_0)^2}{\vec{w}^T \Sigma_1 \vec{w} + \vec{w}^T \Sigma_0 \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T (\Sigma_0 + \Sigma_1) \vec{w}}$$

Наведена міра порівнює відмінності між двома наборами спостережень із дисперсією всередині кожного з наборів для оцінки ефективності розділення класів. Коли використовується певна лінійна комбінація ознак, вектор, що є нормаллю до дискримінантної гіперплощини, визначає максимальне відокремлення.

$$\vec{w} \propto (\Sigma_0 + \Sigma_1)^{-1} (\vec{\mu}_1 - \vec{\mu}_0)$$

Ця гіперплощина у двовимірному випадку зображується як пряма, перпендикулярна до цього вектора. Після проєкції точок даних на цю гіперплощину, поріг вибирають, досліджуючи одномірний розподіл проєкцій. Одним із варіантів є встановлення цього порогу між проєкціями середніх значень двох наборів спостережень.

$$c = \vec{w} \cdot \frac{1}{2} (\vec{\mu}_0 + \vec{\mu}_1) = \frac{1}{2} \vec{\mu}_1^T \Sigma_1^{-1} \vec{\mu}_1 - \frac{1}{2} \vec{\mu}_0^T \Sigma_0^{-1} \vec{\mu}_0$$

У такому випадку значення параметра c в умові порогового класифікатора $\vec{w} \cdot \vec{x} > c$ можна явно визначити. Варто зазначити, що припущення оригінального дискримінанта Фішера про класи дещо відрізняються від LDA, оскільки вони не вимагають нормальності розподілів або рівності коваріацій.

Висновки до розділу 2

- Розглянуто та формалізовано статистику Петуніна, яка є основою для подальших доведень і практичних застосувань у задачах виявлення точок переходу.
- Доведено консистентність статистики, що гарантує правильність її асимптотичної поведінки при зростанні обсягу вибірки.
- Показано потужність p -статистики Петуніна, що підтверджує її придатність до використання у задачах статистичного висновування.
- Викладено основи методу головних компонент (PCA), який забезпечує зменшення розмірності даних і виявлення головних напрямків варіації.
- Розглянуто класичні непараметричні критерії — статистику Колмогорова–Смірнова та статистику Вілкоксона, які є фундаментом для перевірки гіпотез і пошуку структурних змін.
- Запропоновано використання еліпсоїдів Петуніна як геометричного узагальнення для багатовимірних випадків, що дозволяє виявляти локальні відхилення та групування даних.
- Детально описано лінійний дискримінант Фішера (FLD) як класичний метод багатовимірної класифікації, що доповнює спектр досліджуваних інструментів.

Таким чином, у розділі сформовано теоретичний фундамент, який поєднує класичні статистичні критерії та сучасні підходи (PCA, еліпсоїди Петуніна, FLD). Отримані результати створюють базу для практичної перевірки методів у наступному розділі та для розробки нових інструментів виявлення точок переходу у багатовимірних даних.

Розділ 3. Практична частина

У цьому розділі подано результати застосування розглянутих у попередньому розділі методів до конкретних задач аналізу даних та пошуку точок переходу. Дослідження проводяться як на одновимірних, так і на багатовимірних даних, що дозволяє оцінити ефективність і обмеження кожного з підходів у різних умовах.

Спершу розглядається одновимірний випадок, який є базовим і дозволяє наочно продемонструвати ключові властивості статистики Петуніна та класичних непараметричних критеріїв. Далі увага приділяється застосуванню еліпсоїдів Петуніна як геометричного інструмента для аналізу багатовимірних вибірок.

Окремий розділ присвячено дослідженню поведінки p -статистики Петуніна у тривимірному випадку, що дозволяє простежити вплив розмірності на її ефективність. Після цього виконується порівняння результатів, отриманих за допомогою методу головних компонент (PCA) та лінійного дискримінанту Фішера (FLD), що є класичними інструментами для зменшення розмірності та класифікації.

Далі здійснюється порівняльний аналіз кількох статистик у контексті багатовимірних часових рядів, де особливо важливим є питання коректності виявлення точок переходу при складній структурі залежностей. Завершальним етапом є симуляційні експерименти на квазіреальних даних, що дозволяють оцінити практичну придатність і стабільність методів у ситуаціях, наближених до реальних прикладних задач.

Спільною метою проведених експериментів є виявлення сильних та слабких сторін кожного з методів і вироблення рекомендацій щодо їх застосування для задач аналізу часових рядів та багатовимірних даних.

3.1. Одновимірний випадок

У цьому підрозділі розглянемо застосування статистики Петуніна до одновимірних часових рядів. Основною метою є перевірка працездатності методу на простіших даних та оцінка його можливостей щодо виявлення точок переходу.

Для оцінки ефективності статистик було проведено серію експериментів із використанням вибірок обсягом $N = 100$ елементів. Розглянемо згенеровану сукупність із 200 спостережень, перші 100 з яких належать до генеральної сукупності G_1 , а наступні 100 — до сукупності G_2 .

На першому етапі фіксується вибірка X , що складається з перших 100 елементів. Далі застосовується ковзне вікно довжиною 100 зі зміщенням на один елемент, починаючи з першого елемента всієї вибірки з 200 точок. Таким чином формується вибірка Y , яка на кожному кроці дедалі більше віддаляється від X . На останньому кроці вибірка Y містить останні 100 елементів.

Для кожної пари вибірок (X, Y) обчислюються значення трьох статистик: p -статистики Петуніна, статистики Колмогорова–Смірнова та статистики Вілкоксона. Отримані значення накопичуються на всіх кроках ковзного вікна, після чого проводиться усереднення.

Процедура (формування вибірок і розрахунок статистик) повторюється 200 разів, що дає змогу отримати стабільні середні результати та побудувати графіки для кожної статистики. На осях таких графіків по осі абсцис відкладено кількість кроків ковзного вікна, а по осі ординат — відповідні усереднені значення статистик. Для візуалізації застосовуються позначення: квадратики відповідають p -статистиці, кола — статистиці Колмогорова–Смірнова, хрестики — статистиці Вілкоксона.

У ході експериментів варіювались початкові розподіли сукупностей G_1 та G_2 : $N(\mu, \sigma^2)$ – нормальний розподіл з математичним сподіванням μ і дисперсією σ^2 , $LogN(\mu, \sigma^2)$ – логнормальний розподіл, $Exp(\lambda)$ – експоненціальний

розподіл, $\text{Gamma}(k, \theta)$ – гамма-розподіл, $U(a, b)$ – рівномірний розподіл на відрізьку $[a; b]$. Для кожного випадку визначалося, наскільки швидко статистика досягає критичного рівня: 0.95 для р-статистики та 0.05 для статистик Колмогорова–Смірнова й Вілкоксона. Це дозволяє порівняти швидкість та точність виявлення точок переходу різними підходами.

У ході експериментальних досліджень розглядалися пари розподілів, в яких змінювався один із параметрів (надалі позначений через α). Для кожної такої пари було згенеровано вибірки, після чого проводився аналіз ефективності трьох статистик: р-статистики Петуніна, статистики Колмогорова–Смірнова та статистики Вілкоксона.

У дослідженні були використані такі комбінації розподілів:

- 1) $N(0,1)$ і $N(\alpha, 1)$
- 2) $N(0,1)$ і $N(0, \alpha)$
- 3) $\text{Log}N(0,1)$ і $\text{Log}N(\alpha, 1)$
- 4) $\text{Log}N(0,1)$ і $\text{Log}N(0, \alpha)$
- 5) $\text{Exp}(1)$ і $\text{Exp}(\alpha)$
- 6) $\text{Gamma}(9,0.5)$ і $\text{Gamma}(\alpha, 2)$
- 7) $\text{Gamma}(9,0.5)$ і $\text{Gamma}(\alpha, 1)$
- 8) $U(0,1)$ і $U(\alpha, \alpha + 1)$

Для кожної пари розподілів будувалися графіки середніх значень трьох статистик залежно від номера кроку ковзного вікна. Це дозволяє оцінити, наскільки швидко кожна статистика реагує на зміну параметра та досягає критичного рівня, що свідчить про виявлення неоднорідності вибірок.

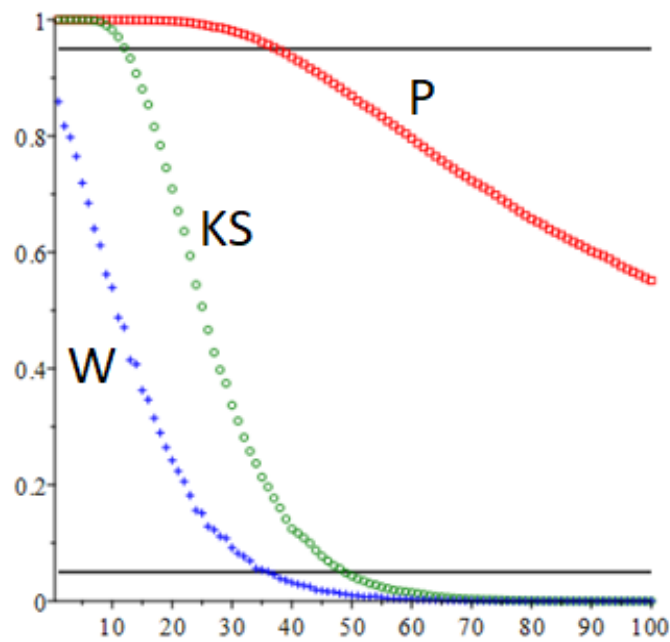


Рис. 3.1.1. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова–Смірнова ($\circ$) та Вілкоксона ($\times$) для порівнянні вибірок із розподілів $N(0,1)$ та $N(1,1)$. Порогові точки: 38 для p -статистики, 49 для Колмогорова–Смірнова та 37 для Вілкоксона.

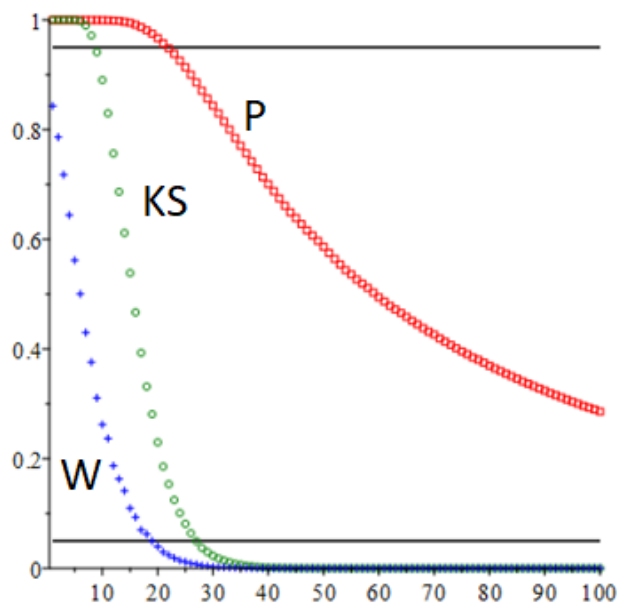


Рис. 3.1.2. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова–Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(2,1)$. Порогові точки: 22 для p -статистики, 27 для Колмогорова–Смірнова і 20 для Вілкоксона

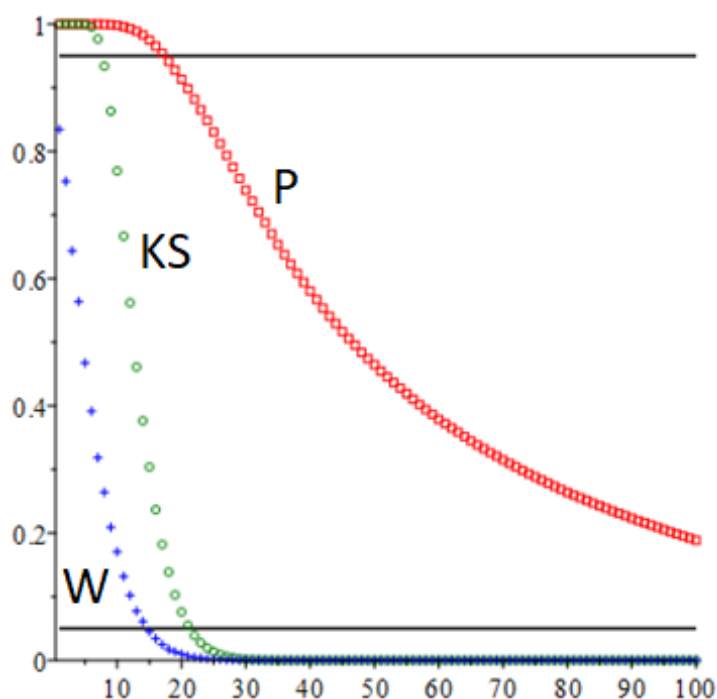


Рис. 3.1.3. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(3,1)$. Порогові точки: 18 для p -статистики, 22 для Колмогорова-Смірнова і 15 для Вілкоксона

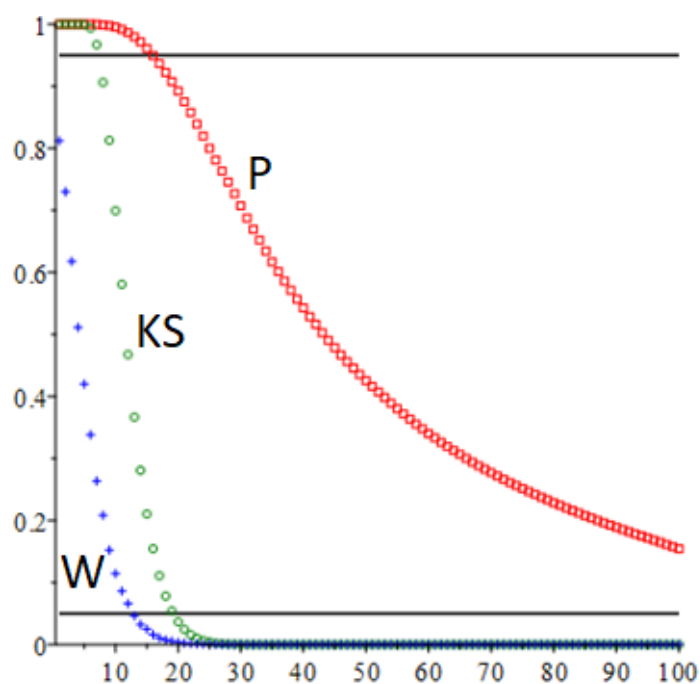


Рис. 3.1.4. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(10,1)$. Порогові точки: 16 для p -статистики, 20 для Колмогорова-Смірнова і 13 для Вілкоксона

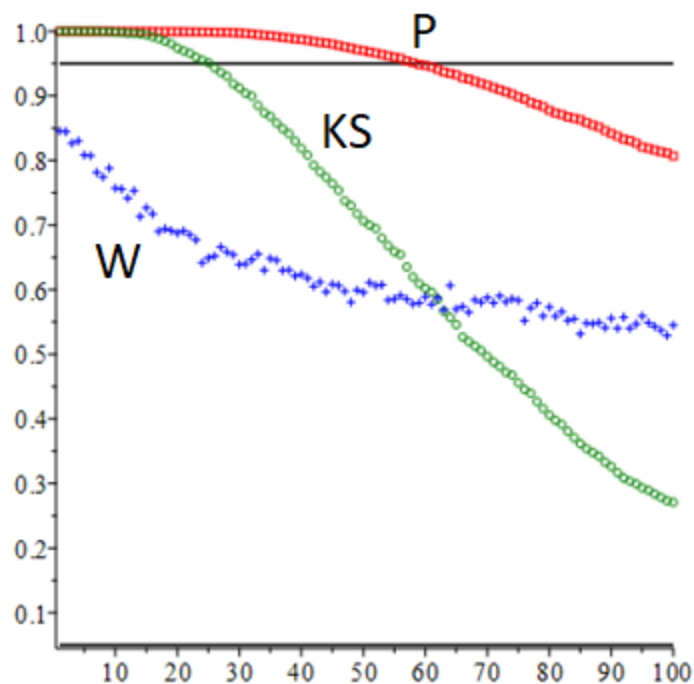


Рис. 3.1.5. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(0,2)$. Порогові точки: 59 для p -статистики

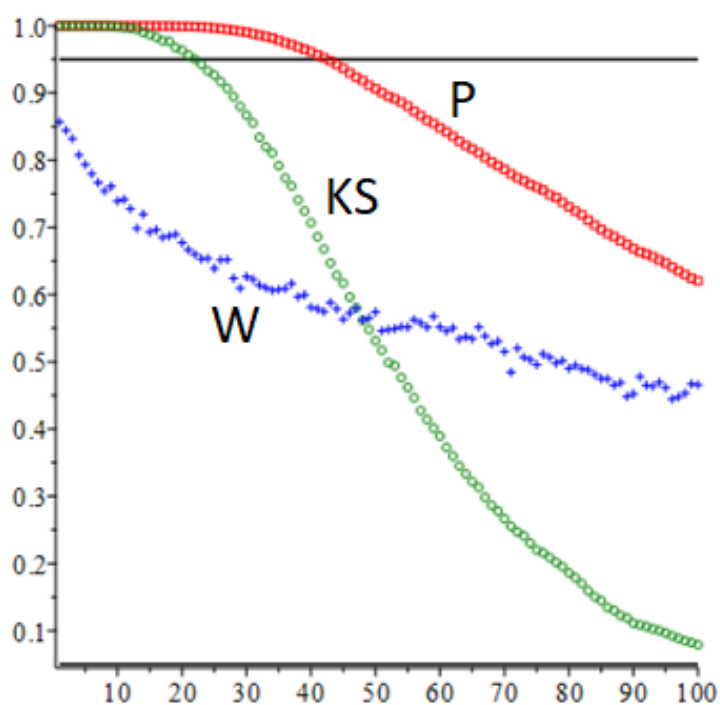


Рис. 3.1.6. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(0,3)$. Порогові точки: 43 для p -статистики

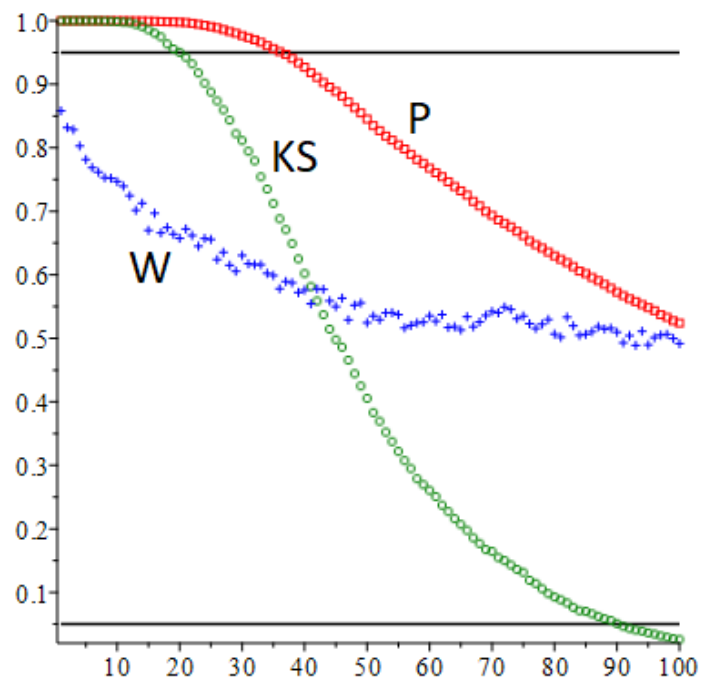


Рис. 3.1.7. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(0,4)$. Порогові точки: 37 для p -статистики, 91 для Колмогорова-Смірнова

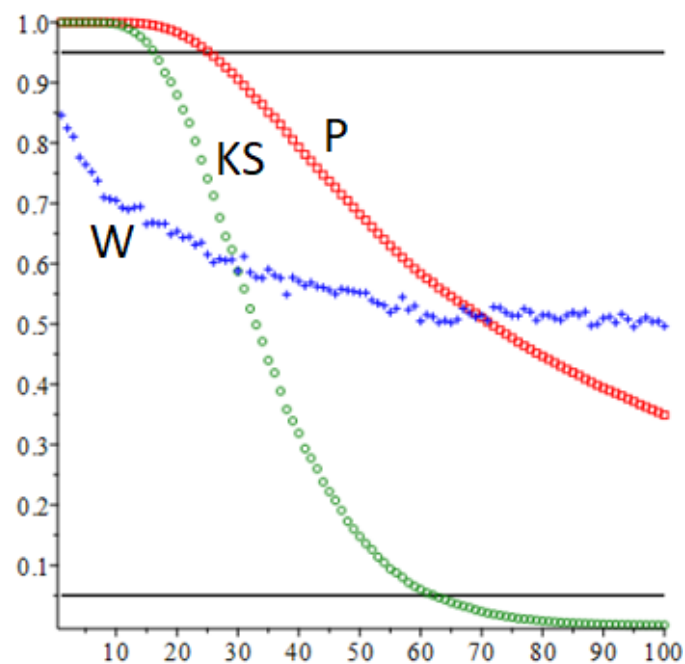


Рис. 3.1.8. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $N(0,1)$ і $N(0,10)$. Порогові точки: 26 для p -статистики, 63 для Колмогорова-Смірнова

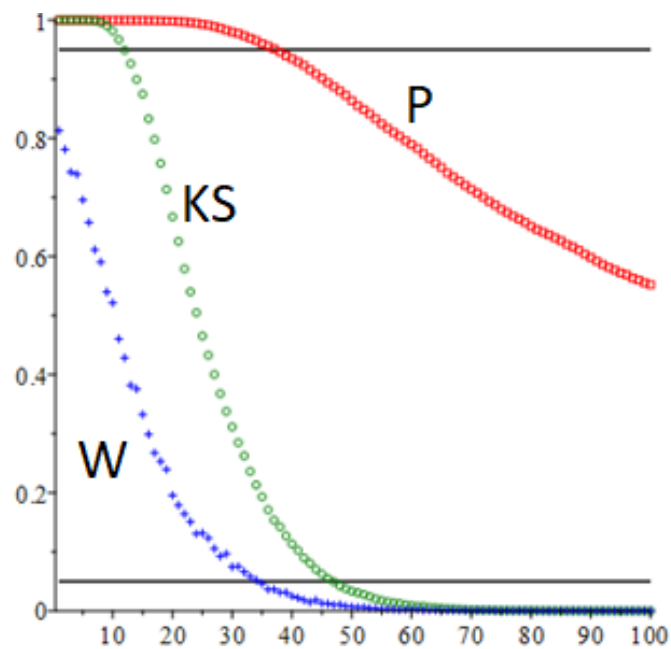


Рис. 3.1.9. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(1,1)$. Порогові точки: 38 для p -статистики, 47 для Колмогорова-Смірнова і 35 для Вілкоксона

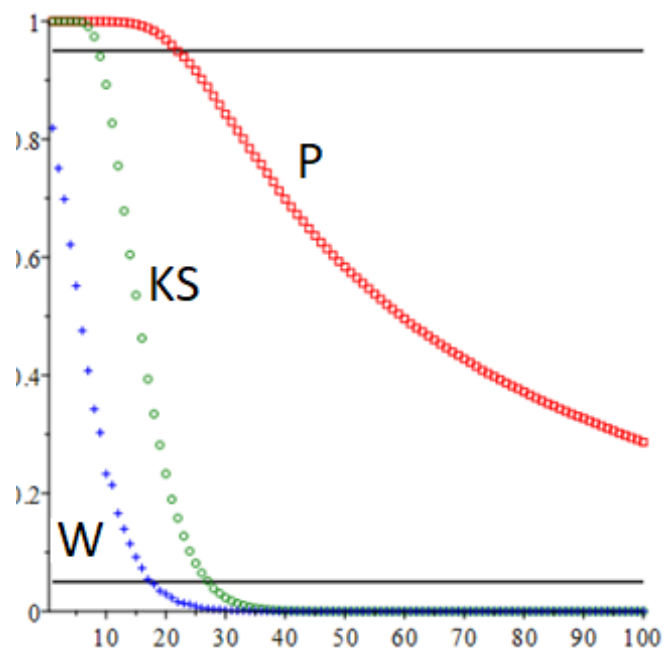


Рис. 3.1.10. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(2,1)$. Порогові точки: 22 для p -статистики, 28 для Колмогорова-Смірнова і 18 для Вілкоксона

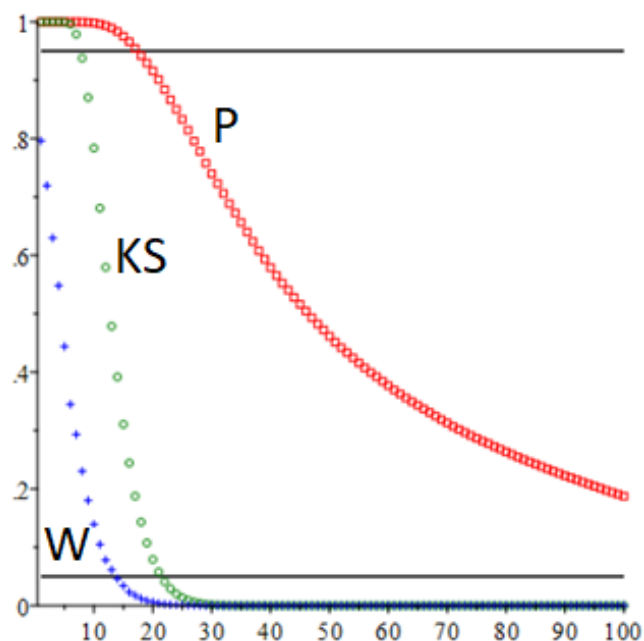


Рис. 3.1.11. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $LogN(0,1)$ і $LogN(3,1)$. Порогові точки: 18 для p -статистики, 22 для Колмогорова-Смірнова і 14 для Вілкоксона

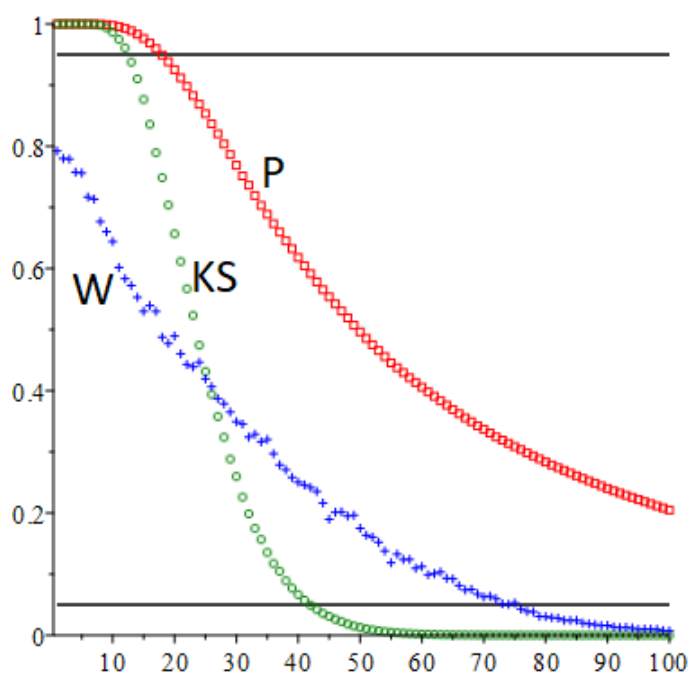


Рис. 3.1.12. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $LogN(0,1)$ і $LogN(0,10)$. Порогові точки: 18 для p -статистики, 42 для Колмогорова-Смірнова і 76 для Вілкоксона

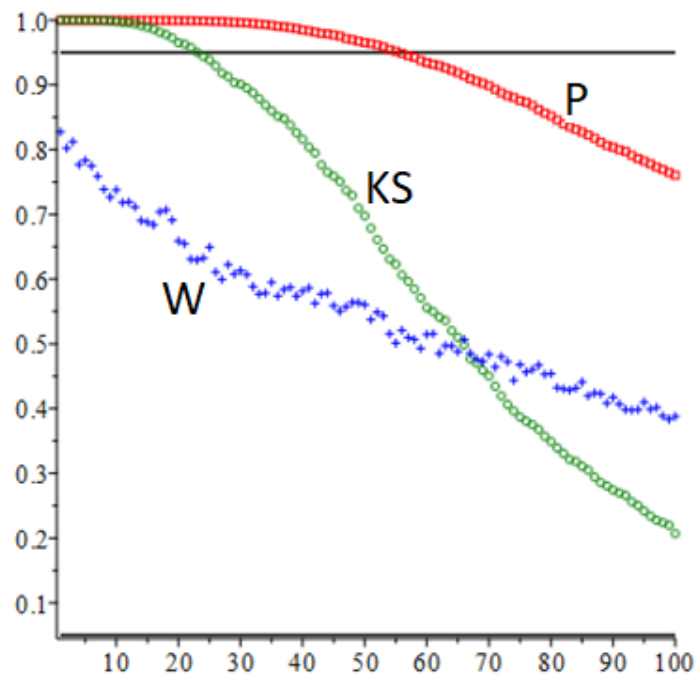


Рис. 3.1.13. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(0,1.5)$. Порогові точки: 56 для p -статистики

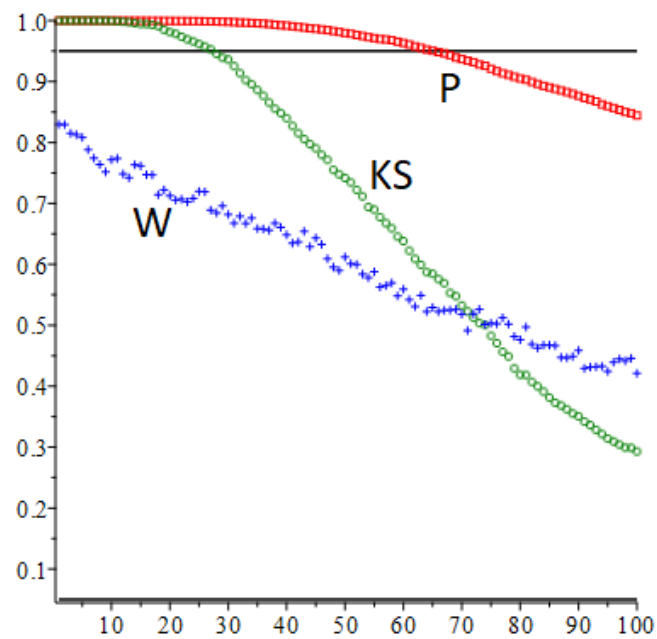


Рис. 3.1.14. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(0,0.75)$. Порогові точки: 66 для p -статистики

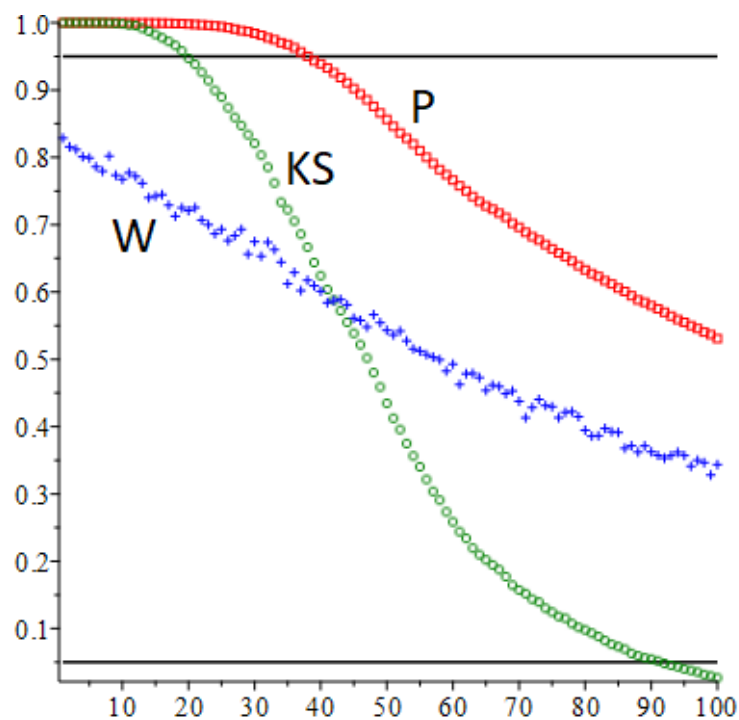


Рис. 3.1.15. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(0,0.5)$. Порогові точки: 39 для p -статистики, 92 для Колмогорова-Смірнова

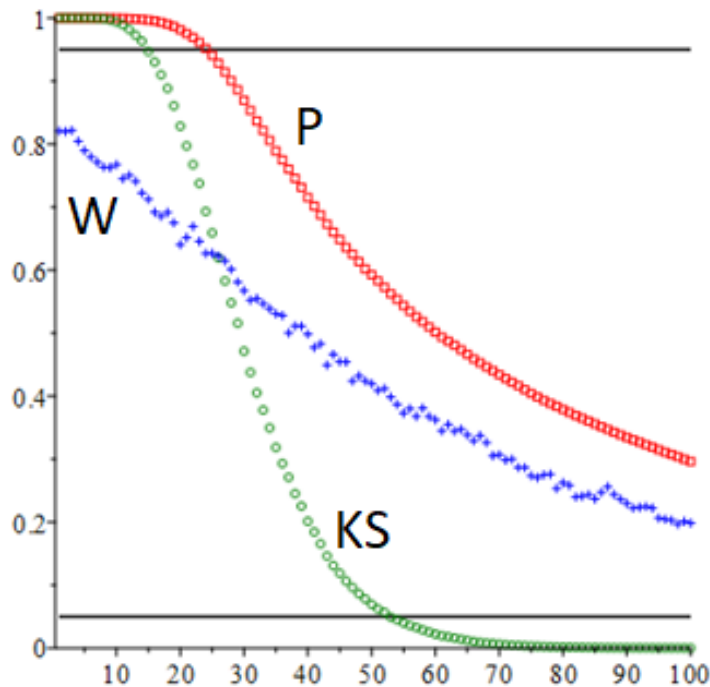


Рис. 3.1.16. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(0,0.25)$. Порогові точки: 25 для p -статистики, 53 для Колмогорова-Смірнова

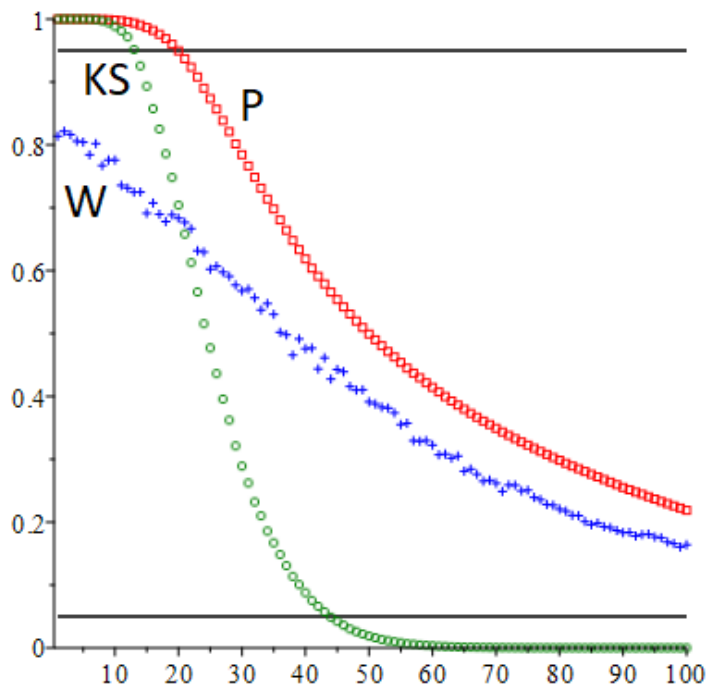


Рис. 3.1.17. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Log}N(0,1)$ і $\text{Log}N(0,0.125)$. Порогові точки: 20 для p -статистики, 44 для Колмогорова-Смірнова

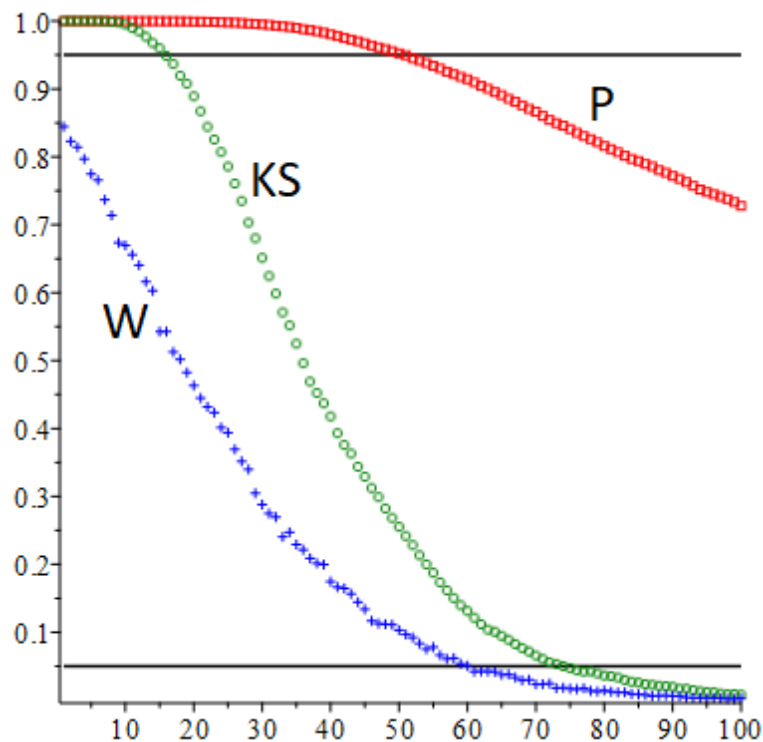


Рис. 3.1.18. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Exp}(1)$ і $\text{Exp}(0.5)$. Порогові точки: 51 для p -статистики, 75 для Колмогорова-Смірнова і 61 для Вілкоксона

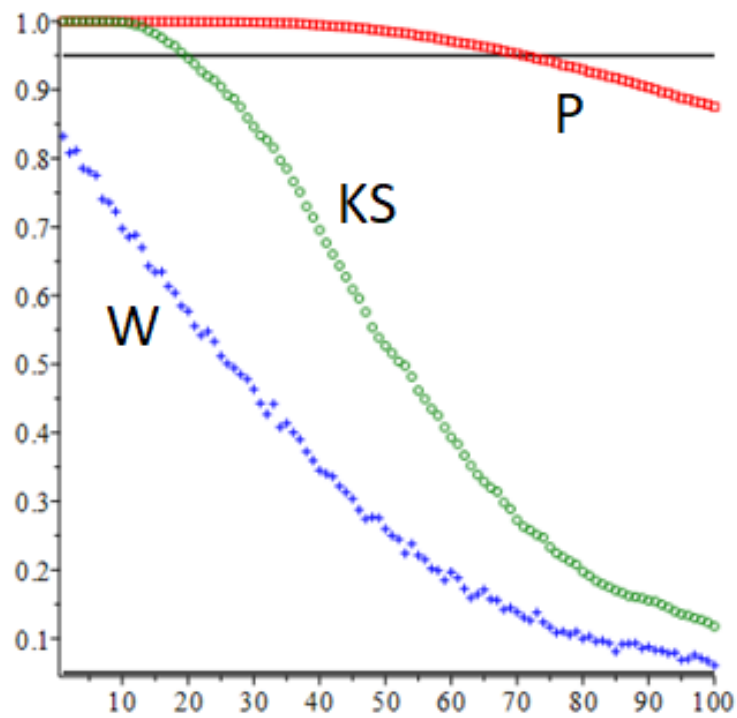


Рис. 3.1.19. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(1.5)$. Порогові точки: 72 для p -статистики

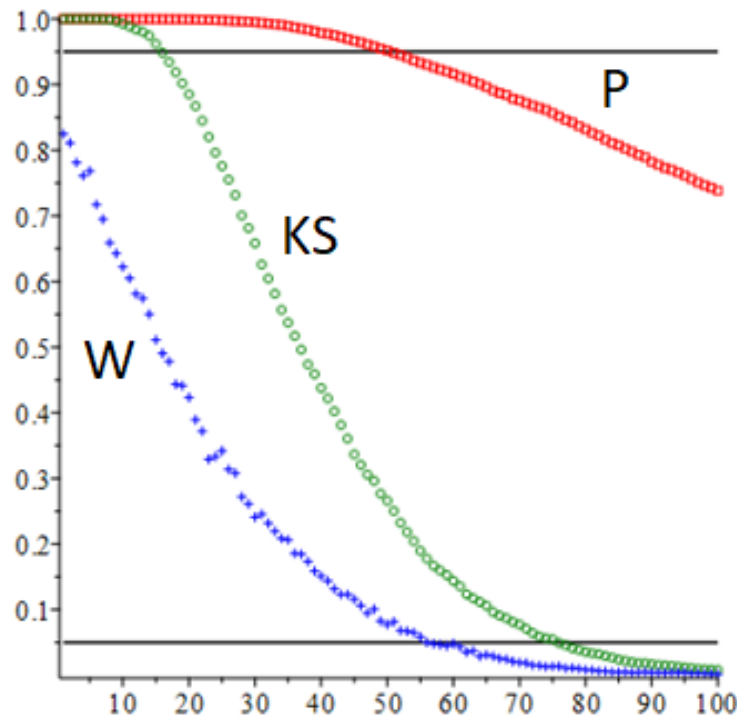


Рис. 3.1.20. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(2)$. Порогові точки: 51 для p -статистики, 76 для Колмогорова-Смірнова і 56 для Вілкоксона

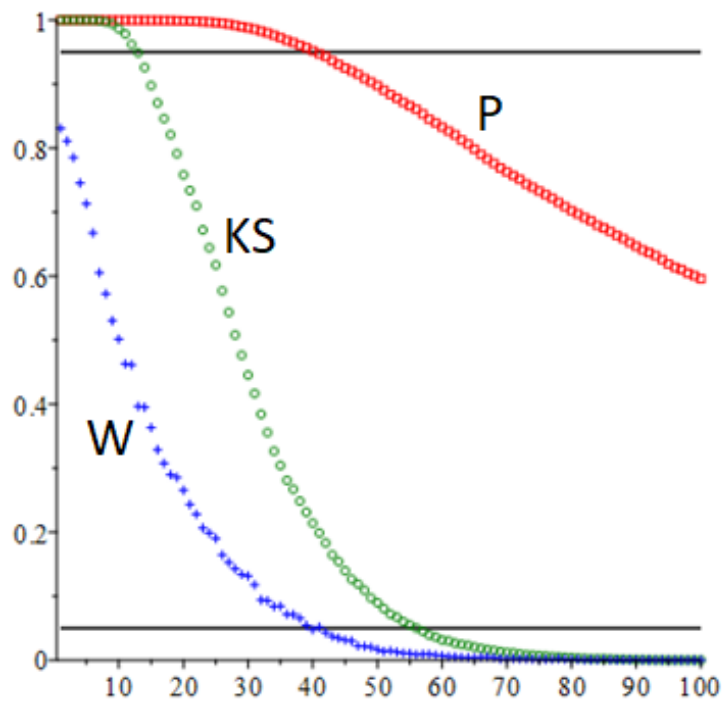


Рис. 3.1.21. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(2.5)$. Порогові точки: 41 для p -статистики, 57 для Колмогорова-Смірнова і 40 для Вілкоксона

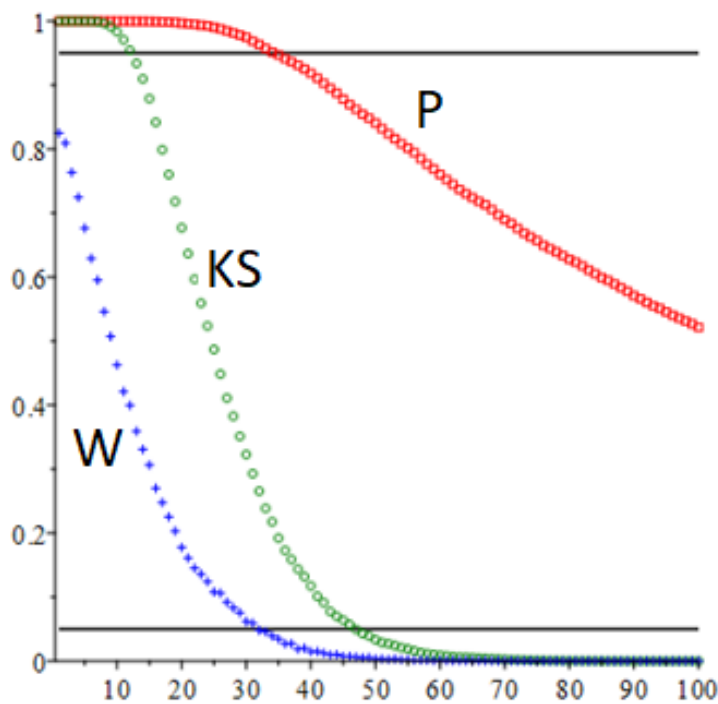


Рис. 3.1.22. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(3)$. Порогові точки: 35 для p -статистики, 47 для Колмогорова-Смірнова і 32 для Вілкоксона

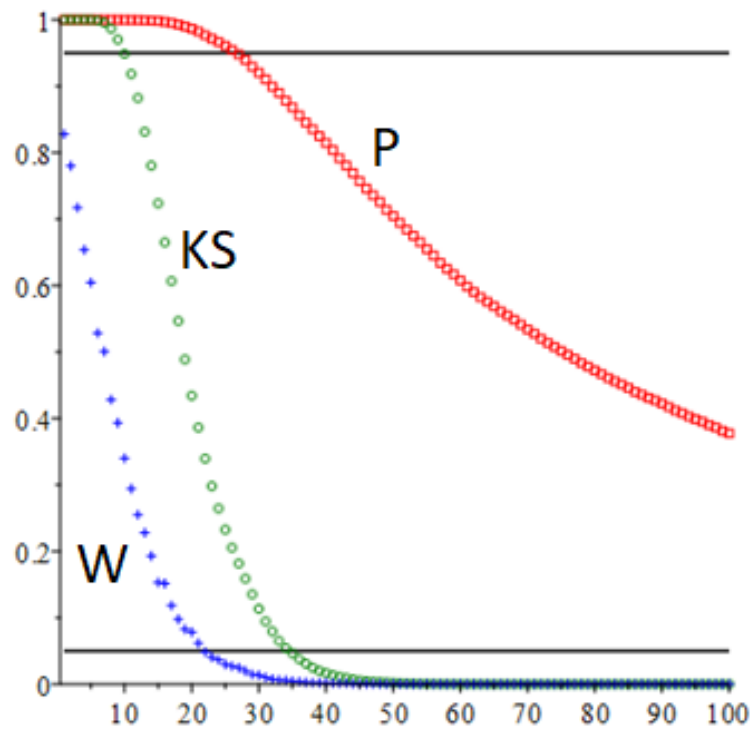


Рис. 3.1.23. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(5)$. Порогові точки: 27 для p -статистики, 35 для Колмогорова-Смірнова і 22 для Вілкоксона

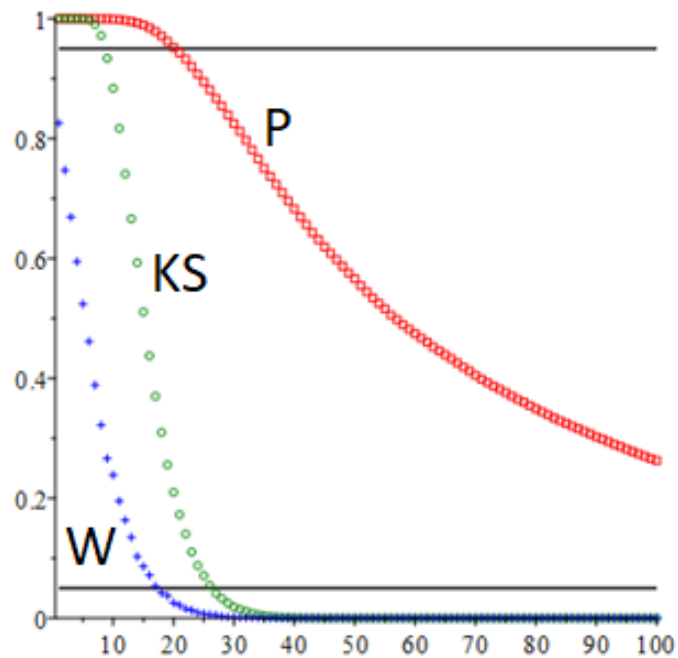


Рис. 3.1.24. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(10)$. Порогові точки: 21 для p -статистики, 27 для Колмогорова-Смірнова і 18 для Вілкоксона

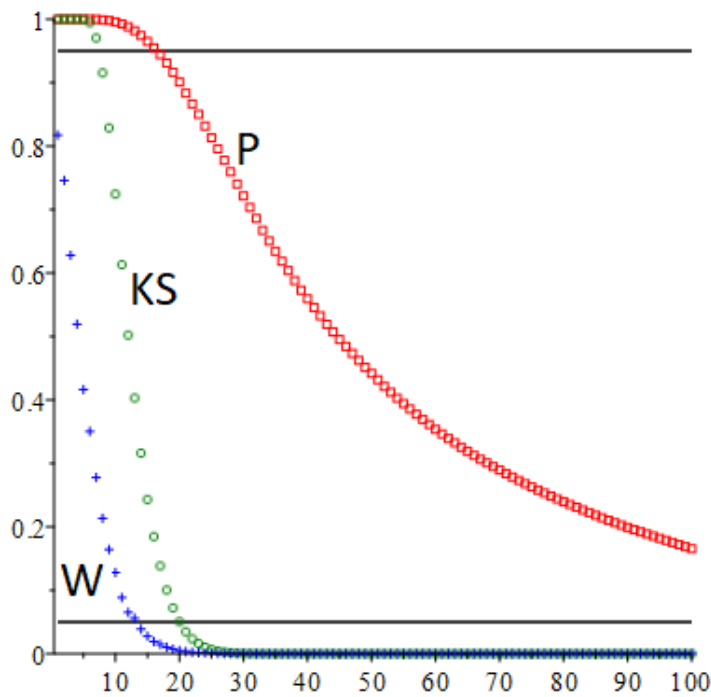


Рис. 3.1.25. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Exp(1)$ і $Exp(100)$. Порогові точки: 17 для p -статистики, 21 для Колмогорова-Смірнова і 14 для Вілкоксона

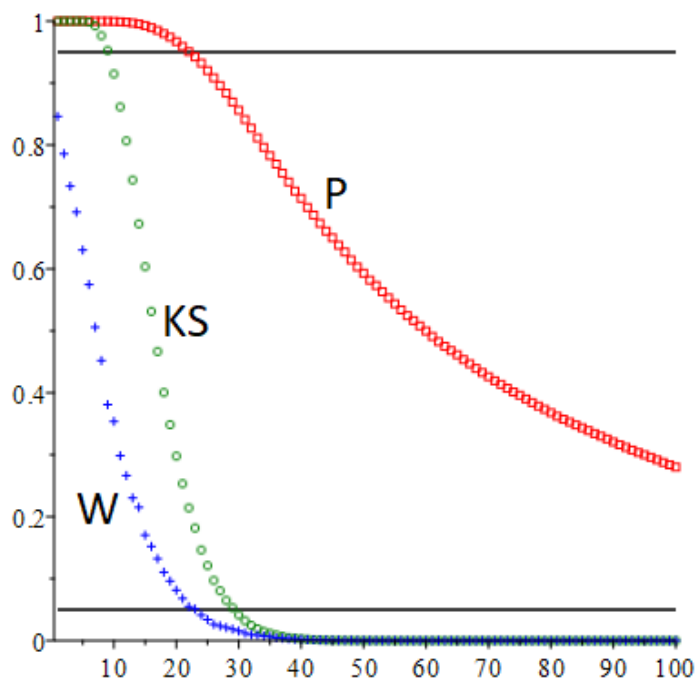


Рис. 3.1.26. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $Gamma(9,0.5)$ і $Gamma(1,2)$. Порогові точки: 23 для p -статистики, 30 для Колмогорова-Смірнова і 24 для Вілкоксона

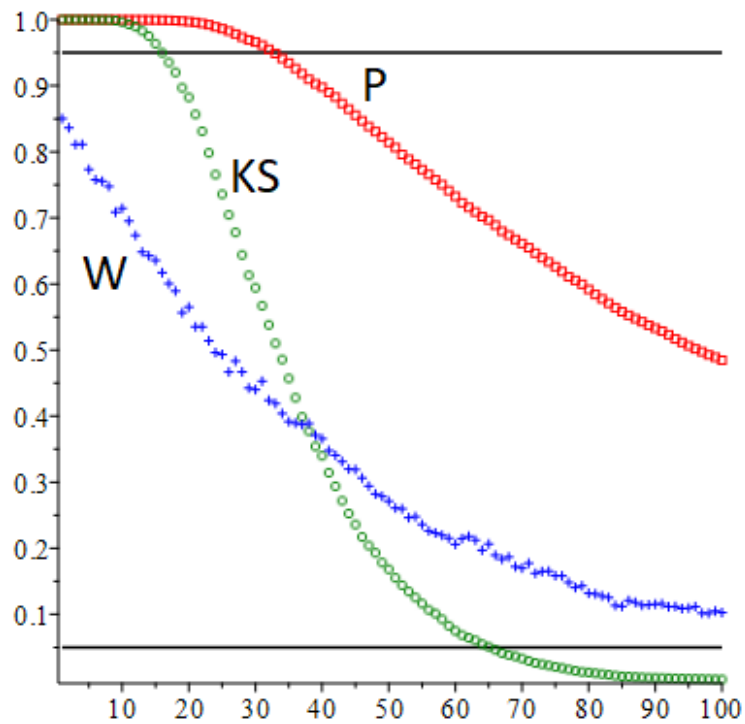


Рис. 3.1.27. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Gamma}(9,0.5)$ і $\text{Gamma}(2,2)$. Порогові точки: 33 для p -статистики, 66 для Колмогорова-Смірнова

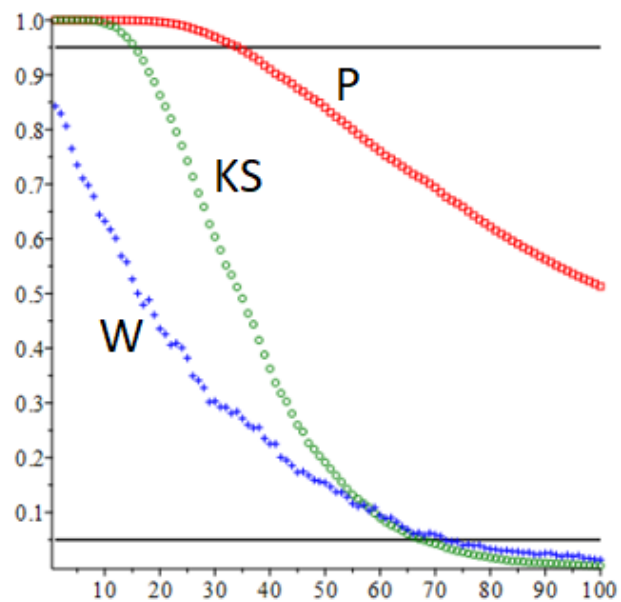


Рис. 3.1.28. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Gamma}(9,0.5)$ і $\text{Gamma}(3,2)$. Порогові точки: 35 для p -статистики, 68 для Колмогорова-Смірнова і 72 для Вілкоксона

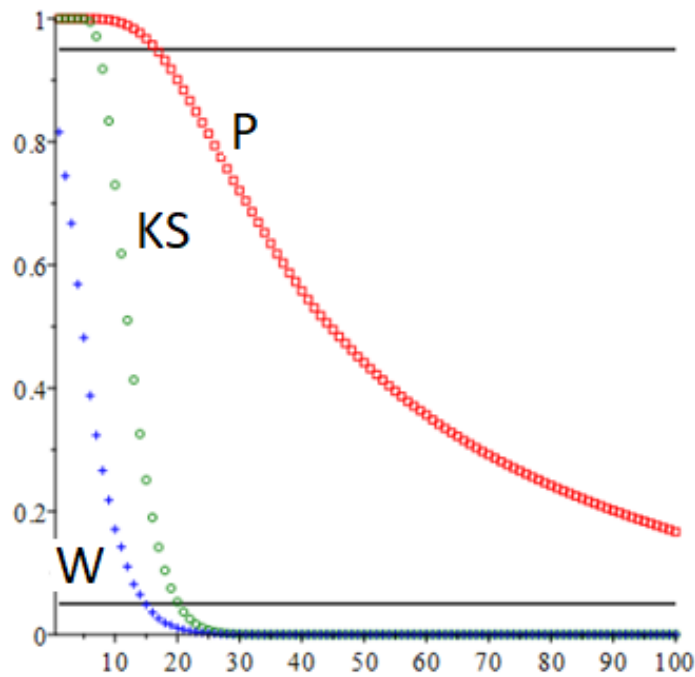


Рис. 3.1.29. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Gamma}(9,0.5)$ і $\text{Gamma}(0.5,1)$. Порогові точки: 17 для p -статистики, 21 для Колмогорова-Смірнова і 15 для Вілкоксона

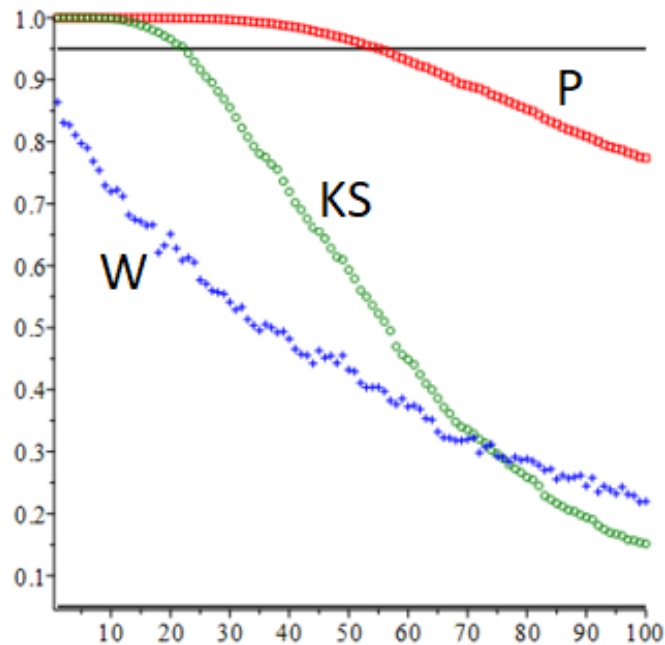


Рис. 3.1.30. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Gamma}(9,0.5)$ і $\text{Gamma}(5,1)$. Порогові точки: 56 для p -статистики

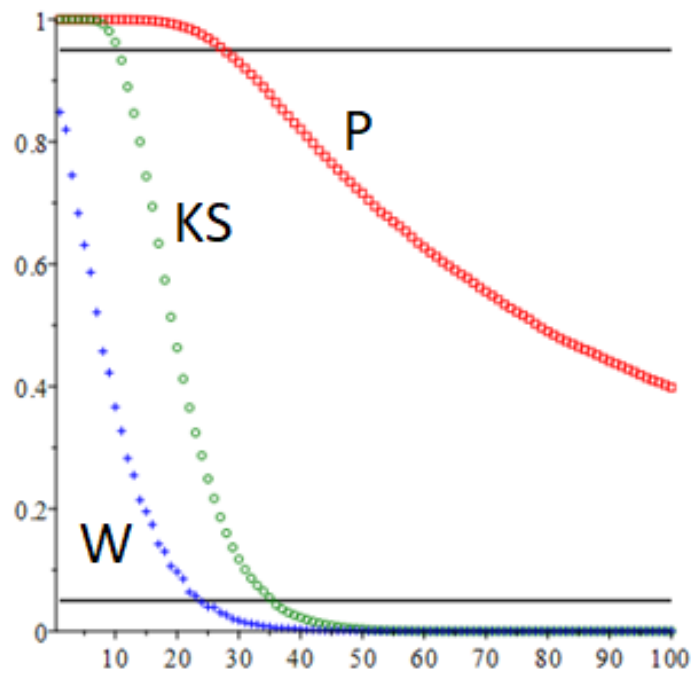


Рис. 3.1.31. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $\text{Gamma}(9,0.5)$ і $\text{Gamma}(7.5,1)$. Порогові точки: 28 для p -статистики, 36 для Колмогорова-Смірнова і 24 для Вілкоксона

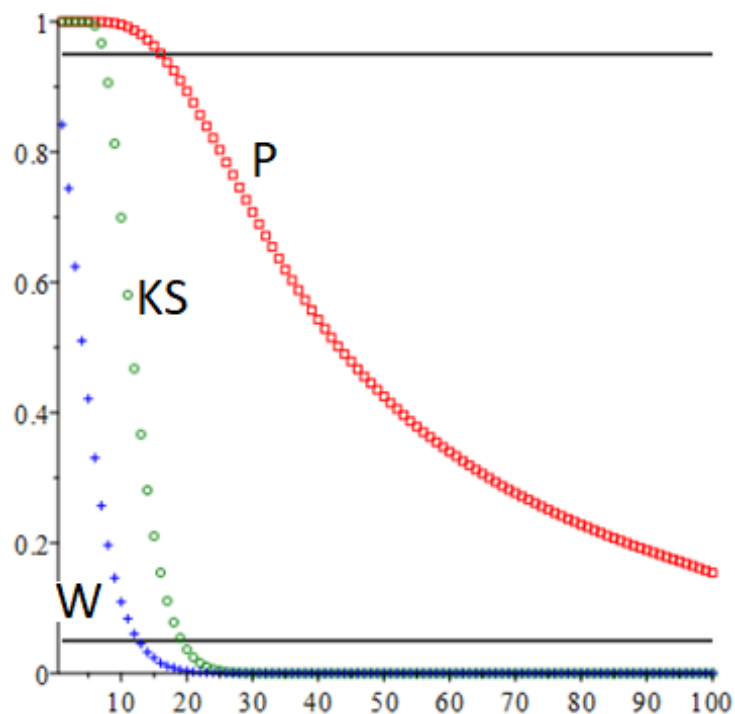


Рис. 3.1.32. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $U(0,1)$ і $U(2,3)$. Порогові точки: 17 для p -статистики, 20 для Колмогорова-Смірнова і 13 для Вілкоксона

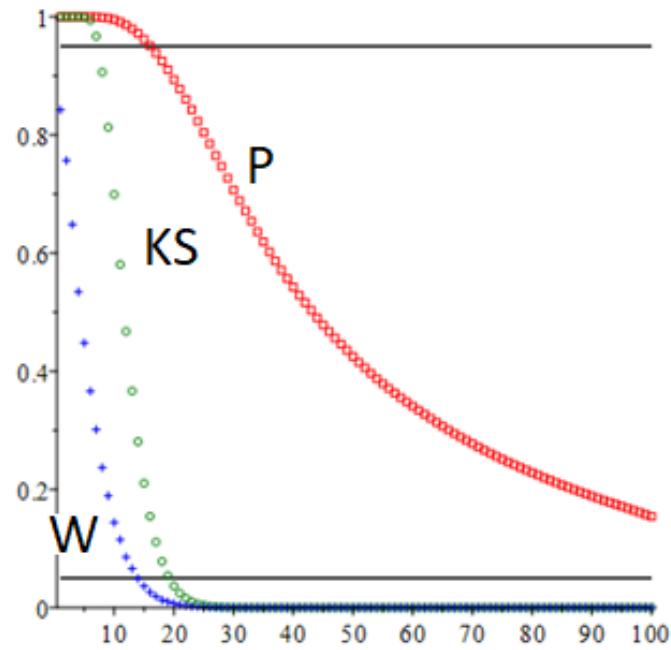


Рис. 3.1.33. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $U(0,1)$ і $U(1,2)$. Порогові точки: 17 для p -статистики, 20 для Колмогорова-Смірнова і 14 для Вілкоксона

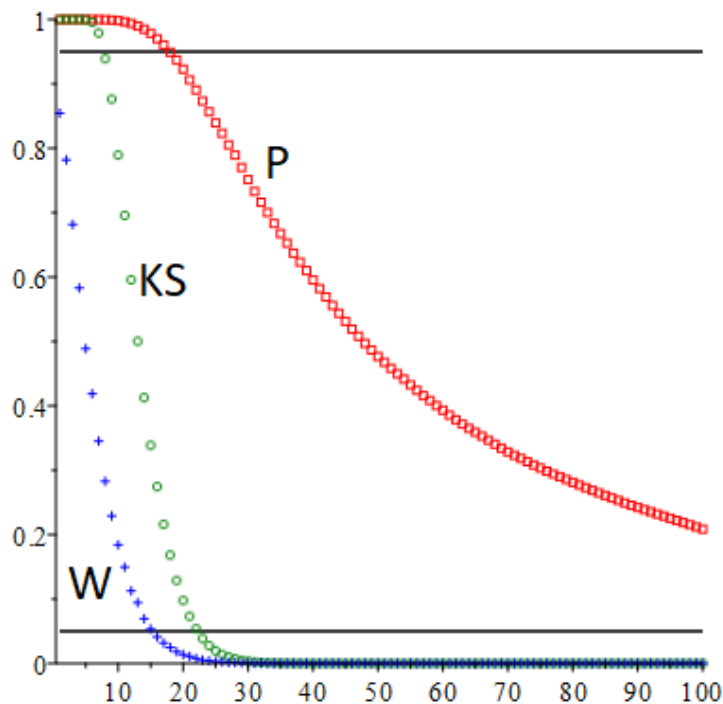


Рис. 3.1.34. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $U(0,1)$ і $U(0.8,1.8)$. Порогові точки: 18 для p -статистики, 23 для Колмогорова-Смірнова і 16 для Вілкоксона

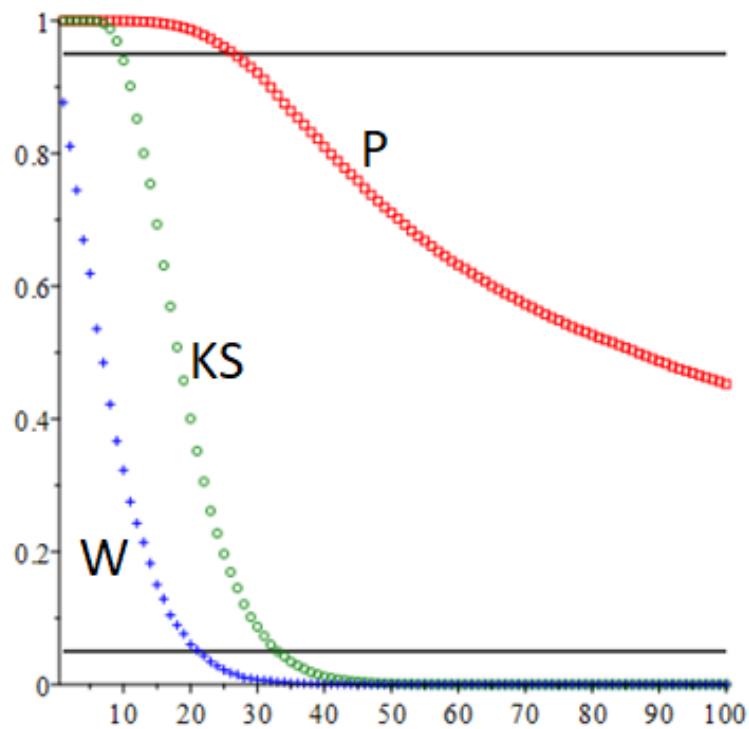


Рис. 3.1.35. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $U(0,1)$ і $U(0.5,1.5)$. Порогові точки: 27 для p -статистики, 33 для Колмогорова-Смірнова і 22 для Вілкоксона

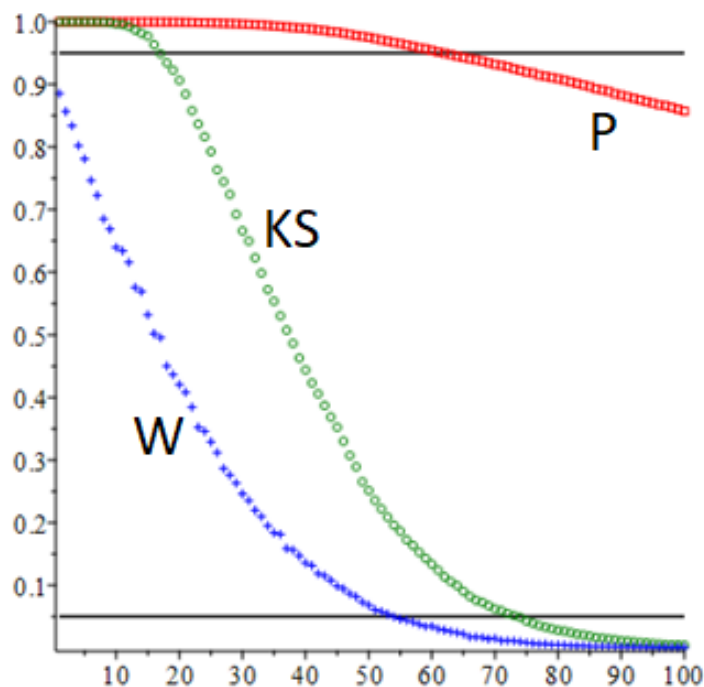


Рис. 3.1.36. Динаміка значень p -статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $U(0,1)$ і $U(0.2,1.2)$. Порогові точки: 62 для p -статистики, 74 для Колмогорова-Смірнова і 54 для Вілкоксона

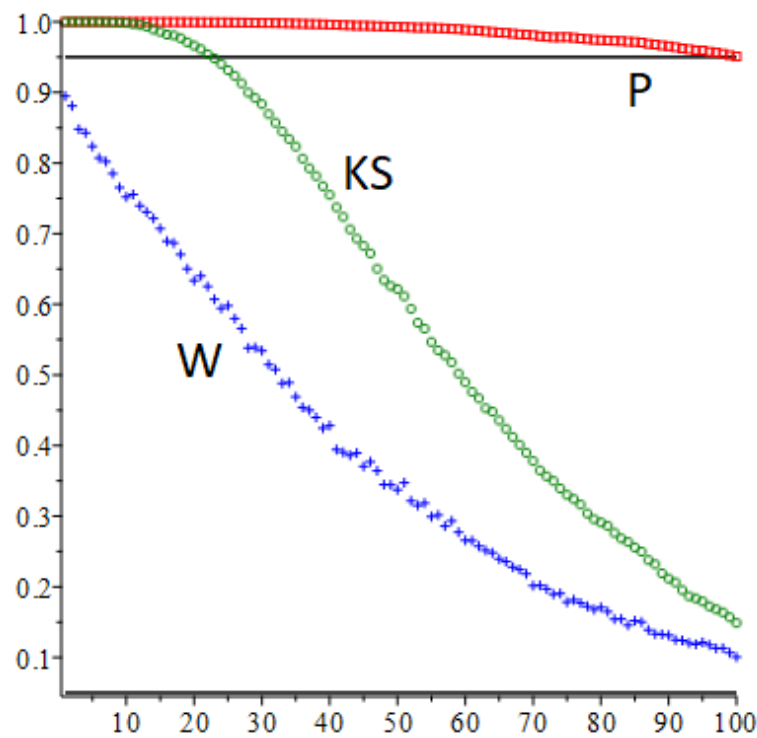


Рис. 3.1.37. Динаміка значень р-статистики Петуніна ($\square$), Колмогорова-Смірнова ($\circ$) та Вілкоксона ($\times$) для порівняння $U(0,1)$ і $U(0.1,1.1)$. Порогових точок немає

Для узагальнення результатів було складено таблицю порогових точок. У ній наведено значення номерів кроків, на яких кожна зі статистик уперше перетинала своє критичне значення (0.95 для р-статистики та 0.05 для двох інших).

№	Розподіл	Перетин P	Перетин KS	Перетин W
1	N(0,1) & N(1,1)	38	49	37
2	N(0,1) & N(2,1)	22	27	20
3	N(0,1) & N(3,1)	18	22	15
4	N(0,1) & N(10,1)	16	20	13
5	N(0,1) & N(0,2)	59	-	-
6	N(0,1) & N(0,3)	43	-	-
7	N(0,1) & N(0,4)	37	91	-
8	N(0,1) & N(0,10)	26	63	-
9	LogN(0,1)&LogN(1,1)	38	47	35
10	LogN(0,1)&LogN(2,1)	22	28	18
11	LogN(0,1)&LogN(3,1)	18	22	14
12	LogN(0,1)&LogN(0,10)	18	42	76
13	LogN(0,1)&LogN(0,1.5)	56	-	-

14	LogN(0,1)&LogN(0,0.75)	66	-	-
15	LogN(0,1) & LogN(0,0.5)	39	92	-
16	LogN(0,1)&LogN(0,0.25)	25	53	-
17	LogN(0,1)&LogN(0,0.125)	20	44	-
18	Exp(1) & Exp(0.5)	51	75	61
19	Exp(1) & Exp(1.5)	72	-	-
20	Exp(1) & Exp(2)	51	76	56
21	Exp(1) & Exp(2.5)	41	57	40
22	Exp(1) & Exp(3)	35	47	32
23	Exp(1) & Exp(5)	27	35	22
24	Exp(1) & Exp(10)	21	27	18
25	Exp(1) & Exp(100)	17	21	14
26	Gamma(9,0.5)&Gamma(1,2)	23	30	24
27	Gamma(9,0.5)&Gamma(2,2)	33	66	-
28	Gamma(9,0.5)&Gamma(3,2)	35	68	72
29	Gamma(9,0.5)&Gamma(0.5,1)	17	21	15
30	Gamma(9,0.5) & Gamma(5,1)	56	-	-
31	Gamma(9,0.5)&Gamma(7.5,1)	28	36	24
32	U(0,1) & U(2,3)	17	20	13
33	U(0,1) & U(1,2)	17	20	14
34	U(0,1) & U(0.8,1.8)	19	23	16
35	U(0,1) & U(0.5,1.5)	27	33	22
36	U(0,1) & U(0.2,1.2)	62	74	54
37	U(0,1) & U(0.1,1.1)	-	-	-

Таблиця 3.1.1 Порогові значення кроків, при яких статистики вперше перетинали свої критичні рівні для різних пар розподілів. Якщо у відповідній клітинці стоїть символ «—», це означає, що статистика не досягла критичного значення на жодному з кроків, і, відповідно, неможливо було відкинути гіпотезу про однорідність вибірок.

Висновки до підрозділу 3.1

У проведених експериментах було здійснено порівняння ефективності р-статистики Петуніна з класичними статистиками Колмогорова–Смірнова та Вілкоксона. Побудовані графіки та таблиці порогових точок дали можливість проаналізувати швидкість досягнення критичних значень кожною зі статистик.

Результати показали, що у низці випадків ефективність p -статистики була співставною з ефективністю статистики Вілкоксона. Це підтверджується близькістю порогових значень, наведених у підписах до графіків та у зведених таблиці. Водночас у тих експериментах, де статистики Колмогорова–Смірнова та Вілкоксона не перетинали свої критичні рівні, p -статистика демонструвала здатність досягати порогового значення відносно швидко. Більш того, у деяких випадках саме p -статистика виявлялася єдиною, що дозволяла відкинути нульову гіпотезу про однорідність вибірок.

Окремо слід відзначити стабільність p -статистики: на відміну від статистики Вілкоксона, вона не демонструє осциляцій, що підвищує її надійність у практичних застосуваннях.

Таким чином, можна зробити висновок, що p -статистика Петуніна має низку переваг у порівнянні з класичними критеріями — вона є швидшою у досягненні порогових значень, менш схильною до флуктуацій і здатною виявляти відмінності навіть у тих випадках, коли інші статистики виявляються неефективними. Це дозволяє розглядати її як перспективний інструмент для задач виявлення точок переходу у вибірках.

Водночас отримані результати створюють підґрунтя для подальших досліджень у багатовимірних випадках. Очікується, що ключові переваги p -статистики — відсутність осциляцій та швидке досягнення критичних значень — збережуться і в багатовимірних задачах, що робить її доцільним інструментом для аналізу часових рядів і багатовимірних даних у наступних підрозділах роботи.

3.2. Використання еліпсоїдів Петуніна

У цьому підрозділі розглянемо результати чисельних експериментів, проведених із застосуванням тривимірної версії алгоритму з використанням еліпсоїдів Петуніна для виявлення точок переходу у часових рядах. Метою дослідження є оцінка точності запропонованого підходу на даних, що

складаються з відрізків вибірок різних розподілів. Для цього було згенеровано три часові ряди T_1, T_2, T_3 , кожен з яких побудований як послідовність інтервалів із різними параметрами розподілу. Кожен експеримент повторювався 1000 разів, що дозволило оцінити статистичну стійкість результатів.

Для пошуку точок переходу виконаємо наступну послідовність дій з шириною вікна $window = 40$.

- 1) Візьмемо перші 40 трійок точок і побудуємо за ними еліпсоїд Петуніна.
- 2) Розглянемо точку на одну позицію праворуч і перевіримо, чи потрапила вона у побудований еліпсоїд.
 - 2.1) Якщо так, то просто рухаємось далі.
 - 2.2) Якщо ні, то цю точку називаємо викидом і теж рухаємось далі.
- 3) Коли ми знайшли два викиди поспіль, то другий з них називаємо точкою переходу і зміщуємо наше вікно на цю позицію, після чого продовжуємо наш алгоритм, поки не дійдемо до кінця даних.

Приклад 1. Рівномірні розподіли з різними математичними сподіваннями, що майже не перетинаються

У першому експерименті розглядалися майже неперекривні рівномірні розподіли з різними математичними сподіваннями (табл. 3.2.1). У часових рядах спостерігався чіткий «східчастий» характер зміни значень, що добре відповідає різким переходам між станами. Алгоритм успішно виявив зміни у точках [201, 401, 601], що збігається з реально закладеною структурою даних.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0–199	U(65;75)	U(96.5;97.5)	U(36.4;36.7)	[201, 401, 601]
200–399	U(100;110)	U(97.0;99.0)	U(38.0;39.0)	

400–599	$U(65;75)$	$U(96.5, 97.5)$	$U(36.4;36.7)$	
600–799	$U(70,90)$	$U(97.5;99.0)$	$U(37.0;37.5)$	

Таблиця 3.2.1. Часові інтервали і рівномірні розподіли з різними середніми

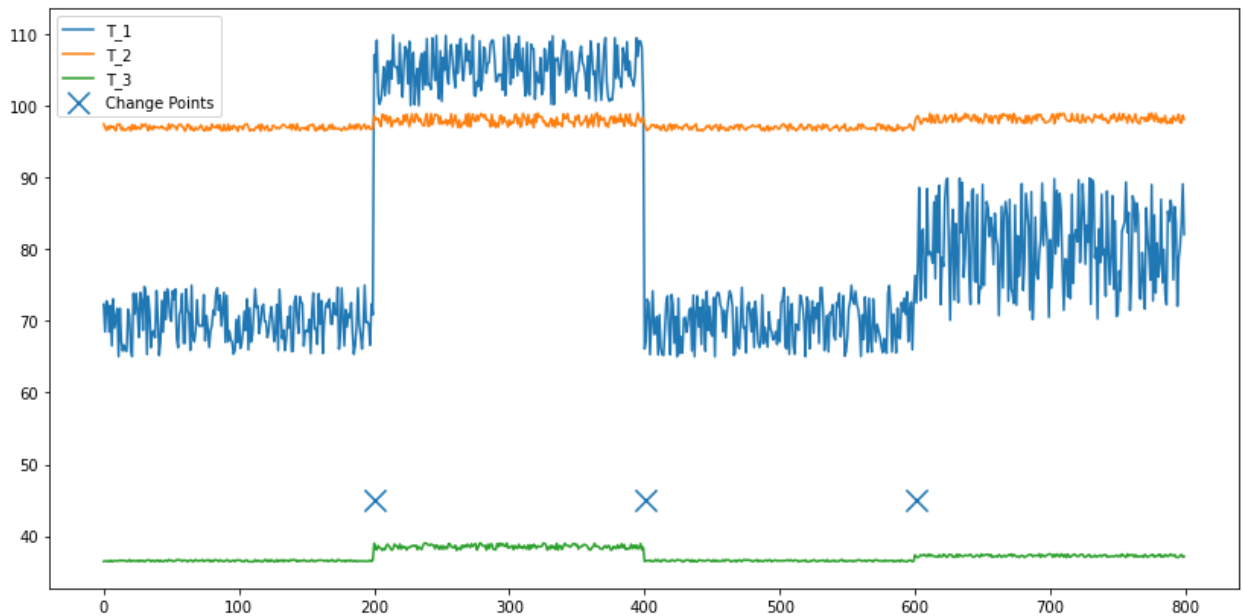


Рис. 3.2.1. Приклад часового ряду, що складається з вибірок майже неперекривних рівномірних розподілів з різними середніми значеннями та їх точками переходів

Приклад 2. Рівномірні розподіли з різними математичними сподіваннями, що спочатку сильно перетинаються, потім трошки перетинаються, а в кінці зовсім не перетинаються

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$U[60; 70]$	$U[96; 97]$	$U[36.4; 36.7]$	[201, 401, 601]
200 – 399	$U[63; 73]$	$U[96.3; 97.3]$	$U[36.5; 36.8]$	
400 – 599	$U[70; 80]$	$U[97; 98]$	$U[36.7; 37]$	
600 – 799	$U[85; 95]$	$U[99; 100]$	$U[37.5; 37.8]$	

Таблиця 3.2.2. Часові інтервали і рівномірні розподіли з різними математичними сподіваннями, що спочатку сильно перетинаються, потім трошки перетинаються, а в кінці зовсім не перетинаються

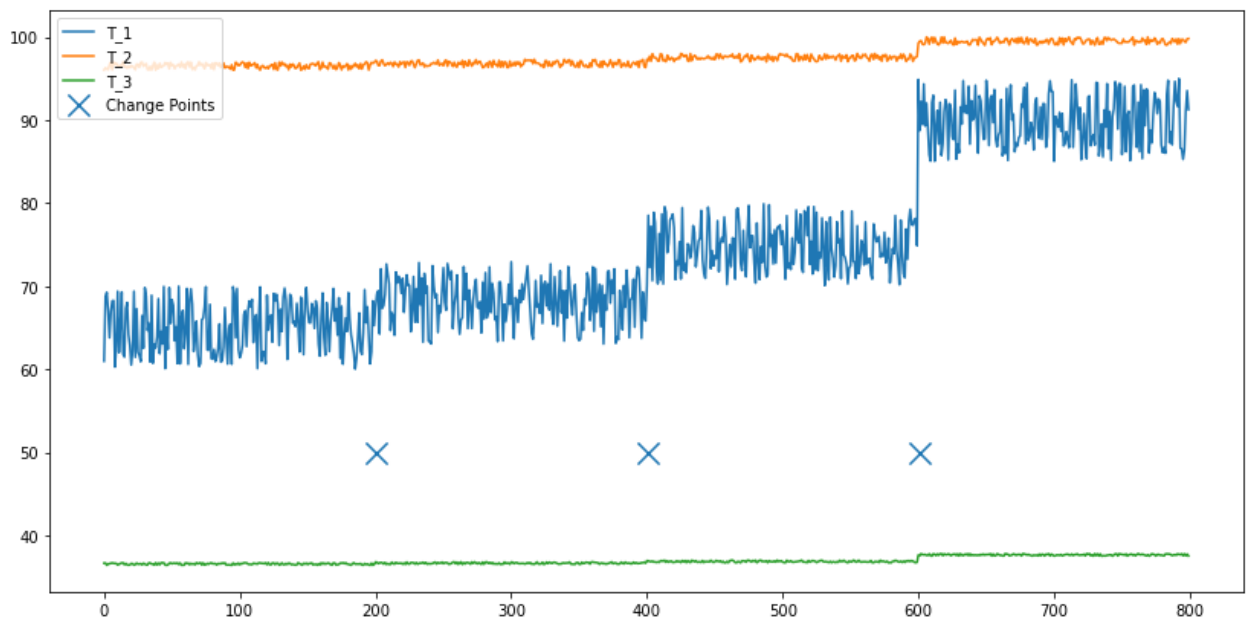


Рис. 3.2.2. Приклад часового ряду, що складається з вибірок майже неперекривних рівномірних розподілів з різними математичними сподіваннями, що спочатку сильно перетинаються, потім трошки перетинаються, а в кінці зовсім не перетинаються та їх точками переходів

Приклад 3. Нормальні розподіли з різними математичними сподіваннями, що майже взагалі не перетинаються

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$N(70; 2)$	$N(96; 0.15)$	$N(36.55; 0.05)$	[201, 401, 601]
200 – 399	$N(105; 2)$	$N(98; 0.33)$	$N(38.5; 0.15)$	
400 – 599	$N(70; 2)$	$N(96; 0.15)$	$N(36.55; 0.05)$	
600 – 799	$N(80; 4)$	$N(98.25; 0.25)$	$N(37.25; 0.08)$	

Таблиця 3.2.3. Часові інтервали і нормальні розподіли з різними математичними сподіваннями, що майже взагалі не перетинаються

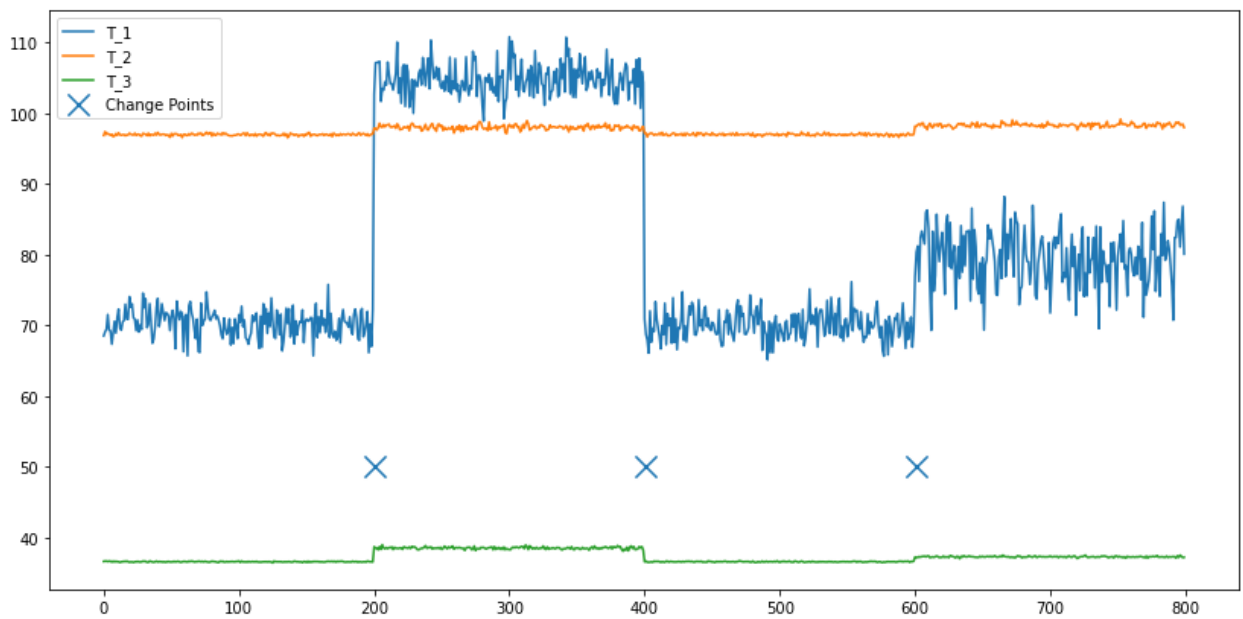


Рис. 3.2.3. Приклад часового ряду, що складається з вибірок нормальних розподілів з різними математичними сподіваннями, що майже взагалі не перетинаються та їх точками переходів

Приклад 4. Нормальні розподіли з однаковими математичними сподіваннями, але з дисперсіями, що поступово починають відрізнятися.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$N(70; 1)$	$N(97; 0.1)$	$N(36.55; 0.05)$	[203, 411, 632]
200 – 399	$N(70; 2)$	$N(97; 0.15)$	$N(36.55; 0.1)$	
400 – 599	$N(70; 3)$	$N(97; 0.2)$	$N(36.55; 0.15)$	
600 – 799	$N(70; 5)$	$N(97; 0.3)$	$N(36.55; 0.2)$	

Таблиця 3.2.4. Часові інтервали і нормальні розподіли з однаковими математичними сподіваннями, але з дисперсіями, що поступово починають відрізнятися

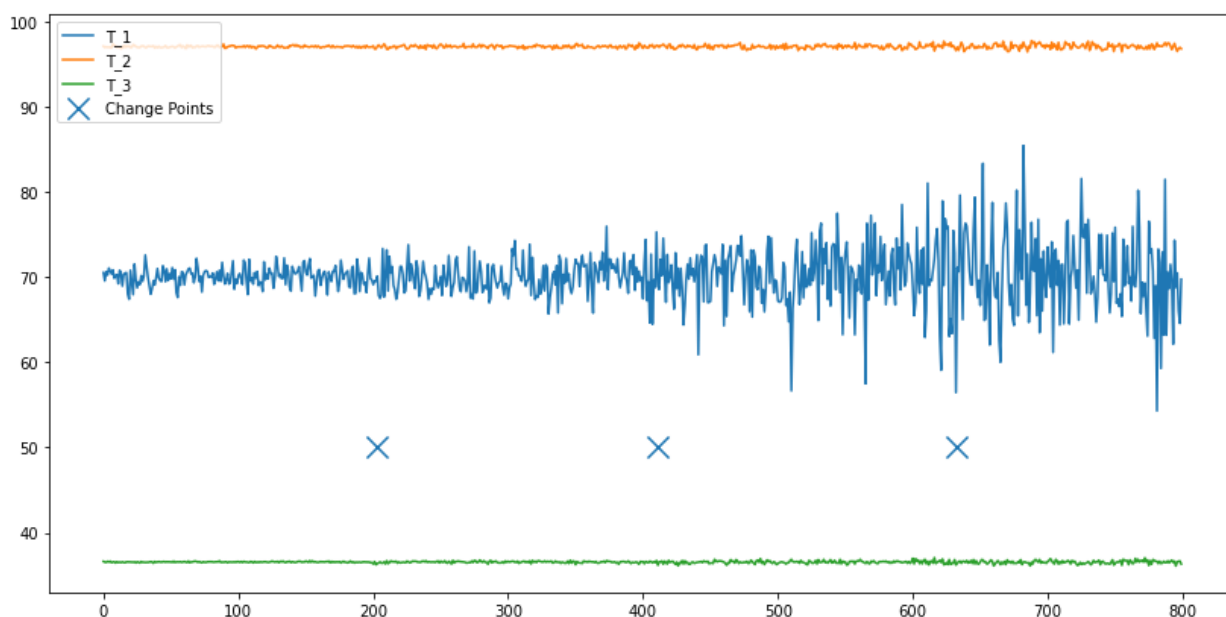


Рис. 3.2.4. Приклад часового ряду, що складається з вибірок нормальних розподілів з однаковими математичними сподіваннями, але з дисперсіями, що поступово починають відрізнятися та їх точками переходів

Приклад 5. Нормальні розподіли з однаковими математичними сподіваннями, але з дисперсіями, що сильніше відрізняються одна від одної

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$N(70; 1)$	$N(97; 0.1)$	$N(36.55; 0.05)$	[203, 401, 615]
200 – 399	$N(70; 5)$	$N(97; 0.5)$	$N(36.55; 0.25)$	
400 – 599	$N(70; 7)$	$N(97; 1)$	$N(36.55; 0.5)$	
600 – 799	$N(70; 10)$	$N(97; 1.5)$	$N(36.55; 0.75)$	

Таблиця 3.2.5. Часові інтервали і нормальні розподіли з однаковими математичними сподіваннями, але з дисперсіями, що сильніше відрізняються одна від одної

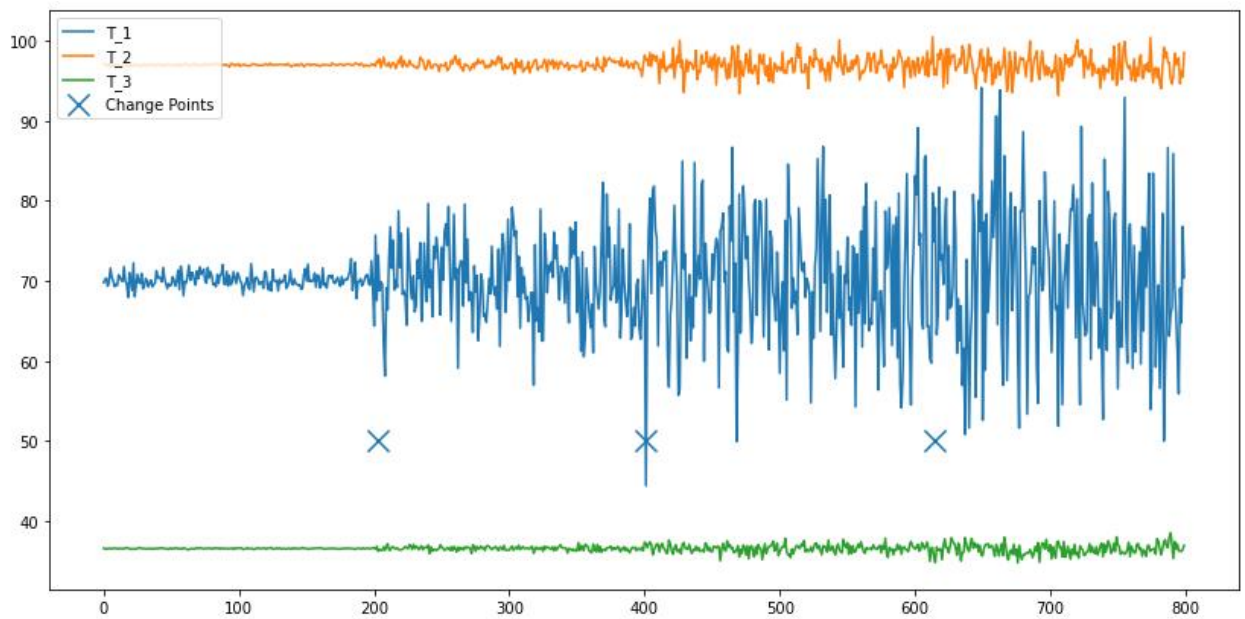


Рис. 3.2.5. Приклад часового ряду, що складається з вибірок нормальних розподілів з однаковими математичними сподіваннями, але з дисперсіями, що сильніше відрізняються одна від одної та їх точками переходів

Приклад 6. Експоненціальні розподіли з різними параметрами

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$Exp(1)$	$Exp(1)$	$Exp(0.1)$	[208, 417, 628]
200 – 399	$Exp(3)$	$Exp(2)$	$Exp(0.2)$	
400 – 599	$Exp(7)$	$Exp(3)$	$Exp(0.3)$	
600 – 799	$Exp(10)$	$Exp(4)$	$Exp(0.4)$	

Таблиця 3.2.6. Часові інтервали і експоненціальні розподіли з різними параметрами

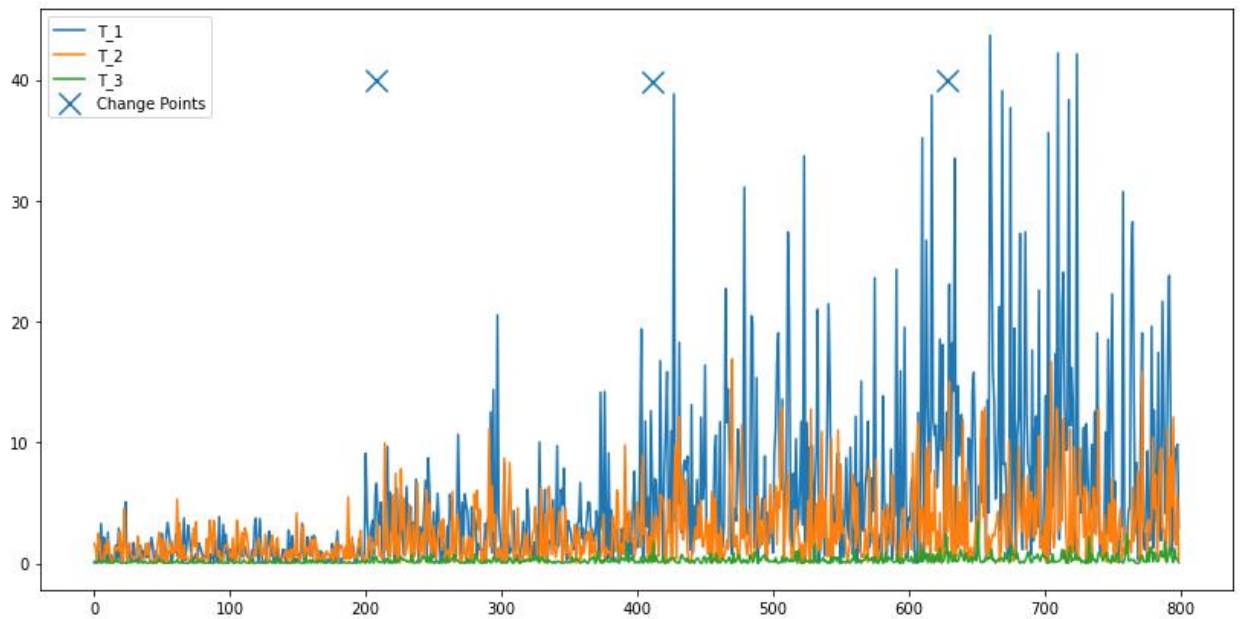


Рис. 3.2.6. Приклад часового ряду, що складається з вибірок експоненціальних розподілів з різними параметрами та їх точками переходів

Приклад 7. Гамма розподіли з однаковими математичними сподіваннями, але різними дисперсіями

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$\Gamma(12,1)$	$\Gamma(6,0.5)$	$\Gamma(10,0.1)$	(232, 430, 603)
200 – 399	$\Gamma(6,2)$	$\Gamma(3,1)$	$\Gamma(5,0.2)$	
400 – 599	$\Gamma(4,3)$	$\Gamma(1,3)$	$\Gamma(2,0.5)$	
600 – 799	$\Gamma(3,4)$	$\Gamma(0.5,6)$	$\Gamma(1,1)$	

Таблиця 3.2.7. Часові інтервали і гамма розподіли з однаковими математичними сподіваннями, але різними дисперсіями

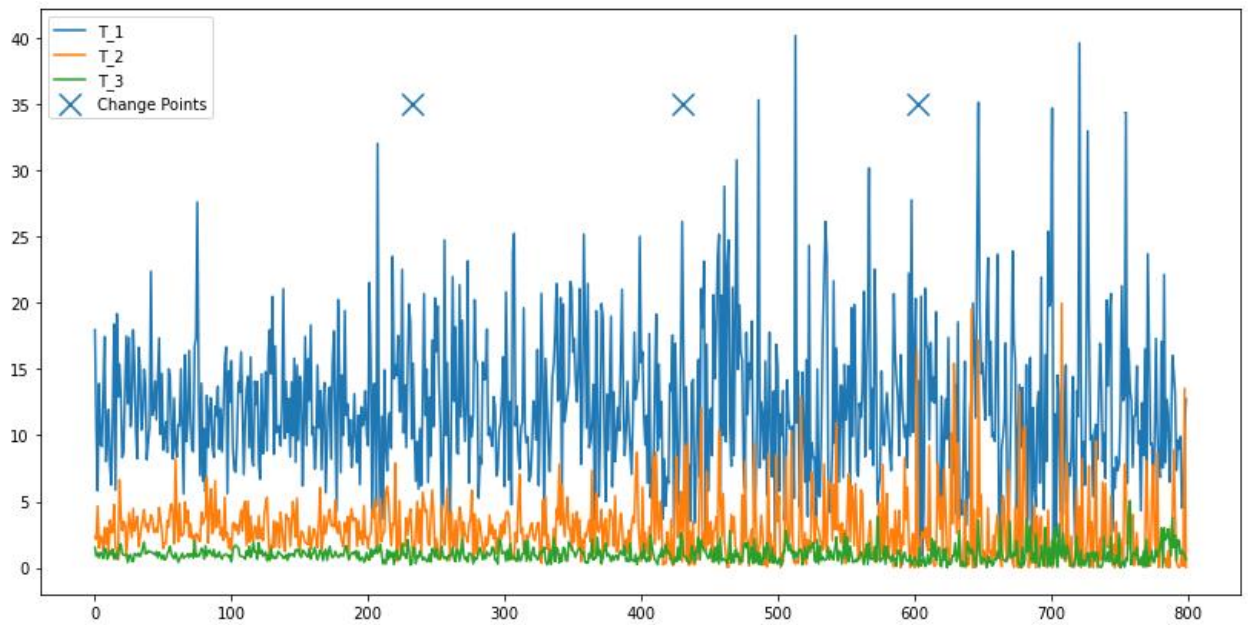


Рис. 3.2.7. Приклад часового ряду, що складається з вибірок гамма розподілів з однаковими математичними сподіваннями, але різними дисперсіями та їх точками переходів

Приклад 8. Гамма розподіли з різними математичними сподіваннями, але однаковими дисперсіями

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$\Gamma(36,1)$	$\Gamma(16,1)$	$\Gamma(4,1)$	[202, 419, 682]
200 – 399	$\Gamma(9,2)$	$\Gamma(4,2)$	$\Gamma(1,2)$	
400 – 599	$\Gamma(4,3)$	$\Gamma(1,4)$	$\Gamma(0.25,4)$	
600 – 799	$\Gamma(1,6)$	$\Gamma(0.25,8)$	$\Gamma(0.0625,8)$	

Таблиця 3.2.8. Часові інтервали і гамма розподіли з різними математичними сподіваннями, але однаковими дисперсіями

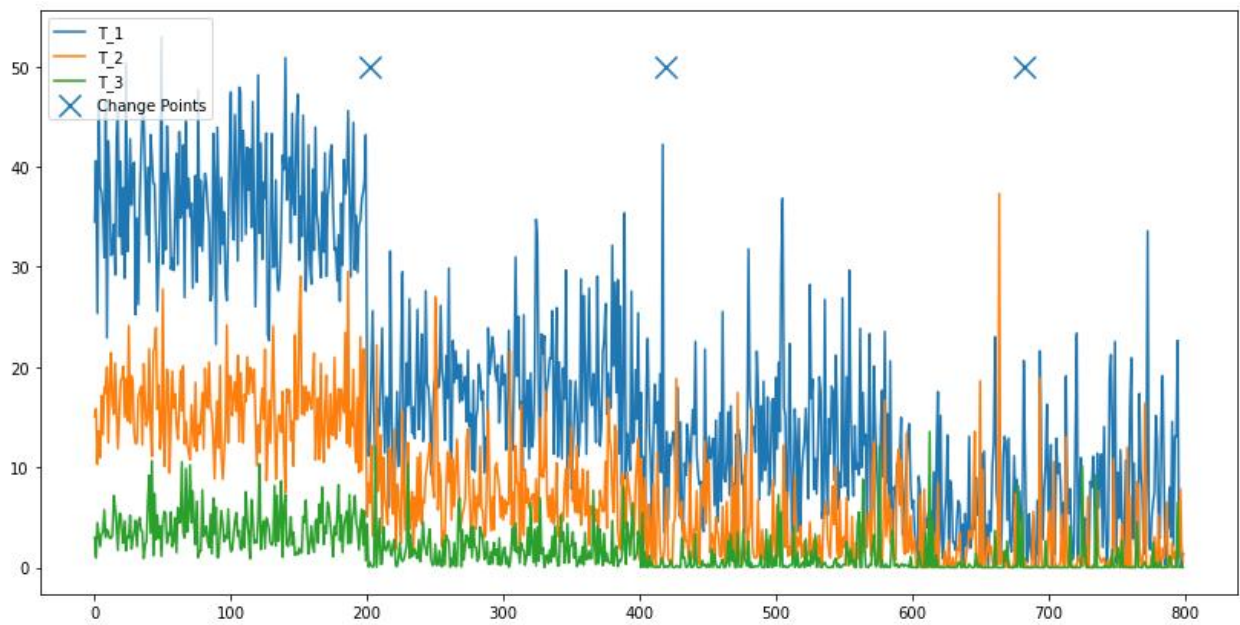


Рис. 3.2.8. Приклад часового ряду, що складається з вибірок гамма розподілів з різними математичними сподіваннями, але однаковими дисперсіями та їх точками переходів

Приклад 9. Розподіли Гумбеля з різними першими параметрами, але однаковими другими

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$G(50,1)$	$G(20,2)$	$G(5,5)$	[201, 401, 601]
200 – 399	$G(60,1)$	$G(25,2)$	$G(7,5)$	
400 – 599	$G(70,1)$	$G(30,2)$	$G(9,5)$	
600 – 799	$G(80,1)$	$G(35,2)$	$G(11,5)$	

Таблиця 3.2.9. Часові інтервали і розподіли Гумбеля з різними першими параметрами, але однаковими другими

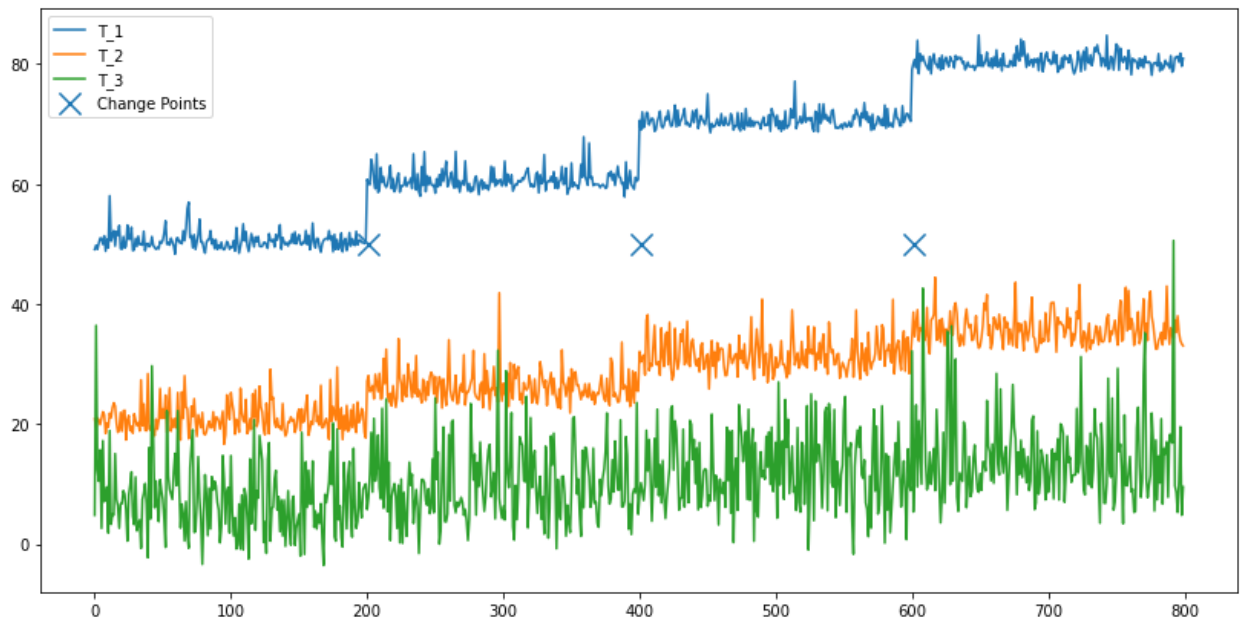


Рис. 3.2.9. Приклад часового ряду, що складається з вибірок розподілів Гумбеля з різними першими параметрами, але однаковими другими та їх точками переходів

Приклад 10. Розподіли Гумбеля з однаковими першими параметрами, але різними другими

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$G(60,1)$	$G(30,1)$	$G(10,1)$	[216, 414, 633]
200 – 399	$G(60,2)$	$G(30,2)$	$G(10,2)$	
400 – 599	$G(60,3)$	$G(30,3)$	$G(10,3)$	
600 – 799	$G(60,4)$	$G(30,4)$	$G(10,4)$	

Таблиця 3.2.10. Часові інтервали і розподіли Гумбеля з однаковими першими параметрами, але різними другими

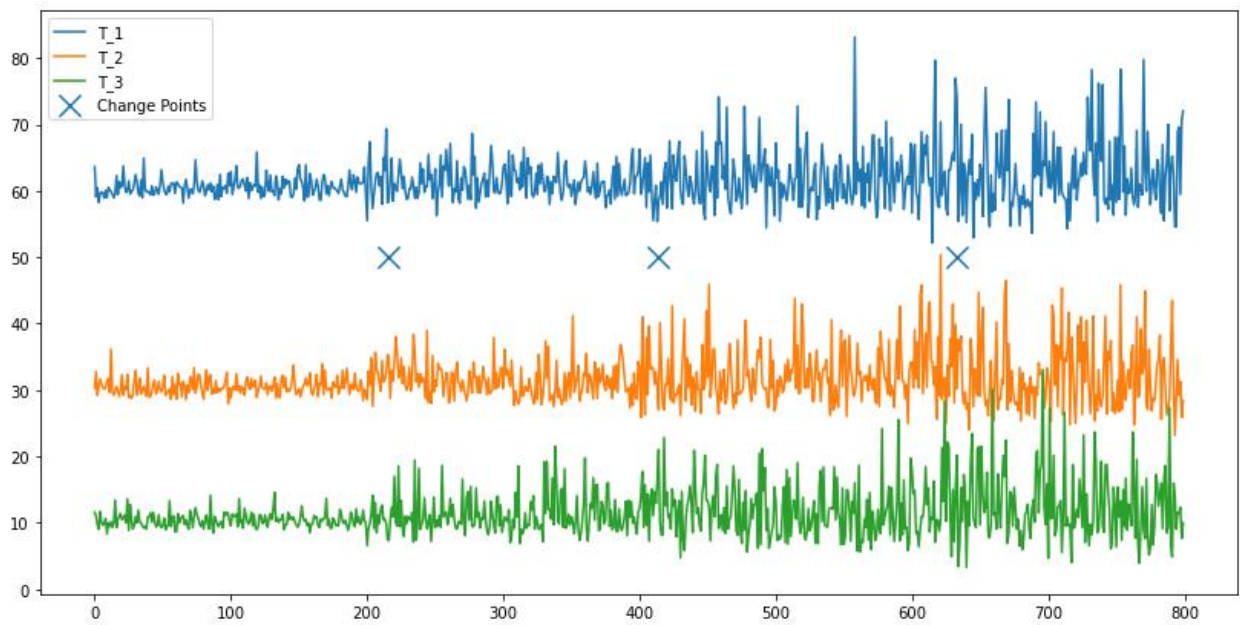


Рис. 3.2.10. Приклад часового ряду, що складається з вибірок розподілів Гумбеля з однаковими першими параметрами, але різними другими та їх точками переходів

Приклад 11. Розподіли Релея з різними параметрами

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$R(20)$	$R(10)$	$R(6)$	[208, 415, 603]
200 – 399	$R(30)$	$R(15)$	$R(8)$	
400 – 599	$R(40)$	$R(20)$	$R(10)$	
600 – 799	$R(50)$	$R(25)$	$R(12)$	

Таблиця 3.2.11. Часові інтервали і розподіли Релея з різними параметрами

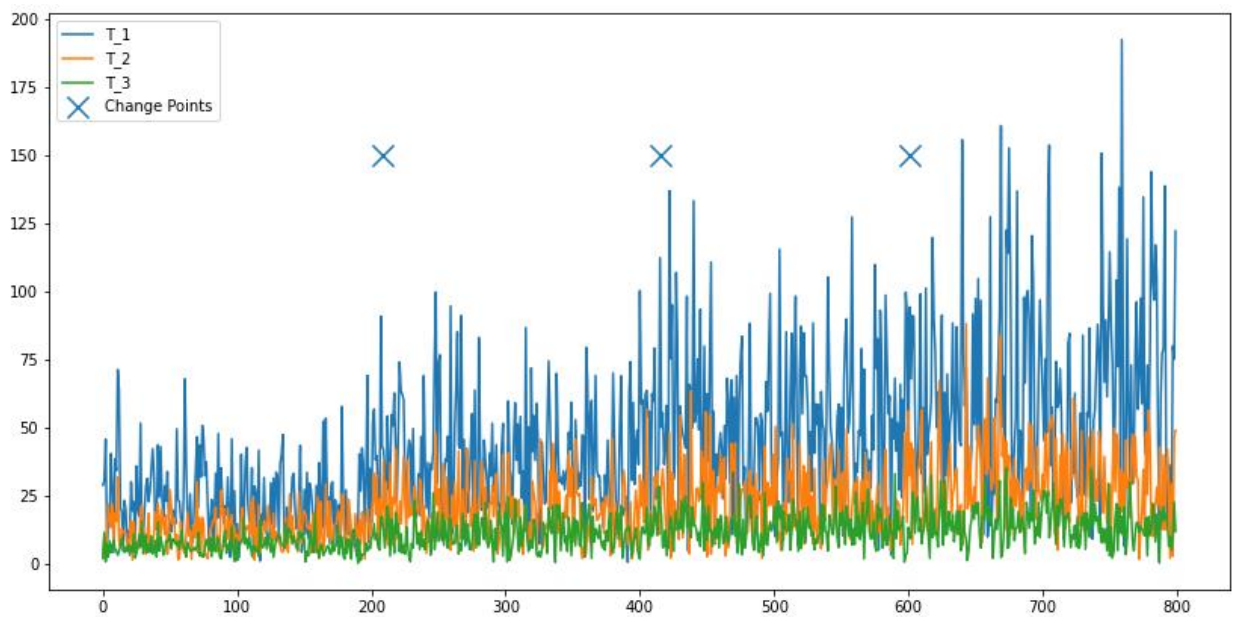


Рис. 3.2.11. Приклад часового ряду, що складається з вибірок розподілів Релея з різними параметрами та їх точками переходів

Приклад 12. Розподіли Лапласа з різними математичними сподіваннями, але однаковими дисперсіями

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$L(50,1)$	$L(20,2)$	$L(10,3)$	[201, 401, 601]
200 – 399	$L(60,1)$	$L(25,2)$	$L(12,3)$	
400 – 599	$L(70,1)$	$L(30,2)$	$L(14,3)$	
600 – 799	$L(80,1)$	$L(35,2)$	$L(16,3)$	

Таблиця 3.2.12. Часові інтервали і розподіли Лапласа з різними математичними сподіваннями, але однаковими дисперсіями

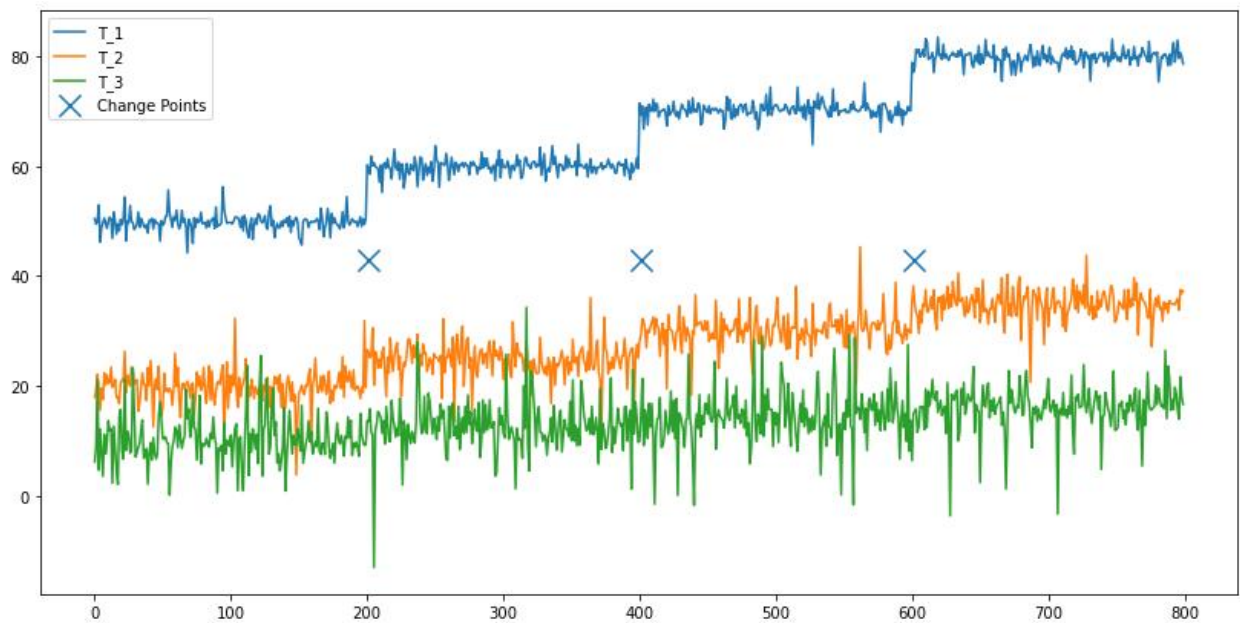


Рис. 3.2.12. Приклад часового ряду, що складається з вибірок розподілів Лапласа з різними математичними сподіваннями, але однаковими дисперсіями та їх точками переходів

Приклад 13. Розподіли Лапласа з однаковими математичними сподіваннями, але різними дисперсіями

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$L(60,1)$	$L(20,2)$	$L(0,3)$	[220, 408, 603]
200 – 399	$L(60,2)$	$L(20,3)$	$L(0,4)$	
400 – 599	$L(60,3)$	$L(20,4)$	$L(0,5)$	
600 – 799	$L(60,4)$	$L(20,5)$	$L(0,6)$	

Таблиця 3.2.13. Часові інтервали і розподіли Лапласа з однаковими математичними сподіваннями, але різними дисперсіями

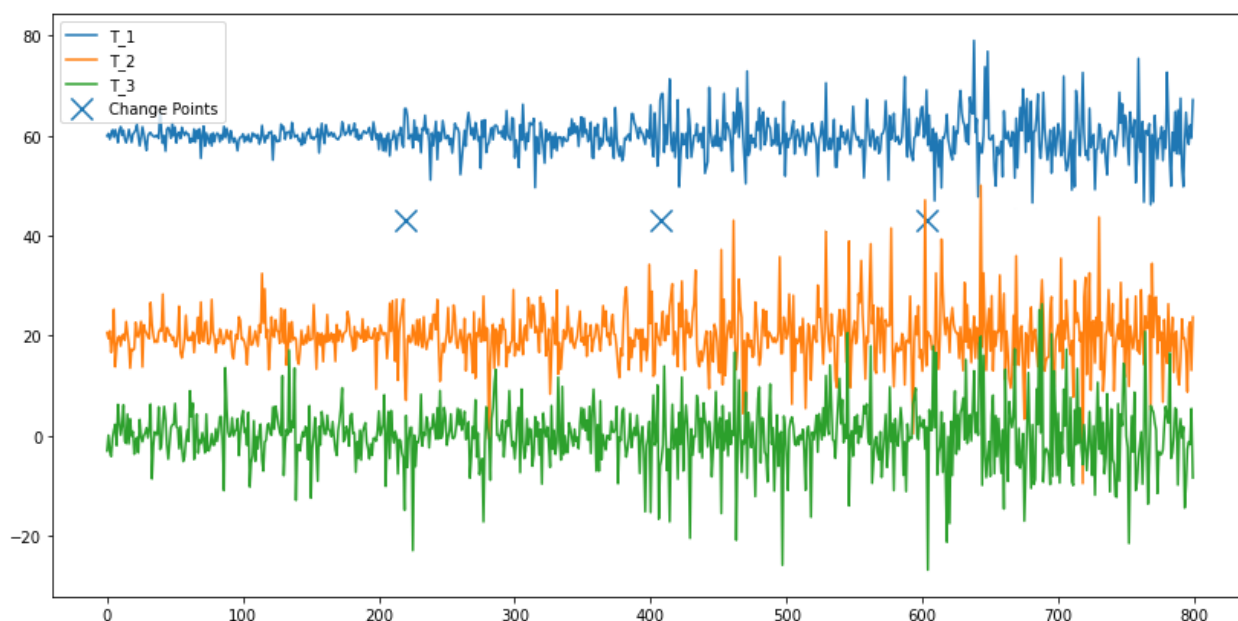


Рис. 3.2.13. Приклад часового ряду, що складається з вибірок розподілів Лапласа з однаковими математичними сподіваннями, але різними дисперсіями та їх точками переходів

Приклад 14. Логістичні розподіли з різними математичними сподіваннями, але однаковими дисперсіями

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$Logis(50,1)$	$Logis(20,2)$	$Logis(0,4)$	[201, 401, 601]
200 – 399	$Logis(60,1)$	$Logis(25,2)$	$Logis(2,4)$	
400 – 599	$Logis(70,1)$	$Logis(30,2)$	$Logis(4,4)$	
600 – 799	$Logis(80,1)$	$Logis(35,2)$	$Logis(6,4)$	

Таблиця 3.2.14. Часові інтервали і логістичні розподіли з різними математичними сподіваннями, але однаковими дисперсіями

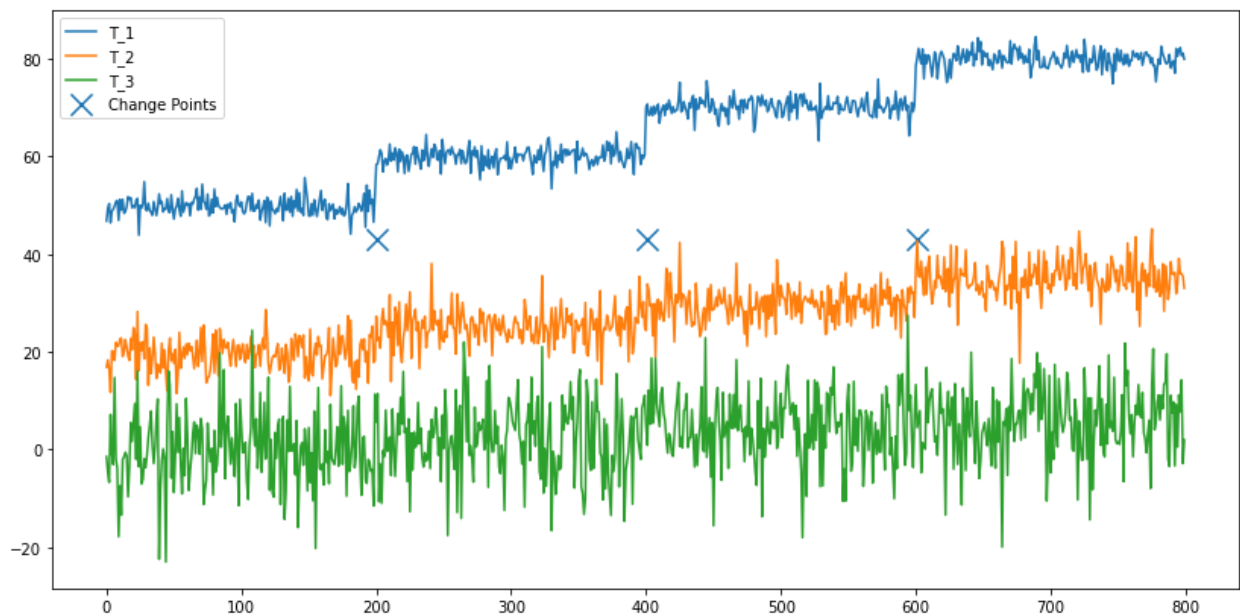


Рис. 3.2.14. Приклад часового ряду, що складається з вибірок логістичних розподілів з різними математичними сподіваннями, але однаковими дисперсіями та їх точками переходів

Приклад 15. Логістичні розподіли з однаковими математичними сподіваннями, але різними дисперсіями

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3	Знайдені точки переходу
0 – 199	$Logis(60,1)$	$Logis(20,2)$	$Logis(0,4)$	[227, 414, 616]
200 – 399	$Logis(60,2)$	$Logis(20,3)$	$Logis(0,5)$	
400 – 599	$Logis(60,3)$	$Logis(20,4)$	$Logis(0,6)$	
600 – 799	$Logis(60,4)$	$Logis(20,5)$	$Logis(0,7)$	

Таблиця 3.2.15. Часові інтервали і логістичні розподіли з однаковими математичними сподіваннями, але різними дисперсіями

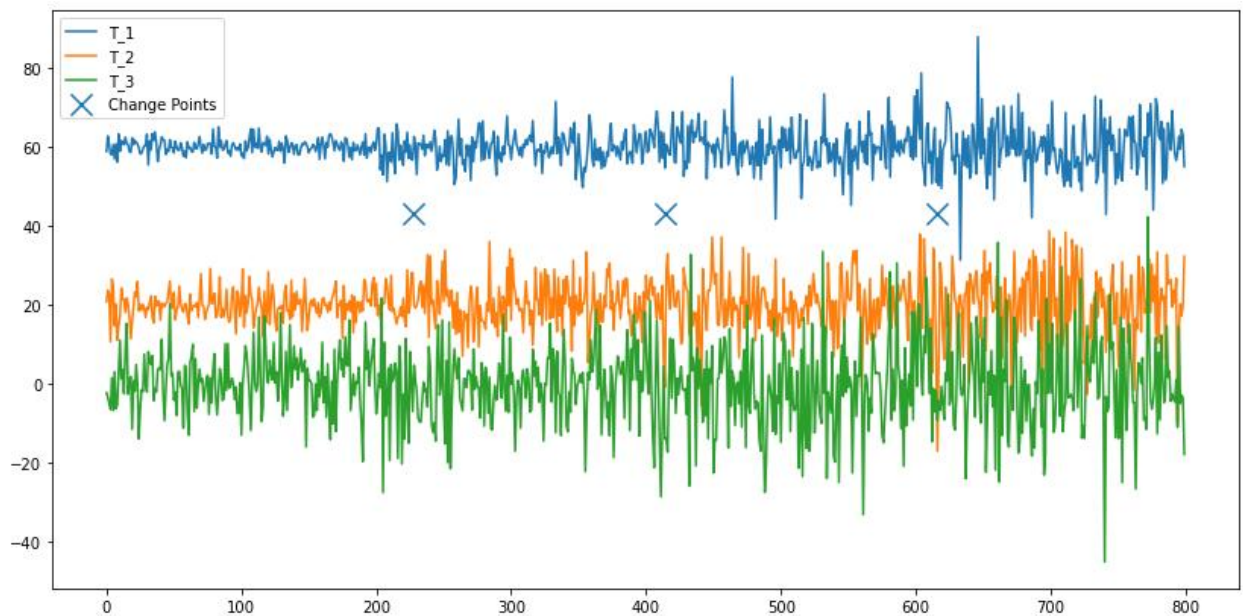


Рис. 3.2.15. Приклад часового ряду, що складається з вибірок логістичних розподілів з однаковими математичними сподіваннями, але різними дисперсіями та їх точками переходів

Висновки до підрозділу 3.2

У даному підрозділі було досліджено можливості використання еліпсоїдів Петуніна для виявлення точок переходу у багатовимірних часових рядах. Проведені числові та квазіреальні експерименти показали, що розроблений алгоритм забезпечує високу точність та стабільність при аналізі вибірок, що походять з широкого класу розподілів. Отримані результати свідчать, що метод залишається ефективним навіть за наявності розподілів з мінімальним перекриттям, різними параметрами та у випадку різкої зміни станів.

Особливо важливою перевагою є те, що метод на основі толерантних еліпсоїдів дозволяє формувати людинозрозумілу інтерпретацію результатів, тобто не лише ідентифікувати момент зміни, а й пояснити логіку побудови меж між класами. Це робить його цінним інструментом для сучасних систем пояснюваного штучного інтелекту (XAI), які застосовуються у практично орієнтованих сферах, зокрема у медицині.

Таким чином, використання еліпсоїдів Петуніна у задачах детекції точок переходу може вважатися ефективним, інтерпретованим і практично корисним інструментом, що поєднує теоретичну строгість із прикладною значущістю.

3.3. Дослідження поведінки p -статистики Петуніна у тривимірному випадку

У цьому підрозділі розглядається поведінка p -статистики Петуніна при застосуванні до тривимірних часових рядів. Основною метою є перевірка здатності алгоритму своєчасно ідентифікувати точки переходу та правильно відрізняти однорідні дані від неоднорідних у багатовимірному середовищі. Для цього було побудовано експерименти з часовими рядами, у яких розподіли змінювалися на різних інтервалах часу, що дозволяє оцінити чутливість методу до структурних змін.

Алгоритм базується на методі ковзного вікна з послідовним тестуванням гіпотези про однорідність вибірок. На кожному кроці будується лінійний дискримінант Фішера для двох послідовних підвибірок, після чого здійснюється проєкція багатовимірних даних на пряму. Отримані одновимірні дані використовуються для обчислення p -статистики Петуніна. Якщо значення статистики перевищує поріг 0.95, вважається, що вибірки належать одному розподілу, інакше фіксується зміна. У випадку, коли після певного елемента зразок стає неоднорідним, точка переходу визначається як наступний елемент ряду.

З метою ілюстрації роботи методу було згенеровано часовий ряд довжиною $N = 400$, поділений на чотири інтервали з різними розподілами. Алгоритм запускався багаторазово з подальшим усередненням результатів. Отримані значення p -статистики візуалізуються у вигляді двоколірного графіка: сині точки відповідають випадкам з однорідними вибірками ($p \geq 0.95$), а червоні точки — випадкам неоднорідності ($p < 0.95$). Такий підхід дозволяє не лише

кількісно оцінити ефективність алгоритму, а й надати наочне графічне підтвердження його роботи.

Приклад 1. Майже неперекривні рівномірні розподіли з різними середніми

Розглянемо стрибкоподібний часовий ряд, утворений рівномірними розподілами, які практично не перекриваються. На такому часовому ряді можна перевірити гіпотезу про зсув.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3
0-99	U(65;75)	U(96.5;97.5)	U(36.4;36.7)
100-199	U(100;110)	U(97.0;99.0)	U(38.0;39.0)
200-299	U(65;75)	U(96.5, 97.5)	U(36.4;36.7)
300-399	U(70,90)	U(97.5;99.0)	U(37.0;37.5)

Таблиця 3.3.1. Часові інтервали і майже неперекривні рівномірні розподіли з різними середніми

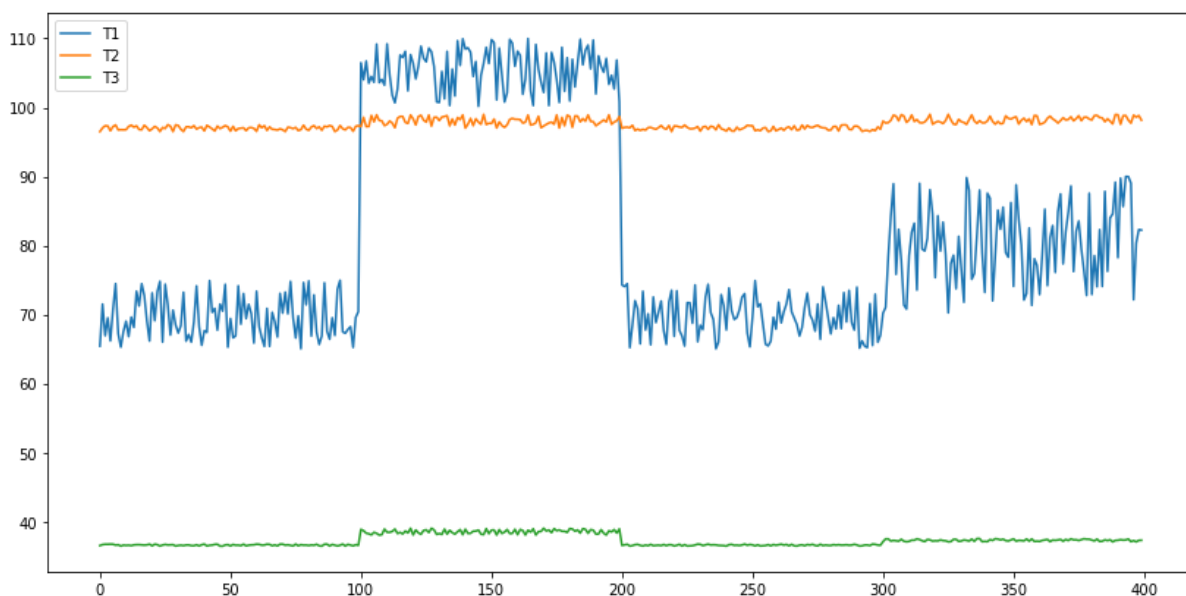


Рис. 3.3.1. Приклад часового ряду, що складається з вибірок майже неперекривних рівномірних розподілів з різними середніми

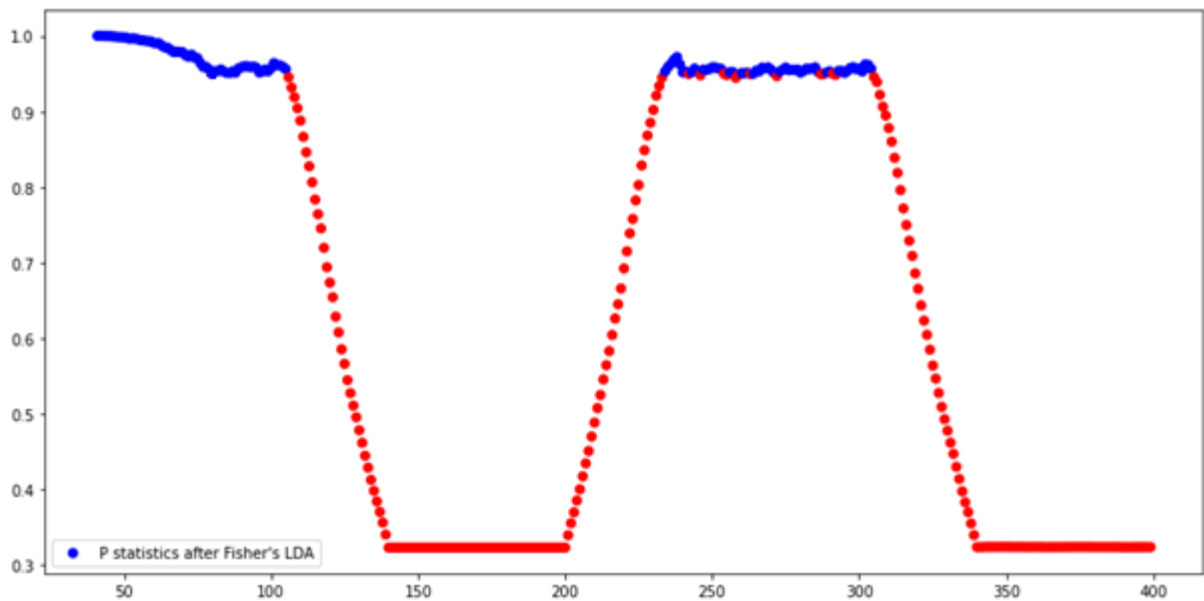


Рис. 3.3.2. Значення р-статистики для майже неперекривних рівномірних розподілів з різними середніми. Сині точки відповідають випадкам, коли значення не менше 0.95, а червоні — протилежним, коли вони є меншими

Приклад 2. Рівномірні розподіли з різними середніми значеннями, які спочатку сильно перекриваються, потім незначно перекриваються і, нарешті, не перекриваються

Розглянемо стрибкоподібний часовий ряд, який складається з рівномірних розподілів, які спочатку сильно перекриваються, потім незначно перекриваються і, нарешті, не перекриваються. На цьому часовому ряді ми зможемо перевірити гіпотезу про зсув.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3
0-99	U(60;70)	U(96.0;97.0)	U(36.4;36.7)
100-199	U(63;73)	U(96.3;97.3)	U(36.5;36.8)
200-299	U(70;80)	U(97.0, 98.0)	U(36.7;37.0)
300-399	U(85,95)	U(99.0;99.9)	U(37.5;37.8)

Таблиця 3.3.2. Часові інтервали і рівномірні розподіли з різними середніми значеннями, які спочатку сильно перекриваються, потім незначно перекриваються і, нарешті, не перекриваються

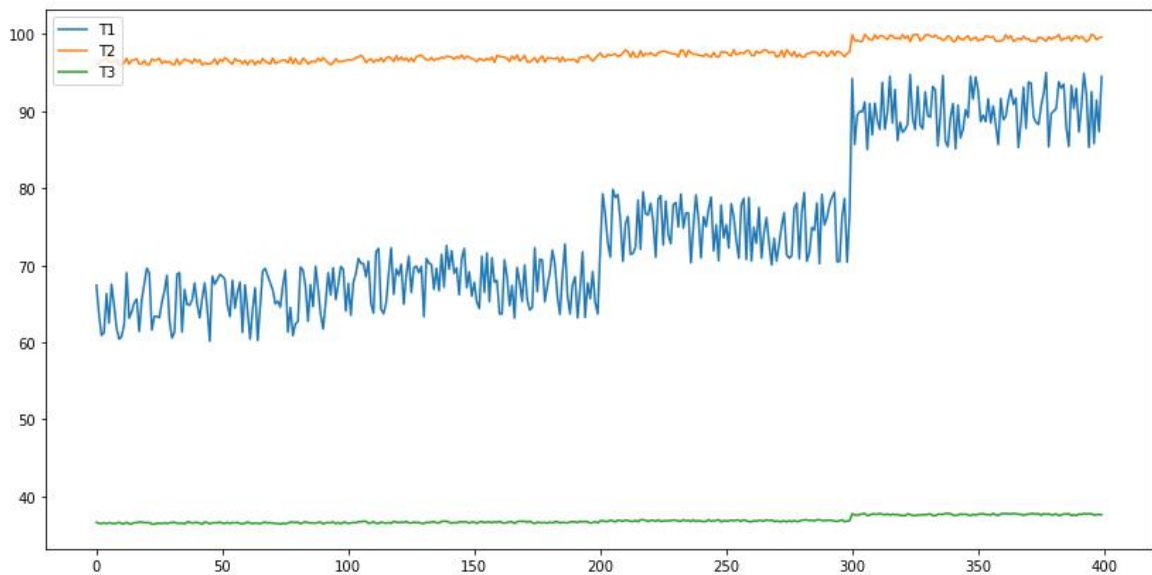


Рис. 3.3.3. Приклад часового ряду, що складається з вибірок рівномірних розподілів з різними середніми значеннями, які спочатку сильно перекриваються, потім незначно перекриваються і, нарешті, не перекриваються

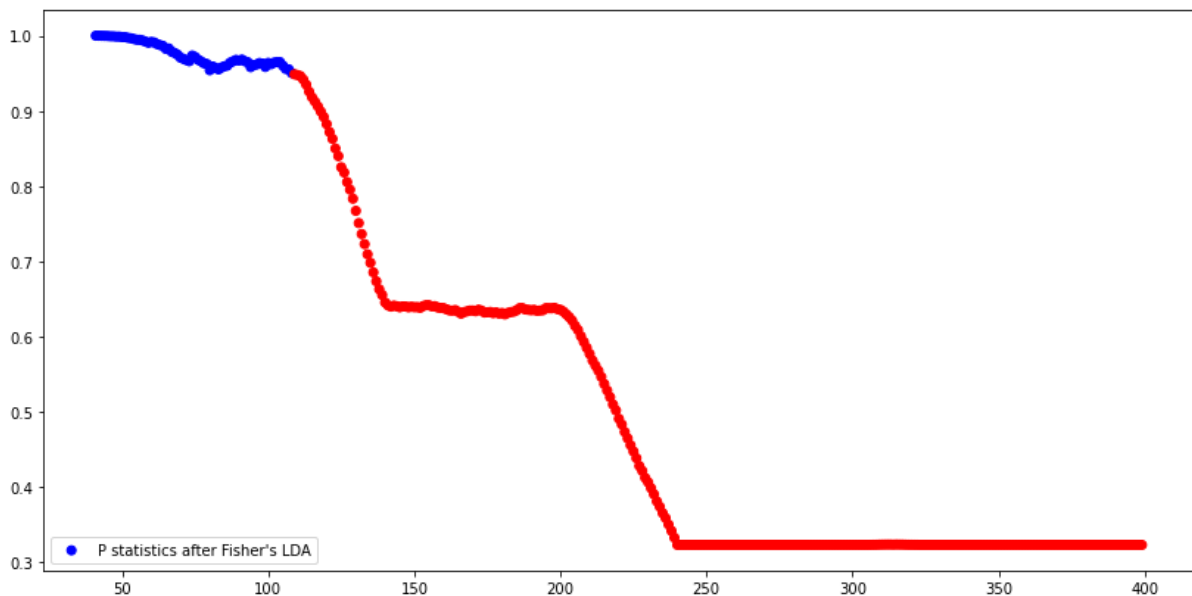


Рис. 3.3.4. Значення р-статистики для рівномірних розподілів з різними середніми значеннями, які спочатку сильно перекриваються, потім незначно перекриваються і, нарешті, не перекриваються. Сині точки відповідають випадкам, коли значення не менше 0.95, а червоні — протилежним, коли вони є меншими

Приклад 3. Нормальні розподіли з різними середніми значеннями, які майже не перетинаються

Розглянемо стрибкоподібний часовий ряд, який складається з нормальних розподілів з різними середніми значеннями, які майже не перетинаються. На цьому часовому ряді ми зможемо перевірити гіпотезу про зсув.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3
0-99	$N(70;2)$	$N(96.0;0.15)$	$N(36.5;0.05)$
100-199	$N(105;2)$	$N(96.3;0.33)$	$N(38.5;0.15)$
200-299	$N(70;2)$	$N(97.0, 0.15)$	$N(36.5;0.05)$
300-399	$N(80,4)$	$N(99.0;0.25)$	$N(37.3;0.98)$

Таблиця 3.3.3. Часові інтервали і нормальні розподіли з різними середніми значеннями, які майже не перетинаються

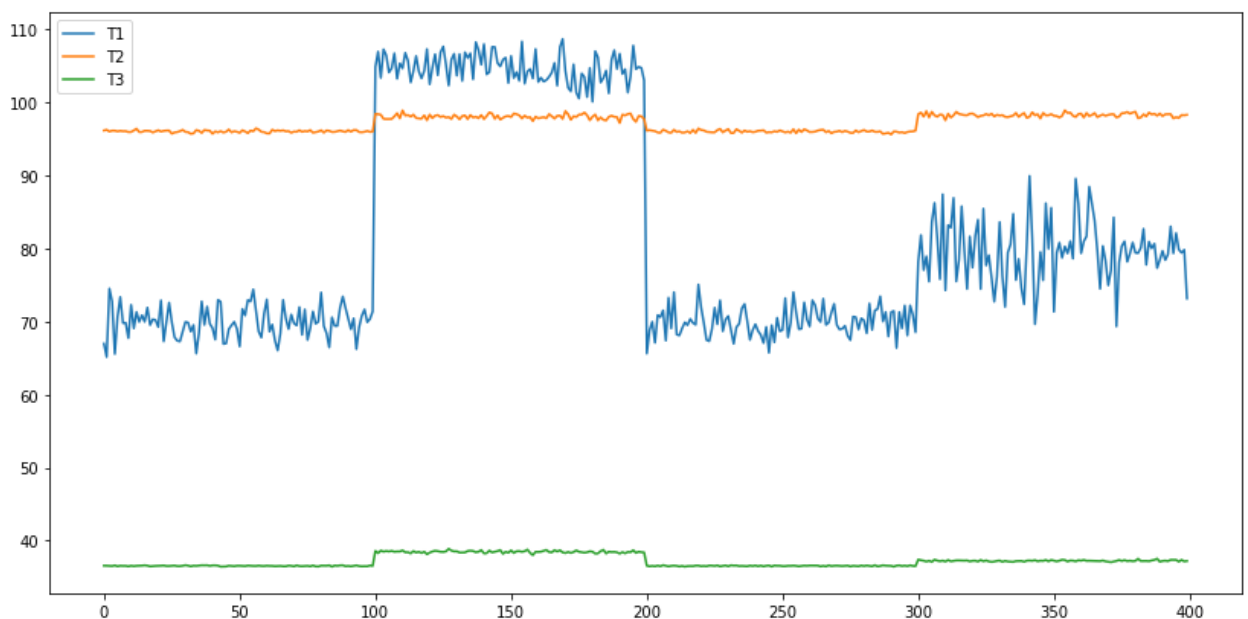


Рис. 3.3.5. Приклад часового ряду, що складається з вибірок нормальних розподілів з різними середніми значеннями, які майже не перетинаються

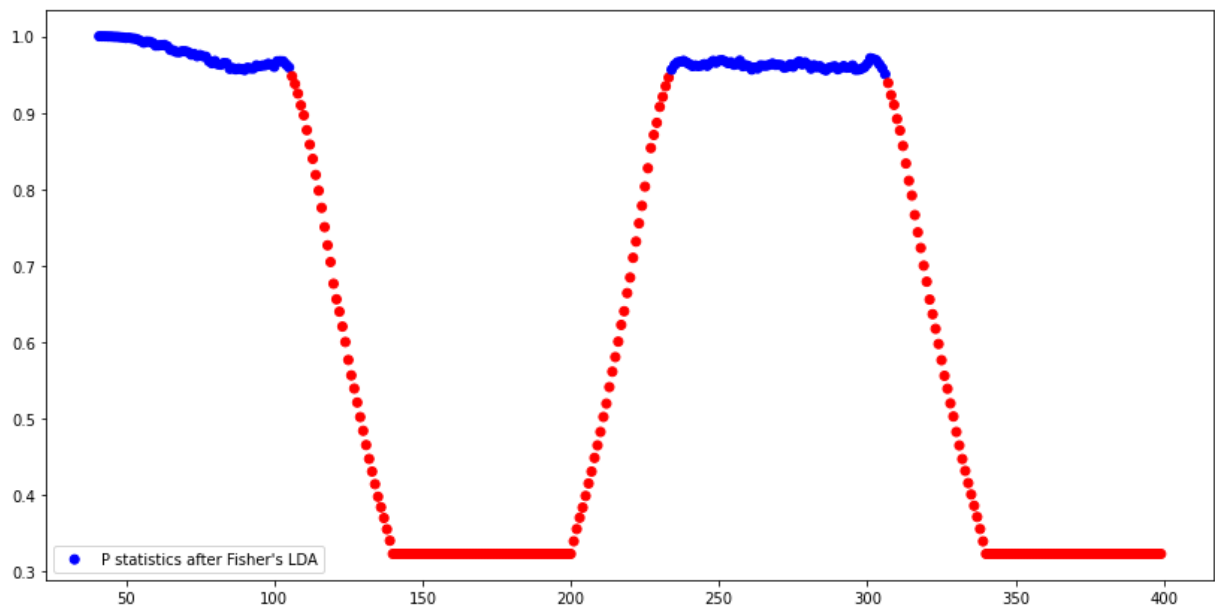


Рис. 3.3.6. Значення р-статистики для нормальних розподілів з різними середніми значеннями, які майже не перетинаються. Сині точки відповідають випадкам, коли значення не менше 0.95, а червоні — протилежним, коли вони є меншими

Приклад 4. Нормальні розподіли з однаковими середніми значеннями, але з дисперсіями, які поступово починають відрізнятися

Розглянемо стрибкоподібний часовий ряд, який складається з нормальних розподілів з однаковими середніми значеннями, але з дисперсіями, які поступово починають відрізнятися. На цьому часовому ряді ми зможемо перевірити гіпотезу про масштаб.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3
0-99	$N(70;1)$	$N(97.0;0.10)$	$N(36.55;0.05)$
100-199	$N(70;2)$	$N(97.0;0.15)$	$N(36.55;0.10)$
200-299	$N(70;3)$	$N(97.0;0.20)$	$N(36.55;0.15)$
300-399	$N(70;5)$	$N(97.0;0.30)$	$N(36.55;0.20)$

Таблиця 3.3.4. Часові інтервали і нормальні розподіли з однаковими середніми значеннями, але з дисперсіями, які поступово починають відрізнятися

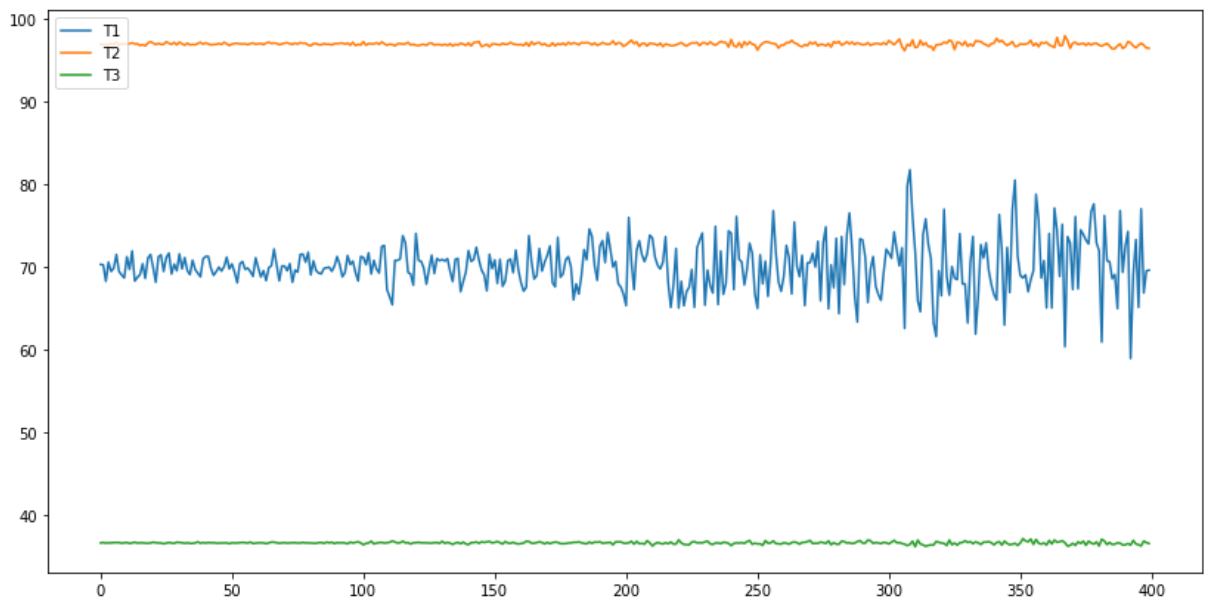


Рис. 3.3.7. Приклад часового ряду, що складається з вибірок нормальних розподілів з однаковими середніми значеннями, але з дисперсіями, які поступово починають відрізнятися

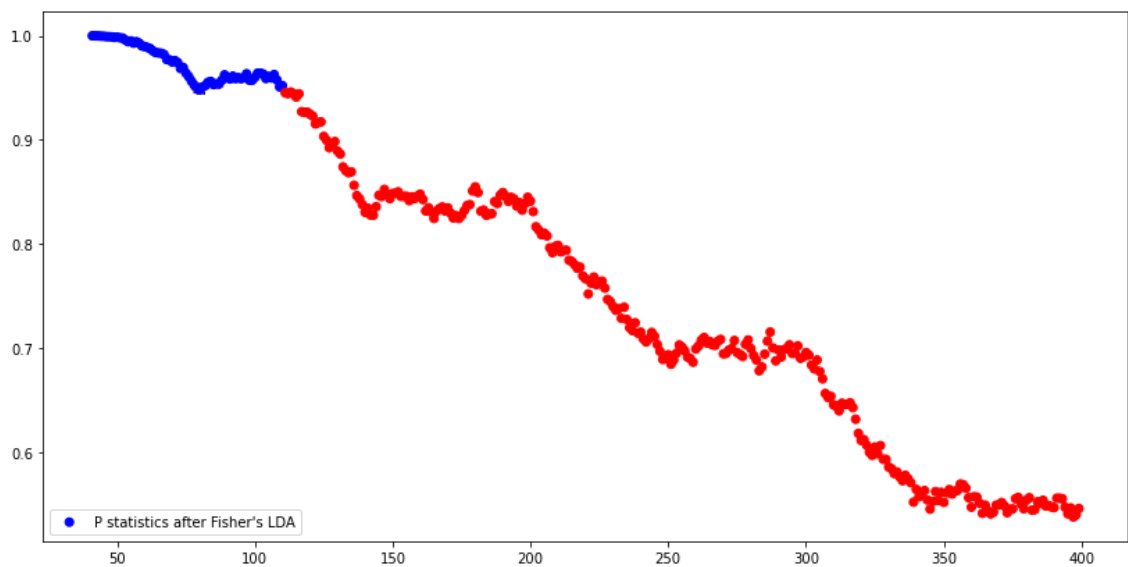


Рис. 3.3.8. Значення р-статистики для нормальних розподілів з однаковими середніми значеннями, але з дисперсіями, які поступово починають відрізнятися. Сині точки відповідають випадкам, коли значення не менше 0.95, а червоні — протилежним, коли вони є меншими

Приклад 5. Нормальні розподіли з однаковими середніми значеннями, але з дисперсіями, які відрізняються сильніше

Розглянемо стрибкоподібний часовий ряд, який складається з нормальних розподілів з однаковими середніми значеннями, але з дисперсіями, які відрізняються сильніше. На цьому часовому ряді ми зможемо перевірити гіпотезу про масштаб.

Часовий інтервал	Розподіл T_1	Розподіл T_2	Розподіл T_3
0-99	$N(70;1)$	$N(97.0;0.10)$	$N(36.55;0.05)$
100-199	$N(70;5)$	$N(97.0;0.50)$	$N(36.55;0.25)$
200-299	$N(70;7)$	$N(97.0,1.00)$	$N(36.55;0.5)$
300-399	$N(70;10)$	$N(97.0,1.50)$	$N(36.55;0.75)$

Таблиця 3.3.4. Часові інтервали і нормальні розподіли з однаковими середніми значеннями, але з дисперсіями, які поступово починають відрізнятися

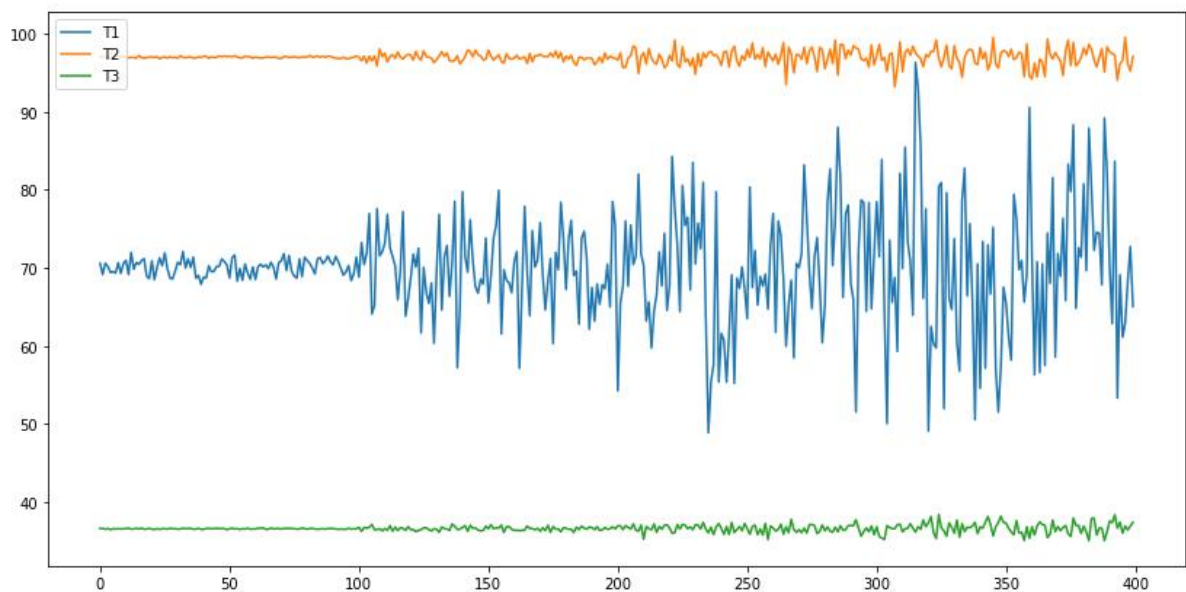


Рис. 3.3.9. Приклад часового ряду, що складається з вибірок нормальних розподілів з однаковими середніми значеннями, але з дисперсіями, які відрізняються сильніше

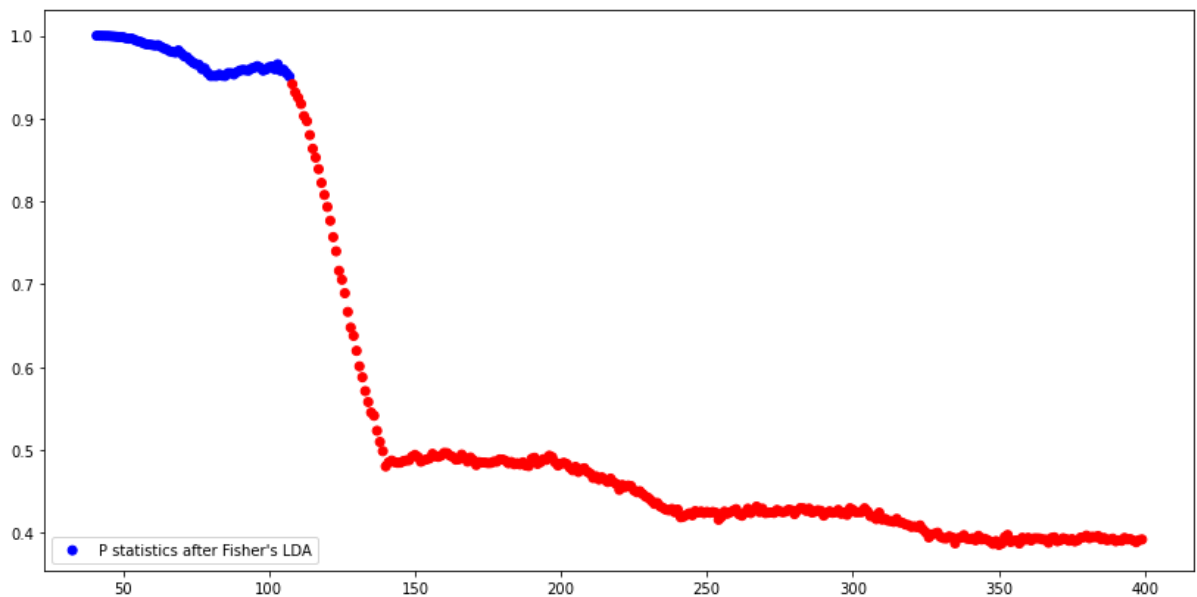


Рис. 3.3.10. Значення р-статистики для нормальних розподілів з однаковими середніми значеннями, але з дисперсіями, які відрізняються сильніше. Сині точки відповідають випадкам, коли значення не менше 0.95, а червоні — протилежним, коли вони є меншими

Висновки до підрозділу 3.3

У цьому підрозділі було розглянуто застосування алгоритму пошуку точок переходу на основі лінійного дискримінанта Фішера та р-статистики Петуніна для часових рядів. Експерименти продемонстрували швидке та точне виявлення змін у функції розподілу, що підтверджує ефективність методу для широкого класу задач. Отримані результати свідчать про можливість практичного використання цього алгоритму в інтелектуальних системах підтримки прийняття рішень, зокрема у медичних застосуваннях, де необхідний своєчасний контроль стану пацієнтів та надійне виявлення різких змін у даних.

3.4 Порівняння методів FLD і PCA

У цьому підрозділі ми досліджуємо застосування двох методів зниження розмірності — лінійного дискримінанта Фішера (FLD) та методу головних компонент (PCA) — для задачі виявлення точок переходу у стаціонарних часових рядах. Обидва підходи дозволяють відобразити багатовимірні дані на

пряму та далі перевірити однорідність вибірок за допомогою р-статистики Петуніна.

Алгоритм роботи побудований на методі рухомого вікна: на кожному кроці здійснюється порівняння початкового інтервалу з новою вибіркою, для яких обчислюються проєкції за допомогою FLD та PCA. Для отриманих одномірних проєкцій незалежно рахується р-статистика, після чого перевіряється гіпотеза про однорідність розподілів. Якщо статистика перевищує 0.95, вибірки вважаються однорідними, інакше — фіксується точка переходу.

Мета експериментів полягає у порівнянні ефективності FLD та PCA в умовах, коли часовий ряд зазнає змін у функції розподілу. Для цього береться ряд довжини $N = 200$, поділений на два інтервали з різними законами розподілу. Алгоритм багаторазово повторюється, після чого усереднені значення р-статистики візуалізуються: синіми точками відмічаються випадки зберігання однорідності, а червоними — виявлення зміни.

Приклад 1. Стрибок середнього значення рівномірних розподілів

Часовий інтервал	Розподіл
0-99	$U(0;1)$
100-199	$U(1;2)$

Таблиця 3.4.1. Часові інтервали і рівномірні розподіли зі стрибком у середньому значенню

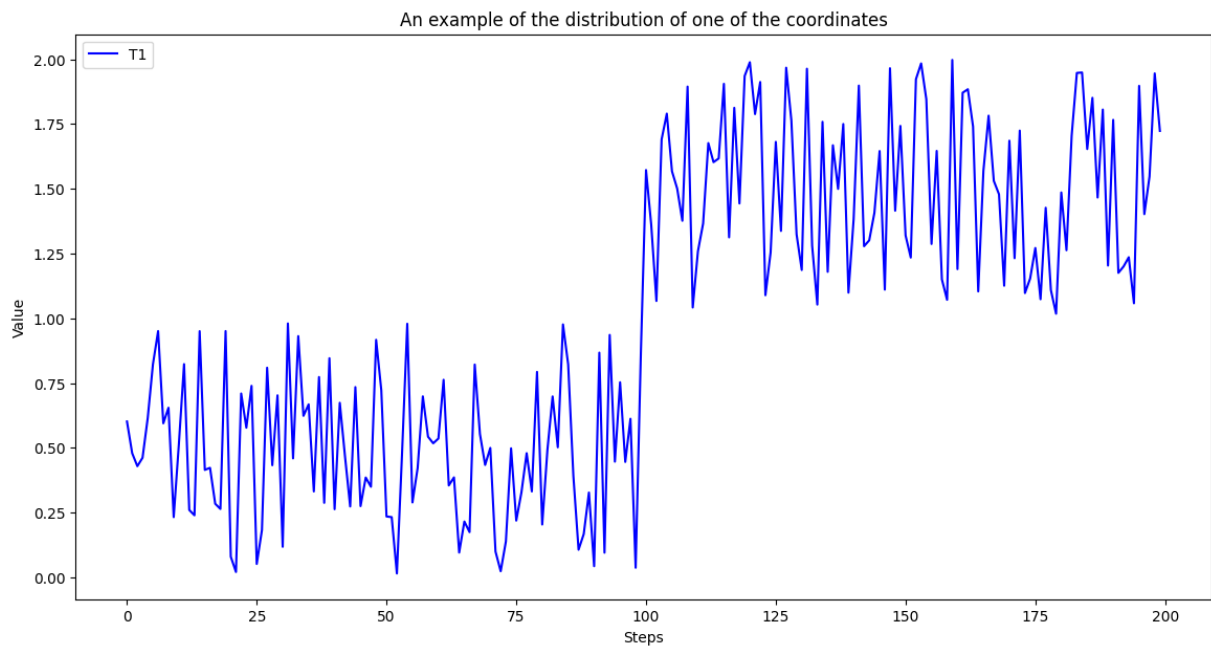


Рис. 3.4.1. Приклад розподілу однієї з координат для рівномірних розподілів зі стрибком середнього значення

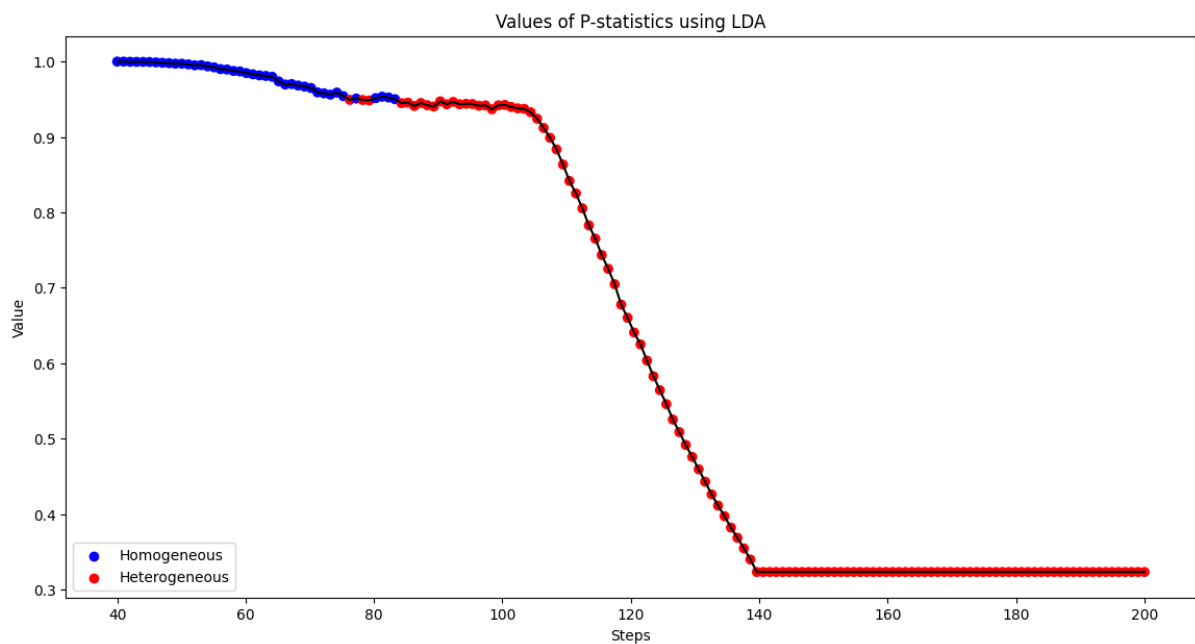


Рис. 3.4.2. Значення Р-статистики для координат розмірності 3 зі стрибком середнього значення після використання LDA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

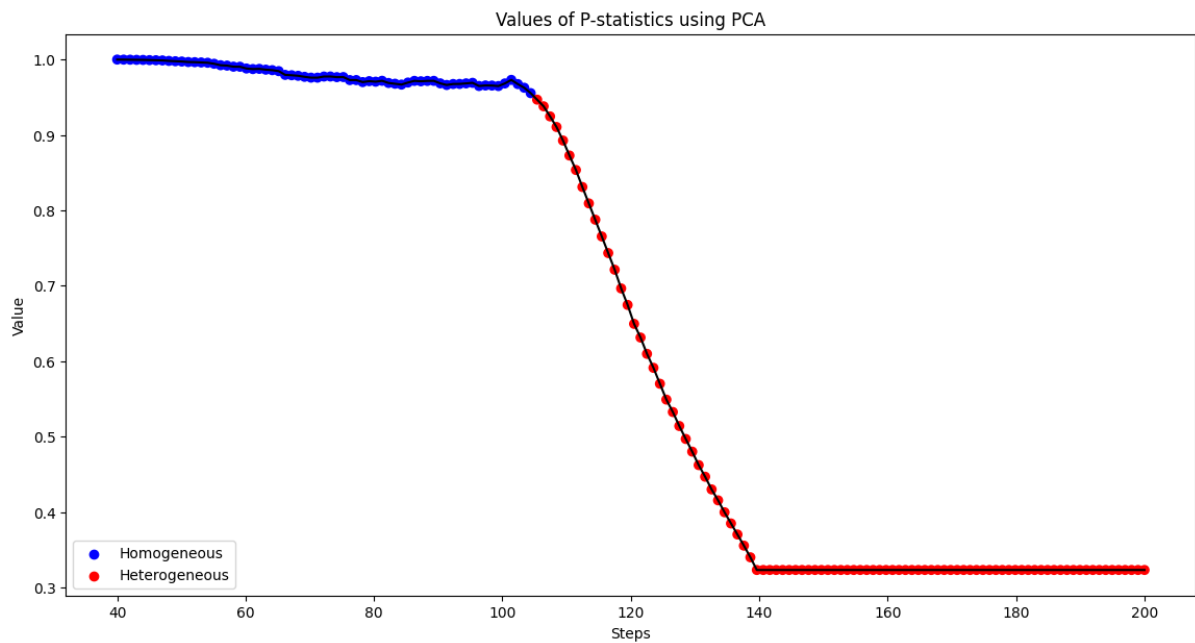


Рис. 3.4.3. Значення Р-статистики для координат розмірності 3 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

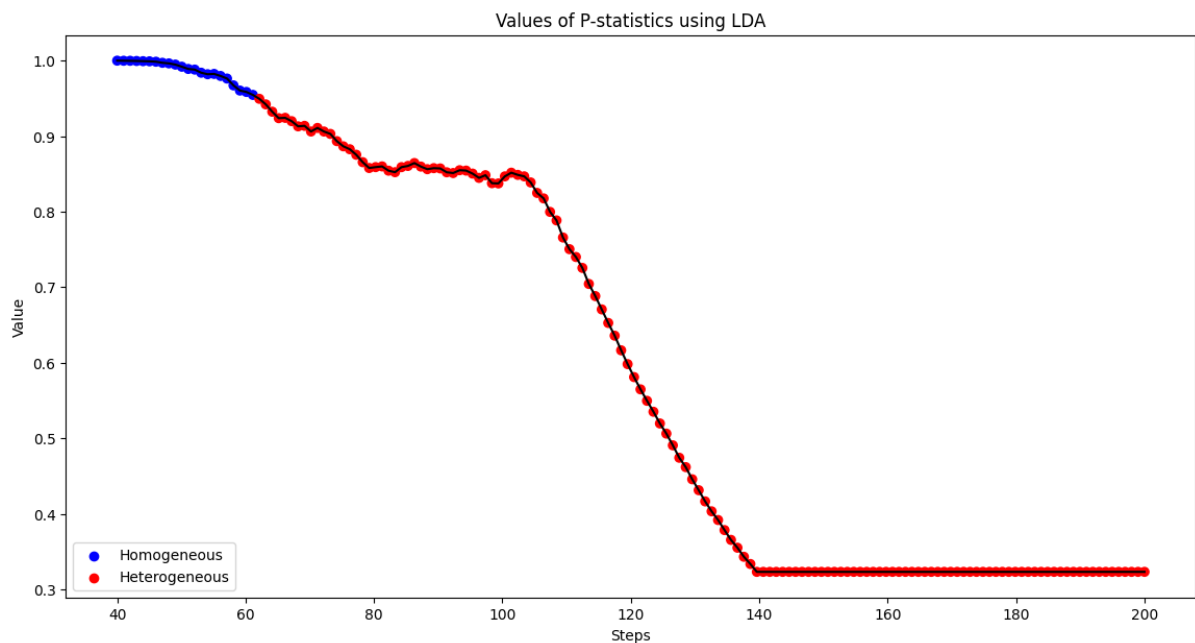


Рис. 3.4.4. Значення Р-статистики для координат розмірності 10 зі стрибком середнього значення після використання LDA. Сині точки вказують на час,

коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

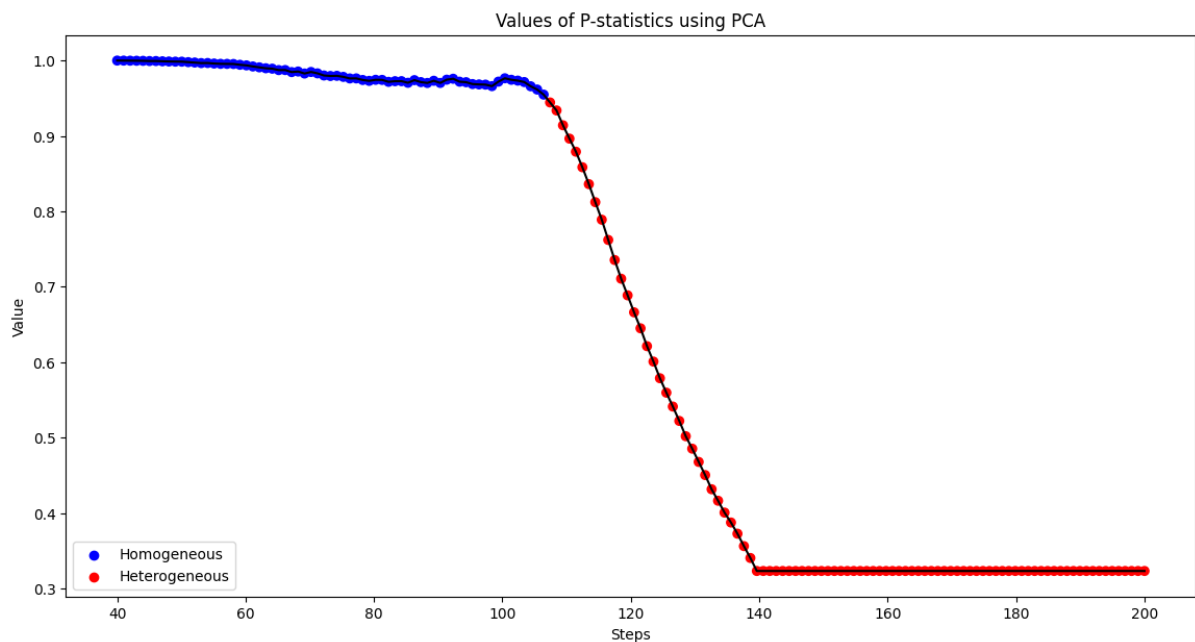


Рис. 3.4.5. Значення Р-статистики для координат розмірності 10 зі стрибком середнього значення після використання РСА. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

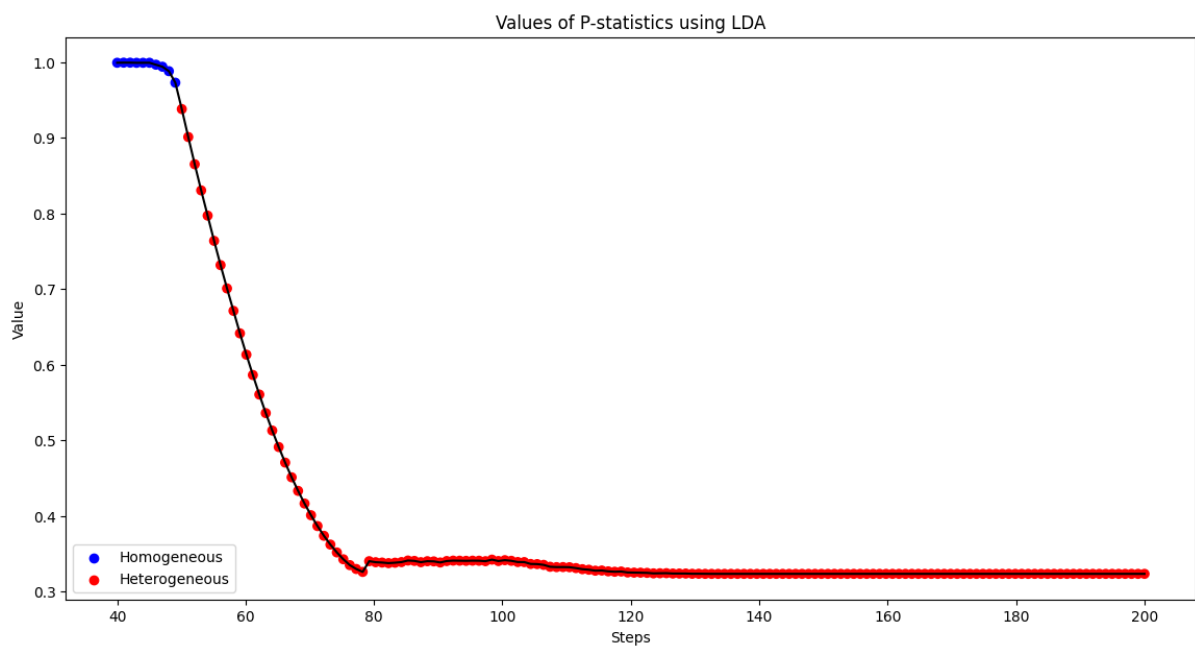


Рис. 3.4.6. Значення Р-статистики для координат розмірності 100 зі стрибком середнього значення після використання LDA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

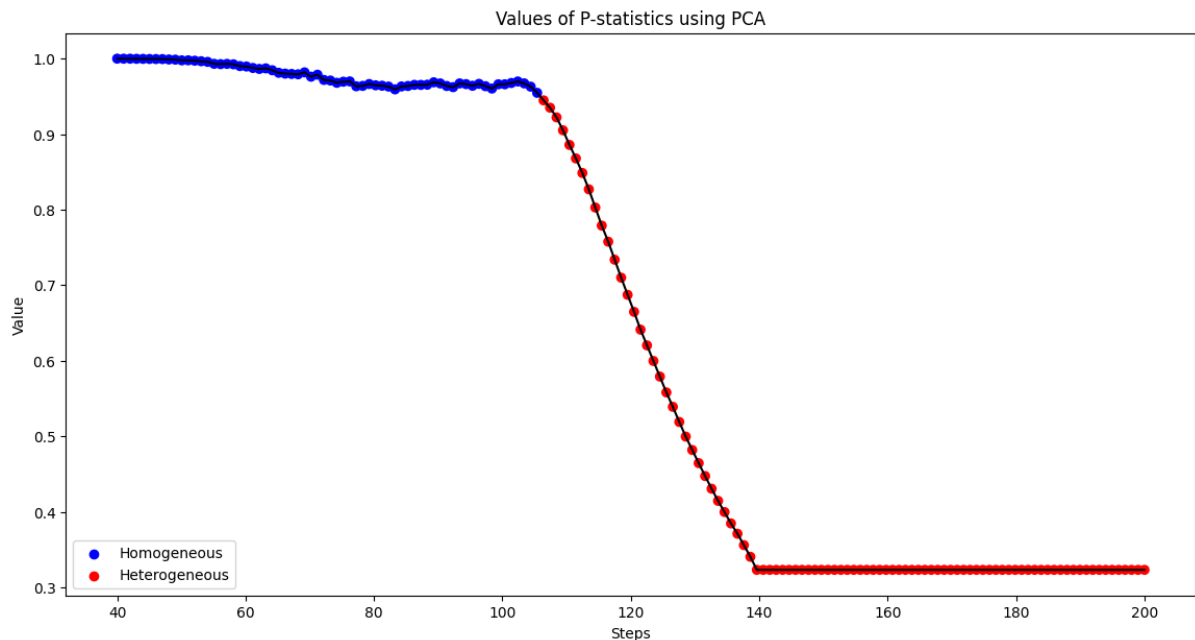


Рис. 3.4.7. Значення Р-статистики для координат розмірності 100 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

Приклад 2. Стрибок дисперсії рівномірних розподілів

Часовий інтервал	Розподіл
0-99	$N(0;1)$
100-199	$N(0;2)$

Таблиця 3.4.2. Часові інтервали і нормальні розподіли зі стрибком у дисперсії

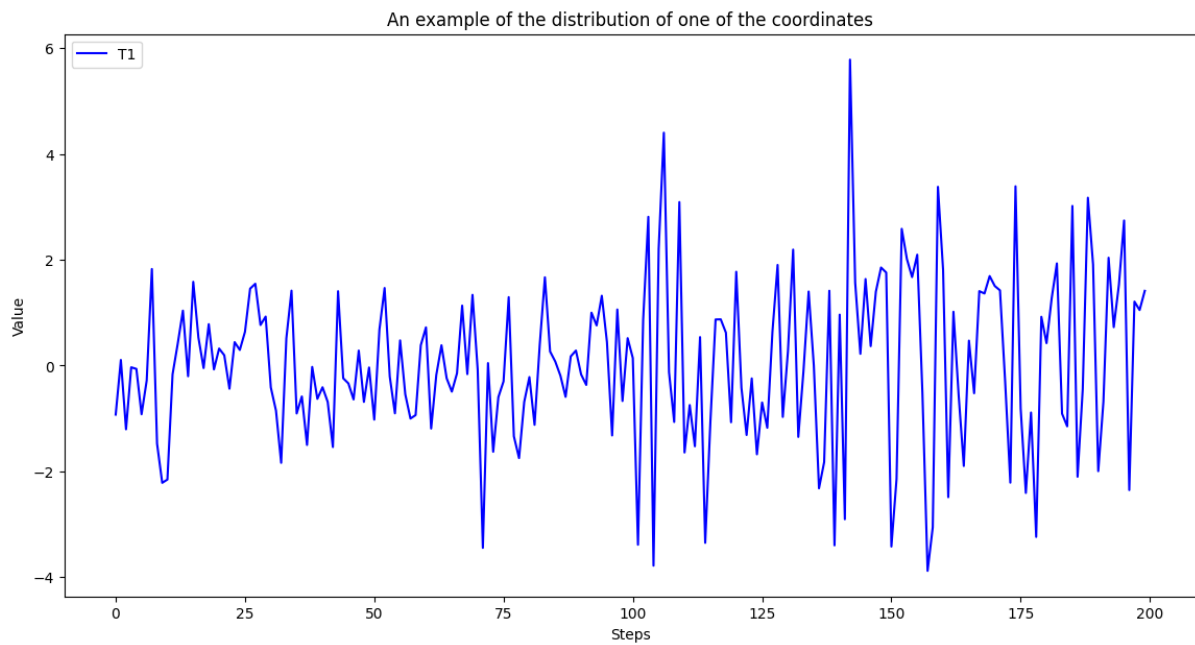


Рис. 3.4.8. Приклад розподілу однієї з координат для рівномірних розподілів зі стрибком дисперсії

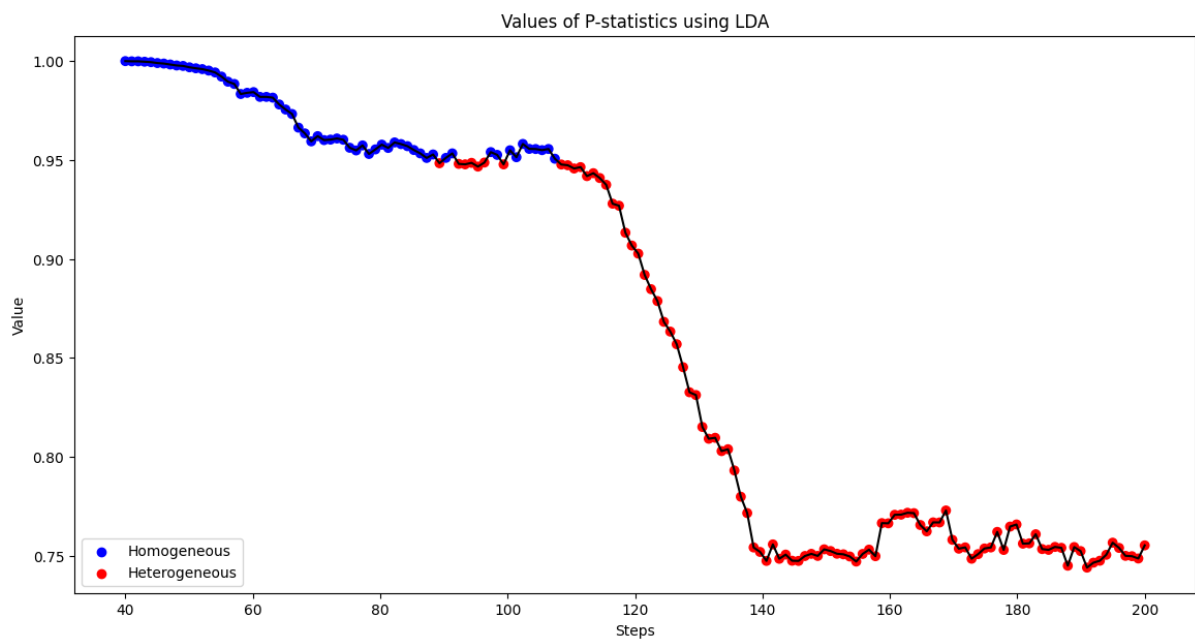


Рис. 3.4.9. Значення Р-статистики для координат розмірності 3 зі стрибком середнього значення після використання LDA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

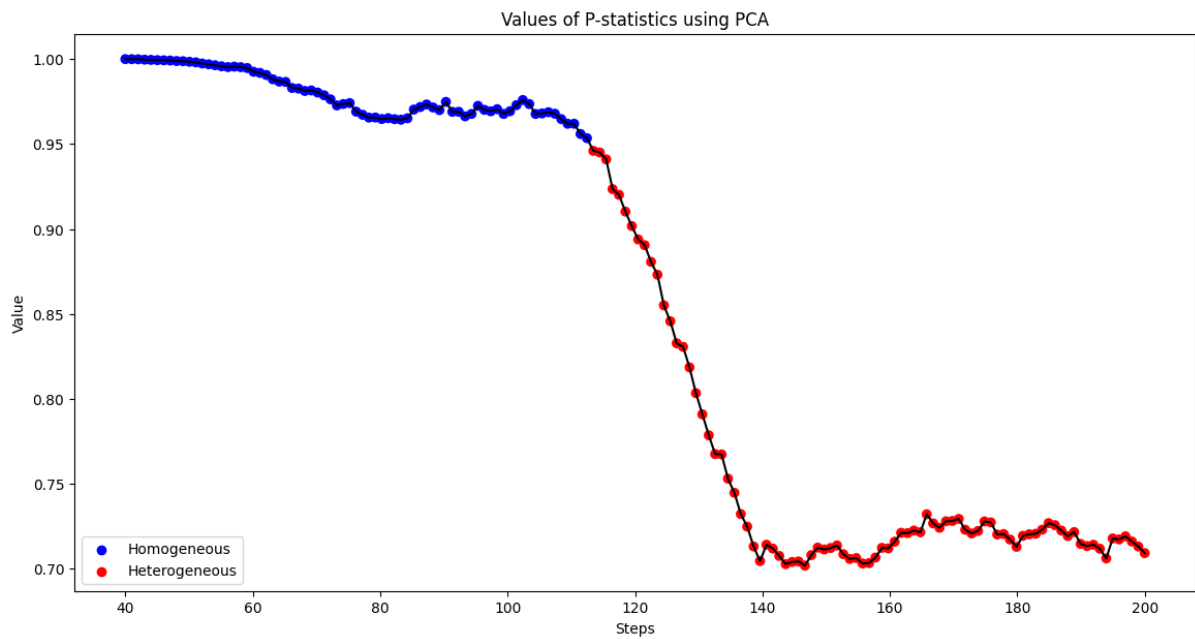


Рис. 3.4.10. Значення Р-статистики для координат розмірності 3 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

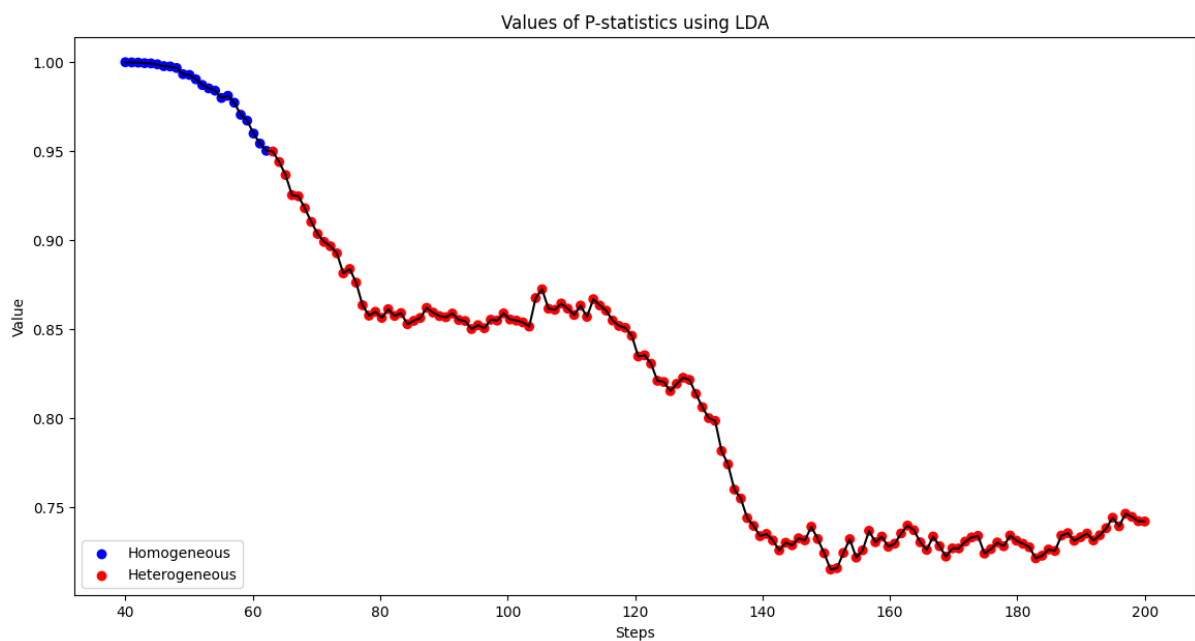


Рис. 3.4.11. Значення Р-статистики для координат розмірності 10 зі стрибком середнього значення після використання LDA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

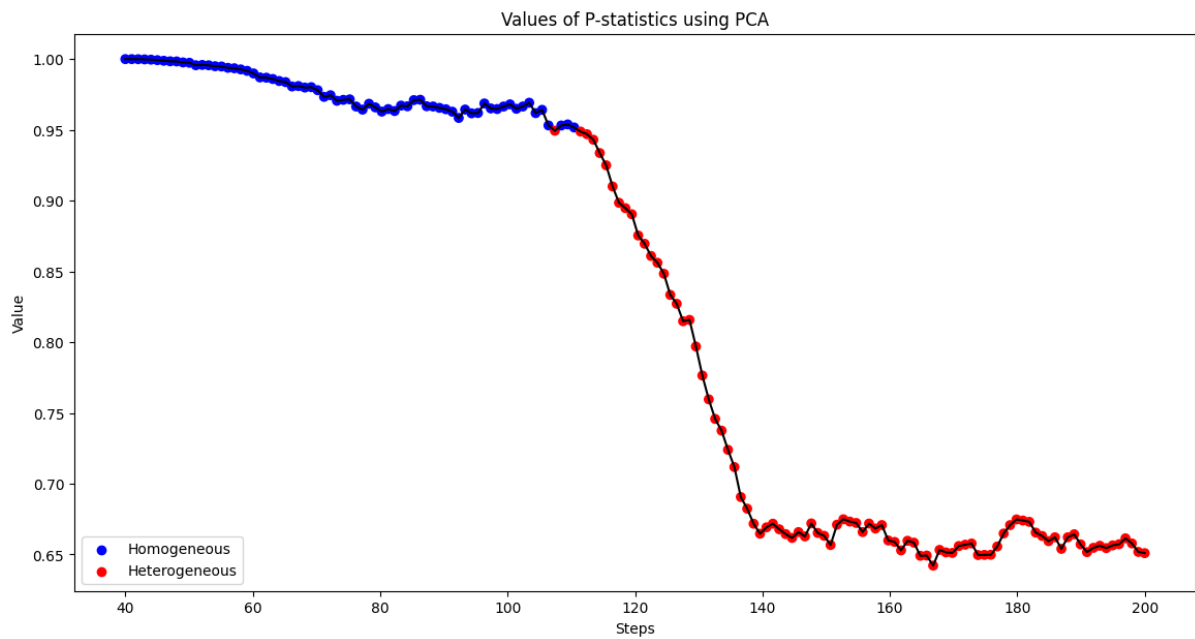


Рис. 3.4.12. Значення Р-статистики для координат розмірності 10 зі стрибком середнього значення після використання РСА. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

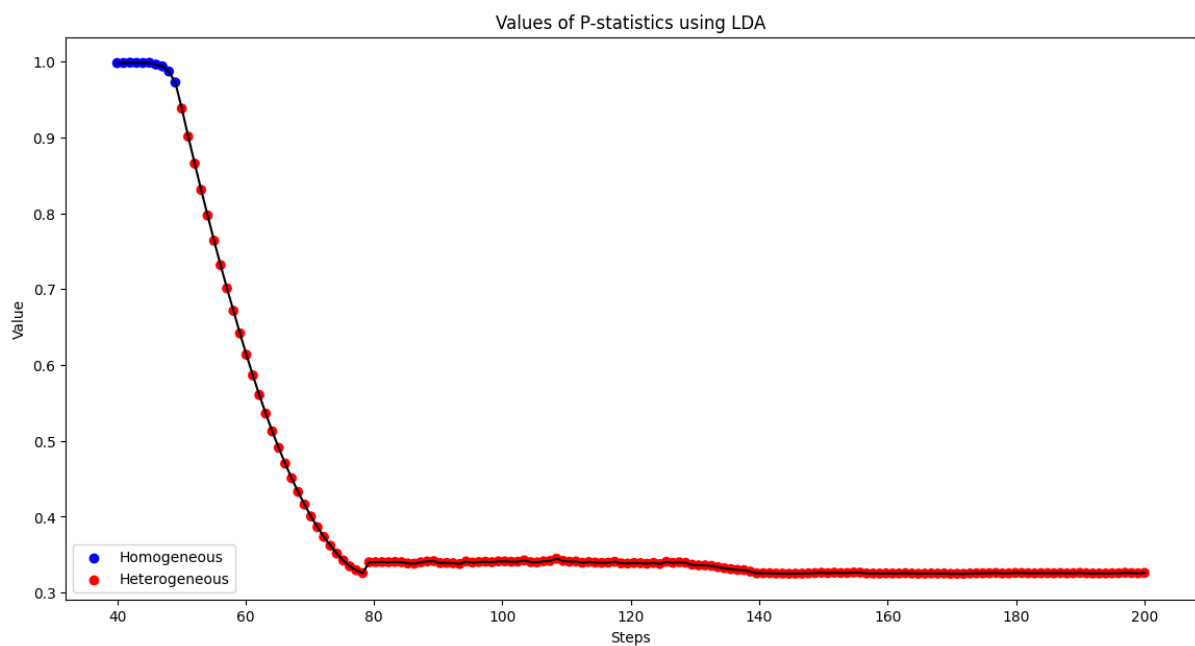


Рис. 3.4.13. Значення Р-статистики для координат розмірності 100 зі стрибком середнього значення після використання LDA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

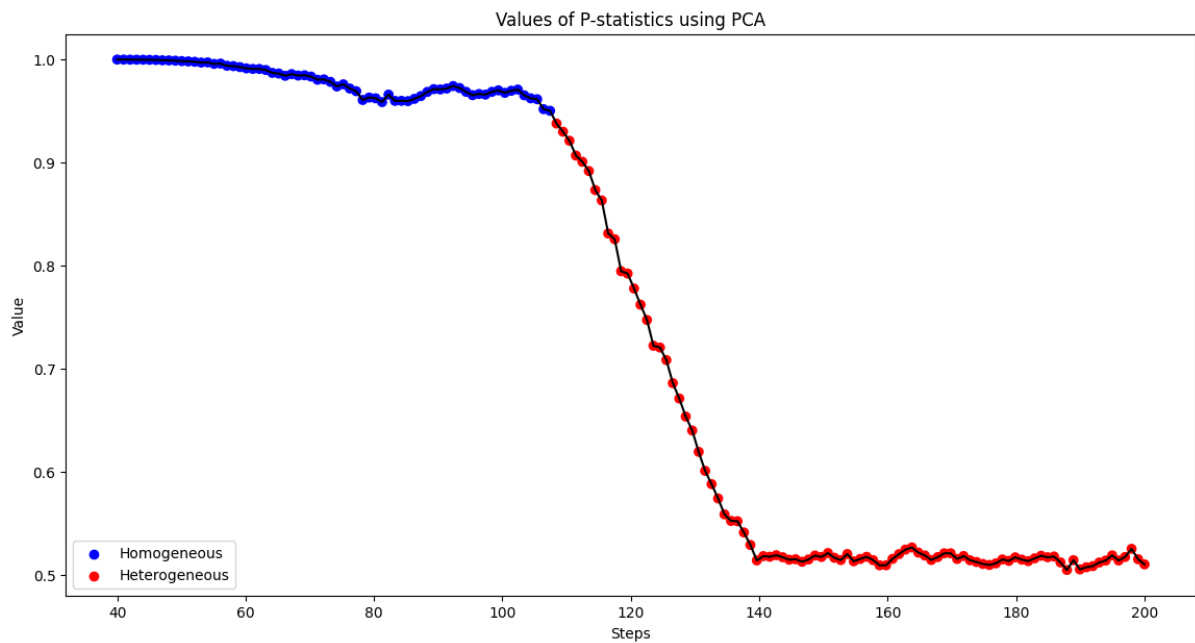


Рис. 3.4.14. Значення Р-статистики для координат розмірності 3 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші

Висновки до підрозділу 3.4

Порівняння використання лінійного дискримінанта Фішера та методу головних компонент у поєднанні з р-статистикою Петуніна показало, що обидва підходи здатні виявляти точки переходу у багатовимірних часових рядах, проте відрізняються за чутливістю та точністю. FLD забезпечує швидке реагування на зміни, але в окремих випадках схильний до передчасного виявлення переходів. Натомість PCA демонструє більш стабільну та узгоджену поведінку, що дозволяє точніше локалізувати точку зміни навіть за зростання розмірності. Таким чином, результати підтверджують доцільність комбінованого використання цих методів залежно від характеру даних: FLD ефективний при різко виражених відмінностях, тоді як PCA виявляється більш надійним у складніших багатовимірних випадках.

3.5. Порівняння статистик у багатовимірних часових рядах

Метою цього підрозділу є порівняння ефективності різних статистичних критеріїв при аналізі багатовимірних часових рядів. Особливу увагу зосереджено на перевірці гіпотези про однорідність вибірок та визначенні першої точки переходу. Для цього були використані три статистики:

- p -статистика Петуніна,
- статистика Колмогорова–Смірнова,
- W -статистика Вілкоксона.

Алгоритм побудований на принципі ковзного вікна: початкова вибірка визначає базовий розподіл, після чого послідовно перевіряються наступні підвибірки. Для кожної з них обчислюються значення трьох статистик, які потім порівнюються з відповідними критичними порогами. Якщо значення перевищує порогове, вважаємо вибірки однорідними; у протилежному випадку фіксується неоднорідність.

Результати представлені у вигляді графіків, де синім кольором позначені ділянки однорідності, а червоним — відмінності у розподілах. Крім того, для кожної статистики вказано номер елемента, починаючи з якого було зафіксовано відхилення від початкового розподілу. Це дозволяє оцінити точність локалізації першої точки переходу та порівняти між собою розглянуті критерії.

Опис алгоритму

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 $(x_{i+1}, \dots, x_{i+w})$ виконується наступне:

- Аналіз головних компонент. Проводиться проєкція вибірок методом головних компонент (Principal Component Analysis).

- Зменшення розмірності. Отримуються проєкції для обох підвибірок: $(p_1, \dots, p_w)$ та $(p_{i+1}, \dots, p_{i+w})$.

- Розрахунок статистики Петуніна. Обчислюється p -статистика Петуніна ($pstat$) для кожної з підвибірок.

- Перевірка однорідності. Якщо значення $pstat \geq 0.95$, то вважаємо, що розподіли співпадають. Якщо $pstat < 0.95$ — розподіли відмінні.

3. *Зсув вікна.* Вікно $(x_{i+1}, \dots, x_{i+w})$ зсувається на один елемент вправо. Алгоритм повторюється з кроку 2.

4. *Завершення.* Процедура повторюється доти, доки не буде проаналізовано всю послідовність. Якщо підвибірка 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. Процедuru 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.

Приклад 1. Стрибок по середньому для рівномірного розподілу

У цьому прикладі розглядається часова послідовність, яка складається з двох інтервалів з різними розподілами.

Часовий інтервал	Розподіл
0-99	$U(0;1)$
100-199	$U(1;2)$

Таблиця 3.5.1. Часові інтервали і рівномірні розподіли зі стрибком у середньому

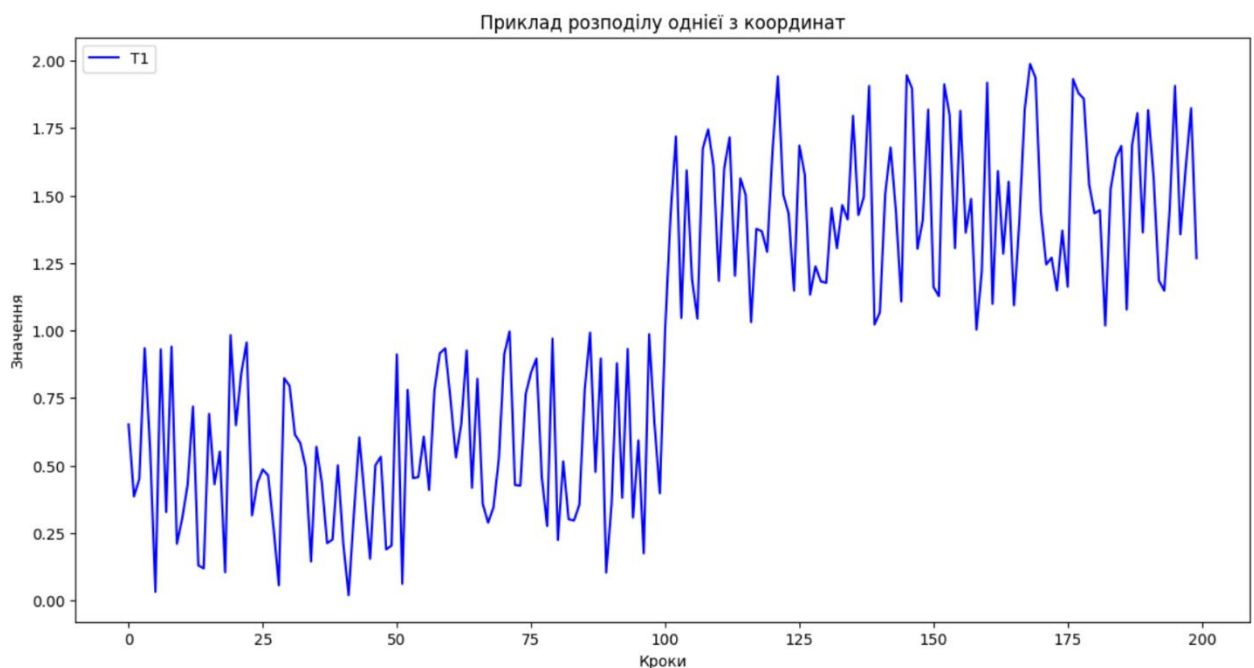


Рис. 3.5.1. Приклад розподілу однієї з координат для рівномірних розподілів зі стрибком середнього

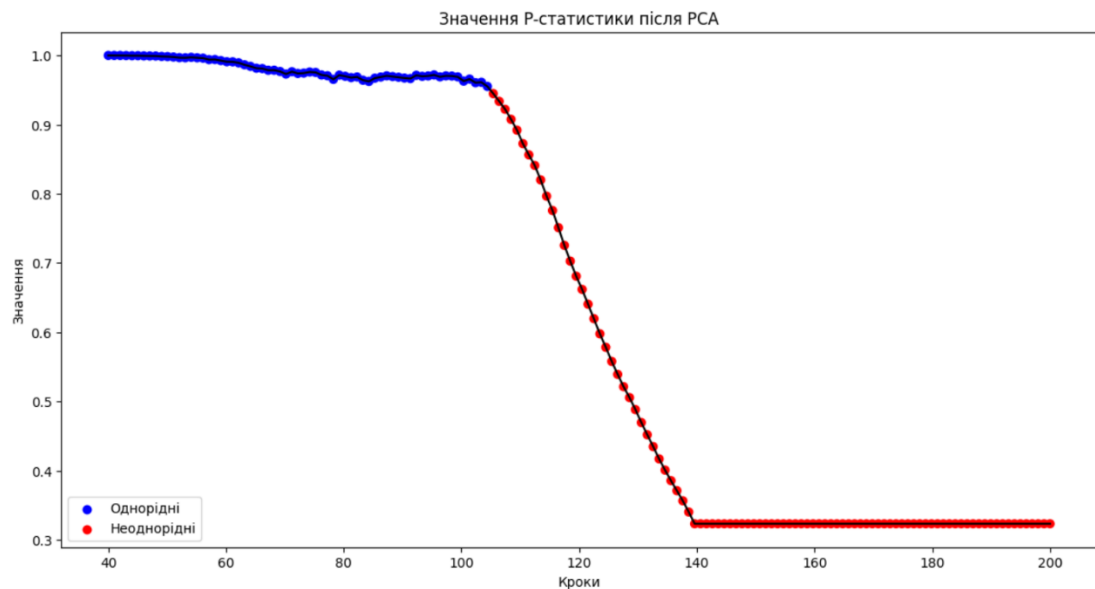


Рис. 3.5.2. Значення Р-статистики для рівномірних розподілів розмірності 100 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 105

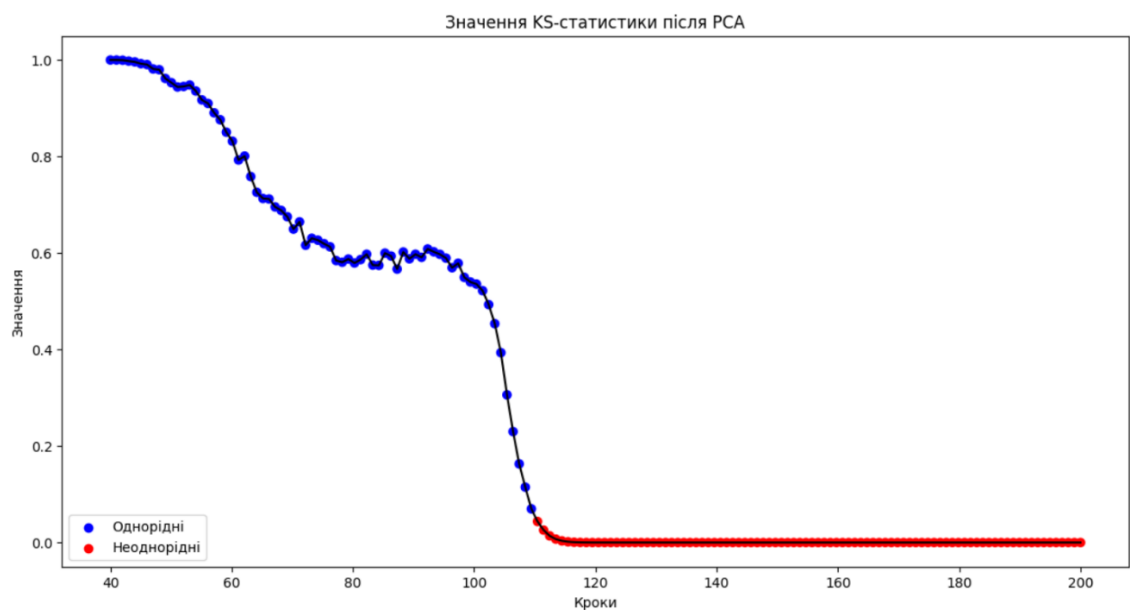


Рис. 3.5.3. Значення KS-статистики для рівномірних розподілів розмірності 100 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 110

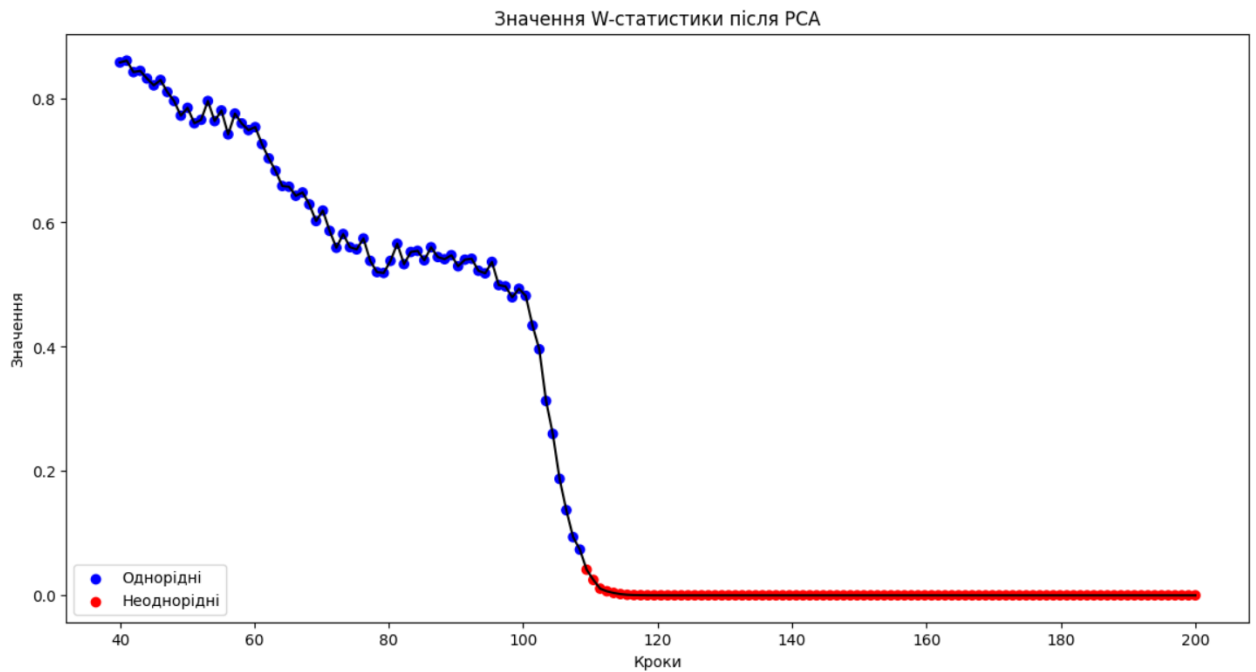


Рис. 3.5.4. Значення W-статистики для рівномірних розподілів розмірності 100 зі стрибком середнього значення після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 109

На графіках чітко видно момент, коли значення статистик починають відхилятися від порогового рівня, що відповідає точці переходу між двома розподілами. Для кожного методу визначається перший індекс, з якого було зафіксовано відмінність у розподілах. У цьому прикладі P-статистика визначила першу точку переходу на 105-му елементі, KS-статистика — на 110-му, а W-статистика — на 109-му. Таким чином, усі три методи коректно зафіксували зміну, проте p-статистика показала найшвидшу реакцію та більш точну локалізацію.

Приклад 2. Стрибок по дисперсії для нормального розподілу

У цьому прикладі розглядається часова послідовність, яка складається з двох інтервалів з різними розподілами.

Часовий інтервал	Розподіл
0-99	$N(0;1)$
100-199	$N(0;2)$

Таблиця 3.5.2. Часові інтервали і нормальні розподіли зі стрибком у дисперсії

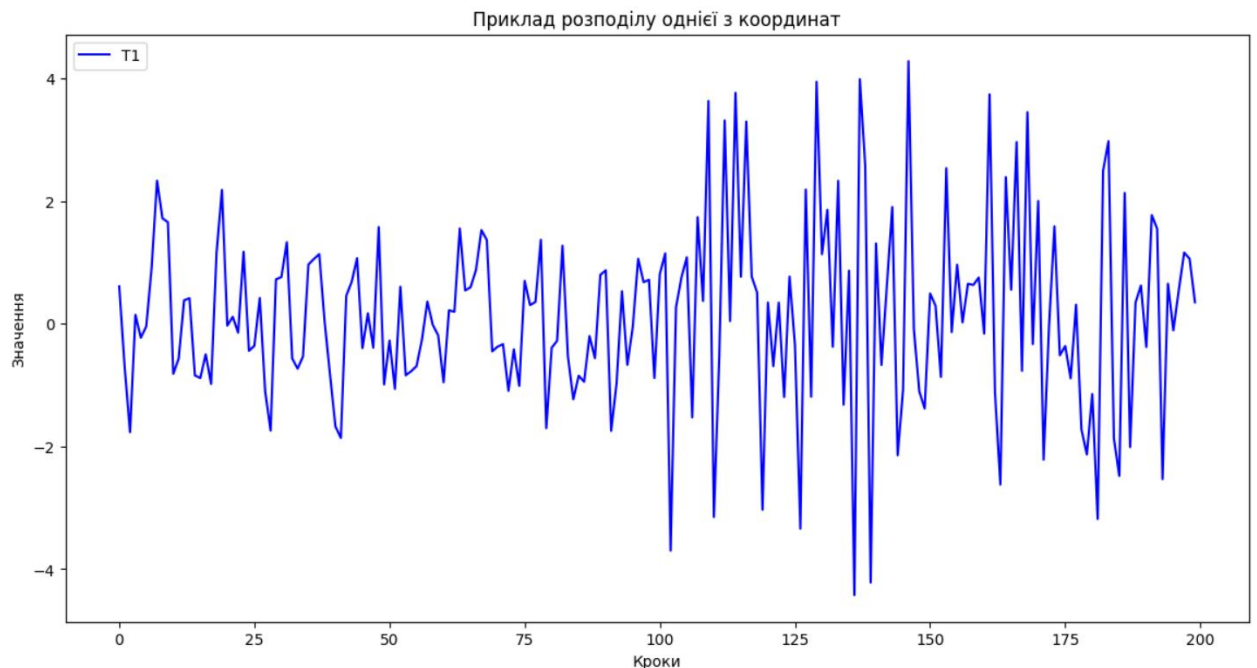


Рис. 3.5.5. Приклад розподілу однієї з координат для нормальних розподілів зі стрибком у дисперсії

Як видно з Таблиці 3.5.2, шукана перша точка переходу дорівнює 100, а потім відбулося відхилення від початкового розподілу. З Рисунку 3.5.6 видно, що перша знайдена точка переходу дорівнює 108, тобто через 8 кроків після початку стрибка. З Рисунку 3.5.7 видно, що перша знайдена точка переходу дорівнює 132, тобто через 32 кроки після початку стрибка. З Рисунку 3.5.8 видно, статистика Вілкоксона не знайшла жодної точки переходу. В цьому експерименті вже видно, що в гіпотезі про змінення масштабу Р-статистика показує кращі результати за інші статистики.

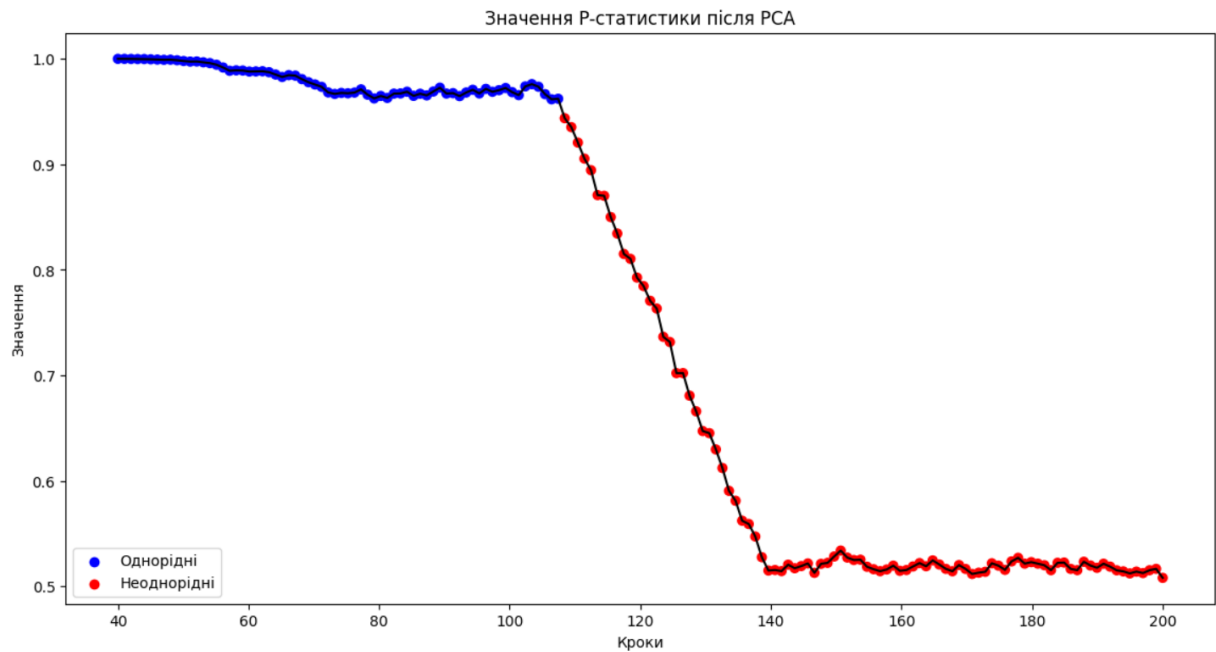


Рис. 3.5.6. Значення Р-статистики для нормальних розподілів розмірності 100 зі стрибком у дисперсії після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 108

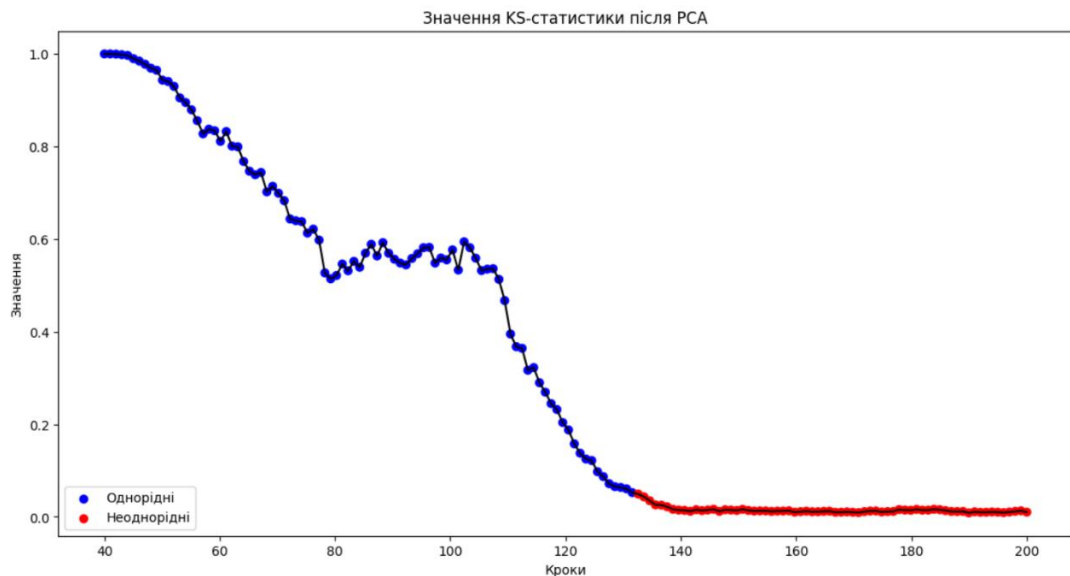


Рис. 3.5.7. Значення KS-статистики для нормальних розподілів розмірності 100 зі стрибком у дисперсії після використання PCA. Сині точки вказують на час,

коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 132

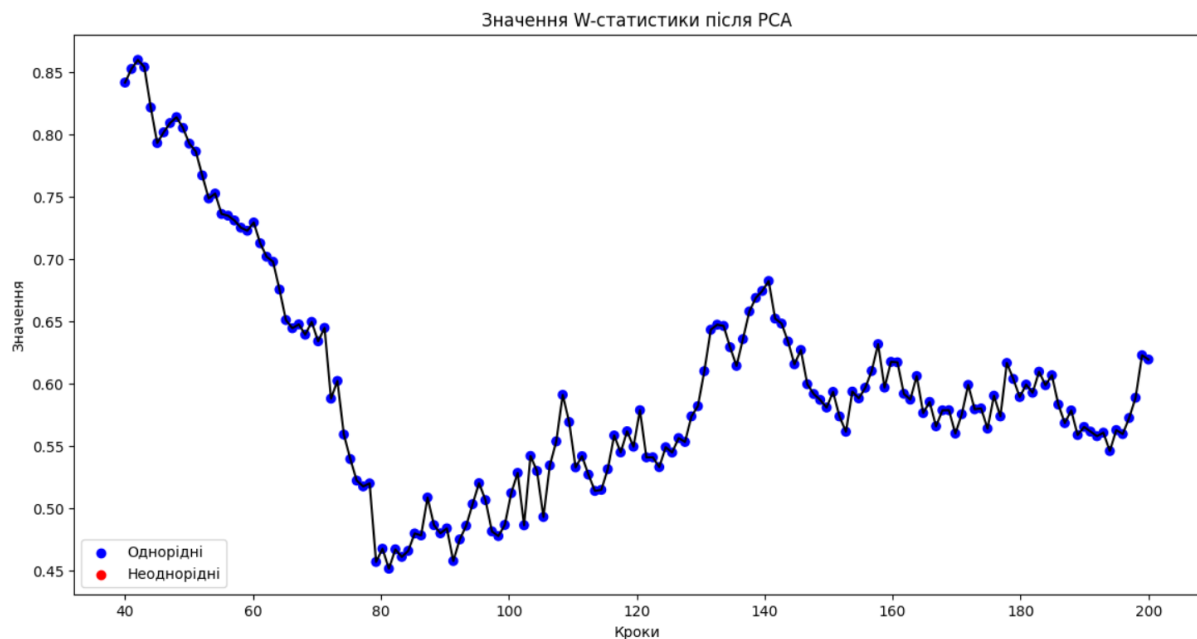


Рис. 3.5.8. Значення W-статистики для нормальних розподілів розмірності 100 зі стрибком у дисперсії після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Не виявлено жодної точки переходу

Приклад 3. Стрибок по дисперсії для логнормального розподілу

У цьому прикладі розглядається часова послідовність, яка складається з двох інтервалів з різними розподілами.

Часовий інтервал	Розподіл
0-99	$\text{LogN}(0;1)$
100-199	$\text{LogN}(0;0.5)$

Таблиця 3.5.3. Часові інтервали і логнормальні розподіли зі стрибком у дисперсії

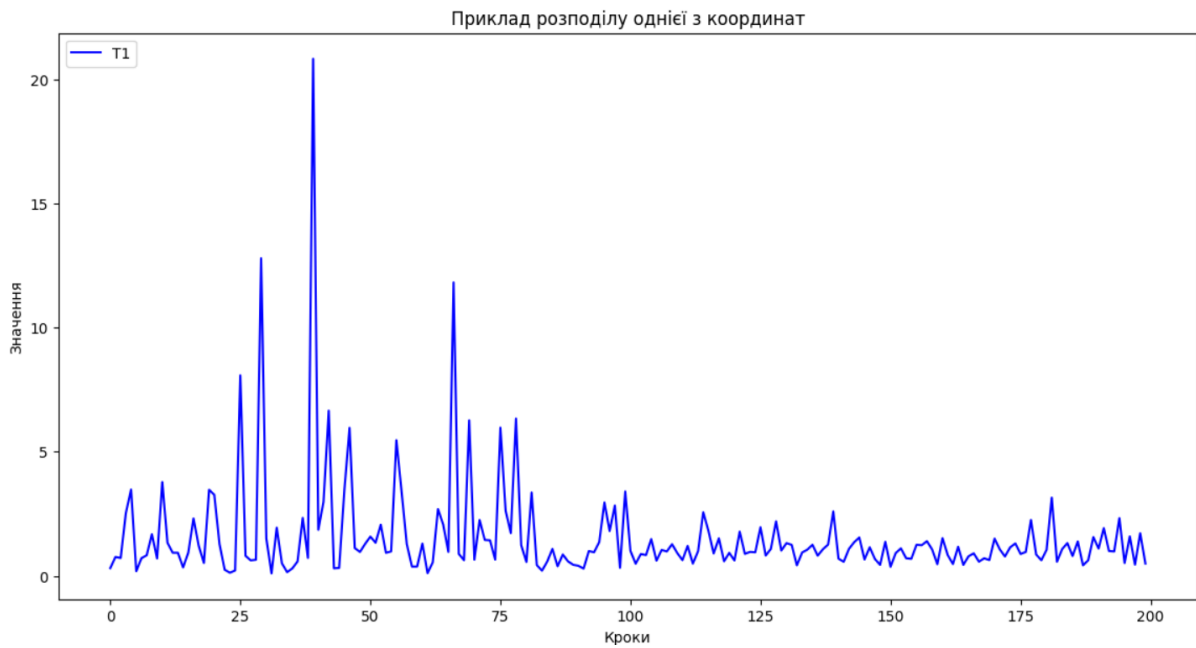


Рис. 3.5.9. Приклад розподілу однієї з координат для логнормальних розподілів зі стрибком у дисперсії

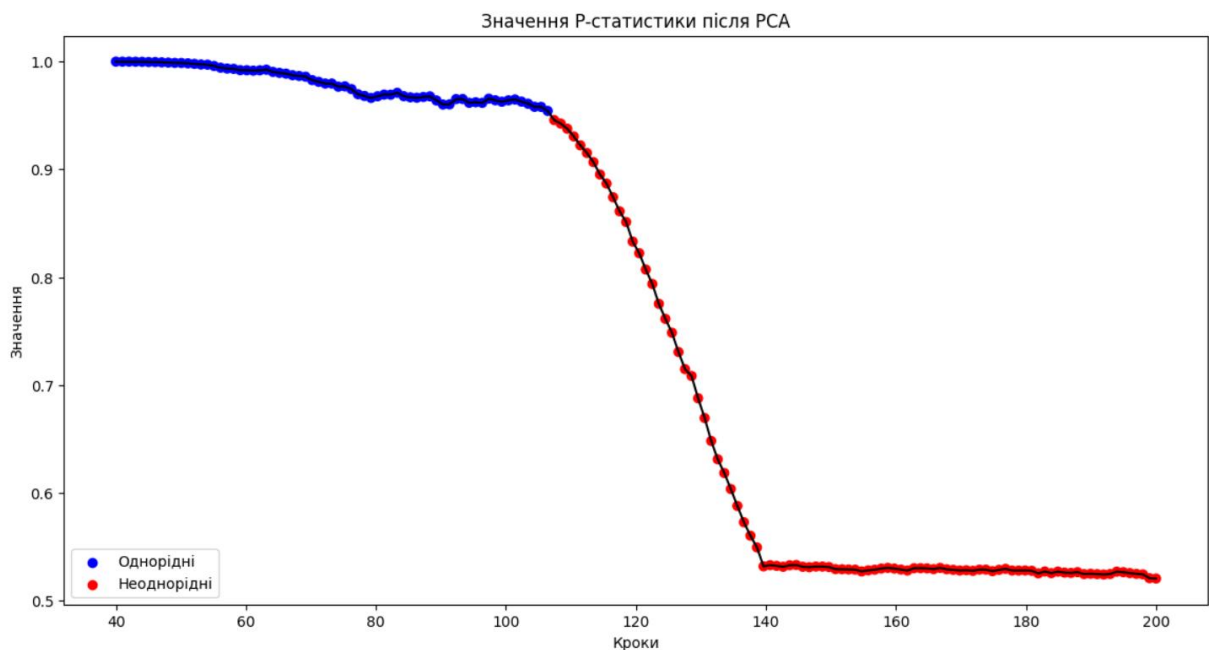


Рис. 3.5.10. Значення Р-статистики для логнормальних розподілів розмірності 100 зі стрибком у дисперсії після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 107

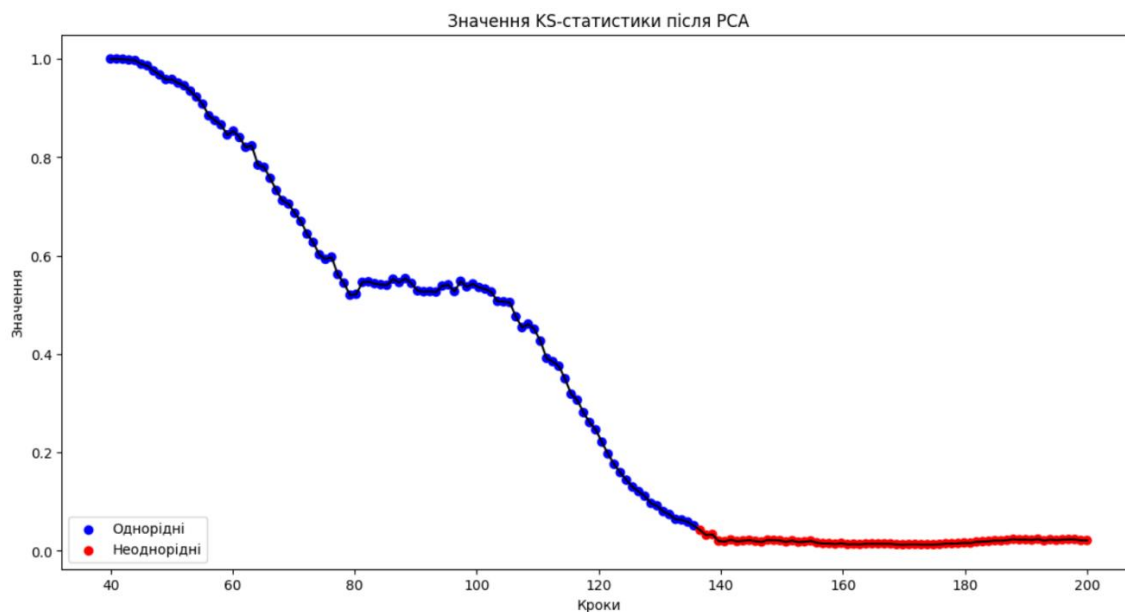


Рис. 3.5.11. Значення KS-статистики для логнормальних розподілів розмірності 100 зі стрибком у дисперсії після використання PCA. Сині точки вказують на час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Перша точка переходу — 136

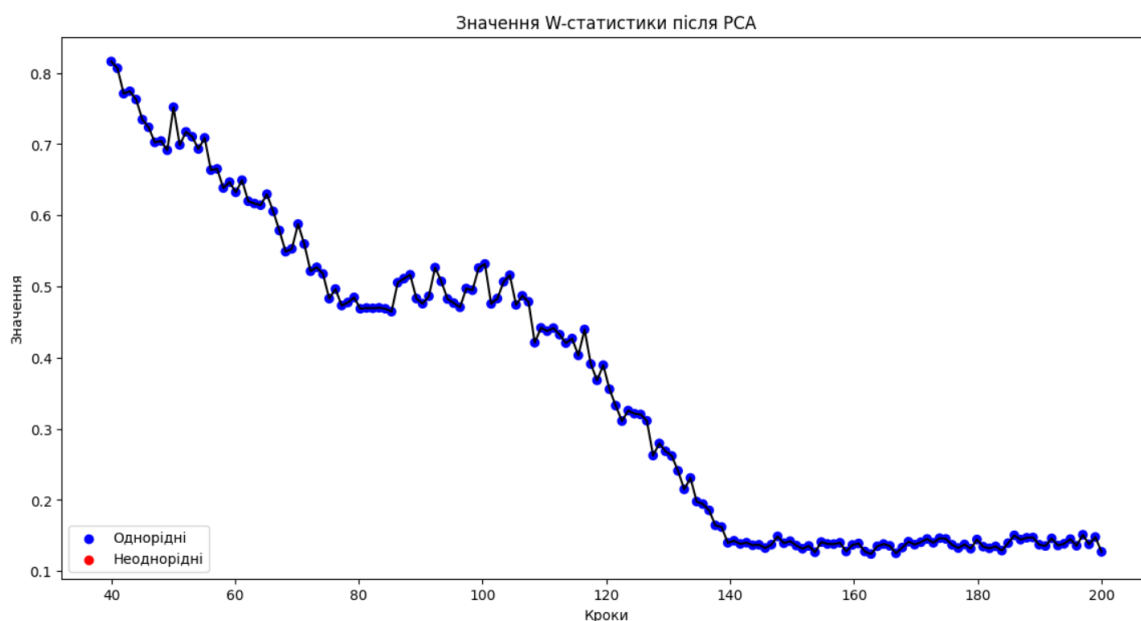


Рис. 3.5.12. Значення W-статистики для логнормальних розподілів розмірності 100 зі стрибком у дисперсії після використання PCA. Сині точки вказують на

час, коли значення не менші за 0.95, а червоні точки вказують на протилежне, коли вони менші. Не виявлено жодної точки переходу

Як видно з Таблиці 3.5.3, шукана перша точка переходу дорівнює 100, а потім відбулося відхилення від початкового розподілу. З Рисунку 3.5.10 видно, що перша знайдена точка переходу дорівнює 107, тобто через 7 кроків після початку стрибка. З Рисунку 3.5.11 видно, що перша знайдена точка переходу дорівнює 136, тобто через 36 кроки після початку стрибка. З Рисунку 3.5.12 видно, що статистика Вілкоксона не знайшла жодної точки переходу. Так само як і в попередньому експерименті, Р-статистика показує кращі результати за інші статистики.

Висновки до підрозділу 3.5

У цьому підрозділі було проведено порівняльний аналіз трьох статистик — р-статистики Петуніна, KS-статистики Колмогорова–Смірнова та W-статистики Манна–Уїтні–Вілкоксона — на багатовимірних часових рядах, сформованих із різних розподілів.

Результати експериментів показали, що при перевірці гіпотези про зсув середнього значення усі три статистики змогли коректно визначити момент зміни. У цьому випадку р-статистика не поступалася конкурентам і забезпечила точну локалізацію точки переходу.

Проте у випадках перевірки гіпотези про зміну масштабу або дисперсії переваги р-статистики стали очевидними. Лише вона та KS-статистика були здатні виявити зміну, але KS-критерій виявився надмірно чутливим до викидів і давав коливні результати. Натомість р-статистика працювала стабільніше, точніше та виявляла першу точку переходу швидше за інші методи. W-статистика в умовах зміни масштабу взагалі не продемонструвала задовільних результатів.

Таким чином, саме р-статистика Петуніна виявила себе найбільш ефективною у задачах детектування аномалій, зумовлених варіативністю чи розсіюванням даних. Її поєднання з методами зменшення розмірності (зокрема, PCA) відкриває нові можливості у виявленні точок переходу у складних динамічних системах, а також в ідентифікації змін режимів роботи кіберфізичних чи клінічних систем. Це робить її перспективним інструментом для сучасних задач аналізу багатовимірних часових рядів.

3.6. Симуляційні експерименти на квазіреальних даних

Для перевірки універсальності та практичної цінності розробленого підходу були проведені симуляційні експерименти, що імітують роботу складних технічних і біомедичних систем у квазіреальних умовах. На відміну від чисто математичних прикладів, ці сценарії відображають можливі критичні ситуації, які можуть виникати в інженерній практиці чи клінічній діагностиці.

Мета цих досліджень полягає у тому, щоб перевірити здатність алгоритму виявляти приховані зміни станів систем, що супроводжуються появою аномалій у багатовимірних часових рядах. Такий підхід дозволяє оцінити ефективність алгоритму не лише на модельних даних, а й у випадках, наближених до реальних.

У межах роботи були змодельовані такі сценарії:

- **Перегрів ядерного реактора** — для перевірки виявлення техногенних аварійних ситуацій.
- **Захоплення безпілотного літального апарата** — для моделювання кіберфізичних атак.
- **Пробойна в корпусі корабля та потрапляння судна у шторм** — для демонстрації морських аварій.
- **Імітація тахікардії, коронавірусної пневмонії та раку легенів** — для перевірки медичних застосувань алгоритму.

- **Імітація фізичної активності та стресу/паніки** — для демонстрації можливостей у сфері моніторингу психофізіологічного стану людини.

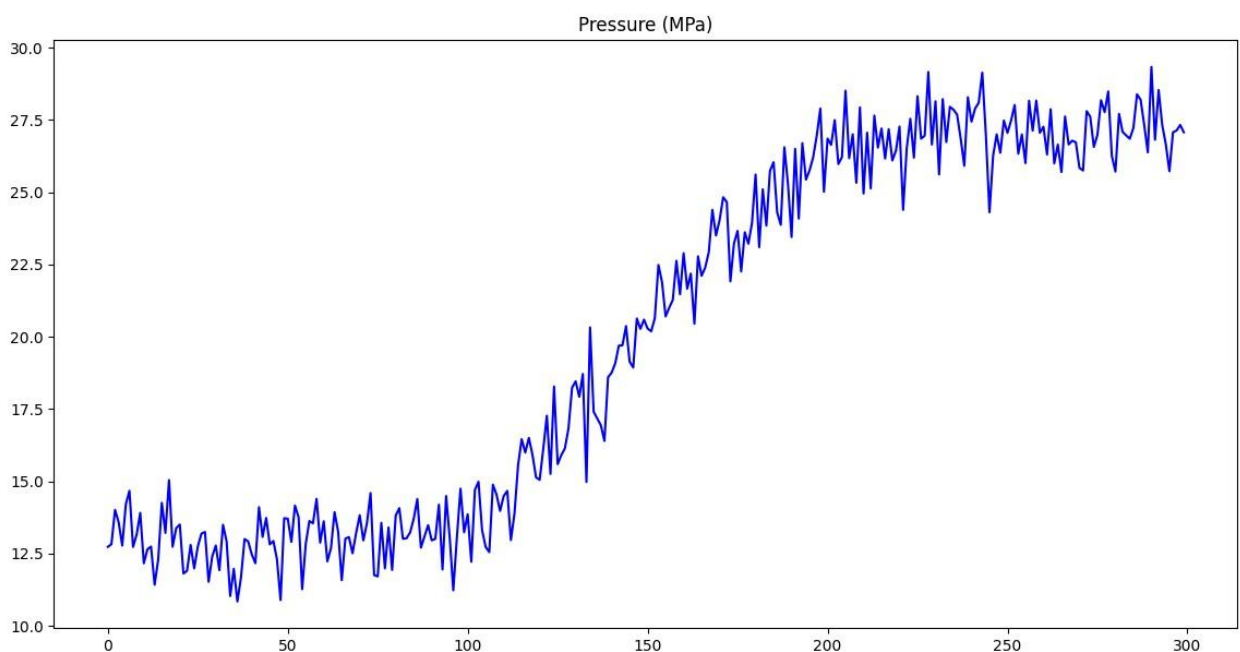
Таким чином, проведені симуляції дозволяють дослідити роботу алгоритму у міждисциплінарних задачах: від моніторингу технічних систем до клінічних застосувань у медицині. Це суттєво підсилює практичну значущість розробленого підходу.

Приклад 1. Перегрів ядерного реактора

Спочатку розглянемо поширену проблему, яка спричиняє аварії на атомних електростанціях: перегрів реактора та пожежа з подальшим розплавленням активної зони.

Часовий інтервал	Тиск МПа	Температура, °C	Потужність реактора, %
0-99 Нормальний стан	N(13,1)	N(290,10)	N(60,1)
100-199 Поломка	перехід		
200-299 Критичний стан	N(27,1)	N(650,10)	N(95,1)

Таблиця 3.5.1. Часові інтервали та розподіли ймовірностей для моделювання перегріву реактора



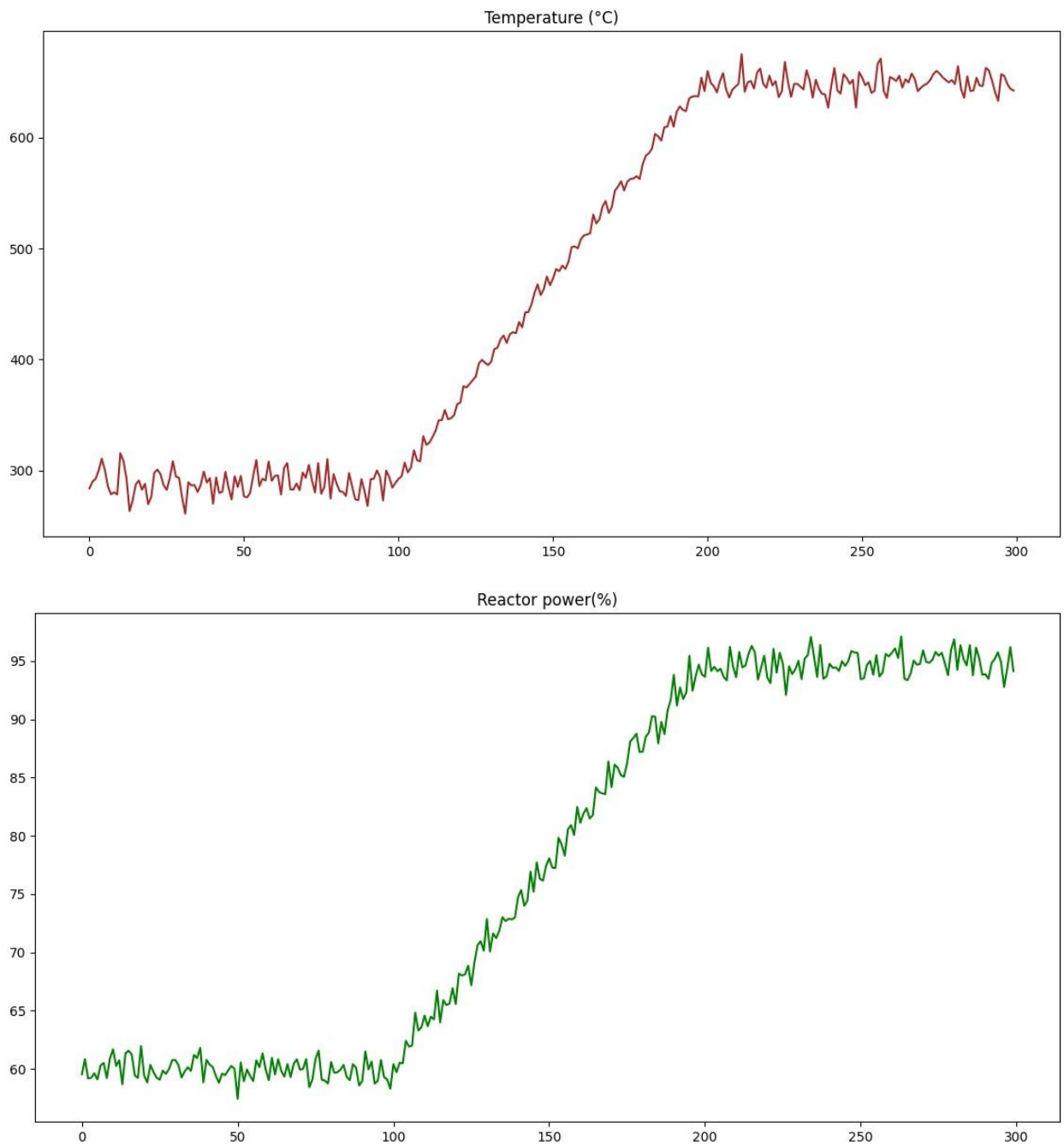


Рис. 3.5.1. Часові ряди тиску, температури та потужності реактора в симуляції перегріву реактора

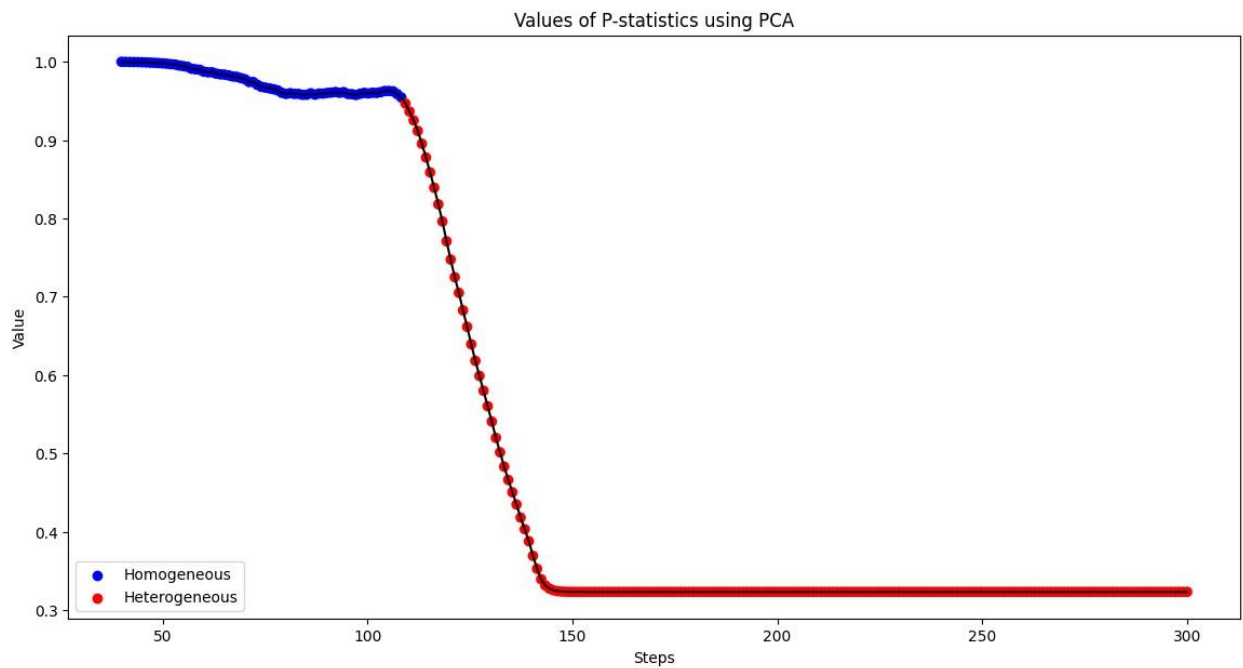


Рис. 3.5.2. Р-значення для часових рядів тиску, температури та потужності реактора в симуляції перегріву. Сині точки представляють моменти часу зі значеннями, більшими або рівними 0,95, тоді як червоні точки вказують на моменти, коли значення менші за 0,95.

У таблиці 3.5.1 та на рисунку 3.5.1 чітко показано цільову точку переходу на 100. Рисунок 3.5.2 показує, що виявлена точка зміни близько наближається до фактичної, а саме на 109, що відповідає 10 секундам після початку аварії.

Як показано на графіку параметрів (рисунок 3.5.1), в межах робочої зони реактора було визначено критичну точку (рисунок 3.5.2), де тиск, температура та потужність реактора почали швидко зростати. Визнаючи це, можна вжити оперативних заходів, таких як зупинка або охолодження реактора, щоб запобігти перегріву та зберегти цілісність системи.

Приклад 2. Симуляція викрадення безпілотної літальної апарату

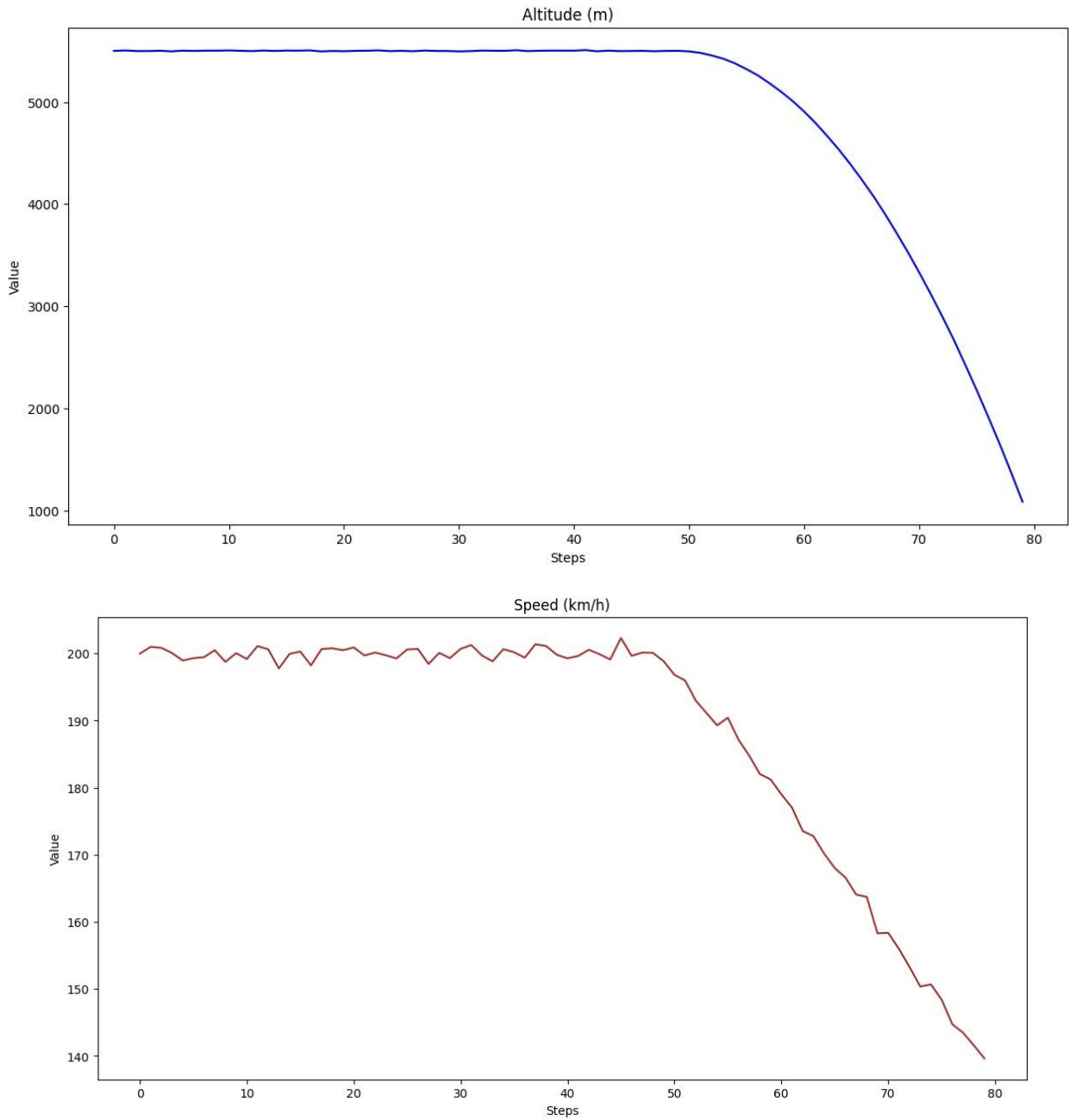
Розглянемо проблему викрадення безпілотної апарату, тобто коли хакер дистанційно унеможливує керування дроном через оператора та хоче або знищити його до досягнення цілі, або перехопити та прийняти на озброєння.

Ця модель відстежуватиме 4 характеристики: висоту, швидкість,

горизонтальне та вертикальне відхилення. Почнемо з того, що наш дрон летить прямо до цілі, і в певний момент його перехоплюють і намагаються знищити, падаючи на землю. Подивимося, як швидко внутрішній ШІ дрона зможе визначити, що його викрали.

Часовий інтервал	Висота m	Швидкість km/h	Гор. Відх. °	Верт. Відх. °
0-49 Нормальний стан	N(5500,3)	N(200,1)	N(0,2)	N(0,1)
50-79 Викрадення	перехід			

Таблиця 3.5.2. Часові інтервали та розподіли ймовірностей для симуляції викрадення безпілотної літальної апарату



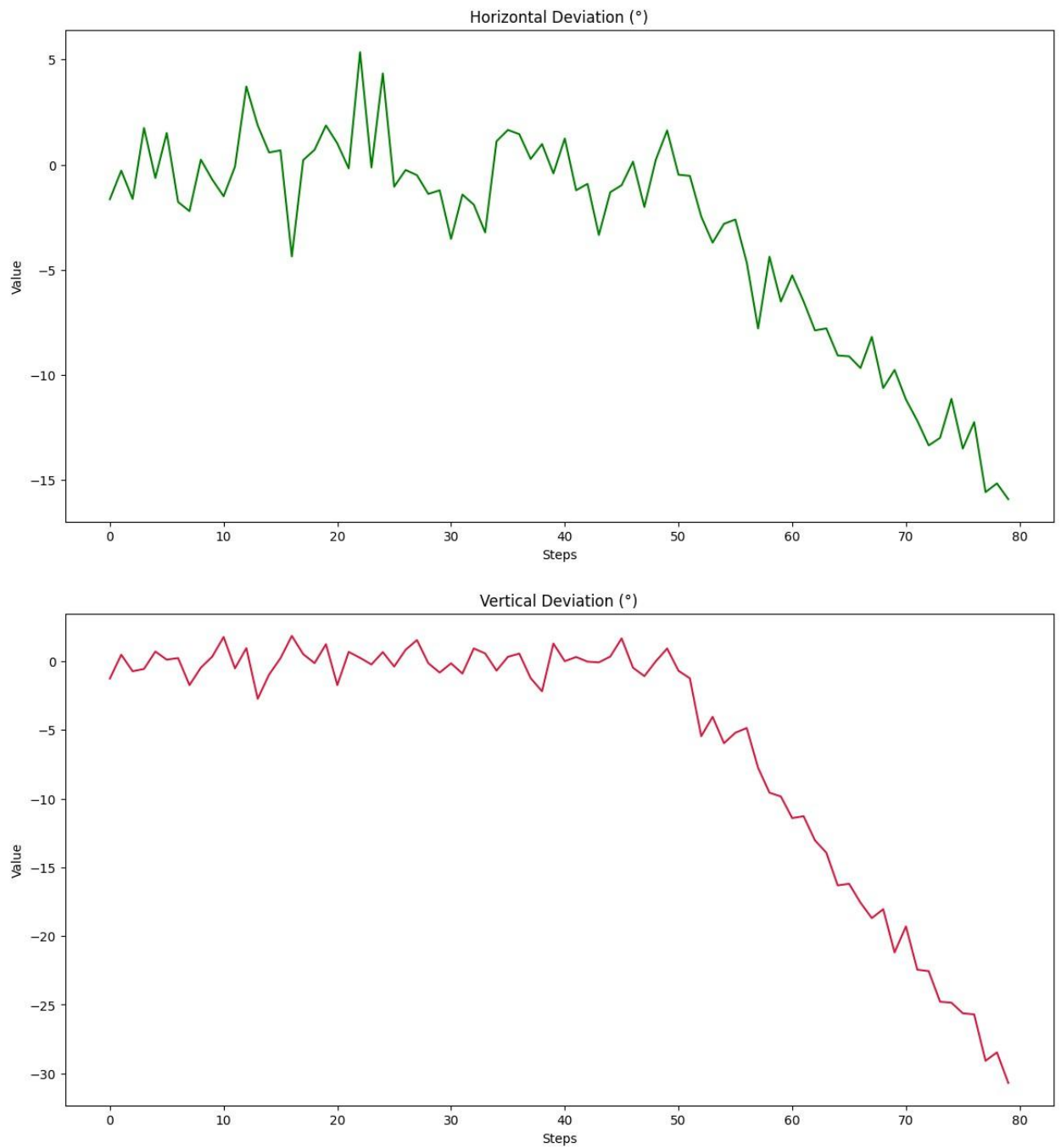


Рис. 3.5.3. Часові ряди висоти, швидкості, горизонтального та вертикального відхилення під час симуляції захоплення безпілотної літального апарату

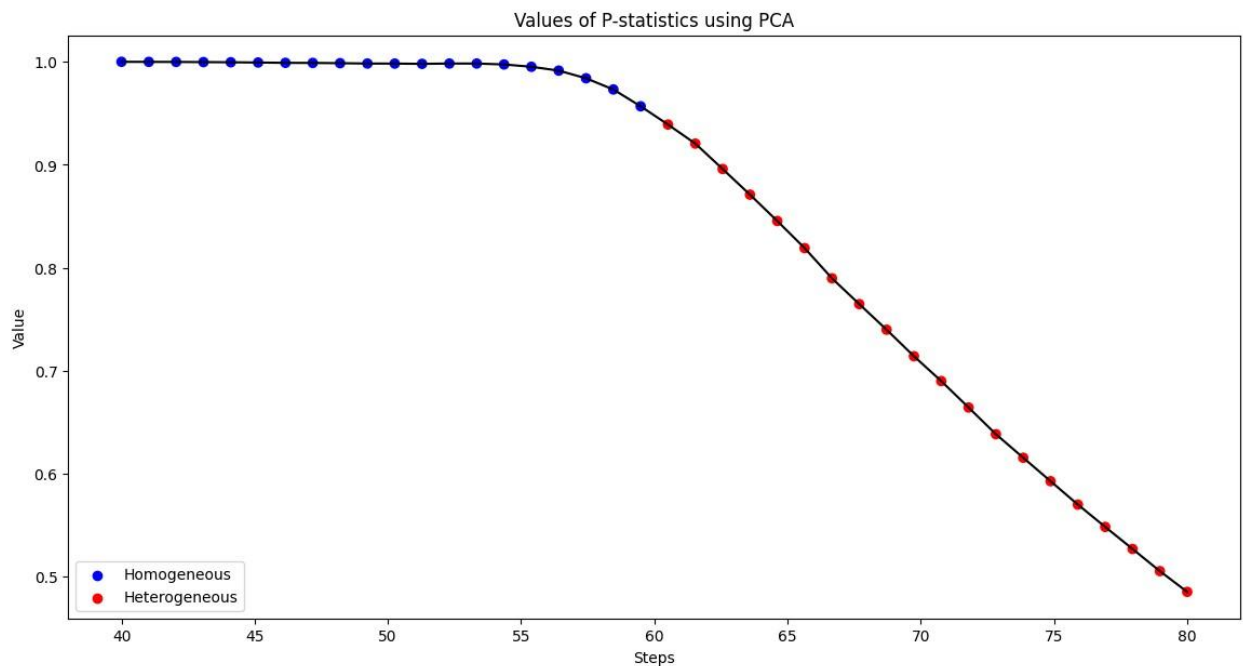


Рис. 3.5.4. Р-значення для висоти, швидкості, горизонтального та вертикального відхилення в симуляції захоплення безпілотного літального апарату. Сині точки представляють моменти часу зі значеннями, більшими або рівними 0,95, тоді як червоні точки вказують на моменти, коли значення менші за 0,95

У таблиці 3.5.2 та на рисунку 3.5.3 чітко показано цільову точку переходу на 50. Рисунок 3.5.4 показує, що виявлена точка зміни близько наближається до фактичної, а саме на 60, що відповідає 10 секундам після початку аварії.

Динаміка параметрів керування, зображена на (Рис. 3.5.3), виявляє чітку закономірність під час імітації захоплення БПЛА. Виявлення цих закономірностей (Рис. 3.5.4) дає можливість втрутитися, потенційно відновити контроль або вжити контрзаходів для запобігання несанкціонованому доступу та маніпуляціям.

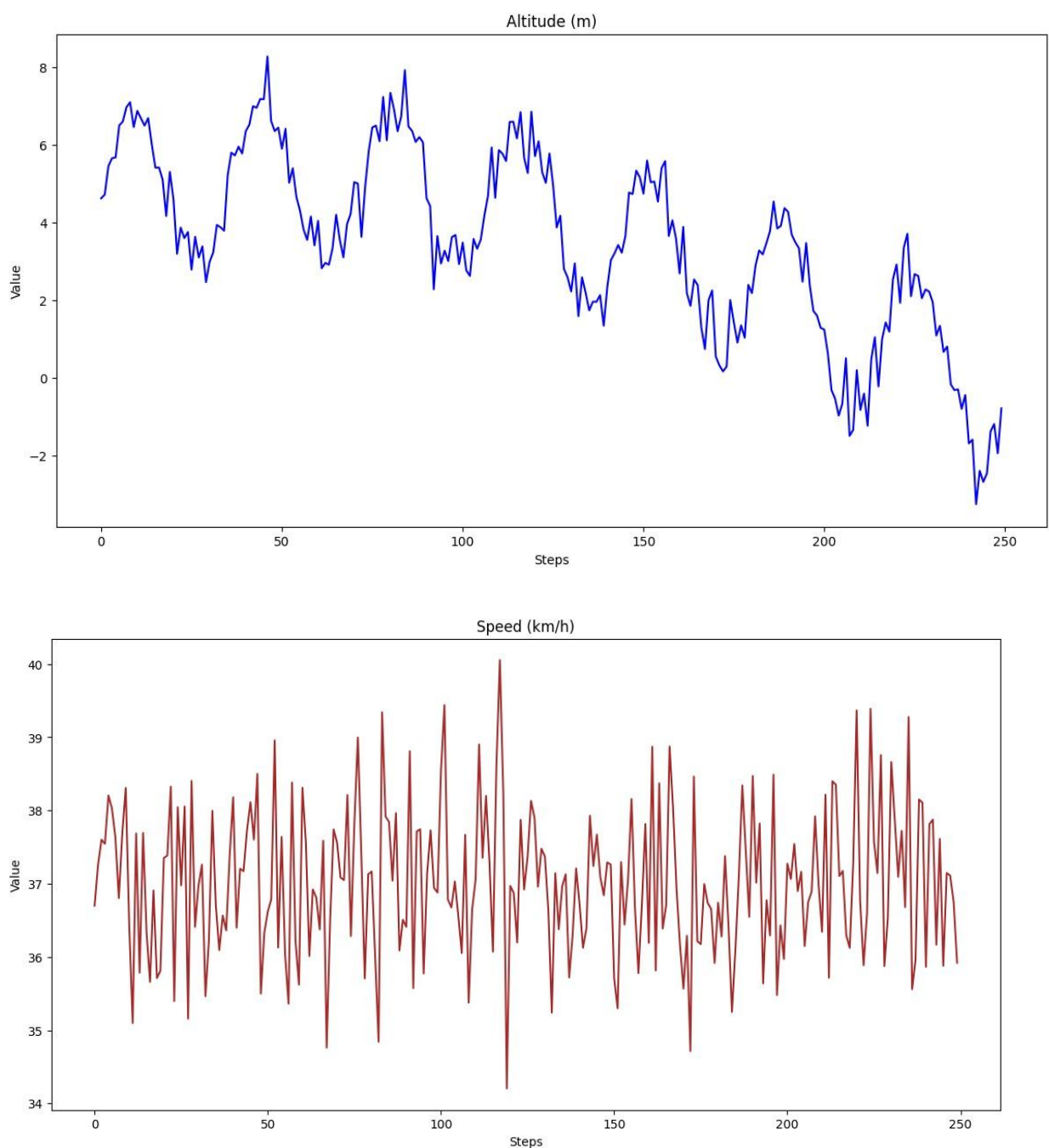
Приклад 3. Моделювання пробойни в кораблі

Розглянемо симуляцію пробойни в кораблі, коли на кораблі змінюється лише висота встановленого датчика над рівнем моря. Щоб ускладнити ситуацію, залишимо незмінними ще два значення – швидкість і кут відхилення.

Подивимося, наскільки глибоко зануриться наш корабель, перш ніж пробоїну буде помічено.

Часовий інтервал	Висота m	Швидкість km/h	Кут відх. °
0-99 Нормальний стан	$2 \cdot \sin(10t) + 5 + N(0,0.5)$	N(37,1)	N(0,2)
100-249 Пробоїна та затоплення	перехід	N(37,1)	N(0,2)

Таблиця 3.5.3. Часові інтервали та розподіли ймовірностей для симуляції судна, що потрапило в шторм



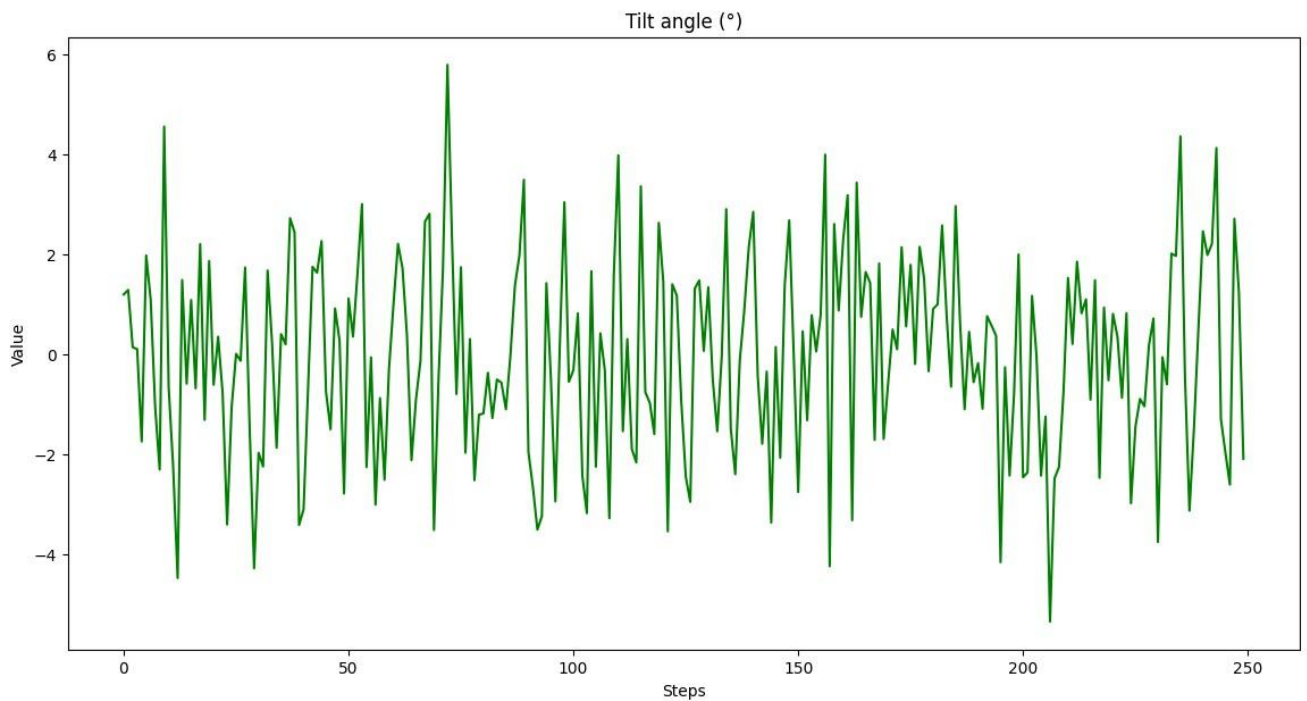


Рис. 3.5.5. Часові ряди висоти, швидкості та кута нахилу при моделюванні пробоїни в кораблі

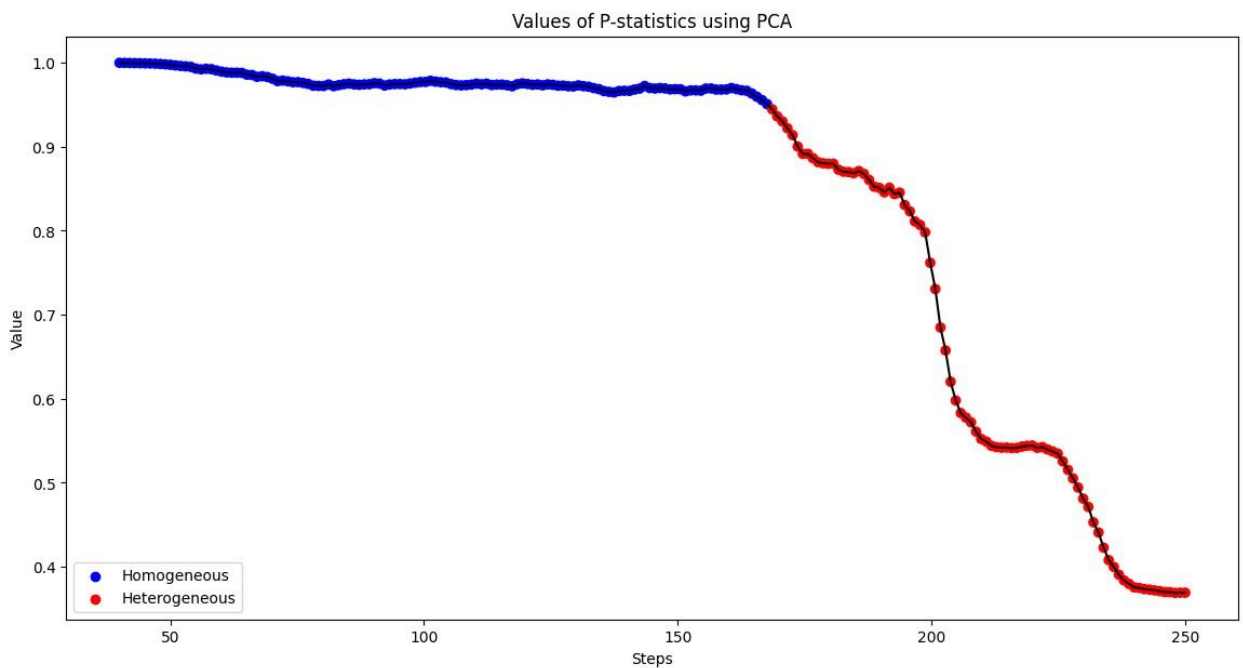


Рис. 3.5.6. Значення Р-статистики висоти, швидкості та кута нахилу при моделюванні пробоїни в кораблі. Сині точки вказують на час, коли значення не менше 0,95, а червоні точки вказують на протилежне, коли вони менше

У таблиці 3.5.3 та на рисунку 3.5.5 чітко показано точку зміни цільового показника на 100. На рисунку 3.5.6 показано, що виявлена точка зміни дорівнює 168, що відповідає 68 секундам після початку затоплення. Досліджуючи дані часових рядів щодо цілісності корпусу судна (рисунок 3.5.5), можна помітити помітне відхилення, що вказує на потенційний сценарій пошкодження (рисунок 3.5.6). Це спостереження спонукає до подальших досліджень щодо розробки систем моніторингу, які могли б виявляти такі аномалії на ранній стадії, що дозволяло б своєчасно вживати превентивних або коригувальних заходів.

Приклад 4. Моделювання корабля, що потрапив у шторм

Давайте змоделюємо ситуацію, коли судно, на якому встановлено датчик, що вимірює його висоту над рівнем моря, швидкість та кут нахилу, потрапляє у шторм. Спочатку відбудеться незначне погіршення погоди, що призведе до сильнішого підйому та падіння нашого судна на хвилях, а також збільшить розкид швидкості та кута нахилу. Потім погода ще більше погіршиться, що призведе до подальшого відхилення показань. Давайте розглянемо, як швидко виявити погіршення погоди за допомогою цих датчиків.

Часовий інтервал	Висота m	Швидкість km/h	Кут відх. °
0-99 Нормальний стан	$2*\sin(10t)+5+N(0,0.5)$	$N(37,1)$	$N(0,2)$
100-299 Погіршення	$3.5*\sin(10t)+5+N(0,0.5)$	$N(37,2)$	$N(0,4)$
300-499 Шторм	$5*\sin(10t)+5+N(0,0.5)$	$N(37,3)$	$N(0,6)$

Таблиця 3.5.4. Часові інтервали та розподіли ймовірностей для симуляції судна, що потрапило в шторм

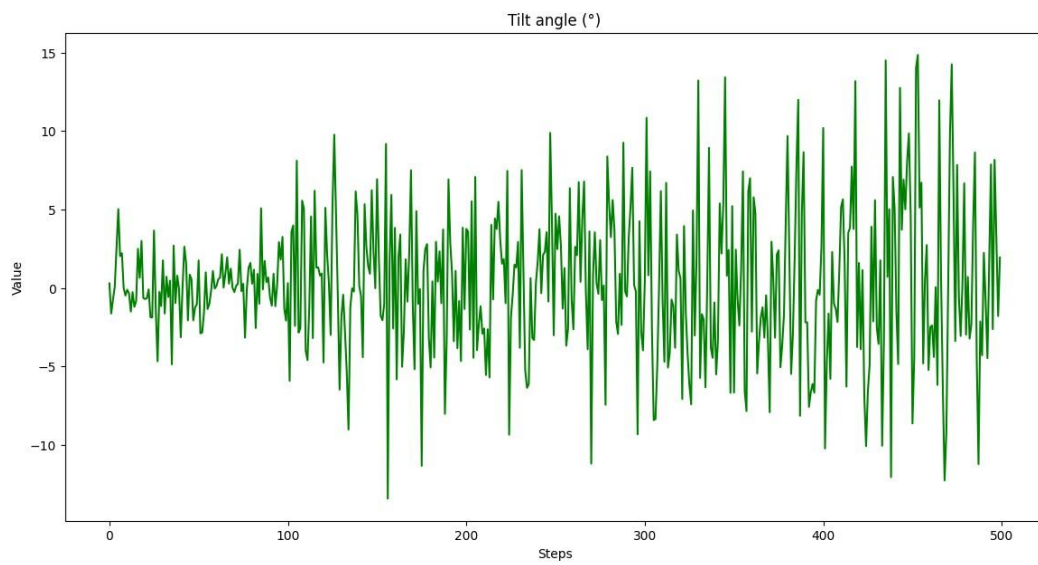
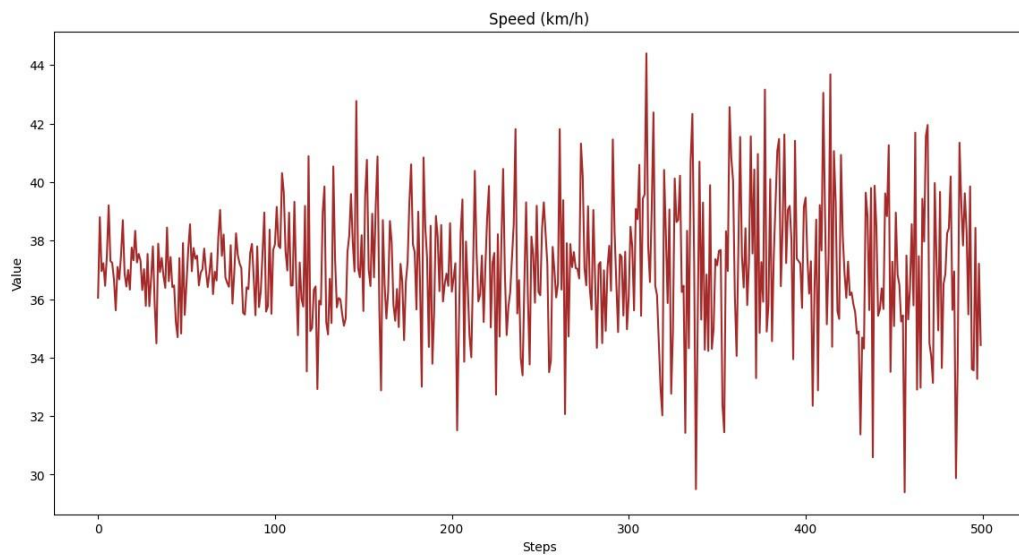
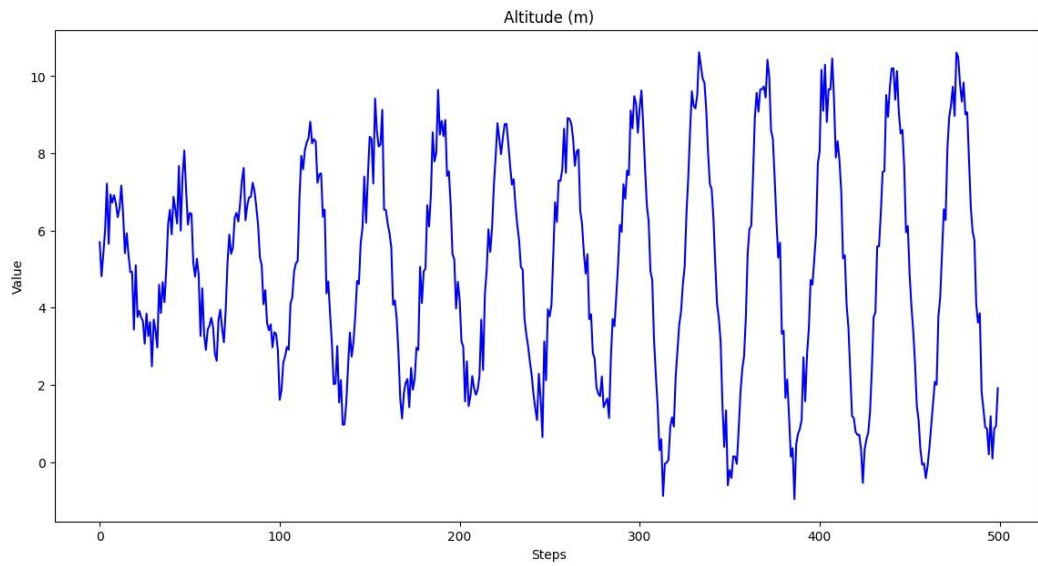


Рис. 3.5.7. Часові ряди висоти, швидкості та кута нахилу для симуляції судна, що потрапило в шторм

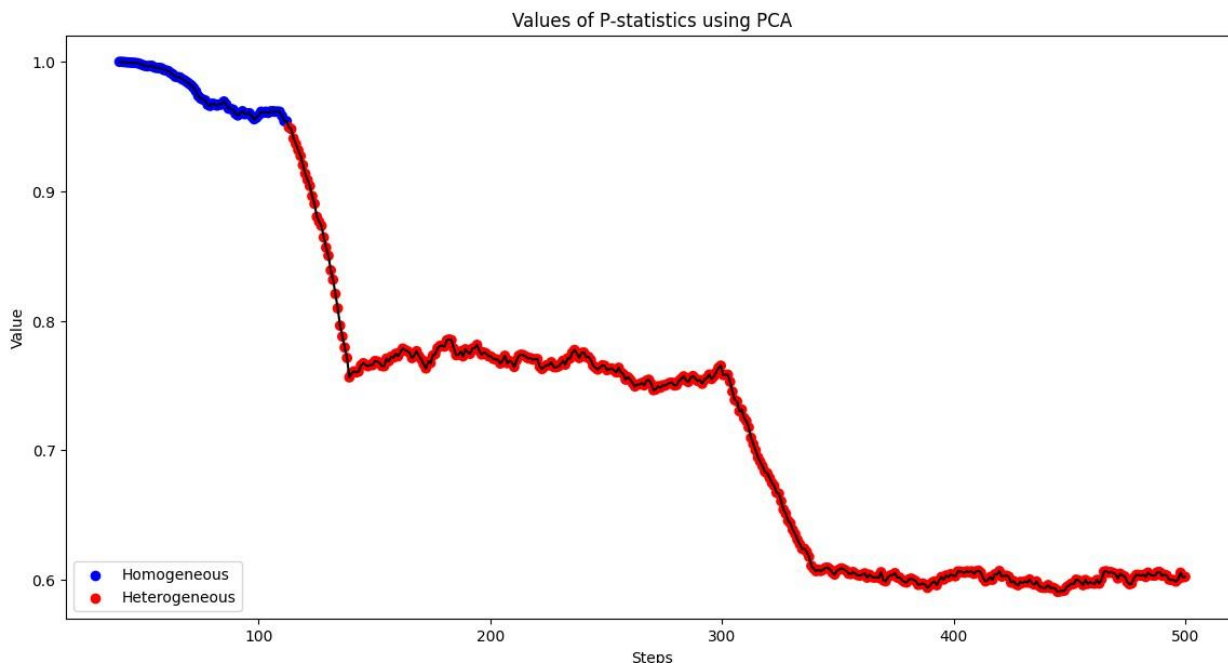


Рис. 3.5.8. Значення Р-статистики висоти, швидкості та кута нахилу судна, що потрапило в симуляцію шторму. Сині точки представляють моменти часу зі значеннями, більшими або рівними 0,95, тоді як червоні точки вказують на моменти, коли значення менші за 0,95

У таблиці 3.5.4 та на рисунку 3.5.7 чітко показано цільову точку зміни на 100. На рисунку 3.5.8 видно, що виявлена точка зміни – 113, що відповідає 13 секундам після початку погіршення, що дозволить нам вжити заходів для зменшення шкоди від погоди, таких як тимчасова зміна курсу, видача наказів екіпажу або повна зупинка, поки шторм не мине. Графічне представлення параметрів суден під час імітованого шторму (рисунок 3.5.7) підкреслює проблеми, з якими стикаються судна в екстремальних погодних умовах. Це вимагає додаткових досліджень удосконалених систем попередження та контролю для підвищення безпеки на морі, надання екіпажам ефективних інструментів для навігації та пом'якшення наслідків сильних штормів. Своєчасне виявлення змін (рисунок 3.5.8) дозволяє вжити відповідних заходів на борту судна, щоб уникнути аварій та травмування персоналу.

Приклад 5. Імітація тахікардії

Проміжок часу	Розподіл HR	Розподіл SpO_2	Розподіл T	Знайдені точки переходу
0 – 199 (здоровий стан)	$U(60,80)$	$U(96,100)$	$U(36.5,36.7)$	[201,401,601]
200 – 399 (тахікардія)	$U(100,120)$	$U(92,95)$	$U(37,37.2)$	
400 – 599 (здоровий стан)	$U(60,80)$	$U(96,100)$	$U(36.5,36.7)$	
600 – 799 (ще гірша тахікардія)	$U(120,140)$	$U(91,93)$	$U(37.2,37.5)$	

Таблиця 3.5.5. Часові інтервали і розподіли медичних показників для симуляції пацієнта з тахікардією

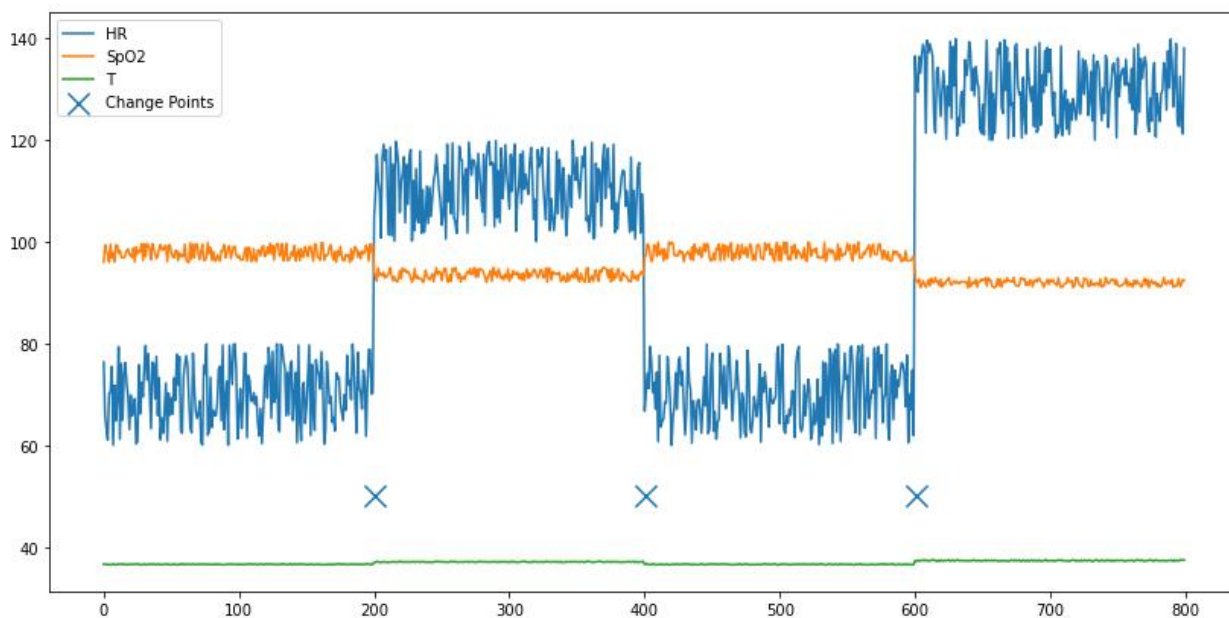


Рис. 3.5.9. Приклад часового ряду, що складається з медичних показників для симуляції пацієнта з тахікардією та їх точками переходів

Приклад 6. Імітація коронавірусної пневмонії

Проміжок часу	Розподіл HR	Розподіл SpO_2	Розподіл T	Знайдені точки переходу
0 – 199 (здоровий стан)	$U(60,80)$	$U(96,100)$	$U(35.5,37.2)$	[201,401,601]
200 – 399 (лихорадка і гіпоксія)	$U(60,80)$	$U(90,95)$	$U(37.8,39.0)$	
400 – 599 (здоровий стан)	$U(60,80)$	$U(96,100)$	$U(35.5,37.2)$	
600 – 799 (ще гірша лихорадка і гіпоксія)	$U(60,80)$	$U(88,90)$	$U(38.5,39.5)$	

Таблиця 3.5.6. Часові інтервали і розподіли медичних показників для симуляції пацієнта з коронавірусною пневмонією

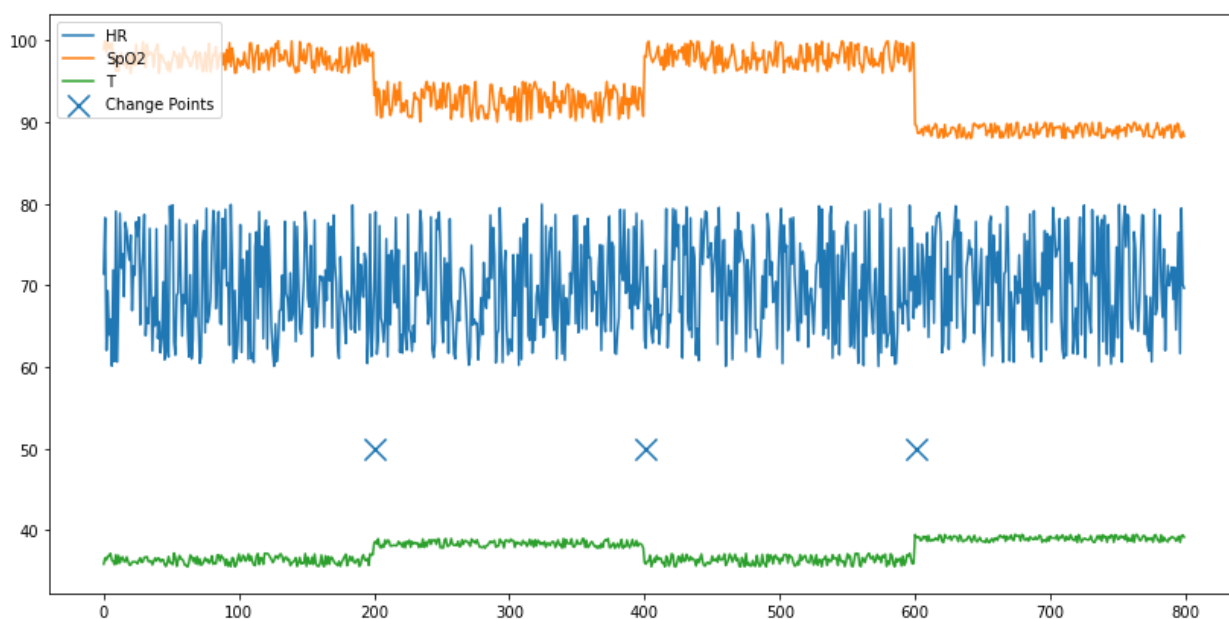


Рис. 3.5.10. Приклад часового ряду, що складається з медичних показників для симуляції пацієнта з коронавірусною пневмонією та їх точками переходів

Приклад 7. Імітація раку легенів

Проміжок часу	Розподіл HR	Розподіл SpO_2	Розподіл T	Знайдені точки переходу
0 – 199 (здоровий стан)	$U(60,80)$	$U(96,100)$	$U(36.5,36.7)$	[201, 401, 601]
200 – 399 (рак легенів)	$U(65,93)$	$U(94.5,95.5)$	$U(37.0,37.2)$	
400 – 599 (здоровий стан)	$U(60,80)$	$U(96,100)$	$U(36.5,36.7)$	
600 – 799 (рак легенів гірша стадія)	$U(70,90)$	$U(94,95)$	$U(37,38)$	

Таблиця 3.5.7. Часові інтервали і розподіли медичних показників для симуляції пацієнта з раком легенів

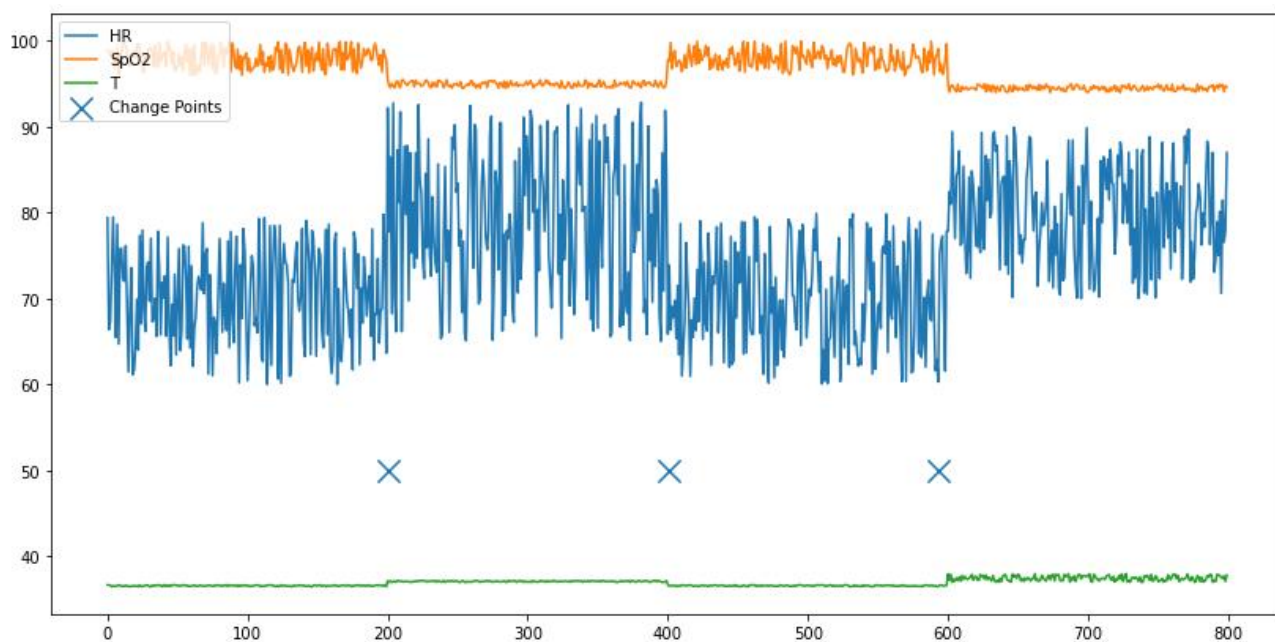


Рис. 3.5.11. Приклад часового ряду, що складається з медичних показників для симуляції пацієнта з раком легенів та їх точками переходів

Приклад 8. Імітація фізичної активності

Проміжок часу	Розподіл HR	Розподіл SpO_2	Розподіл T	Знайдені точки переходу
0 – 199 (спокійний стан)	$U(70,77)$	$U(99,99.4)$	$U(36.5,36.7)$	[201,401,601]
200 – 399 (фізична активність)	$U(95,120)$	$U(92,94)$	$U(37,37.3)$	
400 – 599 (спокійний стан)	$U(70,77)$	$U(99,99.4)$	$U(36.5,36.7)$	
600 – 799 (інтенсивніше тренування)	$U(100,130)$	$U(91.5,93.5)$	$U(37.1,37.4)$	

Таблиця 3.5.8. Часові інтервали і розподіли медичних показників для симуляції пацієнта з фізичною активністю

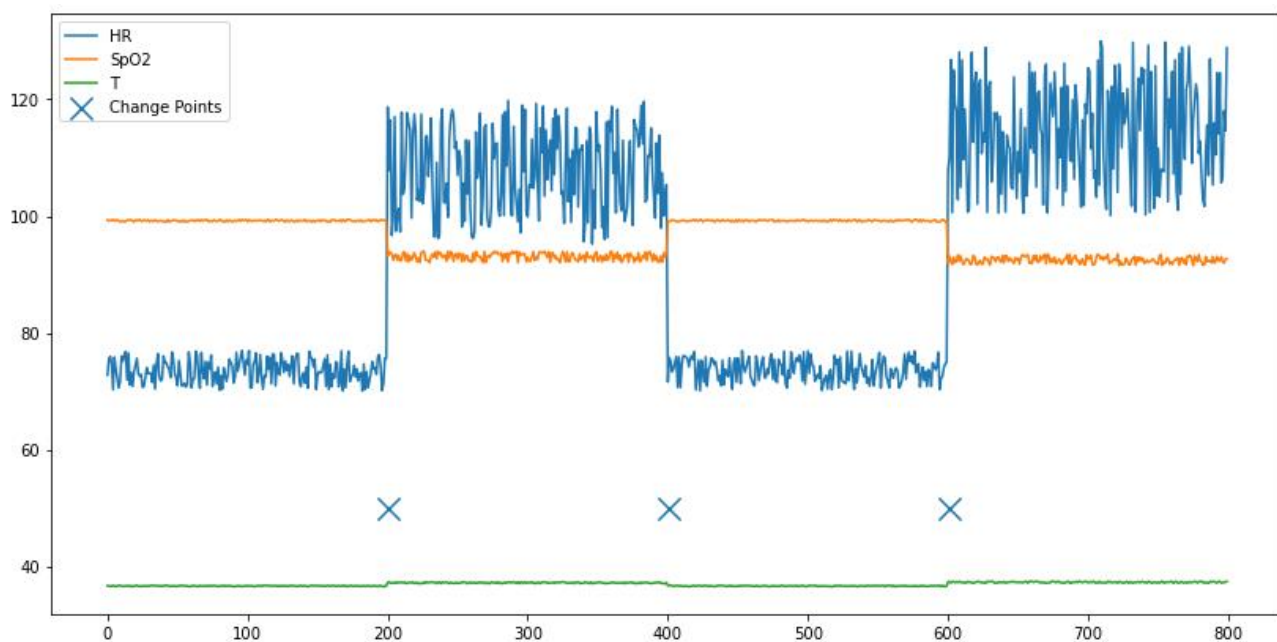


Рис. 3.5.12. Приклад часового ряду, що складається з медичних показників для симуляції пацієнта з фізичною активністю та їх точками переходів

Приклад 9. Імітація стресу\паніки

Проміжок часу	Розподіл HR	Розподіл SpO_2	Розподіл T	Знайдені точки переходу
0 – 199 (спокійний стан)	$U(70.9,77.3)$	$U(99,99.4)$	$U(36.5,36.7)$	[201,401,601]
200 – 399 (стресовий стан)	$U(87,89.8)$	$U(98.6,98.8)$	$U(37,37.3)$	
400 – 599 (спокійний стан)	$U(70.9,77.3)$	$U(99,99.4)$	$U(36.5,36.7)$	
600 – 799 (тривалий стресовий стан)	$U(89.1,91.5)$	$U(97.9,98.1)$	$U(37.1,37.4)$	

Таблиця 3.5.9. Часові інтервали і розподіли медичних показників для симуляції пацієнта зі стресом і панікою

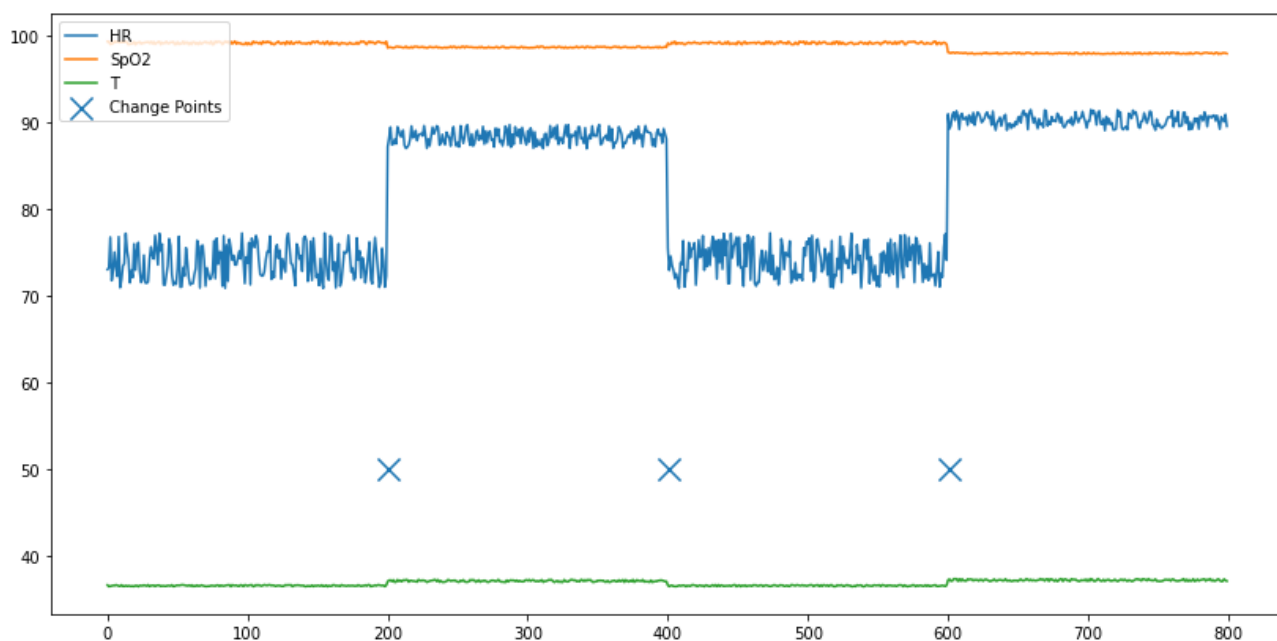


Рис. 3.5.13. Приклад часового ряду, що складається з медичних показників для симуляції пацієнта зі стресом і панікою та їх точками переходів

Висновки до підрозділу 3.6

У даному підрозділі було проведено серію симуляційних експериментів, спрямованих на оцінювання ефективності запропонованого непараметричного методу виявлення точок переходу у багатовимірних часових рядах, що моделюють поведінку кіберфізичних систем. Отримані результати підтвердили здатність алгоритму забезпечувати точну ідентифікацію аномалій та змін станів системи за умов широкого спектра розподілів (нормальних, рівномірних, експоненціальних, гамма-, гumbелівських та розподілів Релея) і при різних сценаріях зміни їхніх параметрів.

Особливістю розглянутих експериментів є використання квазіреальних даних, що відтворюють критичні сценарії як техногенного, так і медичного характеру: перегрів ядерного реактора, викрадення безпілотного літального апарата, пробоїна у корпусі корабля, потрапляння судна у шторм, тахікардія, коронавірусна пневмонія, рак легенів, підвищена фізична активність, стрес чи панічні стани. У всіх випадках метод на основі Р-статистики продемонстрував стійкість до шуму та викидів, здатність коректно локалізувати точки переходу й забезпечувати інтерпретовані результати.

Таким чином, проведені дослідження підтвердили високу теоретичну та прикладну цінність розробленого підходу. Його універсальність та адаптивність створюють передумови для інтеграції в системи підтримки прийняття рішень, моніторингу складних динамічних об'єктів та клінічної діагностики. Крім того, поєднання статистичного аналізу з техніками зменшення розмірності забезпечує пояснюваність результатів (XAI), що є важливою вимогою до сучасних інтелектуальних систем.

Висновки дисертації

- Вперше обґрунтовано та доведено консистентність і потужність Р-статистики Петуніна, що дало змогу забезпечити строгий теоретичний фундамент для подальшого використання цієї статистики у задачах виявлення точок переходу.
- Побудовано уніфіковану схему застосування методів зменшення розмірності (PCA, FLD) у поєднанні з р-статистикою, яка дозволяє ефективно працювати з багатовимірними часовими рядами.
- Розроблено та адаптовано метод побудови еліпсоїдів Петуніна для багатовимірних випадків, що дало змогу підвищити точність і стабільність при аналізі складних даних.
- Проведено порівняльний аналіз трьох критеріїв — Р-статистики, KS-статистики (Колмогорова–Смірнова) та W-статистики (Манна–Уїтні–Вілкоксона). Експериментально доведено, що саме Р-статистика забезпечує найкращий результат при зміні масштабу та дисперсії даних, тоді як альтернативні критерії або надмірно чутливі, або втрачають ефективність.
- Вперше проведено чисельні та квазіреальні експерименти із застосуванням запропонованого алгоритму для моделювання змін станів у складних системах. Досліджено широкий спектр розподілів (нормальні, логнормальні, рівномірні, експоненціальні, гамма- та гumbелівські, розподіли Релея, Лапласа і логістичні) та реалістичних сценаріїв: перегрів реактора, викрадення безпілотного літального апарата, пробоїна в кораблі, шторм, тахікардія, коронавірусна пневмонія, рак легенів, фізичне навантаження, стрес/паніка.
- Експерименти підтвердили, що запропонований підхід не тільки забезпечує високу точність і стабільність, а й дозволяє отримати зрозумілу та інтерпретовану (XAI) картину змін у багатовимірних часових рядах.

- Запропонований алгоритм має високу практичну цінність для систем підтримки прийняття рішень, зокрема у сфері моніторингу техногенних процесів та медичної діагностики, де своєчасне виявлення аномалій є критично важливим.

Список використаних джерел:

1. Klyushin D., Urazovskyi A. Nonparametric Test for Change-Point Detection of IoT Time-Series Data // *Chapter in: Kumar P., Obaid A., Cengiz K., Balas A. (Eds.) A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems*, Intelligent Systems Reference Library, vol. 210, Springer, 2022, pp. 99-122.
2. Klyushin D., Urazovskyi A. Non-Parametric Methods of Finding Changepoints in Multidimensional Time Series // *In: Proceedings of the 3rd International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk, Neubiberg, Germany, January 12, 2023)* Pickl S. et al (Eds.), pp. 106–118.
3. Klyushin D., Urazovskyi A. Locating changepoints in multidimensional time series using non-parametric methods // *In: Proceedings of the The Sixth International Workshop on Computer Modeling and Intelligent Systems, (CMIS, Zaporizhzhia, Ukraine, May 3, 2023)*
4. Klyushin D., Urazovskyi A. Statistical Algorithm for Change Point Detection in Multivariate Time Series of Medicine Data Based on Principles of Explainable Artificial Intelligence, 2023, *Medical Data Analysis and Processing using Explainable Artificial Intelligence*, pp. 58-90.
5. Klyushin D., Urazovskyi A. Locating Changepoints in Multidimensional Time Series Using Non-Parametric Methods // *In: Proceedings of the Modern Machine Learning Technologies and Data Science Workshop, (MoML&T&DS 2023)*, Emmerich M., Vysotska V., Lytvynenko V. (Eds.) Lviv, Ukraine, June 3, 2023, pp. 12–24.
6. Klyushin D., Urazovskyi A. Comparison of Dimensionality Reduction Methods for Finding Changepoints in Multidimensional Real-time Series // *In: Proceedings of the 4rd International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk, Neubiberg, Germany, January 26, 2024)* Pickl S. et al (Eds.).

7. Klyushin D., Urazovskiy A. A non-parametric statistical technique for anomaly detection in CPS. *Cyber Physical System 2.0: Communication and Computational Technologies. Volume 2*, Routledge, 2025.
8. Urazovskiy A. A non-parametric statistical technique for changepoint detection in cyber-physical systems. *Journal of Numerical and Applied Mathematics*, 2025, pp. 101-122.
9. 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.
10. 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. Pp. 1368–1374. doi:10.48550/arXiv.1510.04850.
11. 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.
12. Sorba O., Geissler C. Online Bayesian inference for multiple changepoints and risk assessment. 2021. arXiv preprint arXiv:2106.05834v1. doi:10.48550/arXiv.2106.05834.
13. 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.
14. Tickle S.O., Eckley I.A., Fearneadh P. A computationally efficient, highdimensional multiple change point procedure with application to global terrorism incidence. 2020. arXiv:2011.03599v2. doi:10.1111/rssc.12695.
15. Fearneadh P., Rigai G. Change point Detection in the Presence of Outliers. *Journal of the American Statistical Association*. 2018. Vol. 114. Pp. 169–183. doi:10.1080/016214512.2017.1385466.

16. Romano G., Eckley I., Fearnadh P., Rigai G. Fast Online Change point Detection via Functional Pruning CUSUM statistics. 2021. arXiv:2110.08205v2. doi:10.48550/arXiv.2110.08205.
17. Wang Z., Lin X., Mishra A., Sriharsha R. Online Change point Detection on a Budget. International Conference on Data Mining Workshops (ICDMW), 2022. Pp. 414–420. doi:10.1109/ICDMW53433.2021.00057.
18. 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.
19. Steude H., Windmann A., Niggemann O. Learning Physical Concepts in Cyber-Physical Systems: A Case Study. 2021. doi:10.48550/arXiv.2111.14151.
20. Aminikhanghahi S., Cook D.J. A survey of methods for time series change point detection. Knowledge and Information Systems. 2017. Vol. 51. Pp. 339–367.
21. 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.
22. 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.
23. Adams P., Mackay D. Bayesian Online Change point Detection. 2007. arXiv preprint arXiv:0710.3742v1. doi:10.48550/arXiv.0710.3742.
24. Cooney P., White A. Change-point Detection for Piecewise Exponential Models. 2021. arXiv preprint arXiv:2112.03962v1. doi:10.48550/arXiv.2112.03962.
25. 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.

26. 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. doi:10.1016/j.compbimed.
27. Renz K., Stache N., Fox G., Varol G., Albanie S. Sign Segmentation with Change point-Modulated Pseudo-Labelling. 2021. arXiv:2104.13817v1. doi:10.48550/arXiv.2104.13817.
28. 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
29. Zsolnai C., Lörch N., Arnold J. Neural Total Variation Distance Estimators for Changepoint Detection in News Data. *arXiv preprint arXiv:2506.18764*, 2025.
30. Murillo-Gonzalez A., Liu L. Situationally-Aware Dynamics Learning. *arXiv preprint arXiv:2505.19574*, 2025.
31. Sempowski A. C. et al. Microscopic constraints for the equation of state and structure of neutron stars: a Bayesian model mixing framework. *arXiv preprint arXiv:2505.18921*, 2025.
32. Nguyen T. L., Hocking T. Learning Penalty for Optimal Partitioning via Automatic Feature Extraction. *arXiv preprint arXiv:2505.07413*, 2025.
33. Prinster D. et al. WATCH: Adaptive Monitoring for AI Deployments via Weighted-Conformal Martingales. *arXiv preprint arXiv:2505.04608*, 2025.
34. Dandapanthula S., Ramdas A. Conformal changepoint localization. *arXiv preprint arXiv:2505.00292*, 2025.
35. Liu J. et al. Spatial-Sign based High dimensional Change Point Inference. *arXiv preprint arXiv:2504.19306*, 2025.
36. Moen P. A. J. A general methodology for fast online changepoint detection. *arXiv preprint arXiv:2504.09573*, 2025.

37. Moutonnet N., Scott G., Mandic D. P. Augmentation of EEG and ECG Time Series for Deep Learning Applications: Integrating Changepoint Detection into the iAAFT Surrogates. *arXiv preprint arXiv:2504.03761*, 2025.
38. Park J. et al. Bottom-up Iterative Anomalous Diffusion Detector (BI-ADD). *arXiv preprint arXiv:2503.11529*, 2025.
39. Cheng J. et al. Modeling Behavior Change for Multi-model At-Risk Students Early Prediction (extended version). *arXiv preprint arXiv:2503.05734*, 2025.
40. Grundy T., Killick R., Svetunkov I. Online detection of forecast model inadequacies using forecast errors. *Journal of Time Series Analysis*, 2025.
41. Saha A., Ramdas A. Post-detection inference for sequential changepoint localization. *arXiv preprint arXiv:2502.06096*, 2025.
42. Heaton T. J., Al-assam S., Bard E. A New Approach to Radiocarbon Summarisation: Rigorous Identification of Variations/Changepoints in the Occurrence Rate of Radiocarbon Samples using a Poisson Process. *arXiv preprint arXiv:2501.15980*, 2025.
43. Mameche S. et al. SPACETIME: Causal Discovery from Non-Stationary Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
44. Cui X. et al. ART: Distribution-Free and Model-Agnostic Changepoint Detection with Finite-Sample Guarantees. *arXiv preprint arXiv:2501.04475*, 2025.
45. Cook K. J. et al. Considering experimental frame rates and robust segmentation analysis of piecewise-linear microparticle trajectories. *arXiv preprint arXiv:2412.21025*, 2024.
46. Naik C., Chew E. Tipping points, pulse elasticity and tonal tension: An empirical study on what generates tipping points. *arXiv preprint arXiv:2412.10481*, 2024.
47. Duran-Martin G. et al. A unifying framework for generalised Bayesian online learning in non-stationary environments. *arXiv preprint arXiv:2411.10153*, 2024.

48. Qian C. et al. Changepoint Detection in Complex Models: Cross-Fitting Is Needed. *arXiv preprint arXiv:2411.07874*, 2024.
49. Wu W., Veitch D., Zhou Z. Asynchronous Jump Testing and Estimation in High Dimensions Under Complex Temporal Dynamics. *arXiv preprint arXiv:2410.23706*, 2024.
50. Li M. changepointGA: An R package for Fast Changepoint Detection via Genetic Algorithm. *arXiv preprint arXiv:2410.15571*, 2024.
51. Jang J., Walther G. Fast and optimal changepoint detection and localization using Bonferroni triplets. *arXiv preprint arXiv:2410.14866*, 2024.
52. Baum M., Passino F. S., Gandy A. Doubly unfolded adjacency spectral embedding of dynamic multiplex graphs. *arXiv preprint arXiv:2410.09810*, 2024.
53. Hocking T. D. Finite Sample Complexity Analysis of Binary Segmentation. *arXiv preprint arXiv:2410.08654*, 2024.
54. Fowler J., Hocking T. D. Efficient line search for optimizing Area Under the ROC Curve in gradient descent. *arXiv preprint arXiv:2410.08635*, 2024.
55. Hou Y., Tan V., Zhong Z. Almost minimax optimal best arm identification in piecewise stationary linear bandits. *Advances in Neural Information Processing Systems*, 2024. – T. 37. – C. 128967-129041.
56. Fiecas M. B., Cullen K. R., Killick R. Generalised mixed effects models for changepoint analysis of biomedical time series data. *arXiv preprint arXiv:2410.00183*, 2024.
57. Watkins J. et al. Entropy-based feature selection for capturing impacts in Earth system models with abrupt forcing. *Journal of Computational and Applied Mathematics*, 2024. – T. 471. – C. 116724.
58. Jia Y. et al. TUNE: Algorithm-Agnostic Inference after Changepoint Detection. *arXiv preprint arXiv:2409.15676*, 2024.
59. Shi-Jun S., Shand L., Li B. Tracing the impacts of Mount Pinatubo eruption on regional climate using spatially-varying changepoint detection. *The Annals of Applied Statistics*, 2025. – T. 19. – №. 1. – C. 465-484.

60. Biswas S., Banerjee B., Laha A. K. Angular Co-variance using intrinsic geometry of torus: Non-parametric change points detection in meteorological data. *arXiv preprint arXiv:2409.08838*, 2024.
61. D'Angelo N. Testing for a general changepoint in psychometric studies: changes detection and sample size planning. *arXiv preprint arXiv:2408.04056*, 2024.
62. Nguyen T. L., Hocking T. D. Enhancing Changepoint Detection: Penalty Learning through Deep Learning Techniques. *arXiv preprint arXiv:2408.00856*, 2024.
63. Giampino A. et al. Local level dynamic random partition models for changepoint detection. *arXiv preprint arXiv:2407.20085*, 2024.
64. Baumer B. S., Sierra B. M. S. tidy-changepoint: a unified framework for analyzing changepoint detection in univariate time series. *arXiv preprint arXiv:2407.14369*, 2024.
65. Penaloza E., Stevens N. Changepoint Detection in Highly-Attributed Dynamic Graphs. *arXiv preprint arXiv:2407.06998*, 2024.
66. Corneek J. et al. Online Bayesian changepoint detection for network Poisson processes with community structure. *Statistics and Computing*, 2025. – T. 35. – №. 3. – C. 1-29.
67. Whalen M. C. R., Matthews G. J., Mills B. M. Empirical Determination of Baseball Eras: Multivariate Changepoint Analysis in Major League Baseball. *arXiv preprint arXiv:2407.01797*, 2024.
68. Ricci F. Z. et al. Bayesian temporal biclustering with applications to multi-subject neuroscience studies. *arXiv preprint arXiv:2406.17131*, 2024.
69. Carrington R., Fearnhead P. Post-selection inference for quantifying uncertainty in changes in variance. *arXiv preprint arXiv:2405.15670*, 2024.
70. Khamesi V. K. et al. Online Changepoint Detection via Dynamic Mode Decomposition. *arXiv preprint arXiv:2405.15576*, 2024.
71. Chen T. et al. Euclidean mirrors and first-order changepoints in network time series. *arXiv preprint arXiv:2405.11111*, 2024.

72. Liu C. et al. Bayesian functional graphical models with change-point detection. *Computational Statistics & Data Analysis*, 2024. – T. 206. – C. 108-122.
73. Horvath L., Trapani L., Wang S. Sequential monitoring for explosive volatility regimes. *arXiv preprint arXiv:2404.17885*, 2024.
74. Sun S. et al. SNSeg: An R Package for Time Series Segmentation via Self-Normalization. *arXiv preprint arXiv:2404.07451*, 2024.
75. Yang Z., Eckley I. A., Fearnhead P. A communication-efficient, online changepoint detection method for monitoring distributed sensor networks. *Statistics and Computing*, 2024. – T. 34. – №. 3. – C. 115.
76. Köhne J., Mies F. At the edge of Donsker's Theorem: Asymptotics of multiscale scan statistics. *arXiv preprint arXiv:2506.05112*, 2025.
77. Du H., Gong S., Xu J. A Proof of The Changepoint Detection Threshold Conjecture in Preferential Attachment Models. *arXiv preprint arXiv:2502.00514*, 2025.
78. Bhattacharyya S., Chattopadhyay S., Basu S. Application of Random Matrix Theory in High-Dimensional Statistics. *arXiv preprint arXiv:2412.06848*, 2024.
79. Moen P. A. J. Minimax rates in variance and covariance changepoint testing. *arXiv preprint arXiv:2405.07757*, 2024.
80. Cakmak A. et al. Benchmarking changepoint detection algorithms on cardiac time series. *arXiv preprint arXiv:2404.12408*, 2024.
81. Brodsky, B. E., Darkhovsky, B. S. Nonparametric Methods in Change Point Problems. *Springer*, 1993.
82. Pepelyshev A., Polunchenko A. S. Real-time financial surveillance via quickest change-point detection methods. *arXiv preprint arXiv:1509.01570*, 2015.
83. Minimax methods for multihypothesis sequential testing and change-point detection problems. *Sequential Analysis*, 2008, 27, no. 2, pp. 141–173.

84. Brodsky, B. E., Darkhovsky, B. S. Nonparametric Statistical Diagnosis: Problems and Methods. Mathematics and Its Applications. Kluwer Academic, Dordrecht. MR1862475, 2000
85. Niu Y. S., Hao N., Zhang H. Multiple change-point detection: a selective overview. *Statistical Science*, 2016, pp. 611-623.
86. Madreimov I., Petunin Yu.I Characterization of a uniform distribution using order statistics. *Teor Ver Mat Statist*, 1982, 27, pp. 96–102.
87. Matveichuk S.A., Petunin Yu.I. Generalization of Bernoulli schemes that arise in order statistics, I, *Ukrainian. Math. J.*, 1990. vol. 42, no. 4, pp. 459–466.
88. Matveichuk S.A., Petunin Yu.I. Generalization of Bernoulli schemes that arise in order statistics. II, *Ukrainian Math. J.*, 1991, vol. 43, no. 6, pp. 728–734.
89. Hill B.M. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *J Am Stat Assoc*, 1968, 63, pp. 677–691
90. Waerden B.L. van der. Mathematische Statistik. *Springer-Verlag*, Berlin, 1957; English transl. of 2nd ed., *Springer-Verlag*, Berlin and New York, 1969.
91. 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, pp. 58–68.
92. Klyushin D.A., Petunin Y.I., Nonparametric population equivalence test based on measure of closeness between samples, *Ukrainian Mathematical Journal*, 2003, 2nd. ed., pp. 147-163.
93. 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>
94. Jolliffe I. Principal Component Analysis, *Springer Series in Statistics*, *Springer-Verlag*, New York, 2002. <https://doi.org/10.1007/b98835>

95. Orlov A. Non-parametric criteria for the agreement of Kolmogorov, Smirnov, omega-square and errors in their application. *Науковий журнал КьбГАУ*, 2014. — Вип. 3. — С. 1-29.
96. Conover, W. Practical nonparametric statistics. *John Wiley & Sons*, 1999, p. 350
97. Klyushin D., Lyashko S., Tymoshenko A. Optimal Petunin Ellipsoids in the Design Space. *26th International Conference on Automatic Control*, 2020, pp. 40-41
98. Fisher R. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, volume 7, 2nd. ed., 1936, pp. 179–188.