17 lutego 2026 r.

prof. dr hab. inż. Rafał Scherer
Katedra Sztucznej Inteligencji
Wydział Informatyki i Sztucznej Inteligencji
Politechnika Częstochowska

Wydział Informatyki
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Recenzja

rozprawy doktorskiej mgr. Piotra Tempczyka pt.: Estimating Local Intrinsic Dimension via Density Estimation (pol. Estymacja lokalnej wymiarowości rozmaitości danych za pomocą modeli estymacji gęstości).

Niniejszą recenzję opracowano zgodnie z uchwałą Rady Naukowej Dyscypliny Informatyka Uniwersytetu Warszawskiego z dnia 30 października 2025 r. Promotorem jest dr hab. Marek Cygan, prof. ucz.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Rozprawa wpisuje się w dynamicznie rozwijający się obszar badań nad strukturą geometryczną danych wysokowymiarowych, ze szczególnym uwzględnieniem lokalnego wymiaru wewnętrznego rozmaitości danych oraz jego zastosowań w nowoczesnych metodach uczenia maszynowego. Tematyka pracy ma istotne znaczenie zarówno teoretyczne, jak i praktyczne: lokalny wymiar wewnętrzny stanowi pomost między abstrakcyjną geometrią danych a konkretnymi decyzjami modelowymi, takimi jak dobór architektury, rozmiaru przestrzeni ukrytej czy strategii uczenia modeli generatywnych i klasyfikatorów. W dobie coraz większych i bardziej złożonych zbiorów danych potrzeba narzędzi pozwalających wiarygodnie opisywać lokalną złożoność rozkładu staje się kluczowa dla oceny generalizacji modeli, projektowania benchmarków oraz diagnozowania trudnych regionów przestrzeni cech, co uzasadnia wagę podjętego w rozprawie problemu.

Recenzowana rozprawa doktorska podejmuje aktualny i istotny problem estymacji lokalnego wymiaru wewnętrznego danych wysokowymiarowych, osadzony w szerszym kontekście hipotezy rozmaitości i nowoczesnych metod modelowania gęstości. Celem pracy jest zaproponowanie i dogłębna analiza metod szacowania lokalnej struktury wymiarowej rozkładów danych, ze szczególnym uwzględnieniem podejść opartych na sieciach neuronowych i procesie Wienera, oraz ocena ich zachowania na realistycznych zbiorach danych. Tezą badawczą rozprawy jest stwierdzenie, że lokalny wymiar wewnętrzny można wykryć z zachowania rozkładu danych pod działaniem kontrolowanego szumu Gaussa, a nowo zaproponowane algorytmy i benchmarki pozwalają zarówno na skalowalne szacowanie lokalnego wymiaru, jak i na bardziej krytyczną ocenę istniejących metod.

2. Zawartość rozprawy

Recenzowana praca mgr Piotra Tempczyka składa się z siedmiu rozdziałów oraz dodatku. Dokument liczy 122 strony.

Rozdział 1 stanowi wprowadzenie do problematyki estymacji lokalnego wymiaru wewnętrznego (Local Intrinsic Dimension, LID) danych wysokowymiarowych, osadza badania w kontekście hipotezy rozmaitości (manifold hypothesis), definiuje problem lokalnego wymiaru wewnętrznego jako własności punktowej i zależnej od skali, a następnie przybliża tło historyczne, wkład własny autora, pokrewną literaturę oraz plan dalszych rozdziałów. Autor podkreśla, że efektywna wymiarowość danych w praktyce nie jest stałą globalną, lecz zmienia się w zależności od położenia i skali obserwacji oraz zależy m.in. od krzywizny, brzegów, szumu i niejednorodności próbkowania, co motywuje perspektywę lokalną zamiast uśredniania po całym zbiorze. Rozdział definiuje intuicyjnie rozmaitość danych: zbiór obserwacji jest traktowany jako (lokalnie) d-wymiarowa struktura zanurzona w przestrzeni o wymiarze D, z rozróżnieniem kierunków stycznych (poruszających się „po" rozmaitości) i normalnych (oddalających od niej). Autor podaje też proste przykłady obiektów o różnym wymiarze wewnętrznym i wykorzystuje je do uzasadnienia, że wymiar ma sens jako liczba lokalnie dostępnych stopni swobody. LID jest definiowany jako miara opisująca, ile „aktywnych" kierunków zmienności występuje w małym sąsiedztwie punktu zapytania, przy czym parametr skali kontroluje, jak lokalnie mierzymy geometrię danych. Następnie rozdział omawia zastosowania LID, traktując go jako narzędzie diagnostyczne łączące geometrię danych z praktyką uczenia maszynowego. Wśród przywoływanych obszarów pojawiają się m.in. analiza dynamiki uczenia i reprezentacji w sieciach (śledzenie zmian zmniejszania się reprezentacji), dobór docelowego wymiaru w redukcji wymiarowości, weryfikacja założeń w uczeniu rozmaitości i estymacji gęstości, dobór rozmiaru przestrzeni ukrytej w autoenkoderach, uogólnianie, analiza zapamiętywania przykładów w modelach generatywnych, detekcja obiektów spoza dystrybucji oraz związek LID z błędem rekonstrukcji i niepewnością predykcji. Wspólnym mianownikiem tych przykładów jest teza, że lokalna wymiarowość pozwala analizować niejednorodność złożoności danych, co może pomagać zarówno w doborze architektury i hiperparametrów, jak i interpretacji zachowania modeli. Dalej autor przybliża historię estymacji wymiaru wewnętrznego: od liniowych miar w rodzaju PCA, poprzez estymatory oparte o najbliższych sąsiadów i idee fraktalne (skalowanie liczności par w promieniu), aż po metody maksymalizacji wiarygodności i nowsze lokalne estymatory uwzględniające zjawiska koncentracji miary. Jest też formalizacja LID jako obiektu ściśle lokalnego, z powiązaniami do teorii wartości ekstremalnych i zastosowań w analizie anomalii czy wyszukiwaniu podobieństwa, a także przykłady użycia LID w kontekście podatności na ataki adwersarialne. Widać, że klasyczne, nieparametryczne techniki (kNN i pokrewne) mają ograniczenia przy bardzo wysokiej wymiarowości i złożonosci danych, co stanowi punkt wyjścia dla rozwiązań neuronowych rozwijanych w pracy.

Sekcja „My Contribution" streszcza trzy główne elementy nowości rozprawy. Po pierwsze, autor stworzył LIDL - estymator LID oparty o neuronową estymację gęstości (w praktyce z użyciem przepływów normalizujących (normalizing flows)), który bazuje na kontrolowanym zaburzaniu danych szumem Gaussa i odczytywaniu wymiaru z nachylenia zależności logarytmu gęstości od szumu, co daje sterowanie granulacją próbkowania i lepszą skalowalność do danych wysokowymiarowych. Po drugie, autor proponuje zastosowanie procesu Wienera (dyfuzji) jako wspólną metodę dla wielu współczesnych metod neuronowych estymacji LID, porządkując je (m.in. na metody „izolowane" i „holistyczne") oraz wyjaśniając, że różne algorytmy w istocie odczytują ten sam sygnał krótkoczasowego zachowania rozmytej (zdyfundowanej) gęstości, tylko w innych współrzędnych: przez nachylenia logarytmu wiarygodności, widma Jacobianu lub geometria funkcji score. Po trzecie, praca proponuje nowoczesny zestaw benchmarków i transformacji testowych, które mają testować estymatory LID na cechach zwykle powodujących błędy (niejednorodność, krzywizna, brzegi, „cienkie" struktury, bliskość komponentów, efekty liczności próby) oraz łączyć narzędzia analityczne z danymi domenowymi poprzez transformacje zachowujące strukturę.

W „Related Work” autor grupuje literaturę tematycznie: przeglądy i benchmarki, prace podstawowe i wczesne estymatory, metody oparte o wiarygodność i *k*NN, estymatory wykorzystujące kąty i/lub widma, podejścia minimalistyczne i wieloskalowe, formalizację LID, zastosowania w uczeniu reprezentacji oraz najnowsze estymatory neuronowe (m.in. opartych o przepływy i dyfuzję), które wpisują się w proponowaną w rozprawie metodę dodawania szumu w celu analizy lokalnej geometrii. Rozdział kończy omówienie zawartości pracy.

Rozdział 2 przedstawia metodę LIDL (Local Intrinsic Dimension estimation using approximate Likelihood) wraz z jej uzasadnieniem teoretycznym, analizą zachowania w przykładach o znanej postaci rozkładu oraz serią kontrolowanych eksperymentów numerycznych, w których zakłada się idealny dostęp do zagęszczonej (po zaszumieniu) gęstości, tak aby obserwowane błędy wynikały z samej konstrukcji estymatora, a nie z niedoskonałości modelu gęstości. Autor zaczyna od intuicji: dodanie izotropowego szumu Gaussa o odchyleniu standardowym σ do danych leżących lokalnie na *d*-wymiarowej rozmaitości w *D*-wymiarowej przestrzeni powoduje, że logarytm gęstości w punkcie na rozmaitości zmienia się w przybliżeniu liniowo względem log σ, a nachylenie tej zależności niesie informację o *d* (dokładniej: o *d*-*D*), co pozwala odzyskać lokalny wymiar przez prostą regresję liniową po kilku poziomach szumu.

W części formalnej rozdział precyzuje założenia: rozważana jest gładka miara prawdopodobieństwa wsparta na d-wymiarowej, gładkiej i (początkowo) spójnej, zanurzonej podrozmaitości *S* zawarty w *D*-wymiarowej przestrzeni euklidesowej, a następnie analizowana jest gęstość rozkładu po splocie z $N(0, \sigma^2 I)$. Kluczowa konstrukcja dowodu opiera się na lokalnym rozkładzie przestrzeni w punkcie $x \in S$ na składową styczną i normalną (dekompozycja na przestrzeń styczną i normalną), reprezentacji rozmaitości jako wykresu funkcji o zerowej pochodnej w zerze (stąd dominacja składnika drugiego rzędu) oraz na oszacowaniach całki splotu po małej kuli wokół x. W efekcie otrzymuje się twierdzenie główne (2.2.7), które mówi, że dla małych σ zachodzi przybliżenie $\log p_\sigma(x) \approx (d - D)\log \sigma + O(1)$; praktycznie oznacza to, że dopasowanie prostej do punktów $\left(\log \sigma_i, \log \hat{p}_{\sigma_i}(x)\right)$ daje estymatę *d* po przekształceniu nachylenia.

Autor następnie omawia procedurę obliczeniową: dla wybranej listy poziomów szumu $\sigma_1, \ldots, \sigma_n$ tworzy się zaszumione zbiory danych, uczy się dla każdego z nich model gęstości (w pracy jako przykład wskazane są przepływy normalizujące, ale podkreśla się „agnostyczność” metody względem wyboru estymatora gęstości), a potem dla każdego punktu zapytania wykonuje się regresję liniową po skali i odczytuje wymiar. Ważnym elementem jest interpretacja σ jako jawnego parametru skali: rozdział pokazuje, że zwiększanie σ prowadzi do „ignorowania” struktur cieńszych niż dana skala (np. w rozkładach o bardzo anizotropowych wariancjach), co czyni z LIDL narzędzie pozwalające świadomie ustawić rozdzielczość geometrii, a jednocześnie ułatwia tłumienie drobnoskalowego szumu obserwacyjnego. Autor zestawia tę własność z klasycznymi metodami sąsiedztwa (*k*NN), gdzie skala jest pośrednio kontrolowana przez liczbę sąsiadów lub promień i silnie zależy od lokalnej gęstości oraz liczebności próby.

Rozdział rozszerza też zakres założeń geometrycznych: omawia przypadek rozmaitości niespójnych oraz sytuacje z samo-przecięciami, przechodząc od zanurzeń do tzw. „dobrych immersji” (good immersions) spełniających warunek skończoności przeciwobrazu w lokalnym otoczeniu. Pokazane jest, że w punkcie przecięcia (lub w otoczeniu, gdzie nakładają się obrazy kilku składowych) dominować będzie wkład tej składowej, która ma najmniejszy wymiar (formalnie pojawia się minimum po wymiarach składowych odwzorowujących się w dane x), co stanowi ważną wskazówkę interpretacyjną dla danych będących unią rozmaitości.

W części z przykładami analitycznymi autor pokazuje przypadki, w których można policzyć zachowanie estymatora wprost. Dla rozkładu jednostajnego na otwartym podzbiorze

przestrzeni euklidesowej zanurzonym w tej przestrzeni okazuje się, że dla punktów wewnętrznych (z dala od brzegu) zachowanie logarytmu gęstości po zaszumieniu prowadzi do właściwego nachylenia i tym samym poprawnej estymacji wymiaru. Dla anizotropowego Gaussa w przestrz. euklidesowej, Autor wyprowadza wynik pokazujący, że w przedziale σ pomiędzy kolejnymi odchyleniami standardowymi estymata odpowiada w przybliżeniu liczbie „grubych" kierunków (tych o wariancji większej niż σ^2), co formalizuje obserwowane wcześniej „odcinanie" cienkich wymiarów przez parametr skali. Analizowany jest też przypadek dyskretnych punktów na prostej (0-wymiarowa rozmaitość), gdzie otrzymuje się warunki na σ zapewniające kontrolę dodatkowego składnika w logarytmie (zależnego od sumy wykładników), oraz wariant „idealnej linii" z rozkładem normalnym wzdłuż jednego wymiaru, gdzie błąd estymacji zależy od położenia punktu (rośnie z kwadratem odległości od środka), ale jego wartość oczekiwana może być kontrolowana.

Ostatnia duża część rozdziału bada zachowanie LIDL w różnych sytuacjach, przy czym gęstość po zaszumieniu jest liczona analitycznie lub numerycznie (bez błędu uczenia modelu gęstości). Dla rozkładu jednostajnego na odcinku autor pokazuje typowy efekt brzegowy: w pobliżu końców nośnika pojawia się błąd estymacji, a obszar problematyczny skaluje się z σ. Dla rozkł.Gaussa na linii potwierdza się narastanie błędu wraz z oddalaniem się od środka (spójne z analizą „idealnej linii"), a następnie omawia się wpływ krzywizny na przykładzie rozkładu jednostajnego na okręgu, gdzie dla zbyt dużych σ estymata zaczyna „spadać" w stronę zera (interpretowane jako utrata rozdzielczości i uśrednianie struktury przez zbyt silne rozmycie). Analizowany jest też przypadek dwóch bliskich, równoległych składowych (sąsiednich komponentów rozmaitości): gdy σ staje się porównywalne z ich odległością, pojawia się dodatnie obciążenie, a przy jeszcze większej skali algorytm zaczyna „widzieć" obie składowe jako jedną strukturę. Rozdział kończy się krótką informacją, że na dodatkowych syntetycznych zbiorach przy idealnych gęstościach uzyskiwano praktycznie dokładne estymacje, oraz zebraniem wniosków.

Rozdział 3 proponuje zastosowanie procesu Wienera (dyfuzji Browna) jako wspólny język do analizy współczesnych estymatorów lokalnego wymiaru wewnętrznego, w których pierwszym krokiem jest dodawanie kontrolowanego szumu Gaussa o wariancji zależnej od „czasu" t. Autor pokazuje, że taka perturbacja danych jest równoważna obserwowaniu ewolucji gęstości w równaniu dyfuzji (drugie prawo Ficka - równanie ciepła), co pozwala przeformułować obiekty używane w algorytmach LID tak, aby zamiast pochodnych po czasie (po skali szumu) pojawiły się pochodne przestrzenne, czyli operatory różniczkowe działające na gęstości w przestrzeni danych. W efekcie rozdział unifikuje etap zaszumiania wielu metod (m.in. LIDL, FLIPD oraz podejść opartych o dyfuzję i modele generatywne) i dostarcza narzędzi do analitycznego ich wyjaśniania. Autor formalizuje dyfuzję poprzez splot początkowej miary prawdopodobieństwa p_S (wspieranej na rozmaitości) z jądrem Gaussa o kowariancji tI, co daje gęstość spełniającą równanie ciepła. Aby ograniczyć złożoność krzywizny, analiza w tym rozdziale koncentruje się na rozmaitościach płaskich: bez straty ogólności rozmaitość jest utożsamiana z D-wymiarowa przestrzeń euklidesowa zanurzoną jako pierwszy czynnik w rozkładzie $\mathbb{R}^D = \mathbb{R}^d \times \mathbb{R}^{D-d}$, co umożliwia rozdzielenie zmiennych i jawne obliczenia Laplasjanu dla gęstości po dyfuzji. Kluczowy wynik to wyprowadzenie wzoru na Laplasjan gęstości zdyfundowanej zarówno poza rozmaitością, jak i na niej, co staje się później „łącznikiem" między dynamiką po skali szumu a geometrią (wymiarem) nośnika danych.

Następnie, w sekcji 3.3, rozdział przechodzi od opisu PDE do estymacji LID, reinterpretując LIDL jako aproksymację nachylenia krzywej $(\log\, t, \log\, \rho_t(x))$ dla $t \to 0$. Zamiast estymować to nachylenie regresją po wielu t, autor wprowadza „reparametryzowane nachylenie", które dzięki równaniu ciepła stanowi lokalny (różniczkowy) odpowiednik nachylenia używanego w LIDL/FLIPD. Pokazane jest, że dla funkcji o odpowiednim zachowaniu przy $t \to 0$ różne sposoby definiowania asymptotycznego nachylenia (przez $\log\, \rho_t$ względem $\log\, t$ albo przez

granicę ilorazu pochodnych) są równoważne, co uzasadnia zamianę regresji na analizę pochodnej i otwiera drogę do obliczeń zamkniętych postaci. Co istotne, rozdział podaje elegancką formułę na $\tau_t(x)$ w terminach gęstości na $\mathbb{R}^d$ po dyfuzji (czyli ρ_t rozpisanej na czynnik „styczny"), co pozwala explicite badać wpływ niejednorodności gęstości oraz położenia punktu.

W sekcji 3.4 autor ilustruje powyższy punkt widzenia na zestawie kanonicznych przykładów, które odsłaniają źródła obciążeń estymatora w skończonym t. Dla „stałej" gęstości na $\mathbb{R}^d$ (traktowanej jako przypadek teoretyczny) otrzymuje się brak obciążenia: $\tau_t(x) = d - D$ niezależnie od t, co pokazuje, że w idealnie jednorodnym wnętrzu płaskiej rozmaitości mechanizm LIDL działa perfekcyjnie. Dla rozkładów normalnych rozdział wyprowadza postacie, które wyjaśniają dwa typowe zjawiska z eksperymentów: (i) „paraboliczny" wzrost dodatniego obciążenia w rejonach małego prawdopodobieństwa (duże $\|x\|$ lub skrajne współrzędne) oraz (ii) „schodkowe" zachowanie przy anizotropii (różne wariancje w różnych kierunkach), gdzie zmiana skali t powoduje, że kolejne kierunki zaczynają dominować w wyrażeniu na nachylenie. Dla rozkładu jednostajnego na przedziale i hipersześcianie analiza redukuje problem do wersji jednowymiarowej i pokazuje, jak pojawiają się efekty brzegowe (w pobliżu granicy nośnika), co formalnie uzasadnia obserwowane w rozdziale 2 błędy LIDL przy punktach blisko brzegów.

Rozdział omawia też przypadki wieloskładowe: unie rozmaitości równoległych, przecinających się oraz ogólnie wypukłe kombinacje rozkładów. Kluczowy wniosek ilościowy jest taki, że wpływ odległych komponentów zanika wykładniczo wraz z kwadratem separacji w skali t, co pozwala precyzyjnie opisać, kiedy komponent „zaczyna przeszkadzać" w estymacji (gdy t jest na tyle duże, że dyfuzja „przenosi masę" między komponentami). Dla przecięć rozmaitości rozdział pokazuje, że estymowane nachylenie znajduje się pomiędzy wymiarami składowych, a o jego wartości decydują wagi asymptotyczne (czyli relatywny wkład gęstości po dyfuzji z każdej składowej w punkcie przecięcia).

W konkluzji autor podkreśla, że ujęcie w kategoriach procesu Wienera i równania ciepła dostarcza spójnego aparatu do analizy wielu współczesnych estymatorów LID: zamienia problem „jak zmienia się logarytm wiarygodności z t" na problem ilorazów pochodnych przestrzennych, które mogą być pozyskiwane z dowolnego modelu gęstości/score. Jednocześnie, podejście to jasno identyfikuje tryby działania i źródła błędów przy skończonym t: niejednorodność gęstości, punkty z obszarów małego prawdopodobieństwa, brzegi nośnika oraz interferencję między składowymi w uniach rozmaitości. Rozdział sugeruje też, że analogiczne narzędzia mogą zostać rozszerzone na przypadki zakrzywione (zależność od geometrii) i bardziej złożone układy wieloskładowe, stanowiąc fundament pod korekty obciążeń i warianty estymatorów LID, które bardziej uwzględniają PDE.

Rozdział 4 przedstawia empiryczne porównanie LIDL z klasycznymi estymatorami lokalnego wymiaru wewnętrznego na syntetycznych zbiorach danych o znanej wymiarowości, z naciskiem na trzy aspekty: skalowalność do bardzo wysokich wymiarów, zachowanie na rozmaitościach wieloskalowych (silnie anizotropowych) oraz działanie na rozmaitościach zakrzywionych i będących unią składowych o różnych wymiarach. Autor podkreśla, że analiza jest prowadzona na danych syntetycznych, ponieważ tylko tam dostępne są wartości referencyjne (*ground truth*), a wyniki są raportowane jako względne obciążenie (*relative bias*), relative MAE oraz wariancja estymat w wielu powtórzeniach losowania próby.

Najpierw, rozdział krótko przypomina rolę modeli normalizing flows jako estymatorów gęstości, które są kluczowe dla LIDL w praktyce: uczone są odwracalne transformacje mapujące prosty rozkład bazowy na rozkład danych, a trening polega na maksymalizacji log-wiarygodności. W eksperymentach wykorzystywane są trzy architektury przepływów (MAF, RQ-NSF, Glow). Jednocześnie w części porównawczej jako „klasyczne" punkty

odniesienia autor przyjmuje zestaw metod dostępnych w bibliotece scikit-dimension, zaznaczając, że niektóre algorytmy (w szczególności FisherS i DANCo) zostały wykluczone z testów w wysokich wymiarach ze względu na problemy z pamięcią lub nieakceptowalny czas działania.

W sekcji 4.2 autor omawia protokół ewaluacji i wyniki trzech bloków eksperymentów. Po pierwsze, w teście skalowalności uruchomiono metody na rozkładach jednostajnych i normalnych o wymiarowości sięgającej 4000 (przy stałej liczności próby rzędu 10 tys. punktów), a następnie oceniano, jak błąd rośnie wraz z wymiarem. Z tabel i wykresów wynika, że LIDL utrzymuje małe względne błędy nawet dla wymiarów liczonych w tysiącach (rzędu kilku procent), podczas gdy większość klasycznych estymatorów zaczyna wyraźnie degradować już powyżej około 100 wymiarów, a przy około tysiącach wymiarów ich estymaty stają się skrajnie zaniżone lub niestabilne; na tle metod klasycznych najlepiej wypada ESS, ale i ona pozostaje słabsza od LIDL w najwyższych wymiarach. Autor ilustruje to m.in. na przykładzie rozkładu jednostajnego na hiperześcianie, gdzie na wykresie zestawia się estymaty wielu algorytmów w funkcji prawdziwego d, pokazując, że krzywa LIDL pozostaje blisko przekątnej także dla bardzo dużych d.

Po drugie, rozdział analizuje wpływ anizotropii i wieloskalowości (multiscale manifolds), gdzie dane mają „cienkie” i „grube” kierunki zmienności oraz gdzie klasyczne metody oparte o sąsiedztwa mogą naruszać założenia o lokalnej jednorodności i w efekcie ujawniać zależność estymaty od rozmiaru próby. Autor pokazuje, że wiele estymatorów daje inne wyniki dla różnych liczności zbioru (co jest problematyczne interpretacyjnie, bo nie wiadomo, czy obserwowany „dryf” to jeszcze błąd skończonej próby, czy już trwałe obciążenie), natomiast LIDL dzięki jawnemu parametrowi skali σ utrzymuje stabilne estymaty w funkcji liczebności danych w dwóch rozważanych ustawieniach skali. Jako metody relatywnie stabilne w tym scenariuszu wskazywane są też niektóre podejścia sąsiedztwowe (np. KNN, MiND-ML), ale ogólna konkluzja jest taka, że możliwość świadomego ustawienia skali w LIDL stanowi praktyczną przewagę przy danych wieloskalowych.

Po trzecie, rozdział przechodzi do rozmaitości zakrzywionych i unii rozmaitości: testowane są klasyczne benchmarki, a także tzw. zbiór „lollipop”, który składa się ze składowych o wymiarze 0, 1 i 2. Na rozmaitościach zakrzywionych LIDL osiąga „przyzwoite” wyniki (małe względne obciążenie i MAE), choć w niektórych przypadkach proste metody lokalne, takie jak lPCA i ESS, potrafią być jeszcze dokładniejsze, co autor wiąże z faktem, że w tym eksperymencie gęstość jest estymowana niedoskonale przez model przepływowy, podczas gdy w rozdziale 2 pokazano niemal perfekcyjne działanie LIDL przy „idealnych” gęstościach. Najbardziej diagnostyczny jest jednak eksperyment „lollipop”: o ile na częściach 1D i 2D wiele metod radzi sobie dobrze, o tyle komponent 0-wymiarowy (wielokrotne kopie tego samego punktu) powoduje, że prawie wszystkie klasyczne algorytmy nie potrafią zbiec do sensownej estymaty; poprawa zbieżności po dodaniu minimalnego jitteru szumem prowadzi z kolei do błędnych estymat bliskich 2. Autor podkreśla, że LIDL potrafi poprawnie odtworzyć wymiar komponentu 0D, a dodatkowo LIDL dzięki doborowi skali działania pozostaje poprawny także po zaszumieniu, ponieważ może „zignorować” sztucznie wprowadzoną mikroskalową grubość danych.

W podsumowaniu autor stwierdza, że LIDL jest konkurencyjny względem klasycznych metod na typowych benchmarkach, a jego kluczową przewagą jest skalowalność do tysięcy wymiarów i możliwość jawnego sterowania skalą, co stabilizuje zachowanie na danych anizotropowych i wieloskalowych. Autor zaznacza też, że ponieważ LIDL estymuje nachylenie zależności log-wiarygodności od skali, systematyczny błąd (bias) w samej log-wiarygodności w dużej mierze przesuwa wyraz wolny regresji, natomiast szum w estymacji gęstości zwiększa przede wszystkim wariancję, co sugeruje, że dalsze usprawnienia mogą wynikać zarówno z mocniejszych modeli gęstości, jak i z bardziej odpornych procedur regresji.

Rozdział 5 prezentuje eksperymentalną ocenę LIDL na rzeczywistych zbiorach obrazów, pokazując, że estymaty lokalnego wymiaru wewnętrznego liczone per-próbka mają sensowną interpretację jakościową, są wrażliwe na dobór skali działania σ (w tym na szum dekwantyzacji), mogą być stabilizowane przez uśrednianie po wielu modelach gęstości, a także korelują z zachowaniem innych modeli uczenia maszynowego (autoenkoderów i klasyfikatorów). Autor organizuje rozdział w cztery bloki: (i) analiza jakościowa i „strukturalna" na zbiorach MNIST, FMNIST i CelebA, (ii) badanie zakresu roboczego parametru σ oraz wpływu dekwantyzacji, (iii) redukcja błędu przez zwiększanie liczby modeli n w estymacji, oraz (iv) powiązanie LID z miarami trudności zadania w rekonstrukcji i klasyfikacji.

W sekcji 5.1 LIDL uruchamiany jest na trzech popularnych zbiorach: MNIST i FMNIST (oba o wymiarze rzędu 1 tys. pikseli) oraz CelebA (rzędu 12 tys. pikseli), przy czym jako estymator gęstości zastosowano model Glow. Dane są sortowane według estymowanego LID, a autor zauważa, że przykłady wizualnie bardziej złożone (np. bardziej „bogate" w szczegóły, mniej jednoznaczne, z większą zmiennością kształtu/tekstury) systematycznie trafiają na koniec rankingu z wyższym LID, podczas gdy obrazy proste i „łatwe" mają LID niższy. Dodatkowo dla MNIST i FMNIST pokazane są empiryczne dystrybuanty (CDF) estymat per klasa, co ma sugerować, że klasy różnią się rozkładem lokalnej złożoności: niektóre cyfry lub kategorie ubrań tworzą bardziej „skomplikowane" podzbiory niż inne, a LIDL wychwytuje te różnice w sposób uporządkowany.

Sekcja 5.2 koncentruje się na doborze σ jako parametru skali i na tym, że w obrazie (jako domenie silnie skwantowanej) pojawiają się praktyczne ograniczenia wynikające z dekwantyzacji i „cienkości" rozmaitości danych. Autor najpierw pokazuje na kontrolowanym przykładzie wieloskalowego rozkładu jednostajnego w 4 wymiarach, że LIDL zachowuje się zgodnie z interpretacją skali: gdy σ przekracza pewne wewnętrzne „grubości" danych, estymator zaczyna ignorować odpowiadające im wymiary (uzyskuje się charakterystyczne schodkowe przejścia podobne do tych z wcześniejszych rozdziałów). Następnie przenosi tę obserwację na FMNIST: przy zbyt dużej σ pewne podzbiory obrazów (w szczególności „ciemne ubrania" opisane jako cienka rozmaitość) zostają sztucznie „spłaszczone" i otrzymują estymaty bliskie zera, co ilustruje, że zbyt agresywna skala może skleić lub zignorować realną zmienność danych i prowadzić do niepoprawnych wniosków. Rozdział podaje też praktyczną obserwację stabilności rankingów: porównując estymaty uzyskane na dwóch rozłącznych zestawach wartości σ, relacje porządku (kto ma wyższy LID od kogo) są zachowane dla większości par, choć same wartości bezwzględne różnią się przeciętnie o kilkanaście pozycji w rankingu.

W tej samej sekcji omawiany jest wpływ dekwantyzacji (dodania małego szumu jednostajnego do pikseli) na estymaty w funkcji σ. Autor prezentuje krzywe średnich estymat per klasa dla MNIST i FMNIST na szerokim zakresie σ (od bardzo małych do relatywnie dużych), dla danych „oryginalnych" (skwantowanych) oraz po dekwantyzacji, zaznaczając teoretyczny próg związany z poziomem dekwantyzacyjnego szumu. Wnioski są zgodne z intuicją: przy bardzo małej σ estymata jest zdominowana przez rozdzielczość/kwantyzację i zbliża się do wymiaru przestrzeni (ambient dimension), natomiast przy bardzo dużej σ estymaty spadają w stronę zera; w pewnym zakresie nieco powyżej progu dekwantyzacji krzywe dla danych skwantowanych i dekwantyzowanych zaczynają się ze sobą zgadzać, co sugeruje „właściwe" okno skali dla obrazów. Sekcja 5.3 pokazuje prostą technikę redukcji błędu estymacji wynikającego z niedoskonałego modelowania gęstości: zamiast opierać się na kilku poziomach szumu, zwiększa się liczbę dopasowywanych modeli w tym samym przedziale σ (czyli „zagęszcza" próbkowanie skali i uśrednia informację o nachyleniu). Na przykładzie 10-wymiarowego rozkł. Gaussa zanurzonego w 20 wymiarach autor demonstruje monotoniczny spadek MSE estymaty

wraz ze wzrostem n, co interpretuje jako klasyczny efekt ensemblingu/średniowania, gdzie losowe błędy estymacji gęstości mniej zaburzają końcową regresję.

W sekcji 5.4 rozdział wiąże LID z wydajnością modeli ML, argumentując, że lokalna złożoność danych mierzona przez LIDL może służyć jako sygnał „trudności" próbki. W eksperymencie z autoenkoderem wariacyjnym (VAE) trenowanym na MNIST autor pokazuje silną dodatnią korelację (Pearson ok. 0.7) między estymowanym LID a błędem rekonstrukcji MSE, dla dwóch rozmiarów przestrzeni ukrytej (50 i 150), co sugeruje niemal liniową zależność: im wyższy LID próbki, tym trudniej ją dobrze odtworzyć. Analogicznie dla klasyfikacji na MNIST i FMNIST autor obserwuje silną ujemną zależność: średnia dokładność klasyfikatora spada wraz ze wzrostem LID, i co ważne, trend utrzymuje się nie tylko globalnie, ale także wewnątrz większości klas, co wzmacnia interpretację LID jako miary lokalnej trudności/niepewności predykcji. Na tej podstawie rozdział sugeruje potencjalne zastosowania LID jako heurystyki w uczeniu półnadzorowanym, aktywnym, curriculum learning oraz w estymacji niepewności, z zastrzeżeniem, że korelacje nie dowodzą związku przyczynowego, a dobór σ i jakość estymatora gęstości mogą istotnie modulować obserwowany efekt.

Rozdział 6 podejmuje problem testowania współczesnych (w szczególności neuronowych) algorytmów estymacji LID i argumentuje, że dotychczasowe praktyki ewaluacyjne są niewystarczające: z jednej strony dominują proste zbiory analityczne z idealnie znanym LID, które nie odzwierciedlają „złośliwych" własności danych rzeczywistych, a z drugiej strony popularne są testy na danych domenowych (np. obrazach), gdzie prawdziwa wartość LID jest nieznana, więc trudno rozstrzygnąć, czy algorytm mierzy geometrię danych, czy raczej wykorzystuje własności architektury modelu. Autor proponuje więc kompromis przez konstruowanie kontrolowanych transformacji danych, które (i) zachowują LID albo (ii) zmieniają je w przewidywalny sposób, dzięki czemu nawet na danych o nieznanym LID można oceniać stabilność i poprawność algorytmu poprzez porównanie estymat przed i po transformacji. Na tej bazie w rozdziale budowany jest zestaw nowych benchmarków badających konkretne cechy rozmaitości (niejednorodność gęstości, krzywiznę, efekty brzegowe, „cienkość", bliskość składowych, zależność od rozmiaru próby), a następnie porównuje szeroki zestaw algorytmów: jako klasyczny punkt odniesienia ESS oraz metody neuronowe LIDL, FLIPD i NB, podając zarówno mapy estymat, jak i tabele zbiorcze.

W sekcji 6.2 rozdział formalizuje zestaw narzędzi do tworzenia domenowych zbiorów testowych. Kluczową ideą jest Inverse Domain Representation (IDR), czyli osadzanie znanych analitycznie rozmaitości w przestrzeni danych domenowych tak, by powstające próbki „wyglądały" jak dane z danej domeny (np. jak obrazy). Drugim elementem jest Monotonic Embedding (ME), czyli gładkie, monotoniczne przekształcenia współrzędnych (w obrazach: deterministyczne przekształcenia intensywności pikseli, np. potęgowe), które zachowują topologię i w sensie różniczkowym nie zmieniają wymiaru, ale mogą silnie zmieniać „trudność" dla modeli z określonymi biasami architektonicznymi. Trzecia technika to Ambient Space Extension (ASE), która zwiększa wymiar otoczenia przez deterministyczne rozszerzenia (np. upscaling obrazów), formalnie nie dodając nowych stopni swobody, ale potencjalnie zmieniając sposób, w jaki sieci konwolucyjne reprezentują dane. Czwarta to Auxiliary Dimension Injection (ADI), czyli kontrolowane dodawanie wymiarów pomocniczych (np. przez parametryczne transformacje, warianty jasności lub łączenie próbek), które w zamierzeniu zwiększa wymiar w sposób znany lub przynajmniej porównywalny między wersjami danych. Piątym składnikiem jest Manifold Synthesis (MS), służący do generowania „realistycznych" zbiorów o znanym LID przez deterministyczne, ciągłe i odwracalne przekształcenia parametryzowanych rozmaitości do przestrzeni domenowej. W sekcji 6.3 autor wykorzystuje powyższe narzędzia do zaprojektowania szeregu testów diagnostycznych i ilustruje na nich typowe tryby awarii algorytmów. Dla niejednorodnych gęstości i prostych rozmaitości (np. „gaussian blobs" i sfery) pokazuje się, że część metod zachowuje poprawność

średnich estymat, ale inne mają tendencje do systematycznego przeszacowania lub dużej wariancji, a wyniki zależą od tego, czy test jest wykonywany w „oryginalnej" przestrzeni euklidesowej czy po transformacji IDR do domeny obrazowej. Dla efektów brzegowych na danych jednostajnych rozdział buduje miarę „ile współrzędnych jest blisko brzegu" i pokazuje, że estymaty zmieniają się istotnie, gdy punkt leży w pobliżu wielu krawędzi (co w praktyce odpowiada efektom podobnym do tych z rozdziałów 2–3, ale teraz diagnozowanym narzędziami benchmarkowymi). Dla krzywizny i geometrii bardziej „zawiniętej" autor prezentuje benchmarki typu „moon" czy „spiral", gdzie widać, że część algorytmów ma duży rozrzut estymat zależny od położenia na rozmaitości, a inne dają bardziej stabilny obraz, lecz często kosztem obciążenia.

Osobny, bardzo istotny blok testów dotyczy „cienkich" i „bliskich" rozmaitości, gdzie w skończonej próbie punkty z różnych składowych mogą być bliżej siebie w przestrzeni otoczenia niż sąsiedzi na tej samej składowej, co prowadzi do fałszywego „zwiększenia" wymiaru. Autor konstruuje tu m.in. zestawy Funnel i Spiral (po IDR), w których lokalnie promień lub separacja zmienia się wzdłuż rozmaitości: w szerokich fragmentach poprawna odpowiedź powinna być 2D, w wąskich fragmentach algorytmy zaczynają „widzieć" wpływ trzeciego wymiaru (bo składowe stają się zbyt blisko), a w granicznym przypadku struktura zaczyna przypominać 1D, co daje charakterystyczny profil estymaty wzdłuż parametru. Następnie rozdział bada inwariancję względem uprzedzeń architektonicznych: obserwuje, że LIDL i FLIPD (które w praktyce opierają się o modele generatywne z określonymi architekturami, np. konwolucyjnymi) mogą dawać gorsze wyniki na tej samej „matematycznej" rozmaitości, gdy jest ona przeniesiona do innej reprezentacji domenowej (IDR), oraz że dobór typu sieci (np. Glow/U-Net vs MAF/MLP) może istotnie zmienić charakter błędu. Kolejny problem to zależność estymat od rozmiaru próby: na FMNIST autor tworzy serie zbiorów treningowych o różnych licznościach i pokazuje, że część metod wykazuje wyraźną zależność średniej estymaty od n, co jest niepokojące, bo sugeruje „bias zależny od danych" (nie tylko malejący błąd wariancyjny), i utrudnia ocenę, czy osiągnięto już reżim wiarygodnych estymat.

Kluczową częścią rozdziału są też transformacje real-world (RWD) na FMNIST, gdzie prawdziwy LID nie jest znany, ale można oceniać spójność zachowania metod. W eksperymencie ADI autor dodaje „ramkę" przez mirror padding oraz w kontrolowany sposób wprowadza 0, 4 lub 8 „dodatkowych wymiarów" przez losowe zmiany jasności wybranych odbić; następnie kalibruje wyniki odejmując liczbę dodanych wymiarów, tak aby idealnie punkty leżały na prostej $y=x$. Wyniki pokazują, że niektóre algorytmy utrzymują się blisko tej relacji tylko dla próbek o niskiej estymowanej złożoności, a dla próbek o wysokiej estymacie bazowej pojawiają się duże odchylenia, co interpretuje się jako utratę kontroli nad zmianą LID po transformacji.

Końcowa konkluzja rozdziału jest taka, że sensowna ewaluacja estymatorów LID powinna (i) testować konkretne „cechy rozmaitości" w izolacji, (ii) obejmować transformacje domenowe i testy stabilności przed/po transformacji, oraz (iii) używać prawdziwych danych o znanym LID, ponieważ bez tego łatwo o myląco optymistyczne wnioski z klasycznych benchmarków.

Rozdział 7 podsumowuje wkład pracy w trzech powiązanych wątkach: wprowadzeniu LIDL jako teoretycznie ugruntowanego estymatora lokalnego wymiaru wewnętrznego opartego o neuronowe modele gęstości, zaproponowaniu perspektywy procesu Wienera/równania ciepła jako wspólnego języka dla nowoczesnych metod (likelihood-, Jacobian- i score-based) oraz zaprojektowaniu transformacyjnego zestawu benchmarków (IDR, ME, ASE, ADI, MS), W rozdziale 7 autor formułuje również przyszłe kierunki badań, które obejmują przede wszystkim rozwój wieloskalowych, diagnostycznych procedur wyboru skali szumu, korekty brzegowo-krzywiznowe oparte na analizie procesu Wienera oraz wykorzystanie mocniejszych modeli gęstości (głównie dyfuzyjnych), aby zmniejszyć bias LIDL w realistycznych warunkach. innym obszarem jest dalsze rozwijanie i ustandaryzowanie transformacyjnych

benchmarków (IDR, ME, ASE, ADI, MS) tak, by tworzyć dziedzinowo-świadome testy dla różnych typów danych (obrazy, audio, EEG itd.) oraz porównywać i łączyć różne klasy estymatorów LID w spójnych wieloskalowych eksperymentach.

Dalej następują podziękowania, bibliografia oraz dodatek domykający część teoretyczną krótkimi dowodami użytymi w rozdziale o zastosowaniu procesu Wienera oraz podaje szczegóły konfiguracji eksperymentów (zarówno dla LIDL na obrazach i danych syntetycznych, jak i dla szerokiego porównania metod w rozdziale 6), a także zawiera dodatkowe wykresy i tabele.

3. Ocena rozprawy

W ramach rozprawy doktorskiej Pan mgr Piotr Tempczyk zaproponował trzy oryginalne elementy, które razem przesuwają estymację lokalnego wymiaru wewnętrznego (LID) w stronę metod skalowalnych i testowalnych na danych wysokowymiarowych: (1) nowy estymator LIDL oparty o neuronowe modele gęstości (normalizing flows) i analizę LID na podstawie nachylenia zależności logarytmu wiarygodności od skali szumu, (2) ujednolicającą metodę opartą na procesie Wienera/równaniu ciepła, która pozwala analizować i porównywać współczesne metody LID (likelihood-, Jacobian- i score-based) w jednym formalizmie oraz wyprowadzać wyrażenia/intuicje o biasach, i (3) transformacyjny framework benchmarków (IDR, ME, ASE, ADI, MS), który umożliwia „stres-testy" metod na danych domenowych poprzez kontrolowane transformacje zachowujące lub przewidywalnie modyfikujące LID.

Rozprawa doktorska uwidacznia wysoką ogólną wiedzę teoretyczną i praktyczną oraz umiejętność prowadzenia pracy naukowej przez mgr Piotra Tempczyka. Opracował wprowadzenie do tematyki i przegląd literatury na temat poszczególnych obszarów swoich badań oraz zadbał o popularyzację wyników swoich badań publikując liczne prace naukowe w materiałach uznanych konferencji (m.in. AAAI, ICML, ICLR, NeurIPS) oraz w czasopiśmie IEEE Access. Rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego. Zaproponowane metody mają duże znaczenie dla rozwoju uczenia maszynowego, zarówno teoretyczne, jak i aplikacyjne.

Niezależnie od mojej dobrej oceny pracy, nasunęły mi się pytania i uwagi:

- Jaka jest procedura doboru σ w metodzie LIDL w realnych danych bez ground truth i jak jest ona uzasadniona?
- W jakim stopniu błąd LIDL pochodzi z samej idei nachylenia (slope), a w jakim z niedokładności/niekalibracji przepływu (lub innego modelu gęstości), i czy da się to rozdzielić w sposób systematyczny?
- Ile realnie kosztuje (czas/GPU) wytrenowanie wielu modeli dla wielu σ, zwłaszcza w porównaniu do NB/FLIPD/ESS, i czy metoda jest praktyczna w rzeczywistych zastosowaniach?
- Czy proponowane benchmarki nie faworyzują metod podobnych do tych użytych do ich konstrukcji (np. domena obrazowa + konkretne modele generatywne)?

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że rozprawa doktorska *Estimating local intrinsic dimension via density estimation* prezentuje oryginalne rezultaty stanowiące rozwiązanie problemu naukowego oraz wkład w rozwój dyscypliny naukowej informatyka w dziedzinie nauk ścisłych i przyrodniczych. Pan Piotr Tempczyk wykazał się umiejętnością samodzielnej

pracy badawczej, znajomością literatury światowej i wiedzą w zakresie uczenia maszynowego. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata. Recenzowana praca spełnia wymagania Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2024 r. poz. 1571 z późn. zm.) w dyscyplinie naukowej informatyka w dziedzinie nauk ścisłych i przyrodniczych. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postępowania doktorskiego. Ponadto ze względu na ponadprzeciętny poziom rozprawy oraz fakt opublikowania prac związanych bezpośrednio z tematyką rozprawy w materiałach najlepszych konferencji klasy A* w tym obszarze (AAAI, ICML, ICLR), wnioskuję o wyróżnienie pracy.

Elektronicznie podpisany przez:

Rafał Scherer

Data:
2026-2-18 23:11:44