УДК 519.2:004.62
DOI https://doi.org/10.32782/2663-5941/2025.4.2/41

Чепурна О.Є.
Національний університет «Одеська юридична академія»

Черевко Є.В.
Національний університет «Одеська юридична академія»

Лобода Ю.Г.
Національний університет «Одеська юридична академія»

Кухаренко С.В.
Національний університет «Одеська юридична академія»

ЗАСТОСУВАННЯ МАТЕМАТИЧНОЇ СТАТИСТИКИ ДЛЯ АНАЛІЗУ ВЕЛИКИХ ДАНИХ В АЛГОРИТМАХ ОБРОБКИ ІНФОРМАЦІЇ

У статті досліджено застосування методів математичної статистики для аналізу великих даних в алгоритмах обробки інформації. Враховуючи проблеми, пов'язані з обсягом, варіативністю та неповнотою даних, доведено доцільність інтеграції статистичних підходів у сучасні системи обчислювального аналізу. Особлива увага приділяється завданням класифікації, виявлення аномалій та оцінки параметрів розподілу. Проаналізовано використання класичних та байєсівських методів, а також метрик типу «Відстань Магаланобіса» для побудови адаптивних рішень. Показано, що навіть у випадку багатьох ознак статистичні методи дозволяють вести інтерпретацію моделей за допомогою засобів SHAP, аналізу перестановок і часткових залежностей. Окремо обґрунтовано важливість поєднання статистичної формалізації з візуалізацією, що підвищує достовірність висновків та практичну значущість моделювання. Надано рекомендації щодо вибору статистичних процедур для конкретних класів завдань (контроль якості, аналіз ризиків, кіберзахист). Окремо відзначено роль використання сучасних інструментів статистичного прогнозування, які забезпечують можливість випереджального виявлення тенденцій у великих потоках інформації. Розглянуті приклади впровадження статистичних методів у практичних інформаційних системах, зокрема у фінансовому моніторингу, кібербезпеці, біоінформатиці, що засвідчує універсальність та ефективність математичної статистики для вирішення актуальних проблем аналізу даних. Запропоновані напрями подальших досліджень: інтеграція інформаційної геометрії як сучасного інструменту аналізу структури даних, автоматизація статистичних перевірок у складі систем AutoML, апробація методів у потоковій передачі даних, а також розширення пояснюваного штучного інтелекту за допомогою статистичних підходів. Окреслено переваги поєднання класичних статистичних методів з новітніми алгоритмічними рішеннями для побудови надійних, гнучких і масштабованих систем обробки та аналізу великих даних. Робота орієнтована на фахівців в області обробки даних, прикладної математики, інформаційних технологій, а також розробників інтелектуальних систем, які працюють із великими і різнорідними наборами даних.

Ключові слова: *математична статистика, великі дані, алгоритми обробки даних, управління кібербезпекою, аналіз даних, математичні моделі, інформаційна безпека, статистичне прогнозування, моделі кіберзагроз.*

Постановка проблеми. Щоденно генерується, збирається та обробляється колосальний обсяг даних – від банківських транзакцій до соціальних мереж, медичних записів та наукових досліджень. Обробка й аналіз цих даних потребує не тільки потужних обчислювальних ресурсів, а й ефективних математичних інструментів, серед яких ключову роль відіграє математична статистика.

Математична статистика є галуззю математики, що забезпечує інструменти для збору, опису, аналізу та інтерпретації даних, а її застосування

© Чепурна О.Є., Черевко Є.В., Лобода Ю.Г. та ін., 2025
Стаття поширюється на умовах ліцензії CC BY 4.0

в алгоритмах обробки інформації дозволяє виявляти закономірності в хаотичних наборах даних, здійснювати прогнозування, оцінювати ймовірності подій та мінімізувати похибки під час прийняття рішень. Особливої актуальності статистичні методи набувають у сфері великих даних (Big Data), які відзначаються значним обсягом, різнорідністю та швидкістю надходження, адже саме поєднання статистики з алгоритмічними рішеннями створює підґрунтя для розробки інтелектуальних систем, що здатні навчатися, адаптуватися й приймати обґрунтовані рішення на основі аналізу даних.

Метою цієї роботи є дослідження можливостей та ефективності використання математичної статистики у процесі обробки великих масивів інформації, а також аналіз алгоритмів, які інтегрують статистичні методи для оптимізації й автоматизації аналізу даних. Інформаційні системи дедалі частіше функціонують в умовах опрацювання великих обсягів складно структурованих даних, що надходять із гетерогенних джерел. Такі масиви часто характеризуються пропущеними значеннями, внутрішньою корельованістю ознак, нелінійною структурою залежностей та нестійкістю розподілів у часі. У результаті, стандартні аналітичні підходи демонструють зниження точності й обґрунтованості висновків [1, с. 123–129; 2, с. 87–95].

Математична статистика надає формалізовану основу для розв'язання задач класифікації, регресії, побудови довірчих інтервалів та перевірки статистичних гіпотез [3, с. 173–178; 4, с. 201–210]. Проте у випадках високої розмірності даних або значних обсягів потокової інформації, ефективність традиційних процедур суттєво знижується [5, с. 91–100].

Особливої уваги набувають підходи, що дозволяють поєднувати статистичні моделі з геометричними уявленнями про простір параметрів розподілу. Такі підходи застосовуються, зокрема, у побудові метрик простору імовірнісних моделей, визначенні статистичних відстаней, а також у вивченні кривизни простору параметрів [6, с. 132–144; 7, с. 77–89].

Інформаційна геометрія як міждисциплінарна концепція використовує апарат диференціальної геометрії для аналізу статистичних моделей як багатоатомних об'єктів у просторах з кривизною. Простір Фішера–Рао, як приклад такого підходу, дає змогу оцінювати відстані між розподілами, базуючись на їх ентропійних властивостях [8, с. 58–70; 9, с. 4–8].

Застосування цих підходів дозволяє не лише глибше інтерпретувати дані, але й підвищити стійкість моделей до спотворень, збоїв і викривлень, характерних для реальних інформаційних потоків [10, с. 44–52; 11, с. 11–22]. Це особливо важливо у критичних галузях, таких як технічна діагностика, безпека IT-систем, біоінформатика та фінансовий моніторинг [12, с. 6–12; 13, с. 1210–1218].

Поєднання класичних методів статистичного аналізу з інформаційно-геометричними підходами сприяє створенню адаптивних обчислювальних моделей, що демонструють вищу точність та інтерпретованість у складних середовищах [14, с. 5–9; 15, с. 14–16], а також відкриває нові перспективи для розвитку інтелектуальних систем обробки інформації у різних сферах діяльності.

Аналіз сучасних досліджень і публікацій. Зростання інтересу до проблем ефективного аналізу великих даних, що має відображення як у статистичній теорії, так і в прикладних дослідженнях. У працях Джона Райса (J.A. Rice) [1, с. 123–129] та Джорджа Каселла і Роджера Бергера (G. Casella і R.L. Berger) [2, с. 87–95] детально розглянуто основи класичної математичної статистики, включаючи гіпотезне тестування, побудову довірчих інтервалів і оцінювання параметрів, що лежать в основі більшості сучасних методів обробки даних.

З погляду на обсяг і складність обробки і аналізу даних особливу актуальність набувають підходи, орієнтовані на зниження розмірності, відбір релевантних ознак і стійке оцінювання. У роботі Крістофера Бішопа (C.M. Bishop) [3, с. 173–178] докладно аналізуються статистичні моделі, придатні для умов високої вимірності, зокрема методи головних компонент (PCA) та латентні змінні. Кевін Мерфі (K.P. Murphy) [4, с. 201–210] зосереджується на ймовірнісних графових моделях, що дозволяють моделювати складні залежності між змінними при мінімізації обчислювальної складності.

Особливої уваги заслуговують прикладні дослідження, пов'язані з використанням методів машинного навчання у статистичному аналізі. Зокрема, Джеймса Гарта та його співавтори (G. James et al) [5, с. 91–100] демонструють застосування регуляризованих моделей до задач класифікації у великих вибірках. Це узгоджується з висновками Мартіна Вейрайта та Майкла Джордана (M.J. Wainwright і M.I. Jordan) [6, с. 132–144] щодо застосування експоненціальних родин розподілів для аналізу взаємозв'язків у даних.

Проблематика аналітики у високорозмірних просторах тісно пов'язана з методами зниження розмірності, обчислення ентропій та інформаційних дивергенцій. Ряд публікацій розглядає інтеграцію таких методів у поточні фреймворки аналізу даних, наприклад, Ліам Панінскі (Paninski L.) [13, с. 1210–1218], зокрема у виявленні аномалій, нечіткій класифікації та роботі зі спотвореними вибірками.

На перетині математичної статистики та геометрії виникають нові концепти – інформаційна геометрія. Праці Шун'їчі Амарі (S. Amari) [7, с. 77–89], Томаса Ковера (T.M. Cover) і Джой Томаса (J.A. Thomas) [8, с. 58–70], а також Ніко Ай зі співавторами (N. Ay et al) [10, с. 44–52] заклали фундамент цього напряму, у якому імовірнісні моделі інтерпретуються як многовиди з ріманової метрикою, де статистичні властивості відображаються у кривизні простору. Це відкриває нові можливості для адаптивного моделювання, зокрема в умовах змінної статистичної структури даних – Євген Черевко, Олена Чепурна, Євгенія Кулешова (Cherevko Y., Chepurna O., Kuleshova Y.) [11, с. 11–22].

У практичному вимірі застосування інформаційної геометрії, наприклад, у задачах кіберзахисту, які висвітлені у Fortinet – Глобальний звіт про ландшафт загроз 2025 [12, с. 6–12] та оптимізації систем виявлення аномалій, дозволяє підвищити чутливість моделей до малопомітних відхилень у даних. Це важливо для галузей, де якість виявлення відхилень критично впливає на ефективність рішень – наприклад, у фінансовому секторі, медицині та розподілених інформаційних системах.

Таким чином, сучасний напрям досліджень зосереджено навколо інтеграції статистичних, геометричних і алгоритмічних методів, що забезпечують інтерпретованість, адаптивність і стійкість аналітичних моделей до складних властивостей даних.

Постановка завдання. Однією з основних проблем роботи з масивами даних є обмежена придатність класичних статистичних процедур для аналізу великих, неоднорідних та неструктурованих даних, особливо у випадках високої розмірності та змінних розподілів. Тому актуальним є дослідження можливостей використання геометризованих статистичних моделей для підвищення стійкості й інтерпретованості аналізу даних. У цьому дослідженні передбачається систематизувати статистичні методи, що ефективні для Big Data (SelectKBest, PCA, Mahalanobis distance, LOF, GMM), оцінити їхню придатність до задач класифікації, виявлення аномалій та зменшення розмірності, проаналізувати сучасні алгоритми, які використовують геометричні структури (ріманові многовиди, дивергенції Фішера–Рао, інформаційну метрику), запропонувати приклад застосування на сінтезованих наборах даних та розробити критерії ефективності моделей (AUC, Precision, Recall, інформаційна дивергенція), а також запропонувати гібридний підхід, у якому статистичні й інформаційно-геометричні інструменти комбінуються.

В цьому контексті потрібно виконати наступні завдання: проаналізувати основні поняття та методи математичної статистики; визначити особливості великих даних (Big Data) та проблеми їх обробки; дослідити алгоритми, що інтегрують статистичні методи для аналізу даних; оцінити переваги та обмеження таких алгоритмів у реальних прикладних сценаріях; запропонувати напрямки вдосконалення методів статистичного аналізу даних.

Таким чином, мета дослідження полягає у поєднанні класичних статистичних методів із новітніми геометричними підходами задля підвищення ефективності аналізу даних у розподілених, змінних та потенційно зашумлених середовищах.

Виклад основного матеріалу. Застосування математичної статистики в аналізі великих даних (Big Data) є фундаментальним для алгоритмів обробки інформації, оскільки воно дозволяє виявляти патерни, робити прогнози, тестувати гіпотези та зменшувати розмірність даних у масштабах, де традиційні методи неефективні. Статистика надає інструменти для оцінки невизначеності, моделювання та оптимізації, особливо в розподілених системах на кшталт Hadoop чи Spark. На основі аналізу сучасних джерел, ключовими алгоритмами, які варто описати, є ті, що поєднують статистичні принципи з обчислювальними техніками для великих обсягів даних. Опишемо найпоширеніших і впливових методи математичної статистики, таких як регресійний аналіз, кластеризація, класифікація, зменшення розмірності та інші. Вони часто реалізуються з використанням стохастичних методів (наприклад, SGD – Stochastic Gradient Descent) для масштабованості. Серед ключових алгоритмів математичної статистики, які застосовуються в аналізі великих даних, варто виділити такі підходи. Лінійна регресія використовується для моделювання залежності між змінними за умови лінійного зв'язку, базуючись на

методі найменших квадратів для мінімізації помилок. Цей підхід дозволяє прогнозувати тренди у великих масивах даних, наприклад у сфері продажів чи кліматичних спостережень; у Big Data часто використовується стохастичний градієнтний спуск для обробки потокових даних, зокрема в бібліотеці Spark MLlib.

Сучасні інформаційні системи дедалі частіше поєднують класичні алгоритмічні процедури з методами математичної статистики для забезпечення ефективного аналізу даних у складних цифрових середовищах. Особливо це стосується алгоритмів сортування, пошуку, обходу графів і дерев, які виступають інфраструктурною основою обробки інформації у високопродуктивних системах (ERP, SCM, IoT-аналітика тощо).

Однією з типових задач є попередня обробка даних, де алгоритми сортування (наприклад, QuickSort, MergeSort) використовуються для підготовки вибірок до статистичного аналізу, зокрема для обчислення медіани, квантилів або емпіричних розподілів. У таких випадках статистичні оцінки (середнє, дисперсія, коефіцієнти асиметрії) вимагають впорядкованих структур, що зменшують складність обчислень з O(n log n) до O(n) у деяких оптимізованих випадках [14, с. 127–132].

У задачах класифікації та прогнозування, що базуються на статистичних моделях (наприклад, логістична регресія, дерева рішень), широко застосовуються деревоподібні структури даних. Зокрема, дерева CART (Classification and Regression Trees) дозволяють реалізовувати рекурсивне ділення простору ознак, а статистичні критерії, як Gini-індекс або ентропія Шеннона, виступають евристиками для побудови оптимальних вузлів [15, с. 264–270].

Алгоритми на основі графів (наприклад, Dijkstra або Bellman–Ford) дедалі частіше інтегруються зі статистичними підходами для зважування ребер. Наприклад, в аналізі транспортних мереж або логістичних ланцюгів, статистичні метрики, як середній час доставки, варіація затримок або довірчі інтервали, використовуються для формування ваг у графі, що підвищує адаптивність маршрутів до реальних умов [16, с. 318–324].

Крім того, кластеризація у системах рекомендацій (наприклад, у маркетингу або енергетиці), структури дерев пошуку (KD-дерева, Ball-дерева) використовуються для прискорення статистичних обчислень, зокрема для знаходження найближчих сусідів (k-NN) чи виявлення аномалій на основі відстаней (Mahalanobis, LOF) [17, с. 291–295].

Таким чином, поєднання структур алгоритмічної ефективності з статистичною обґрунтованістю створює підґрунтя для формування масштабованих і достовірних систем обробки даних. Це особливо важливо у високонавантажених інформаційних середовищах, де обсяг даних, їхня складність і потреба в інтерпретованості зростають одночасно.

Окремої уваги та дослідження заслуговує інформаційна геометрія як інструмент аналізу великих даних. Інформаційна геометрія – це напрям міждисциплінарного аналізу, що поєднує диференціальну геометрію з математичною статистикою для дослідження властивостей сімейств розподілів ймовірностей. Вона базується на трактуванні множини параметризованих розподілів як многовиду, на якому визначено метричну структуру, зокрема метрику Фішера-Рао. Такий підхід забезпечує геометричну інтерпретацію статистичних відстаней, ентропій, дивергенцій та кривизни ймовірнісного простору [18, с. 77–89].

У задачах обробки великих даних, де класичні евклідові уявлення втрачають свою інтерпретативну потужність через високу розмірність, корельованість або розрідженість даних, інформаційна геометрія дає змогу використовувати геометричні характеристики – такі як геодезичні, ріманову кривизну, відстані між розподілами (наприклад, дивергенції Кульбака-Лейблера або α-дивергенції) – для оцінки подібності, ієрархічної класифікації та аналізу аномалій [18, с. 44–52].

У практичному вимірі така методологія застосовується до високорозмірних потоків даних, типових для біоінформатики, фінансової аналітики, IoT-систем та кіберфізичних мереж. Наприклад, аналіз відстаней між розподілами ймовірностей спостережень у багатовимірному просторі (наприклад, за допомогою геодезичної відстані в моделі експоненціальної родини) дозволяє точно відокремлювати кластери спостережень із різною статистичною структурою, навіть якщо класичні методи кластеризації (наприклад, k-means) демонструють низьку роздільну здатність [19, с. 1210–1218].

У роботах Шун'їчі Amari наведено підстави для використання інформаційної геометрії у побудові статистично ефективних адаптивних моделей у середовищах із нестабільною розподільчою структурою. Наприклад, простір експоненціальних розподілів задається через параметри, для яких метрика Фішера-Рао має інваріантність відносно перетворень координат. Це дозволяє будувати процедури узагальненого логарифмічного градієнтного спуску на ріманових многовидах.

Окремої уваги заслуговує використання інформаційно-геометричних методів у задачах виявлення аномалій, наприклад, у кіберзахисті. Замість простого порівняння поточних спостережень із середнім значенням застосовується обчислення геодезичної відстані на просторі ймовірнісних моделей між поточним вектором розподілу характеристик і моделлю «норми».

У середовищах на кшталт Apache Spark чи Dask можливе паралельне обчислення дивергенцій Кульбака-Лейблера, ентропій Шеннона або відстаней Фішера між розподілами з використанням моделей Gaussian Mixture Model або Variational Autoencoders [19, с. 5–9].

Візуалізація застосування інформаційно-геометричної інтерпретації моделей у просторі розподілів може бути представлена у вигляді багатовимірної 3D-моделі, де кожен розподіл зображено як окрему точку на ріманановому многовиді, а геодезичні демонструють, яким чином модель адаптується до змін у потоках даних у динамічному середовищі. Такий підхід дає змогу не лише аналізувати розміщення розподілів, а й відстежувати їхні зміни у часі, що критично важливо для систем із високою частотою оновлення інформації, наприклад, у сферах фінансової аналітики, медичного моніторингу або розумних мереж енергоспоживання.

До основних переваг інформаційно-геометричного підходу належить здатність ураховувати як точкові, так і структурні властивості розподілів, а також гнучка адаптація до динамічних змін статистичних характеристик. Це дає змогу більш формалізовано і точно оцінювати «інформаційну близькість» між різними станами системи, що особливо корисно для побудови ефективних класифікаторів, автоматичних фільтрів і сучасних енкодерів даних.

Водночас варто враховувати певні обмеження: складність реалізації математичного апарату для великих, нестабільних і розріджених вибірок, потребу в знаннях диференціальної геометрії для коректного визначення метрик, кривизни та візуалізації таких просторів, як це детально розглянуто в джерелі [20, с. 193–198].

Наприклад, для аналізу багатовимірних нормальних розподілів у межах інформаційної геометрії доцільно використовувати розширені методи тривимірної візуалізації, де кожна точка може відображати як положення, так і форму розподілу, а траєкторії між ними – шляхи мінімальної інформаційної «відстані». Подібний підхід дозволяє отримати глибше розуміння структури даних і швидко ідентифікувати аномалії чи нові кластери у складних потоках інформації. Завдяки цьому дослідники можуть аналізувати не лише статичні характеристики розподілів, а й відстежувати їхню трансформацію в умовах постійної зміни параметрів, що властиво динамічним системам великого обсягу даних.

У практичному сенсі це означає, що процеси виявлення аномалій або класифікації можуть бути автоматизовані на основі геометричних характеристик, включно з кривизною, геодезичними та інформаційною відстанню, що дозволяє максимально ефективно реагувати на нетипові ситуації у потоках даних. Водночас побудова подібних моделей відкриває нові можливості для створення адаптивних прогнозних систем, де алгоритми не лише аналізують зміни центрів кластерів, а й враховують трансформацію їхніх геометричних властивостей, зокрема розтягнення, стиснення та ротацію у статистичному просторі.

Таким чином, інформаційно-геометричний підхід стає потужним інструментом як для глибокого аналізу внутрішньої структури даних, так і для оперативного виявлення нових тенденцій і закономірностей, що є вкрай важливим у сучасних інформаційних системах з високою мінливістю та складністю потоків. Це, у свою чергу, відкриває широкі можливості для подальшого аналізу динаміки розподілів у часі, а також дає змогу розробляти прогностичні моделі, що враховують не лише зміну центрів кластерів, а й трансформацію їхніх геометричних властивостей (Рис. 1).

Перспективним напрямком використання статистичних методів є інтегроване застосування

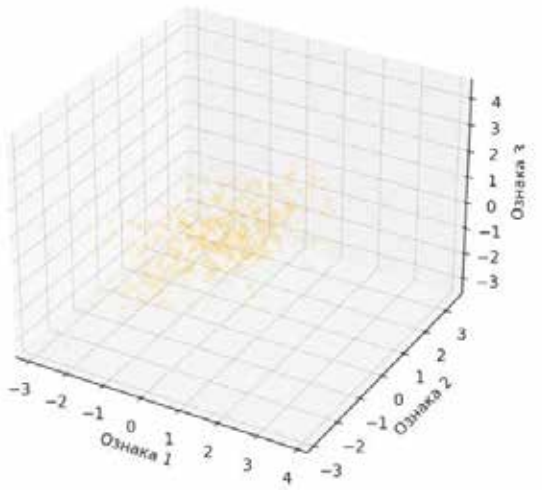

Рис. 1. Візуалізація інформаційно-геометричної інтерпретації структури ймовірнісних розподілів

PCA, GMM та інформаційної геометрії для аналізу даних енергоспоживання. Наприклад, синтетична вибірка, що імітує тиждень споживання електроенергії із 96 точками на добу (15-хвилинний інтервал), обробляється за допомогою PCA, зменшуючи розмірність з 672 до 2 головних компонент (≈87% дисперсії), а GMM дозволяє виділити кластери з власними коваріаційними ядрами. Інформаційна геометрія реалізована через підрахунок міжкластерних відстаней за метрикою Фішера-Рао у просторі головних компонент, що дає змогу аналізувати подібність між кластерами. На рисунку 2 наведено кластери й геометричні зв'язки між ними, що інтерпретуються як геодезичні наближення у метричному просторі Фішера-Рао, дозволяючи оцінювати не лише класи, а й їхню близькість у споживанні. Такий підхід забезпечує не лише глибше розуміння структури даних, але й підвищує точність виявлення аномальних або неефективних патернів у динамічному середовищі енергоспоживання.

Серед ключових переваг такого підходу – інтерпретованість, адже PCA дозволяє виявити часові закономірності, наприклад, денні чи нічні піки та типові шаблони навантаження. Гнучкість кластеризації забезпечується за рахунок того, що GMM дає змогу враховувати неоднорідність дисперсій та кореляцій усередині даних. Використання інформаційної геометрії дозволяє не лише аналізувати центри кластерів, а й порівнювати їхню «форму», що є критично важливим для моніторингу змін структури споживання, зокрема в умовах війни або енергетичних збоїв. На практиці такий інтегрований підхід може бути застосований для виявлення неефективних чи аномальних патернів споживання, оцінки ризиків перевантажень, динамічного перепланування енергопостачання, а також для впровадження адаптивного ціноутворення (Demand Response) на основі кластерної приналежності.

Висновки. Методи математичної статистики, особливо такі сучасні напрями як інформаційна геометрія, стають фундаментом нової парадигми аналізу великих даних. Вони дозволяють дослідникам не лише глибоко зануритися у структуру інформаційних потоків, а й забезпечують ефективніші механізми для своєчасного виявлення аномалій та нестандартних патернів. На відміну від традиційних статистичних підходів, інформаційна геометрія розглядає ймовірнісні розподіли як точки на ріманновому многовиді, що надає змогу оцінювати інформаційну близькість різних станів системи з підвищеною точністю. Це, у свою чергу, значно розширює можливості кластеризації, адаптивного моніторингу та аналізу складних динамічних процесів у реальному часі.

Інтеграція інформаційно-геометричних методів із PCA та Gaussian Mixture Model виявляється особливо цінною: вона не лише зменшує розмірність досліджуваних даних і виявляє часові закономірності, а й дає змогу виділяти різноманітні кластери, враховуючи як їхню інформаційну форму, так і динаміку змін. Це має вагоме значення для прикладних сфер, таких як енергетика,

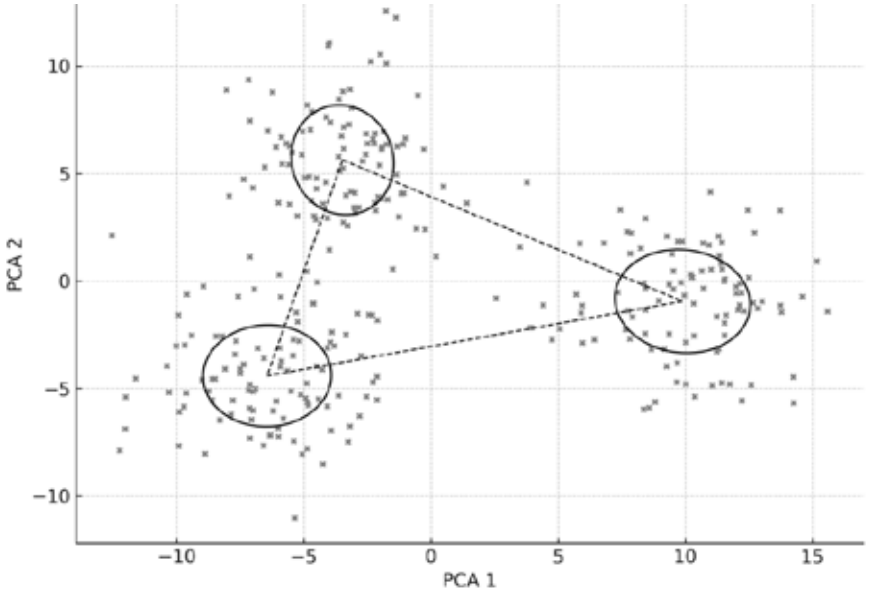

Рис. 2. Кластери енергоспоживання (GMM+PCA) з геометричними зв'язками (Fisher-Rao)
**Умовні позначення: сірі точки – проєкції об'єктів; контури – області кластерів; лінії – геометричні відстані*

фінанси та медицина, де моніторинг поведінки систем потребує високої чутливості до змін параметрів і структури. Водночас залучення інформаційної геометрії вимагає значної підготовки спеціалістів, знань у галузі диференціальної геометрії, математичної статистики та володіння сучасним програмним забезпеченням.

Не менш важливою є розробка й удосконалення відкритих бібліотек, що пришвидшить впровадження цих методів у практику. В цьому сенсі інформаційна геометрія – міждисциплінарна галузь, яка поєднує елементи диференціальної геометрії, теорії інформації та статистичного аналізу – пропонує інноваційний інструментарій для моделювання інформаційних процесів, оцінки ризиків, аналізу кіберзагроз і побудови стійких систем захисту даних та інфраструктур. Перспективи подальшого розвитку цього напрямку полягають у його здатності органічно інтегруватися з іншими статистичними методами, розширюючи спектр прикладних задач та забезпечуючи гнучку, адаптивну аналітику для обробки великих і складних масивів інформації. Саме тому інформаційна геометрія стає дедалі актуальнішою як для сучасної науки, так і для практичних застосувань у різноманітних сферах суспільного життя, де гнучкість та точність аналізу є запорукою ефективних рішень.

Список літератури:

1. Rice J. A. Mathematical Statistics and Data Analysis. 3rd ed. Cengage Learning, 2006. 688 p. URL: https://surl.li/uwsyja

2. Casella G., Berger R. L. Statistical Inference. 2nd ed. Duxbury Press, 2002. 660 p. URL: https://surl.li/liszcs

3. Bishop C.M. Pattern Recognition and Machine Learning. Springer, 2006. 738 p. DOI: https://doi.org/10.1007/978-0-387-45528-0

4. Murphy K. P. Machine learning: a probabilistic perspective. MIT Press, 2012. 1104 p.

5. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Springer, 2013. 426 p. DOI: https://doi.org/10.1007/978-1-4614-7138-7

6. Wainwright M.J., Jordan M.I. Graphical Models, Exponential Families, and Variational Inference // Foundations and Trends® in Machine Learning. 2008. Vol. 1(1–2). P. 1–305 DOI: https://doi.org/10.1561/2200000001

7. Amari S. Information Geometry and Its Applications. Springer, 2016. 360 p. DOI: https://doi.org/10.1007/978-4-431-55978-8

8. Cover T.M., Thomas J.A. Elements of Information Theory. 2nd ed. Wiley-Interscience, 2006. 776 p. DOI: https://doi.org/10.1002/047174882X

9. Nielsen F. An Elementary Introduction to Information Geometry // Entropy. 2020. Vol. 22(10). P. 1100 DOI: https://doi.org/10.3390/e22101100

10. Ay N., Jost J., Lê H.V., Schwachhöfer L. Information Geometry. Springer, 2017. 308 p. DOI: https://doi.org/10.1007/978-3-319-56478-4

11. Cherevko Y., Chepurna O., Kuleshova Y. On Information Geometry Methods for Data Analysis // Geometry, Integrability and Quantization. 2024. Vol. 29. P. 11–22 DOI: https://doi.org/10.7546/giq-29-2024-11-22

12. Fortinet. 2025 Global Threat Landscape Report. URL: https://www.fortinet.com/resources/reports/threat-landscape-report

13. Paninski L. Estimation of entropy and mutual information // Neural Computation. 2003. Vol. 15(6). P. 1191–1253 DOI: https://doi.org/10.1162/089976603321780272

14. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. Introduction to Algorithms. 3rd ed. MIT Press, 2009. 1312 p.

15. Fortinet. 2024 Global Threat Landscape Report. – URL: https://www.fortinet.com/reports

16. Paninski L. Estimation of entropy and mutual information // Neural Computation. 2003. Vol. 15(6). – P. 1191–1253 [с. 1210–1218] DOI: https://doi.org/10.1162/089976603321780272

17. Ay N., Jost J., Lê H.V., Schwachhöfer L. Information Geometry. Springer, 2017. – 308 p. DOI: https://doi.org/10.1007/978-3-319-56478-4

18. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2009. – 745 p. [с. 264–270]. DOI: https://doi.org/10.1007/978-0-387-84858-7

19. Ahuja R.K., Magnanti T.L., Orlin J.B. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993. – 864 p. [с. 318–324].

20. Beyer K., Goldstein J., Ramakrishnan R., Shaft U. When is “nearest neighbor” meaningful? // ICDT. 1999. – P. 217–235.

Chepurna O.Ye., Cherevko Ye.V., Loboda Yu.G., Kukharenko S.V. APPLICATION OF MATHEMATICAL STATISTICS FOR BIG DATA ANALYSIS IN INFORMATION PROCESSING ALGORITHMS

The article investigates the application of methods of mathematical statistics for the analysis of big data in information processing algorithms. Considering the problems associated with volume, variability and incompleteness of data, the expediency of integrating statistical approaches into modern systems of computational analysis has been proved. Particular attention is paid to the tasks of classification, detection of anomalies and evaluation of distribution parameters. The use of classical and Bayesian methods, as well as metrics such as "Magalanobis distance" for building adaptive solutions is analyzed. It is shown that even in the case of many features, statistical methods allow interpreting models using SHAP tools, permutation analysis and partial dependencies. The importance of combining statistical formalization with visualization, which increases the reliability of conclusions and the practical significance of modeling, is substantiated separately. Recommendations are provided for the selection of statistical procedures for specific classes of tasks (quality control, risk analysis, cyber defense). The role of using modern tools of statistical forecasting, which provide the possibility of anticipatory detection of trends in large flows of information, is separately noted. Examples of the implementation of statistical methods in practical information systems, in financial monitoring, cybersecurity, bioinformatics, which testifies to the versatility and efficiency of mathematical statistics for solving topical problems of data analysis, are considered. The proposed areas of further research are integration of information geometry as a modern tool for analyzing data structure, automation of statistical checks as part of AutoML systems, approbation of methods in data streaming, as well as expansion of explainable artificial intelligence using statistical approaches. The advantages of combining classical statistical methods with the latest algorithmic solutions for building reliable, flexible and scalable systems for processing and analyzing big data are outlined. The work is aimed at specialists in the field of data processing, applied mathematics, information technology, as well as developers of intelligent systems who work with large and heterogeneous data sets.

Key words: *mathematical statistics, big data, data processing algorithms, cybersecurity management, data analysis, mathematical models, information security, statistical forecasting, cyber threat models.*

Дата надходження статті: 27.07.2025
Дата прийняття статті: 06.08.2025
Опубліковано: 27.10.2025