МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ В. Н. КАРАЗІНА
ФАКУЛЬТЕТ МАТЕМАТИКИ І ІНФОРМАТИКИ

Кафедра фундаментальної математики

Кваліфікаційна робота

магістр
освітньо-кваліфікаційний рівень

на тему:

«Тополого-алгебраїчний аналіз даних»

Виконав:
студент групи М-162, 2 курсу
другого (магістерського) рівня,
Барабаш Д.В.
спеціальність:
111 Математика
освітня програма: Математика
Керівник: док. фіз.-мат. наук, проф.
Ямпольський О.Л.
Рецензент:

Харків - 2024 рік

Вступ. **2**
Теоретичні основи. **3**
Від хмар точок до діаграм персистентності 3
Порівняння та обчислення діаграм персистентності 8
Інші підходи та останні розробки 11
Застосування топологічного аналізу даних. **12**
Стійка гомологія хмар точок і зображень 12
Пошук сенсу за допомогою абстрактних мір відстані 15
Алгебраїчна геометрія з точки зору топологічного аналізу даних 16
Методи топологічного аналізу даних. **22**
Базовий алгоритм 22
Персистентна гомологія. 23
Майбутні напрямки **24**
Нові методи топологічного аналізу даних 24
Квантовий топологічний аналіз даних 25
Додаток **26**
Бібліографічний список використаної літератури **32**

Вступ.

Топологічний аналіз даних(topological data analysis - TDA) - це новітня галузь, яка швидко розвивається і надає набір нових топологічних і геометричних інструментів для виявлення релевантних ознак у складних структурах даних.

TDA - це напрям сучасної математики, що виник з різних робіт з прикладної (алгебраїчної) топології та обчислювальної геометрії протягом першого десятиліття століття. Хоча геометричні підходи до аналізу даних можна простежити досить далеко в минулому, TDA дійсно почалася як галузь з піонерських робіт Едельсбруннера та ін. (2002) і Зомородяна та Карлссона (2005) у напрямку стійкої гомології, а популяризація методу відбулася у 2009 році у знаковій статті Карлссона (2009). TDA в основному мотивується ідеєю, що топологія та геометрія надають потужний підхід для виведення надійної якісної, а іноді і кількісної, інформації про структуру даних,
TDA має на меті надати добре обґрунтовані математичні, статистичні та алгоритмічні методи для виведення, аналізу та використання складних топологічних та геометричних структур, що лежать в основі даних, які часто представлені у вигляді хмари точок в евклідовому або більш загальних метричних просторах. Протягом останніх кількох років було докладено значних зусиль для створення надійних та ефективних структур даних і алгоритмів для TDA, які зараз реалізовані, доступні і прості у використанні за допомогою стандартних бібліотек, таких як бібліотека Gudhi (C++ і Python) Maria та ін. і її програмний інтерфейс R Fasy, Dionysus, PHAT, DIPHA або Giotto. Незважаючи на те, що TDA активно розвивається, зараз надається набір зрілих та ефективних інструментів, які

можна використовувати в поєднанні з іншими інструментами для аналізу даних або як доповнення до інших інструментів наук про дані.

Теоретичні основи.

Від хмар точок до діаграм персистентності

В якості повчального прикладу розглянемо двовимірні хмари точок, показані на рисунку 1. Кожна точка може відповідати окремому виміру деякого об'єкта, наприклад, розташуванню фотонів, що потрапляють у камеру, або положенням частинок у системі. Неозброєним оком ми можемо чітко бачити, що кожна хмара має різну форму: Точки в хмарах "Circle" і "Figure 8" розподілені навколо однієї і двох петель відповідно. З іншого боку, "Swiss Roll" відповідає зашумленій одновимірній хмарі точок, вбудованій у вищий (двовимірний) простір.

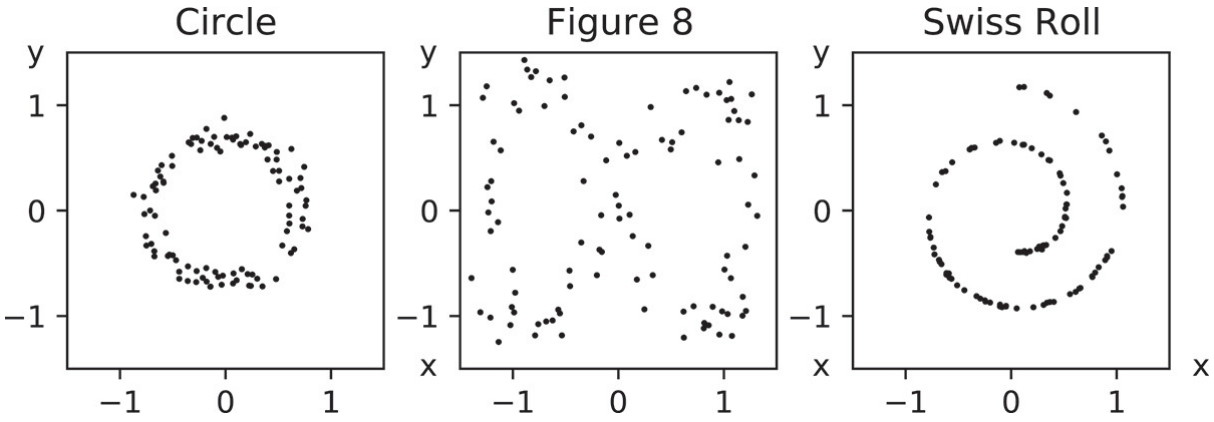

Рис. 1. Приклади зашумлених хмар точок.

Хмари точок, сформовані з об'єктів різної форми і навіть різної розмірності, може бути важко розрізнити за допомогою стандартних зведених статистик, таких як центр мас і дисперсія. У "Circle" та "Figure 8" шум випадковим чином збурює точки у навколишньому двовимірному

просторі. У "Swiss Roll" точки відбираються з одновимірного інтервалу перед тим, як вбудовуються у двовимірний простір (x, y).

Ми хотіли б формалізувати ці якісні спостереження у більш систематичний спосіб, щоб не залежати від безпосереднього побудови графіків даних, що є підходом, обмеженим дво- або тривимірними наборами даних. Як ми можемо кількісно оцінити очевидно різні форми цих хмар точок? Стандартні зведені статистичні показники, такі як центр маси або дисперсія, явно не підходять, оскільки вони не є інваріантними при перетворенні даних зі збереженням форми або при зміні масштабу.

На щастя, теорія графів надає строгі способи кількісної оцінки інтуїтивно зрозумілих форм дискретних наборів даних, включаючи хмари точок. Ідея полягає в тому, щоб побудувати граф, з'єднавши пари точок (вершин), які досить близько розташовані один до одного ребрами, а потім кількісно оцінити його форму, обчисливши топологічні інваріанти графа, його числа Бетті B_k. Число Бетті - це кількість k-вимірних отворів, наприклад, кількість незалежних зв'язних компонент (кластерів B_0) або несумісних циклів (циклів B_1). На практиці оцінка інваріантів графа зводиться до обчислення рангів та нульових просторів лінійних операторів (матриць), що діють на вершинах та ребрах графа. Коротко кажучи, обчислення форми даних хмари точок можна звести до простої лінійної алгебри. Більш вимірні топологічні особливості можуть бути аналогічно отримані шляхом побудови узагальнень графів, відомих як спрощені комплекси, які охоплюють більш вимірні об'єкти (грані, об'єми і т.д.) шляхом тріангуляції. k-симплекс - це комбінація (k+1) вершин; ребра є 1-симплексами, трикутні грані - 2-симплексами, тетраедричні об'єми - 3-симплексами, і так далі. K-симпліціальний комплекс - це сукупність симплексів розмірністю не більше k.

Збільшення k ускладнює ситуацію. По-перше, оскільки k-спрощення є комбінаторними об'єктами, кількість можливих спрощень швидко зростає з k, обмежуючи практичні обчислення низьковимірними топологічними особливостями. По-друге, не існує єдиного способу побудови простого комплексу, маючи лише попарні відстані між точками та масштаб відсікання; різні методи можуть відрізнятися за обчислювальними витратами, властивостями стійкості та здатністю точно відтворювати форми базового простору, з якого вибираються точки.

Є один великий слон у кімнаті, якого ми повинні розглянути: що ми маємо на увазі під "достатньо близько", коли з'єднуємо вершини, щоб утворити граф або найпростіший комплекс? Як визначити, які пари вершин з'єднати ребром, а які залишити не з'єднаними? Кількість циклів і кластерів буде чутливою до вибору відстані відсікання і навіть, можливо, додавання або видалення одного ребра, як показано на рисунку 2. Це здається основною проблемою, яка робить підхід недостатньо стійким до шуму та інших збурень.

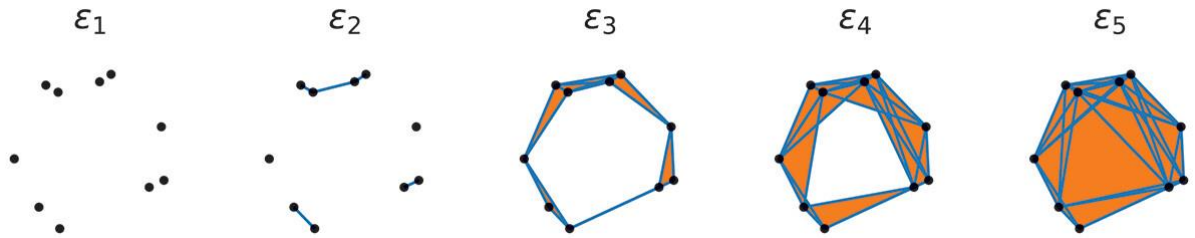

Рис. 2. Прості комплекси, побудовані з хмари точок з використанням різних відстаней відсікання ε_i , де сині лінії та помаранчеві заштриховані області позначають ребра та грані відповідно.

Для малих відстаней відсікання всі точки від'єднуються, утворюючи тривіальний найпростіший комплекс без ребер (ε_1). При збільшенні відстані відсікання сусідні вершини починають з'єднуватися ребрами (ε_2).

При подальшому збільшенні відсікання триплети точок з'єднуються, утворюючи грані. В ε_3 і ε_4 найпростіший комплекс має єдиний зв'язний компонент, що містить нетривіальний цикл. Для достатньо великих відстаней відсікання цикл руйнується додаванням граней, що покривають всю внутрішню частину хмари точок (ε_5).

Адекватним рішенням проблеми залежності інваріантів графіка від масштабу, отриманих з хмари точок, є обчислення форми графіка для всього діапазону масштабів, відомого як фільтрація, тобто вивчення його топології як функції масштабу довжини відсікання. Топологічні особливості (наприклад, кластери та цикли), що зберігаються в широкому діапазоні масштабів, є більш стійкими і повинні забезпечувати змістовну характеристику загальної форми даних. З іншого боку, ознаки, чутливі до невеликих змін масштабу або додавання чи видалення кількох ребер, можна віднести до шуму і відкинути, якщо це необхідно. Вивчаючи стійкість топологічних особливостей, ми зможемо відрізнити стійкі особливості від шуму.

Діаграми персистентності є одним зі стабільних способів представлення залежних від масштабу топологічних особливостей набору даних. На рисунку 3 показано діаграми персистентності, розраховані для кожної з хмар точок на рисунку 1. Найбільш стійкі топологічні особливості не тільки дозволяють зробити висновок про загальну форму даних, але й дають інформацію про геометрію хмари точок. Наприклад, масштаби народження довгоживучих циклів у хмарах "Circle" та "Figure 8" пов'язані з максимальною відстанню між сусідніми точками, що складають цикл, тоді як масштаб смерті буде пов'язаний з діаметром циклу.

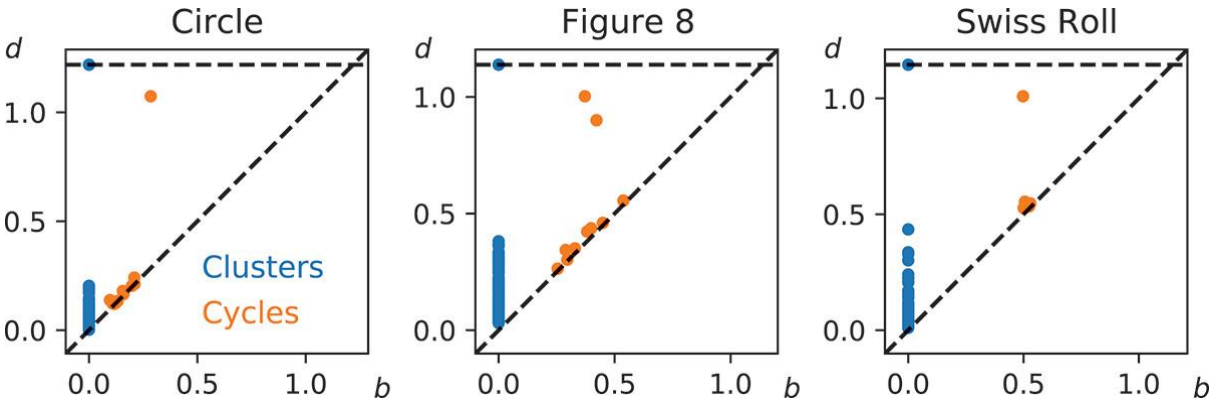

Рис. 3. Діаграми персистентності двовимірних хмар точок, показаних на рисунку 1, розраховані за допомогою комплексу Вітоpіса-Ріпса.

Кожна точка представляє окрему топологічну особливість. Горизонтальні та вертикальні осі позначають масштаби довжини, на яких кожна особливість створюється (b; народження) та знищується (d; смерть) відповідно. Точки, які знаходяться далі від діагональної пунктирної лінії, зберігаються в більшому діапазоні масштабів і, як кажуть, мають довший "час життя" $l = d - b$. Оскільки об'єкти повинні бути створені до того, як вони будуть знищені, жодна точка не лежить нижче діагоналі. На досить великих просторових масштабах всі точки з'єднуються, утворюючи єдиний зв'язний граф, що відповідає одному кластеру з нескінченним часом життя. Зазвичай кластер з нескінченним часом життя або відкидається, або зображується на графіку зі скінченним d і виділяється за допомогою горизонтальної пунктирної лінії.

Діаграми стійкості для хмар "Circle" та "Swiss Roll" мають однакові довготривалі характеристики, незважаючи на їхню очевидну різну форму. Однак при уважному розгляді можна виявити помітні відмінності в їхніх короткочасних характеристиках. Наприклад, цикли, що з'являються в наборі даних "Swiss Roll", мають подібні масштаби народження, що відповідають відстані між внутрішньою і зовнішньою частинами спіралі і вказують на одновимірне вбудовування. Це свідчить про те, що різні

форми цих двоточкових хмар дійсно можна вловити, досліджуючи їхні короткочасні особливості; таким чином, стійка гомологія може також відображати локальні особливості (геометрію) даних.

Порівняння та обчислення діаграм персистентності

Хоча діаграми персистентності надають компактне візуальне узагальнення залежних від масштабу топологічних особливостей одного набору даних, не відразу зрозуміло, як порівнювати діаграми персистентності, обчислені для різних наборів даних; вони, як правило, відрізняються за кількістю ознак і рівнем шуму, що ускладнює встановлення спільного порогу між справжніми ознаками і ознаками, викликаними шумом.

Ці проблеми мотивували розробку стабільних мір відстані та подібності для діаграм персистентності. Тут стабільність означає, що невелика зміна в одному наборі даних призводить, щонайбільше, до такої ж невеликої зміни у схожості з іншими фіксованими діаграмами персистентності. Одним із прикладів стабільної міри відстані є відстань Вассерштейна, яка є найменшою відстанню, на яку потрібно перемістити точки в парі діаграм персистентності, щоб перетворити одну діаграму в іншу. Непарні ознаки (тобто, якщо одна діаграма має більше ознак) переміщуються до діагоналі. Наприклад, на Рисунку 4 показано відповідність між одновимірними циклами кола, Figure 8 і хмарами точок Swiss Roll. Оскільки всі ознаки вносять свій внесок у відстань Вассерштейна, навіть ті, що викликані шумом і розташовані близько до діагоналі, вона може бути менш чутливою до змін у найбільш стійких ознаках. Іншим популярним вибором міри відстані є відстань вузького місця, яка є найбільшою деформацією пари ознак, необхідною для перетворення однієї діаграми в іншу (тобто відстань Вассерштейна при нормі $p = \infty$). Таким чином, відстань вузького місця не залежить від короткоживучих елементів поблизу діагоналі.

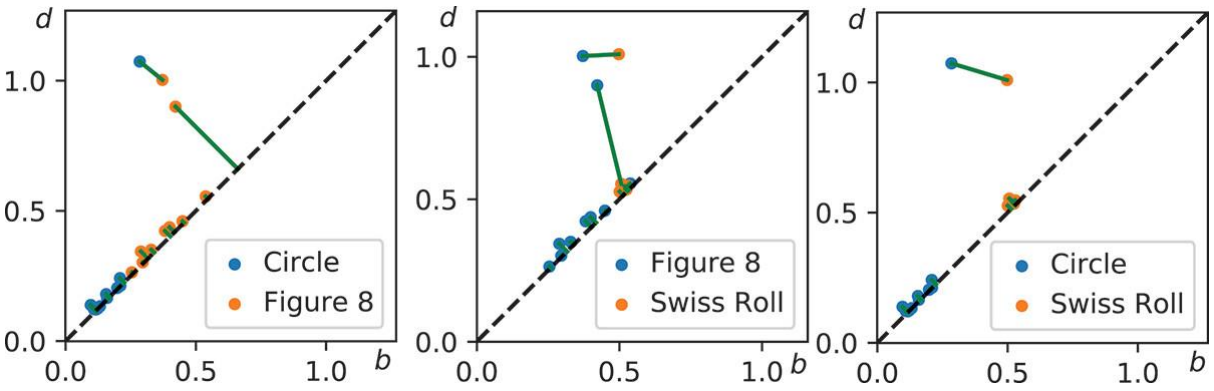

Рис. 4. Відповідність (зелені лінії) одновимірних циклів Circle, Figure 8, і хмари точок Swiss Roll, що використовуються для обчислення відстані Вассерштейна, яка відповідає сумарній довжині зелених ліній.

Альтернативні підходи для характеристики та порівняння інформації, що міститься в діаграмах персистентності, використовують векторизацію: інформація змінної довжини, закодована в парах (b, d) діаграм персистентності, відображається у вектор або вектори у фіксованому просторі; різні діаграми персистентності можна вивчати за допомогою знайомих інструментів, таких як внутрішні продукти векторів. Наприклад, можна обчислити набір зведених статистичних даних, таких як ентропія або моменти часу життя ознак $l = |d - b|$, за умови, що вони є релевантними для поставленої задачі. Часто відповідні характеристики невідомі апріорі, і бажано обчислити векторизацію високої розмірності, щоб мінімізувати втрату релевантної інформації. Одним із прикладів є ландшафт персистентності, стабільна та інверсна (тобто така, що зберігає інформацію) векторизація діаграми персистентності.

Використання міри відстані або векторизації дозволяє поєднати стійку гомологію з потужними методами машинного навчання, такими як штучні нейронні мережі або алгоритми кластеризації, для порівняння топологічних особливостей різних наборів даних і виконання завдань, включаючи ідентифікацію на основі форми та класифікацію різних хмар точок. Однак важливим моментом у застосуванні векторизації або

вимірювання відстаней є те, що вони можуть вводити додаткові гіперпараметри, які можуть впливати на чутливість до різних топологічних особливостей даних.

Існує безліч програмних бібліотек для обчислення діаграм персистентності, їх векторизації та мір відстані. Важливим для додатків є те, що діаграми персистентності можна ефективно обчислювати за умови фільтрації шляхом побудови простого комплексу по одному елементу за раз і виявлення будь-яких змін топологічних особливостей на кожному кроці. Це дає не лише шкали народження та смерті елементів, але й їхні представлення, наприклад, ребра, що утворюють цикл. Тим не менш, через комбінаторну природу простих комплексів, обчислювальні вимоги швидко зростають зі збільшенням розмірності об'єктів k, а більшість практичних застосувань обмежуються $k \leq 2$.
Для побудови діаграми персистентності кінцевий користувач повинен надати як мінімум або точки даних, або матрицю відстаней, що кодує попарні відстані між точками. Також можна розглянути можливість користувацької фільтрації. Наприклад, при роботі з даними зображень можна використовувати значення пікселів у відтінках сірого як параметр фільтрації, будуючи найпростіший комплекс з пікселів, менших (або більших) за заданий поріг. В результаті фільтрації на підрівні (надрівні) множини узагальнюються критичні точки зображення, тобто його локальні мінімуми, максимуми та сідлові точки, а також їх узагальнення у вищій розмірності.

Інші підходи та останні розробки

Вище ми обговорювали стійку гомологію лише у найпростішому випадку найпростіших комплексів, побудованих з двовимірних хмар точок. Існує безліч споріднених методів для вивчення складних наборів даних шляхом

зведення їх до сімейств графів або простих комплексів, які ми лише коротко згадуємо тут через брак місця.

Алгоритм Mapper зводить хмари точок до простіших низьковимірних графів, виконуючи кластеризацію на підмножинах даних, що перетинаються. Локальні аномалії, такі як перетини та випуклості, можна виявити подібним чином, порівнюючи стійку гомологію різних підмножин даних.
Стандартна стійка гомологія будує фільтрації як послідовність вкладених спрощених комплексів; зі збільшенням параметра фільтрації (наприклад, відстані відсікання) до комплексу додаються ребра та більш вимірні спрощення і ніколи не видаляються. У певних ситуаціях, наприклад, при вивченні часової динаміки мережі, спрощення можуть як додаватися, так і видалятися при зміні керуючого параметра. Зигзагоподібна персистентність - це техніка, яка дозволяє ідентифікувати значущі топологічні особливості в цьому випадку.

Іншою важливою проблемою є обчислення стійких топологічних особливостей при зміні декількох керуючих параметрів, що називається багатовимірною стійкістю. Ця проблема набагато складніша, ніж у випадку з одним параметром, через відсутність простих представлень діаграм персистентності.
Ми розглянули приклади, де хмари точок використовуються для побудови неорієнтованих графів та симплексів, що кодуються матрицями з бінарними елементами $\{0, 1\}$, які позначають наявність або відсутність симплекса. Стійку гомологію можна також обчислити відносно інших полів, таких як цілі числа за модулем 3, що описують, наприклад, орієнтовані графи або прості комплекси, які можуть бути корисними для

аналізу даних, із закрутками, що включають точки, відібрані з поверхні смуг Мьобіуса.

Застосування топологічного аналізу даних.

Стійка гомологія хмар точок і зображень

У багатьох додатках побудова чітко визначеної форми з даних є менш простою, або може бути зацікавлена в ідентифікації структур, присутніх у різних просторових масштабах. Наприклад, у випадку даних хмари точок може бути важко визначити розмір або радіус окремих точок. В інших випадках може виникнути потреба застосувати інтуїцію, отриману з простих аналітично розв'язуваних обмежень, до більш реалістичних систем. У таких ситуаціях стійка гомологія стає потужним інструментом для вилучення значущої інформації про форму з необроблених даних.

Наприклад, уявімо, що ми хочемо вивчити мікроскопічну структуру матеріалів. Вихідні дані, природно, враховують положення складових атомів та їхні розміри. Стійка гомологія дозволяє вивчати багатомасштабну структуру матеріалів, використовуючи лише положення атомів у тривимірному просторі (отримані з зображень або симуляцій) разом зі стандартною евклідовою відстанню. Автори використовували діаграми персистентності, розраховані на основі симуляцій молекулярної динаміки різних матеріалів, що мають склоподібні фази, для характеристики їхньої структури.
На рисунку 6 показано приклади діаграм персистентності, отриманих для рідкої, скляної та кристалічної фаз кремнезему. У кристалічній фазі кластеризація народжень і смертей ознак виявляє масштаби, що відповідають довжинам зв'язків у матеріалі, тобто відстаням між

складовими атомами. Більше того, вивчення циклів, що відповідають стійким особливостям, також розкриває природу ближнього порядку, що з'являється у скляній фазі.

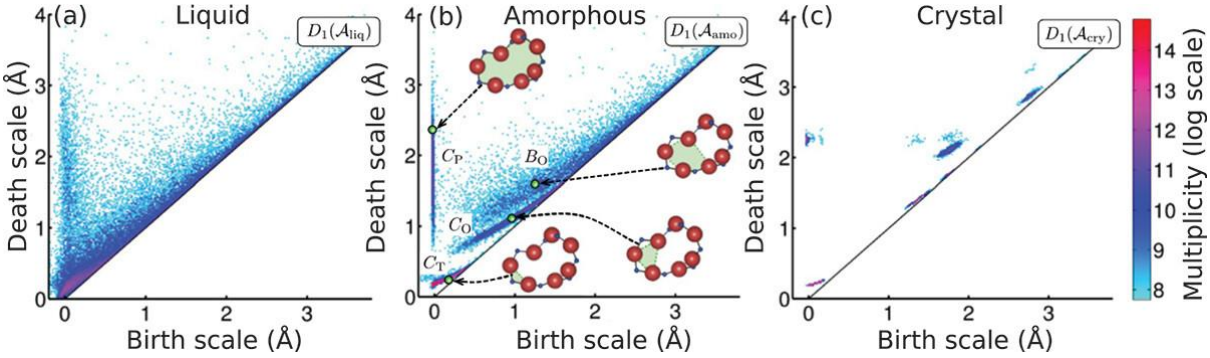

Рис. 5. Діаграми стійкості, отримані в результаті моделювання молекулярної динаміки рідкої (a), аморфної (б) та кристалічної (в) фаз кремнезему. Кольори точок вказують на кратність (у логарифмічному масштабі) одновимірних особливостей. Вставки в (b) ілюструють репрезентативні цикли, що відповідають ближньому і середньому порядку в аморфній фазі.

У наступних роботах подібні методи застосовано до аморфних льодів, гранульованих середовищ, спінових конфігурацій у ґратчастих спінових моделях і калібрувальних теоріях та двовимірних матеріалів, де міри, отримані з використанням стійкої гомології, можна безпосередньо порівнювати з більш стандартними метриками.

Інше застосування формалізму хмари точок стосується аналізу часових рядів сигналів, включаючи виявлення хаотичної динаміки. Вже в 1990-х роках виник інтерес до застосування обчислювальної топології для вивчення форми динаміки у фазовому просторі, включаючи кількісну оцінку форми хаотичних атракторів. Тут ключовим елементом є теорема вбудовування Такенса, яка стверджує, що послідовність спостережень ϕ_t, взятих через регулярні проміжки часу τ, може бути використана для

реконструкції форми динаміки шляхом побудови хмари точкових n-вимірних векторів $v_t = (\phi_t, \phi_{t-\tau}, \phi_{t-2\tau}, \dots)$, за умови, що розмірність вбудовування є достатньо великою.

Стійка гомологія дозволяє систематично вивчати динаміку за формою хмар точок у високорозмірному просторі вбудовування. Наприклад, переходи з подвоєнням періоду можна виявити за появою нових стійких кластерів. Послідовні переходи подвоєння періоду в міру наближення системи до хаотичного режиму призводять до утворення багатьох кластерів, які зливаються в одну лінію або об'єм.

Застосування стійкої гомології до даних зображень дозволяє вивчати форми зображень, в яких може не бути чіткого розмежування між "світлими" і "темними" областями, або в зображеннях, де важлива структурна інформація на декількох шкалах інтенсивності. Великі набори даних хмари точок, для яких прямий розрахунок постійної гомології може зайняти багато часу, можна вивчати за допомогою фільтрації зображень шляхом перетворення хмари в зображення густини.

Ранні роботи з персистентної гомології зображень використовували числа Бетті для характеристики розподілів сонячного магнітного поля і силових мереж у різних типах стиснених гранульованих середовищ, вивчаючи кількість і зв'язність областей на різних масштабах. Пізніше діаграми персистентності, отримані зі знімків, використовували для вивчення негауссівських флуктуацій температури в космічному мікрохвильовому фоні, форми ізочастотних контурів у фотонних кристалах, динаміки багатьох тіл і солітонів у конденсатах Бозе-Ейнштейна, фазових переходів у спінових моделях і переходів порядок-безладдя в нематичних рідких кристалах і оптичних хвилеводних ґратках. Нещодавні роботи спрямовані на те, щоб краще зрозуміти, як пов'язати інформацію про форму, отриману

за допомогою TDA, з фізичними властивостями, зокрема з проникністю матеріалів з тріщинами.

Пошук сенсу за допомогою абстрактних мір відстані

Дотепер ми розглядали приклади, де ми мали певне інтуїтивне уявлення про форму базових даних, і роль TDA полягала в тому, щоб вивчати ці форми більш систематично. Одне з нових цікавих застосувань TDA полягає у вивченні та виявленні структури складних систем, для яких не існує простих візуалізацій (наприклад, зображень або траєкторій у фазовому просторі), включно з сімействами моделей фізики високих енергій. Це, як правило, вимагає визначення відповідної міри відстані для даних.

Наприклад, у випадку квантових систем з багатьма тілами міри заплутаності між парами підсистем, такі як ентропія збігу або ентропія заплутаності, можуть бути використані для вивчення абстрактних форм квантових станів і групування їх у різні класи. Розуміння цієї структури заплутаності може бути корисним для визначення того, коли методи апроксимації, такі як тензорні мережі, можуть бути використані для ефективного моделювання системи, що нас цікавить.

Іншим важливим застосуванням абстрактних мір відстані є вивчення конденсованих систем при скінченних температурах, коли потрібно кількісно оцінити "форму" ансамблю конфігурацій системи, відібраних при певній температурі, щоб виявити фазові переходи та критичні точки. Існують різні поняття відстані, які можна застосувати в цьому контексті, включаючи геодезичну відстань між різними спіновими конфігураціями і квантову відстань, засновану на перекритті власних функцій. Тести моделей Андерсона, Хаббарда і Поттса показують, що TDA може бути

корисним для точного виявлення критичних точок, не вимагаючи обчислювально дорогого аналізу масштабування скінченного розміру.

Алгебраїчна геометрія з точки зору топологічного аналізу даних

Топологічний аналіз даних (TDA) - це історія успіху з широким спектром різноманітних застосувань. Тут я розгляну TDA з точки зору алгебраїчної геометрії.

Алгебра та алгебраїчна геометрія мають певні безпосередні застосування для TDA. Модуль персистентності - структура даних, що лежить в основі персистентної гомології (persistent homology (PH)) - є алгебраїчною концепцією, і спроби розширити PH на множину параметрів використовують поняття з комутативної алгебри. Однак, я хотів би обговорити деякі інші ролі алгебраїчної геометрії в TDA - зокрема, застосування NAG (Numerical Algebraic Geometry) та EAG (Enumerative Algebraic Geometry).

NAG стосується обчислення чисельних розв'язків системи n поліноміальних рівнянь $F(x) = (f_1(x), f_2(x) \dots f_n(x)) = 0$ у n змінних $x = (x_1, \dots, x_n)$ над комплексними числами. Чисельне продовження гомотопії є обчислювальною парадигмою в NAG. Ця парадигма передбачає генерування системи рівнянь $G(x)$ розв'язки якої відомі (так звана стартова система), і продовження розв'язків $G(x) = 0$ вздовж деформації $G(x)$ в напрямку $F(x)$.

З іншого боку, EAG підраховує кількість розв'язків системи поліноміальних рівнянь. Хоча спочатку може здатися, що вони відрізняються, NAG і EAG тісно пов'язані між собою: основна перевага NAG полягає в тому, що можна згенерувати початкові значення для всіх

ізольованих розв'язків $F(x) = 0$ в C^n. Наприклад, якщо степінь i-го полінома дорівнює d_i, то $G(x) = (x^{d_i}{}_i - a_i)_{a_i}{}^n$ - де $a_1, \ldots\ldots, a_n \epsilon C^*$ - може слугувати початковою системою для гомотопії $(1-t)G(x) + t\,F(x), 0 \le t \le 1\,.G(x) = 0$, яка має $D = d_1, \ldots, d_n$ ізольованих розв'язків, а з теореми алгебраїчної геометрії випливає, що $F(x) = 0$ має не більше D ізольованих розв'язків. Якщо продовжити пошук розв'язків $G(x) = 0$ у напрямку $F(x) = 0$, то отримаємо всі ізольовані розв'язки $F(x) = 0$. На практиці, однак, $F(x) = 0$ має значно менше розв'язків, ніж D, і розв'язки, що розходяться, необхідно відкинути. EAG допомагає будувати інші стартові системи, адаптовані до структури $F(x)$ та підвищити ефективність алгоритму.

Повертаючись до TDA, розглянемо ситуацію, в якій $M \subset R^n$ є нульовою множиною s поліномів від n змінних $F(x) = (f_1(x), \ldots, f_8(x))$. В алгебраїчній геометрії такий M називається дійсною алгебраїчною множиною. Прикладом може слугувати конформаційний простір молекули циклооктану. Циклооктани складаються з восьми атомів вуглецю $x_1, x_2 \ldots, x_8 \epsilon R^3$, вирівняних у кільце так, що відстані між сусідніми атомами рівні $c > 0$. Енергія конфігурації $(x_1, \ldots, x_8)$ мінімізується, коли кожен кут між послідовними зв'язками дорівнює $arccos(-\frac{1}{3}) \approx 109,5°$. Поліноміальні рівняння у $3 \cdot 8 = 24$ змінних мають вигляд

$$||x_1 - x_2||^2 = \ldots = ||x_7 - x_8||^2 = ||x_8 - x_1||^2 = c^2$$

і

$$||x_1 - x_3||^2 = \ldots = ||x_6 - x_8||^2 = ||x_7 - x_1||^2 = ||x_8 - x_2||^2 = \frac{8}{3}c^2$$

Множина розв'язків цих рівнянь - до одночасного перекладу та обертання - гомеоморфна об'єднанню пляшки Кляйна та сфери, які перетинаються у двох кільцях.

NAG може генерувати вибірку точок з M, які потім можуть слугувати вхідними даними для PH; на рисунку 6 зображено вибірку з множини циклооктанів. Одна з ідей передбачає вибірку лінійних просторів L розмірності, що дорівнює комірності M, і обчислення точок на перетині M і L. В іншому підході дослідники вибирають точки $q \in R^n$ у навколишньому просторі і визначають точку на *M*, яка мінімізує відстань до q. Обидві обчислювальні задачі можна подати у вигляді системи поліноміальних рівнянь і використовувати NAG для їх розв'язання. Напрямки досліджень в NAG включають в себе те, як робити вибірку відносно розподілу ймовірностей на *M* та створювати вибірки з бажаним рівнем щільності на *M*.

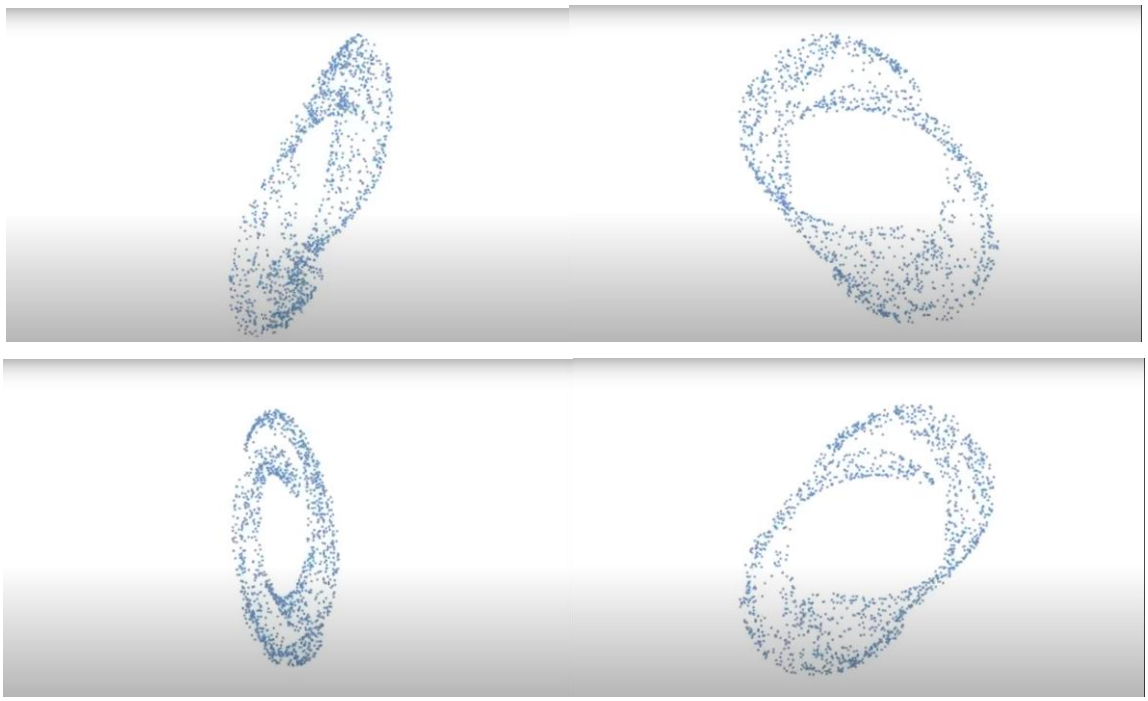

Рис. 6. Приклад з циклооктанової множини для $c^2 = 2$, спроектований на тривимірний простір. Завдяки трансляційній та обертальній інваріантності,

$x_1 = (0,0,0)$ та $x_2 = (c, 0,0)$ фіксовані, а останній елемент x_3 дорівнює нулю. Зображення взяті з робіт Пауля Брейдінга та Саши Тімме.

Дослідники також використовують NAG (Numerical Algebraic Geometry) та EAG (Enumerative Algebraic Geometry) для вивчення двох важливих чисел для TDA: homological feature size $hfs(M)$ та досяжність $\tau(M)$ множини M. Еміль Хоробет та Маделейн Вайнштейн показали, що якщо M є алгебричним многовидом (тобто, реальною алгебричною множиною, яка також є многовидом), заданим поліномами над Q, то як $hfs(M)$ так і $\tau(M)$ є алгебраїчними над Q. Таким чином, їх можна обчислити за допомогою NAG.

Досяжність множини M визначається як відстань від M до її медіальної осі. Еквівалентне визначення:

$$\tau(M) = min\{\frac{1}{\sigma(M)}, \frac{1}{2}\rho(M)\},$$

де $\sigma(M)$ — це максимальна крива геодезичної лінії на M, а $\rho(M)$ — ширина найвужчого “bottleneck” множини M. “bottleneck” — це пара $(x, y)\ \epsilon\ M^2$, така, що $x - y$ перпендикулярний до обох дотичних просторів T_xM та T_yM. Цю задачу можна сформулювати як систему поліноміальних рівнянь (розв'язувану за допомогою NAG) і вилучити реальні розв’язки серед комплексних. Варто зауважити, що під час розрахунків ізольованих розв’язків системи поліномів у NAG тривіальні розв’язки, де $x = y$, не враховуються. Вчені інтенсивно вивчали “bottleneck” як з точки зору NAG, так і EAG, використовуючи полярні класи множини M.

Рівняння для $\sigma(M)$ є менш прямолінійними, але пряма формула для кривини $\gamma(x)$ у точці $x\ \epsilon\ M$ вказує, що:

$\sigma(M) = min_{x\epsilon M}\gamma(x)$, для плоских кривих. У цьому випадку умови першого порядку для $\sigma(M)$, які передбачають, що градієнт $\gamma(x)$ є перпендикулярним до дотичного простору T_xM, породжують систему поліноміальних рівнянь. Розв'язання цієї системи дає $\sigma(M)$.

Таким чином, можна окремо обчислити $\rho(M)$ та $\sigma(M)$ за допомогою NAG і вивести досяжність $\tau(M)$ із цих чисел. На рисунку 7 показано приклад такого обчислення для плоскої кривої.

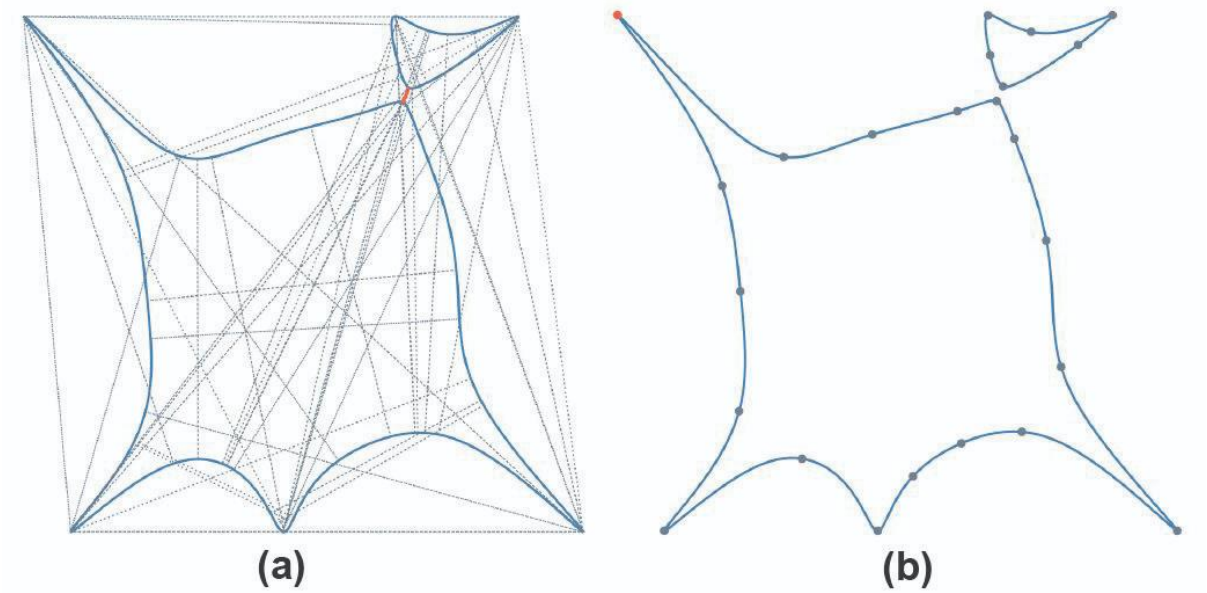

Рис. 7. Компоненти обчислення досяжності плоскої кривої
$C = \{(x^3 - x \cdot y^2 + y + 1)^2(x^2 + y^2 - 1) + y^2 = 5$
7.a. Усі шийки множини C. Найвужчий "bottleneck" виділена із шириною приблизно ≈ 0.138.
7.b. Усі точки критичної кривини множини C. Червона точка у верхньому лівому куті — це точка максимальної кривини, і її значення ≈ 2097.17.
Таким чином,

$$\tau(C) \approx min\{\frac{1}{2097.17}\},\frac{0.138}{2}\} = \frac{1}{2097.17} \text{ [2]}$$

Можна також замінити досяжність нижньою межею, яка включає дійсне число умов системи поліномів [3]. Ця нижня межа справедлива як для дійсних алгебраїчних многовидів, так і для більш загального класу

напівалгебраїчних множин. Дослідники використовують число умов для аналізу складності алгоритму обчислення гомології.

Наостанок, я хотів би запропонувати три можливі майбутні напрямки використання алгебраїчної геометрії в TDA. Перший - це аналіз $\sigma(M)$ в контексті NAG та EAG для пошуку вузьких місць [7, 9]. Це є незамінним при обчисленні досяжності за межами плоских кривих. Другий - PH з використанням еліпсоїдів. Експерименти показують, що цей підхід може значно покращити якість вихідних діаграм у PH [2], але йому бракує теоретичного пояснення. Третій напрямок пов'язаний з дискретизацією. Стандартний підхід до дискретизації нелінійних об'єктів використовує марковські ланцюгові методи Монте-Карло. Поєднання цього підходу з NAG видається перспективним.

Методи топологічного аналізу даних.

Базовий алгоритм

Останнім часом TDA відомий своїми розробками в різних напрямках і сферах застосування. Зараз існує велика різноманітність методів, заснованих на топологічних і геометричних підходах. Надання повного огляду всіх цих існуючих підходів виходить за рамки цього вступного огляду. Однак, багато стандартних підходів спираються на наступний базовий конвеєр, який є основою:

1. Вважається, що вхідними даними є скінченна множина точок з поняттям відстані - або подібності - між ними. або подібності між ними. Ця відстань може бути задана метрикою в навколишньому просторі (наприклад, евклідовою метрикою, коли дані містяться в R^n) або бути внутрішньо притаманною метрикою, визначеною матрицею попарних відстаней. Визначення метрики на даних зазвичай задається на вході або

визначається програмою. Однак важливо зауважити, що що вибір метрики може мати вирішальне значення для виявлення цікавих топологічних і геометричних особливості даних.

2. "Неперервна" фігура будується поверх даних, щоб підкреслити основну топологію або геометрію. Це часто є простим комплексом або вкладеним сімейством простих комплексів, які відображають структуру даних у різних масштабах. Прості комплекси комплекси можна розглядати як вимірні узагальнення сусідніх графів, які класично будуються поверх даних у багатьох стандартних алгоритмах аналізу даних або навчання. Виклик тут полягає в тому, щоб завдання полягає у визначенні таких структур, які, як доведено, відображають релевантну інформацію про структуру даних про структуру даних і якими можна ефективно будувати та маніпулювати на практиці.

3. Топологічна або геометрична інформація витягується зі структур, побудованих поверх даних. Це може призвести або до повної реконструкції, зазвичай тріангуляції, форми, що лежить в основі даних що лежить в основі даних, з якої можна легко виокремити топологічні/геометричні особливості, або у грубих узагальненнях або наближеннях, з яких вилучення відповідної інформації вимагає спеціальних методів, таких як, наприклад, стійка гомологія. Окрім виявлення цікавої топологічної/геометричної інформації цікавої топологічної/геометричної інформації, її візуалізації та інтерпретації, завдання на цьому етапі завдання полягає в тому, щоб показати її релевантність, зокрема її стійкість до збурень або наявності шуму. збурень або наявності шуму у вхідних даних. З цією метою розуміння статистичної поведінки виведених ознак також є важливим питанням.

4. Видобута топологічна та геометрична інформація надає нові сімейства ознак та дескрипторів даних. Їх можна використовувати для кращого розуміння даних - зокрема, за допомогою візуалізації, або поєднувати з

іншими типами ознак для подальшого аналізу та завдань машинного навчання. Ця інформація також може бути використана для розробки відповідних моделей аналізу даних і машинного навчання. Важливим питанням на цьому етапі є демонстрація додаткової цінності та взаємодоповнюваності (по відношенню до інших функцій) інформації, що надається інструментами TDA.

Персистентна гомологія.

По суті, топологія вивчає геометричні та просторові відношення які зберігаються (є стабільними) в умовах безперервних деформацій об'єкта (наприклад, розтягування, скручування та згинання). Така перспектива має низку переваг над іншими методами аналізу даних :

- Топологія вивчає дані у спосіб, який не залежить від обраних координат.
- Топологія вивчає дані таким чином, щоб мінімізувати чутливість до вибору метрики.
- Топологія узагальнює відомі методи теорії графів до просторів високої розмірності.
- Топологія є стійкою до великої кількості шуму.

Основну увагу приділимо методу в області TDA, відомому як стійка гомологія. Метою персистентної гомології є виявлення та кількісна оцінка топологічно домінуючих особливостей у даних у вигляді базових (низьковимірних) топологічних ознак, таких як з'єднані компоненти, отвори, пустоти та їхні узагальнення. Ця інформація може бути використана статистичними методами та методами машинного навчання для виконання завдань регресії, класифікації, перевірки гіпотез та кластеризації.

Майбутні напрямки

Нові методи топологічного аналізу даних

Сферою активних досліджень серед фізиків є застосування інструментів TDA для аналізу структури більш складних систем, включаючи потокові мережі зі спрямованими зв'язками та мережі, що еволюціонують у часі. Один з підходів, який використовується в останніх дослідженнях і сумісний зі стандартними інструментами персистентної гомології, полягає в перетворенні спрямованої мережі в звичайну хмару точок за допомогою карти дифузії, яка будує ребра між парою вершин (i, j) шляхом обчислення ймовірності дифузії між i і j. Буде цікаво дослідити альтернативні підходи, які можуть працювати безпосередньо з односпрямованими системами або системами, що еволюціонують у часі, без використання карт дифузії, таких як персистентність зигзага.

Квантовий топологічний аналіз даних

Всі приклади, розглянуті у фізичній літературі, стосуються вивчення низьковимірних топологічних особливостей за допомогою TDA, здебільшого тому, що більш високовимірні особливості важче інтерпретувати, а обчислення для великих наборів даних стають надзвичайно трудомісткими через експоненціальне масштабування. Очікується, що поява більш ефективних квантових алгоритмів для TDA, включаючи обчислення чисел Бетті та діаграм персистентності, уможливить вивчення топологічних особливостей вищої розмірності у складних наборах даних.

Перший квантовий алгоритм для TDA був запропонований Ллойдом та ін. у 2016 році. Їхній алгоритм продемонстрував експоненціальне прискорення обчислення чисел Бетті завдяки використанню квантової

фазової оцінки для ефективної побудови комбінаторних лапласіанів найпростіших комплексів та ідентифікації циклів шляхом обчислення їхніх нульових модальностей. За цією пропозицією у 2018 році послідував маломасштабний експеримент з квантової оптики, що підтвердив концепцію, розміром у кілька кубітів.

Подальші дослідження почали розглядати обмеження першого квантового алгоритму TDA, пропонуючи більш ефективні варіанти, а також квантові алгоритми для обчислення постійних чисел Бетті та відстані Вассерштейна. Хоча більшість з цих алгоритмів призначені для майбутніх відмовостійких квантових комп'ютерів, існує також потенціал для короткострокового квантового прискорення з використанням неглибоких квантових схем з глибиною, лінійною до кількості точок вхідних даних, експлуатуючи ефективну реалізацію граничного оператора з використанням заплутаних квантових станів. Хоча перспектива експоненціального прискорення порівняно з найкращими класичними алгоритмами TDA є привабливою, питання про те, чи буде досягнута така велика швидкість для практичних задач і коли це станеться, залишається дискусійним, особливо з огляду на те, що нові та вдосконалені класичні алгоритми все ще розробляються.

Додаток

Існує низка пакетів, доступних для TDA; інструментарій швидко розвивається, і різні пакети мають різні сильні сторони. Ми проілюструємо пакет scikit-tda, який є загальнодоступним пакетом scikit-tda для TDA на основі python, доступним за адресою https://scikit-tda.org/libraries.html.

```
# ці команди готують нас до побудови графіків
import numpy as np from ripser
import ripser from persim
import plot_diagrams
import matplotlib.pyplot as plt

# наш перший приклад бере 100 випадкових точок на площині
data = np.random.random((100,2))
plt.scatter(data[:,0], data[:,1])
plt.show()
```

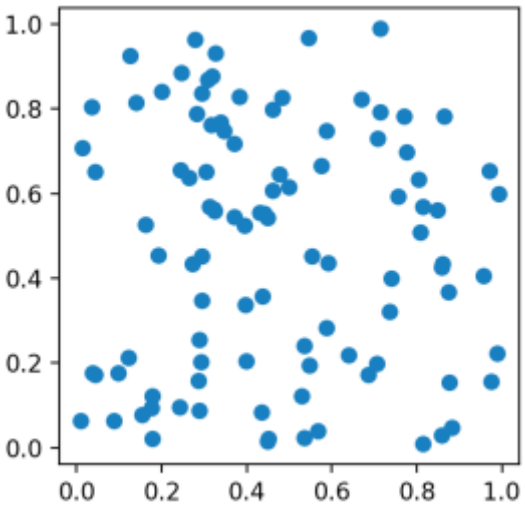

Рис. 8. 100 випадкових точок

 # і потім будуємо діаграму персистентності

$diagrams = ripser(data)['dgms']$

$plot\_diagrams(diagrams, show = True)$

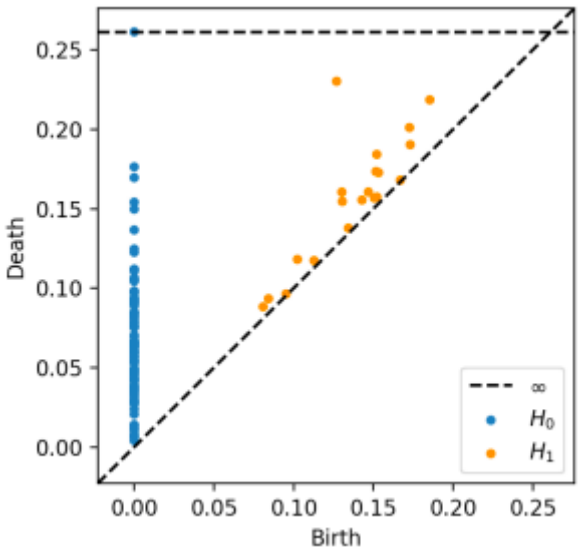

Рис. 9. Відповідна діаграма персистентності

 # цікавіший набір даних

масштабуємо всі точки у нашому випадковому наборі даних до норми 1.

$data = data - 0.5$

$data = data/np.linalg.norm(data, ord = 2, axis = 1).reshape((-1,1))$

$plt.scatter(data[:,0], data[:,1])$

$plt.show()$

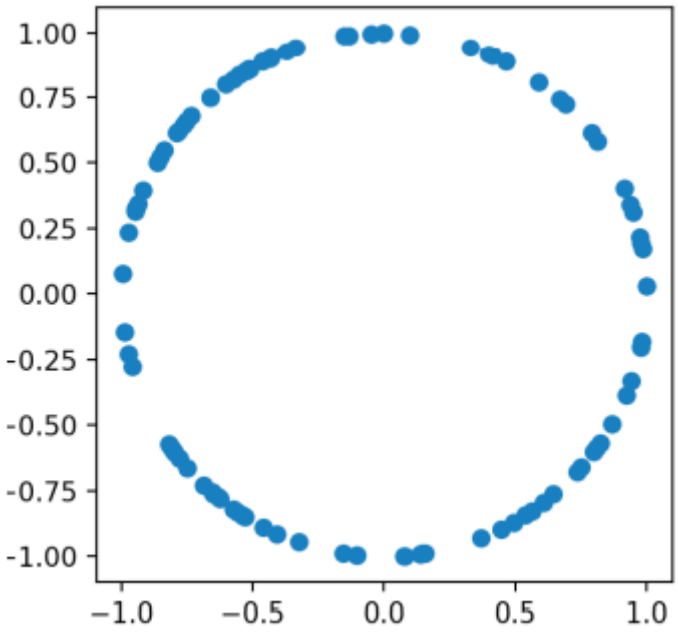

Рис. 10. 100 випадкових точок на колі

 # і потім будуємо діаграму персистентності

$$diagrams = ripser(data)['dgms']$$
$$plot\_diagrams(diagrams, show = True)$$

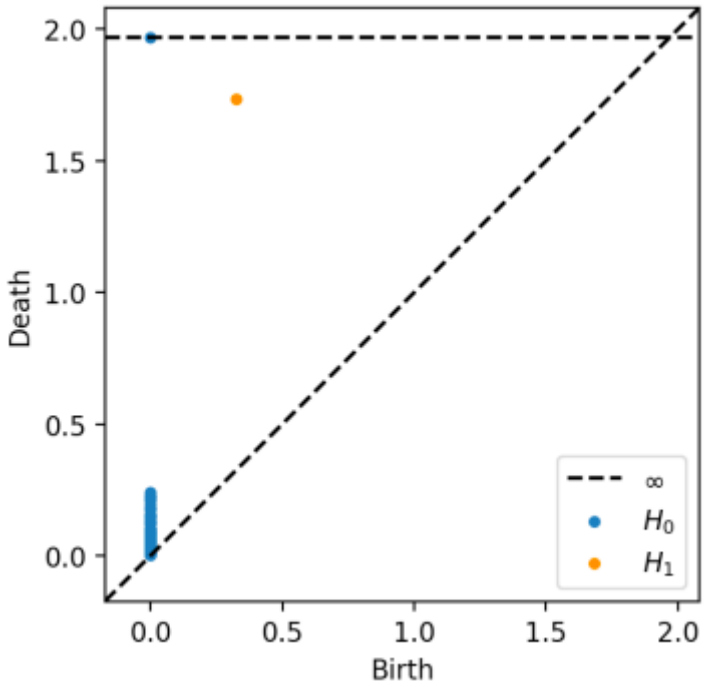

Рис. 11. Відповідна діаграма персистентності

```
# тепер додамо трохи шуму до упорядкованих колових даних
noisydata = np.random.random((10,2)) - 0.5
datapoints = np.vstack([data, noisydata])
plt.scatter(datapoints[:,0], datapoints[:,1])
plt.show()
```

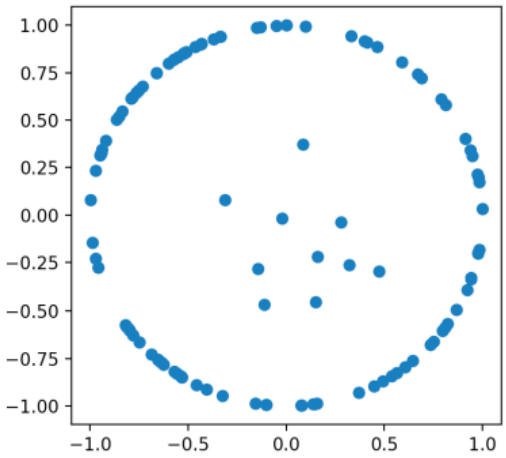

Рис. 12. 100 випадкових точок на колі і 10 випадкових внутрішніх точок

```
# а потім побудуймо діаграму персистентності
diagrams = ripser(datapoints)['dgms']
plot_diagrams(diagrams, show = True)
```

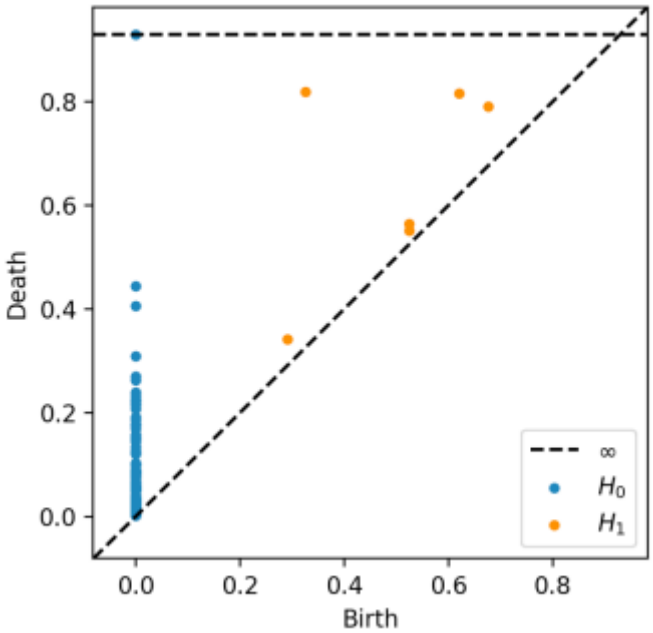

Рис. 13.. Відповідна діаграма персистентності

```
# замість випадкових точок додамо шум до точок кола
rng = np.random.default_rng()
```

$$data = data + rng.normal(scale = 0.1, size = data.shape)$$
$$plt.scatter(data[:,0], data[:,1])$$
$$plt.show()$$

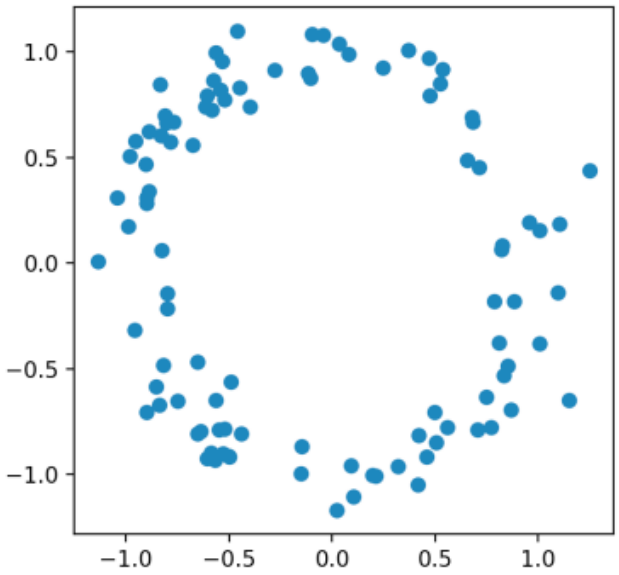

Рис. 14. 100 точок шуму на колі

 # а потім побудуймо діаграму персистентності

$$diagrams = ripser(data)['dgms']$$
$$plot\_diagrams(diagrams, show = True)$$

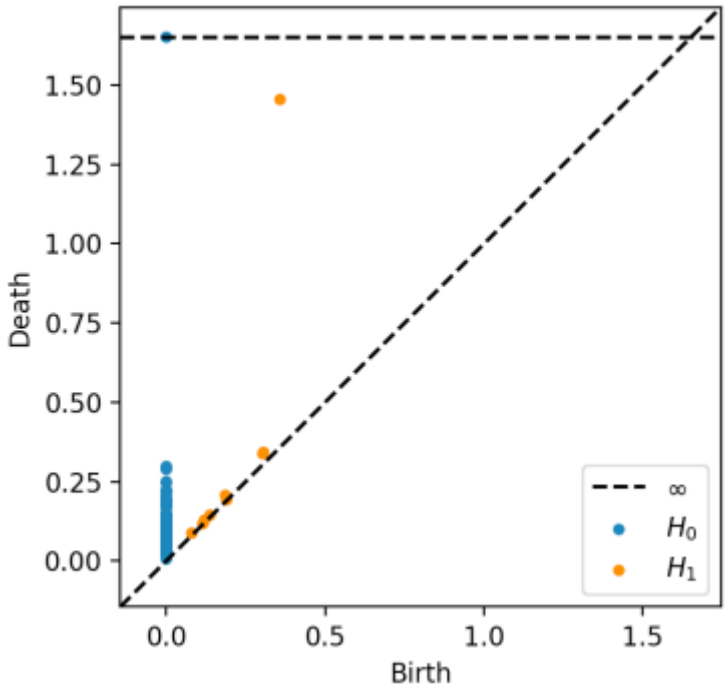

Рис. 15. Відповідна діаграма персистентності

Бібліографічний список використаної літератури

1. Breiding, P., & Marigliano, O. (2019). Random points on an algebraic manifold. Preprint, *arXiv:1810.06271*.
2. Breiding, P., Kalisnik, S., Sturmfels, B., & Weinstein, M. (2018). Learning algebraic varieties from samples. *Revis. Matemát. Complut., 31*, 545-593.
3. Bürgisser, P., Cucker, F., & Lairez, P. (2019). Computing the homology of basic semialgebraic sets in weak exponential time. *J. ACM, 66*(1), 1-30.
4. Carlsson, G., & Zomorodian, A. (2009). The theory of multidimensional persistence. *Discrete Comput. Geom., 42*, 71-93.
5. Cohen-Steiner, D., Edelsbrunner, H., & Harer, J. (2007). Stability of Persistence Diagrams. *Discrete Comput. Geom., 37*, 103-120.
6. Coutsias, E., Martin, S., Thompson, A., & Watson, J. (2010). Topology of cyclo-octane energy landscape. *J. Chem. Phys., 132*, 234115.
7. Di Rocco, S., Eklund, D., & Weinstein, M. (2019). The bottleneck degree of algebraic varieties. Preprint, *arXiv:1904.04502*.
8. Dufresne, E., Edwards, P., Harrington, H., & Hauenstein, J. (2018). Sampling real algebraic varieties for topological data analysis. Preprint, *arXiv:1802.07716*.
9. Eklund, D. (2018). The numerical algebraic geometry of bottlenecks. Preprint, *arXiv:1804.01015*.
10. Harrington, H., Otter, N., Schenck, S., & Tillmann, U. (2017). Stratifying multiparameter persistent homology. *SIAM J. Appl. Alg. Geom., 3*, 439-471.
11. Horobet, E., & Weinstein, M. (2019). Offset hypersurfaces and persistent homology of algebraic varieties. *Comput. Aid. Geom. Des., 74*, 101767.
12. Niyogi, P., Smale, S., & Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom., 39*, 419-441.

13. Frédéric Chazal, Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists.
14. Alexander D. Smith, Paweł Dłotko, Victor M. Zavala. Topological Data Analysis: Concepts, Computation, and Applications in Chemical Engineering.
15. Adams H., Emerson T., Kirby M., Neville R., Peterson C., Shipman P. Persistence Images: a Stable Vector Representation of Persistent Homology.
16. Murugan J, Robertson D. An introduction to topological data analysis for physicists.
17. Carlsson G. Topology and Data.