Математички факултет,
Универзитет у Београду

Мастер рад

Основе тополошке анализе података са применом

Аутор:
Вилдана Бакаревић

Ментор:
др Бојана Милошевић

23. август 2021.

Садржај

Списак слика

Глава 1

Увод

1.1 Кратка историја

Као директна последица широке примене и значаја рачунарске топологије јавља се нова хомолошка теорија - перзистентна хомологија. Она се као званична област први пут појавила недавно. Можемо рећи да је у односу на развој топологије кроз историју, као и математике генерално, ова област веома млада, али брзо се развија.

Оснивачи концепта перзистентне хомологије појављивали су се постепено током блиске прошлости, а у последњих неколико година ова област доживљава велику експанзију. Патрицио Фросини[1] је 1990. године увео функцију која је еквивалентна 0-тој перзистентној хомологији. Скоро деценију касније, Ванеса Робинс[2] проучавала је слике хомолошких хоморфизама изазваних инклузијом. Коначно, убрзо након тога, Херберт Еделсбрунер[3] уводи потпун концепт перзистентне хомологије заједно са алгоритмом и његову визуализацију - перзистентни дијаграм. Гунар Карлсон[4] је преформулисао почетну дефиницију и дао еквивалентну методу визуализације под називом перзистентни бар-кодови. У алгебарској топологији перзистентна хомологија настала је радом Сергеја Бараникова[5] на Морсеовој теорији. Још једна особа, иако не из данашње генерације научника који се баве перзистентном хомологијом, ипак заслужује помен у овом раду. То је Леополд Виеторис[6] који је први дефинисао појам Виеторис-Рипс комплекса (назива се и само Виеторис комплекс или само Рипс комплекс)

[1] Patrizio Frosini - италијански математичар
[2] Vanessa Robins - аустралијска математичарка
[3] Herbert Edelsbrunner - аустријски математичар
[4] Gunnar Carlsson - амерички математичар
[5] Serguei Barannikov - руски математичар
[6] Leopold Vietoris - немачки математичар

чак давне 1927.године. Он се још тада бавио конструкцијом простора виских димензија индуктивно из простора нижих димензија. Тек 2010.године, Афра Зомородијан[7] наставља да се бави том идејом и предлаже оптималне алгоритме конструкције.

1.2 Увод у тополошку анализу података

Важна карактеристика савремене науке и машинског учења је да се подаци различитих врста прикупљају и доступни су у абнормалним количинама. То је делом због нових експерименталних метода, а делом због повећања доступности рачунарске технологије са високим перформансама. Такође је јасно да се природа података до којих долазимо знатно разликује. На пример, сада је често случај да добијамо податке у облику врло дугих вектора, где се све координате, осим неколико, испостављају ирелевантним за питања која нас занимају, а не морамо нужно знати које су то координате оне нама занимљиве. Важна чињеница је и да су подаци често врло високих димензија, што озбиљно ограничава нашу способност да их визуализујемо. Добијени подаци, такође, имају много више шума него раније и садрже више недостајућих података. Наша способност да анализирамо ове податке, како у погледу количине, тако и у погледу природе података, очигледно не иде у корак са подацима које добијамо.
У овом раду ћемо размотрити како се геометрија и топологија могу применити да би се дали корисни доприноси анализи различитих врста података. Геометрија и топологија су врло природни алати за примену у анализи, јер геометрију користимо као грану математике која се заснива на растојањима између тачака у подацима, а топологију користимо као грану математике која се заснива на облику тих података. Математичка област која је развијена за инкорпорирање геометријских и тополошких техника бави се облацима тачака (енг. point clouds), тј. коначним скуповима тачака на којима је дефинисана одређена метрика. Тада се облаци тачака сматрају коначним узорцима тачака узоркованих из геометријског објекта, понекад са додатим шумом.

Споменимо сада најважнија својства, као и проблеме, са којима се сусрећемо приликом примене геометријских метода током анализе података:

- Потребне су квалитативне информације: Један важан циљ анализе података је омогућити кориснику да стекне знање о подацима, односно

[7] Afra Zomorodian - амерички математичар

да разуме како су они у суштини организовани. На пример, ако радимо класификацију, пре свега је битно упознати се са квалитативним особинама класа и њиховим међусобним односима, а потом када је то утврђено развити квантитативне методе за њихово разликовање.

- Метрика није увек теоријски оправдана: У физици, проучавани феномени су често подржани прецизним теоријама које тачно говоре коју метрику треба користити. У другим проблемима, као на пример биолошким, то није веома јасно објашњено. Појмови раздаљине конструишу се често помоћу неких интуитивних мера сличности, али још увек није јасно колики значај треба придати стварним удаљеностима.

- Координате нису природне: Иако често примамо податке у облику вектора реалних бројева, избор координатног система некада није прикладан подацима. Стога, неприкладан избор координата значајно може утицати на погрешну интерпретацију закључака. Један пример је РГБ (енг. Red, Green, Blue) модел боја код кога се комбинацијом црвене, зелене и плаве боје добијају све остале. Те три основне боје су представљене са 256 нивоа на скали од 0 до 255 где сваки ниво представља јачину (осветљеност) те боје. Видимо да у овом случају нема смисла користити еуклидски простор и метрику јер нам не даје смислен осећај за растојање тачака тј. боја.

- Сажетак је вреднији од избора појединачних параметара: Дајемо пример зашто је, генерално, много продуктивније да се развију технике којима се понашање у подацима под променом одређених параметара може ефикасно репрезентовати на целом домену. Један од метода кластеровања је такозвани алгоритам са једноструким повезивањем [1] (енг. single linkage clustering) , агломеративна техника током које се кластери спајају по принципу минималне удаљености њихових најближих чланова. У том случају је много информативније, од самог избора параметра, направити цео дендрограм који даље анализира спајање кластера под свим могућим вредностима удаљености кластера.

У овом раду бавићемо се методама које проучавају горе споменута својства и проблеме. Главна идеја јесте да се користе методе инспирисане топологијом. За сваку од горњих тачака описујемо зашто су тополошке методе право решење:

- Топологија је управо она грана математике која се бави квалитативним геометријским информацијама, а једна од њених важнијих подобласти је алгебарска топологија која повезује алгебарске структуре са тополошким просторима. То укључује проучавање повезаних компоненти простора, али и уопштено проучавање повезаности целог простора у различитим димензијама. Стога природно следи њена употреба у проучавању квалитативних особина података.

- Топологија проучава геометријска својства на начин који је много мање осетљив на избор метрике од директних геометријских метода. У ствари, топологија занемарује квантитативне вредности функција раздаљине и замењује их појмом близине. Ова неосетљивост на метрику корисна је у ситуацијама када имамо само грубу, скоро интуитивну, представу растојања.

- Топологија, заправо, и проучава само својства геометријских објеката која не зависе од изабраних координата, већ од суштински својстава објеката.

- Основна идеја анализе целих домена параметара јесте компарација геометријских објеката конструисаних помоћу различитих вредности параметара из домена. Посматрајмо то као скалу параметара на којој проучавамо односе између параметарских објеката добијених из података. Ти односи укључују непрекидна пресликавања између различитих геометријских објеката и тако долазимо до појма функционалности, тј. идеје да се инваријантност чува не само у посматраним објектима већ и пресликавањима између тих објеката. Функционалност је у алгебарској топологији важна јер омогућава учење глобалних својстава из информација на локалном нивоу, а та особина је једна од занимљивијих у применама математике. Штавише, познато је да се већина информација о тополошким просторима може добити апроксимацијом на локалном нивоу. Једна од њих је и симплицијска апроксимација простора помоћу дискретних скупова, о чему ће бити речи у даљем раду.

Последња тачка је од великог значаја и функционалност треба да игра важну улогу у проучавању облака тачака. Такође, она је битна јер омогућава да уведемо појам перзистенције који ће се испоставити као кључан у применама тополошких метода. Постоји мноштво примена и илустрација описаних својстава. Ми ћемо се у овом раду посветити једној од тих примена, а то је кластеровање.

1.3 Допринос рада

У овом раду ћемо се бавити ToMATo (Topological Mode Analysis Tool) алгоритмом кластеровања који се први пут појављује у раду [2]. Радићемо на подацима узоркованим из непознате функције густине. Алгоритам комбинује две фазе: тражење мода (енг. mode-seeking) и спајање (енг. merging) кластера. Пре свега, посматрамо одређену функцију густине података тј. њену оцену. Детекција мода посматране функције се врши стандардном шемом успона заснованој на графу (graph-based hill-climbing scheme) [5]. Новост овог приступа лежи у употреби тополошке перзистенције за одређивање значајности мода и конструкцију њихове хијерархије, као и спајање кластера у другој фази на основу тих информација. Такође, алгоритам пружа додатне визуализације резултата. Једна од најпознатих је перзистентни дијаграм који осликава значајност мода одређене функције густине. У пракси, ове повратне информације омогућавају да одаберемо релевантне вредности параметара на основу којих ће, под благим условима узорковања, алгоритам дати тачан број кластера. Користећи предности недавног напретка у тополошкој теорији перзистенције, можемо дати теоријске гаранције да су број кластера, али и позиције њихових центроида, уско везани са тачкама глобалних екстрема стварне функције густине. Ове гаранције важе у разним контекстима, штавише, за сваки облак тачака на некој непознатој Римановој многострукости.
Теоријске гаранције под таквим општим условима, користећи само једноставне алате попут графа суседа, имамо захваљујући недавним резултатима о стабилности перзистентних дијаграма [6]. Претходни резултати стабилности [7] захтевали су употребу део по део линеарних апроксимација функција густине. Конструкција таквих апроксимација постаје веома неоптимална када димензионалност података расте. То је и разлог зашто хомолошка перзистенција никада раније није нашла велику примену у анализи података, осим у неким једноставним случајевима нижих димензија.

Алгоритму су довољне чак и само грубе оцене густине и познавање (приближног) међусобног растојања тачака у облаку. Стога следи једно веома важно својство, а то је применљивост алгоритма у било ком метричком простору. Иако величина улазне матрице растојања може бити квадратна у односу на број тачака података, и поред тога, његова сложеност остаје прихватљива. Оптималном имплементацијом алгоритам користи само линеарну количину меморије.

Циљ овог рада је да упозна читаоца са тополошком анализом података на теоријском, али и практичном нивоу. Пре примене и имплементације кластеровања биће дата математичка подлога за све што се примењује, али пре свега идеја и мотивација за саму примену перзистентне хомологије у новом алгоритму, као и генерално у осталим методама машинског учења. Акценат рада је више на интуитивности примене топологије, која је свакако математички оправдана.

Кратак преглед садржаја:

- Глава 1: У овој глави смо се упознајемо са тополошком анализом података генерално, али и са садржином и циљем овог рада; мотивисани читалац може кренути са даљим упознавањем.
- Глава 2: На самом почетку даљег упознавања чекају нас основе хомолошке теорије јер су неопходне као темељ овог рада. Упознаћемо се са појмовима као што су симплицијални комплекси, Бетијеви бројеви[8] и хомолошке групе да бисмо могли увести појам перзистенције у следећој глави.
- Глава 3: Ова глава, иако се директно наставља на претходну, је издвојена због своје важности. У њој дефинишемо и објашњавамо кључну методу - перзистентну хомологију. Такође упознаћемо се са два начина визуализације перзистенције као што су, већ споменути, перзистентни дијаграми, али и перзистентни бар-кодови.
- Глава 4: У овој глави дајемо идеју и мотивацију за примену перзистентне хомологије у кластеровању; упоређујемо тај приступ са неким већ добро познатим методама кластеровања и издвајамо његове предности. Иако је ToMATo алгоритам за кластеровање главна звезда овог рада, покушаћемо читаоцу мало више да приближимо генералну идеју и зато ћемо се упознати са још једним алатом тополошке анализе података, а то је Mapper алгоритам [8].
- Глава 5: Сада већ имамо довољно знања да прођемо кроз све кораке ToMATo кластеровања. Објаснићемо детаљно основне две фазе алгоритма, тражење мода и спајање кластера, њихову имплементацију и комплексност извршавања. Такође, у овом делу рада дајемо и теоријске гаранције алгоритма заједно са неопходном математичком теоријом која их поткрепљује.

[8]Enrico Betti Glaoui - италијански математичар

- Глава 6: Последња глава је посвећена примени идеја и резултата из претходних глава. ToMATo алгоритам примењујемо на разне скупове података у разним димензијама: синтетичи скупови тачака у две и три димензије, сегментација слика у различитим просторима, као и сегментација реалног скупа тачака генерисаног из видео записа просторије.

Глава 2

Основе хомолошке теорије

Хомологија је, уопштено речено, начин повезивања низа алгебарских објеката са другим математичким објектима као што су тополошки простори. Хомолошке групе су првобитно дефинисане у алгебарској топологији. Мотивација за проучавање хомолошких група, које дефинишемо у овој глави, јесте запажање да се два облика могу разликовати посматрањем рупа на тим објектима. У математици, различите теорије хомологије се баве аксиомат ским проучавањем интуитивне геометријске идеје хомологије рупа на тополошким просторима. Ми ћемо се бавити симплицијалном теоријом хомологије.

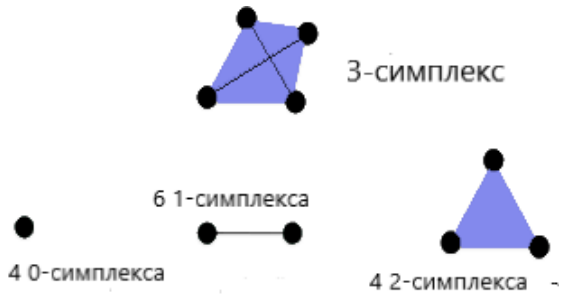
2.1 Симплицијални комплекси

Почнимо прво са дефинисањем основних структура које користимо и на којима се цела теорија заснива.

2.1.1 Геометријски симплицијални комплекси

Дефиниција 2.1.1. Нека је $S = \{v_0, ..., v_k\}$ скуп афино независних тачака у $\mathbb{R}^n$. Најмањи конвексни скуп који садржи те тачке назива се k-симплекс. Тачке скупа S су темена симплекса.

Ако је σ k-симплекс из претходне дефиниције, тада сваки подскуп $T \subset S$ такође дефинише један симплекс τ који називамо страна симплекса σ, ознака $\tau \prec \sigma$. Симплекс σ обележавамо и помоћу његових темена као $\sigma = \{v_0, ..., v_k\}$. Следи да је тада било који подниз низа темена $\{v_0, ..., v_k\}$ страна симплекса σ. Уколико је страна симплекса различита од празног скупа и целог симплекса, она се назива и правом страном тог симплекса.



Слика 2.1: Тетраедар као 3-симплес и све његове стране

Дакле симплекс је један веома велики комбинаторни објекат, али с друге стране, због њихове униформности и једноставности у структури, представљају идеалну репрезентацију за примену у рачунарству.

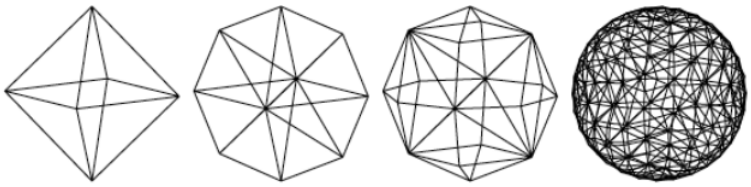
Дефиниција 2.1.2. Геометријски симплицијани комплекс K у $\mathbb{R}^n$ је коначна фамилија симплекса таква да:

- ако је σ симплекс из K, тада је свака страна симплекса такође из K
- ако су σ_1 и σ_2 симплекси из K, тада је и $\sigma_1 \cap \sigma_2$ такође из K.

Под димензијом симплицијалног комплекса K подразумевамо димензију његовог максималног симплекса, тј. $dimK = max\{\sigma|\sigma \in K\}$.

У пракси, да би изучавали тополошки простор, почињемо са његовом тријангулацијом.

Дефиниција 2.1.3. Триангулација тополошког простора X је пар (K, ϕ), где је K симплицијални комплекс, а $\phi : K \to X$ хомеоморфизам.



Слика 2.2: Триангуларизација сфере (в. [10])

2.1.2 Апстрактни симплицијални комплекси

У циљу добијања комбинаторне дефиниције објекта апстрахујемо симплицијални комплекс.

Дефиниција 2.2.1. Непразну фамилију скупова K са колекцијом непразних подскупова S називамо апстрактни симплицијални комплекс (у даљем тексту само симплицијални комплекс) ако:

1. $\{v\} \in S$, за све $v \in K$
2. Ако $\sigma \in S$ и $\tau \subseteq \sigma$ онда $\tau \in K$

Аналогно, елементе симплицијалног комплекса K називамо симплексима, а елементе скупа S теменима. k-симплекс је симплицијални комплекс који има има $k+1$ темена. Симплицијални комплекси се могу разложити на симплексе одређених мањих димензија. Пресек, ако постоји, свих вишедимензионих симплекса мора бити симплекс неке мање димезије.

Сваки геометријски симплицијални комплекс може се апстраховати. Уколико је K геометријски симплицијални комплекс и V скуп његових темена, а Σ фамилија скупова темена свих симплекса из S тада је (V, Σ) апстрактни симплицијални комплекс који одговара комплексу K.

2.2 Циклови и границе

Абелова група C_p придружена симплицијалном комплексу K састоји се од свих комбинација p-симплекса у K. Ако узмемо да су коефицијенти из $\mathbb{Z}_2$ (пре свега због једноставности и симетрије), онда се сваки елемент из групе C_p може написати у облику:

$$\sum_j \sigma_j, \text{ за } \sigma_j \in K.$$

Дефиниција 2.2.1. За дати симплицијални комплекс K, хомоморфизам p-границе ϑ_p је функција која сваком p-симплексу $\sigma = \{v_0, ..., v_p\} \in K$ додељује његову границу:

$$\vartheta_p \sigma = \sum_i \{v_0, ..., \hat{v_i}, ..., v_p\},$$

где $\hat{v_i}$ означава да скуп темена не садржи теме v_i.

Функција $\vartheta_p : C_p \to C_{p-1}$ је заправо хомоморфизам између горе дефинисаних група. Следећом теоремом доказујемо да границе немају границе.

Теорема 2.2.1. $\vartheta_{p-1} \circ \vartheta_p = 0$.
Доказ:

$$\begin{aligned} \vartheta_{p-1} \circ \vartheta_p([v_0, ..., v_p]) = \\ \vartheta_{p-1}(\sum_i (-1)^i [v_0, ..., \hat{v_i}, ..., v_p] = \\ \sum_{j<i} (-1)^i (-1)^j [v_0, ..., \hat{v_j}, ..., \hat{v_i}, ..., v_p] + \\ \sum_{i<j} (-1)^i (-1)^{j-1} [v_0, ..., \hat{v_i}, ..., \hat{v_j}, ..., v_p] = \\ 0 \qquad \star \end{aligned}$$

То води ланчастом комплексу:

$$\xrightarrow{\vartheta_{n+1}} C_n \xrightarrow{\vartheta_n} C_{n-1} \xrightarrow{\vartheta_{n-1}} ... \xrightarrow{\vartheta_2} C_1 \xrightarrow{\vartheta_1} C_0 \xrightarrow{\vartheta_0} 0$$

Дефиниција 2.2.2. Елемент групе C_p зовемо p-ланац.
Елемент групе C_p (p-ланац) који се хомеоморфизмом ϑ_p слика у 0 зовемо p-цикл.
Слику хомеоморфизма ϑ_{p+1} зовемо p-граница.

Дакле:

- Група циклова је дата са: $Z_p = Ker\vartheta_p$
- Група граница је дата са: $B_p = Im\vartheta_{p+1}$

Такође, из горе поменутог својсва $\vartheta_{p-1} \circ \vartheta_p = 0$ следи релација:

$$B_p \subset Z_p \subset C_p$$

тј. свака граница је цикл.

2.3 Хомолошке групе

Како је B_p подгрупа Абелове групе Z_p онда је она и нормална подгрупа од Z_p па можемо дефинисати њене класе еквиваленције као количничку групу коју називамо p-та хомолошка група. Наравно, p-та хомолошка група представља понашање објекта у димензији p.

Дефиниција 2.3.1. p-та хомолошка група ланчастог комплекса (C_p, ϑ_p) је:

$$H_p := Z_p / B_p,$$

где је $Z_p = Ker\vartheta_p$ и $B_p = Im\vartheta_{p+1}$.

С обзиром на то да су групе C_p коначно генерисане Абелове групе, то су и Z_p и B_p, па самим тим и њихова количничка група H_p. Један он најпознатијих резултата алгебре [11] нам каже:

Теорема 2.3.1. Свака коначно генерисана Абелова група изоморфна је групи облика:

$$\mathbb{Z}_{p_1} \otimes ...\mathbb{Z}_{p_n} \otimes \mathbb{Z} \otimes ...\mathbb{Z},$$

где су p_i прости бројеви или степени простих бројева, а $\mathbb{Z}_p$ одговарајуће цикличне групе.

2.4 Бетијеви бројеви

У алгебарској топологији, Бетијеви бројеви се користе за разликовање тополошких простора на основу повезаности n-димензионалних симплицијалних комплекса. Коначно, можемо увести појам Бетијевих бројева као рангове претходно дефинисаних хомолошких група.

Дефиниција 2.4. p-ти Бетијев број је ранг p-те хомолошке групе:

$$\beta_p = r(H_p)$$

На пример, ако је $H_n = \mathbb{Z}^k \times \mathbb{Z}^2$ онда је $\beta_n = k$ јер торзионе (коначне) подгрупе не утичу на вредност Бетијевог броја. Као што можемо прво приметити сви Бетијеви бројеви су коначни, штавише после неке димензије сви су једнаки нула.

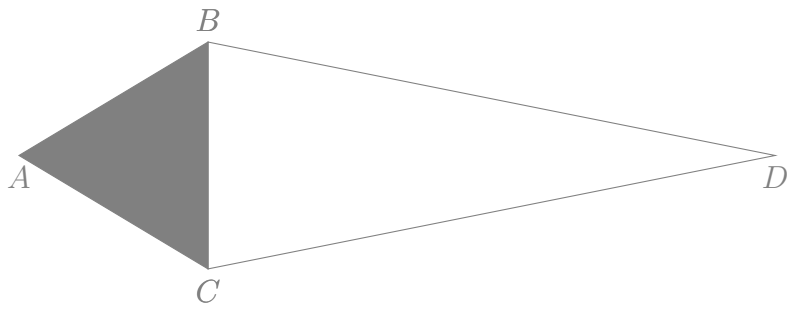
Пример 2.1. Нека је симплицијални комплекс K дат са:

$$K = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}\}.$$

Можемо га замислити као празан троугао. Да бисмо израчунали β_0 треба нам 0-та хомолошка група $H_0 := Z_0/B_0$, односно треба да пронађемо чему су изоморфне групе 0-граница и 0-циклова. Из дефиниције тих група следи:
$Z_0 = Ker\vartheta_0 = span(\{a\}, \{b\}, \{c\})$,
јер се сваки од наведених симплекса слика у 0, зато је $Z_0 = (Z/2Z)^3$
$B_0 = Im\vartheta_1 = span(\{a\} + \{b\}, \{b\} + \{c\}, \{a\} + \{c\})$,
али како су само 2 линеарно независна елемента онда важи $B_0 = (Z/2Z)^2$.
Коначно,

$$H_0 := Z_0/B_0 = (Z/2Z)^3/(Z/2Z)^2 = Z/2Z \text{ и } \beta_0 = r(H_0) = 1. \nabla$$

Пример 2.2. Исте вредности Бетијевих бројева има и следећи, мало компликованији, пример са слике 2.3 који има један попуњен троугао тј. садржи симплекс $\{A, B, C\}$: ∇



Слика 2.3: Симплицијални комплекс за кога важи $\beta_0 = 1$, $\beta_1 = 1$ и $\beta_2 = 0$

2.4.1 Геометријска интуиција

n-ти Бетијев број заправо представља број n-димензионалних рупа ако рупу посматрамо као цикл који није граница у некој већој димензији. То једноставно објашњава дефиницију хомолошке групе и партиције на класе еквиваленције јер бројимо само тополошки нееквивалентне рупе.

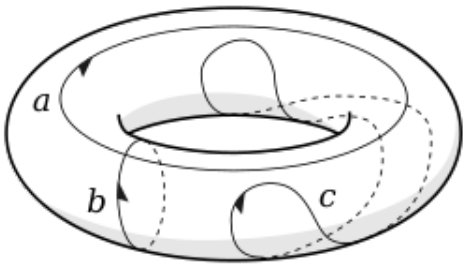
Дајемо интуитивну дефиницију за првих неколико Бетијевих бројева симплицијалног комплекса:

- Бетијев број β_0 једнак је броју повезаних компоненти
- Бетијев број β_1 једнак је броју једнодимензионалних, односно цикличних рупа
- Бетијев број β_2 једнак је броју унутрашњости (енг.void)

Тако смо, интуитивно, могли за претходна два примера 2.1. и 2.2. видети да је $\beta_0 = 1$ јер K има само једну повезану компоненту (односно све је повезано). Слично, без рачунања, следи и $\beta_1 = 1$ јер имамо један цикл (ивице троугла) и $\beta_2 = 0$ јер немамо унутрашњост. Исто тако интуитивно разликујемо сферу и коцку ($\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$) од торуса ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$). Симплицијална хомологија зависи само од симплицијалног комплекса на ком се посматра и као резултат тога можемо разликовати површи. Највећа предност хомологије је што је све израчунљиво и рачунарски применљиво.

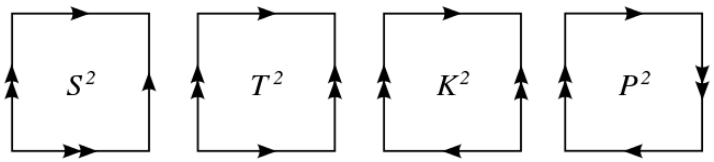
Друга, интуитивна, интерпретација Бетијевих бројева јесте максималан број сечења површи тако да остане повезана тј. n-ти Бетијев број представља максималан број n-циклова по којим можемо сећи површ тако да је не поделимо.

Пример 2.3. Узмимо торус као пример:
Видели смо да за торус важи $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$. Посматрајмо сада следећу слику 2.4 (в. [12]) и објаснимо вредност првог Бетијевог броја.



Слика 2.4: Торус и три хомолошки нееквивалентна цикла

Циклови a, b и c се не могу непрекидно деформисати један у други што значи да нису тополошки еквивалентни. Међутим цикл c се може непрекидно трансформисати (скупити) у тачку што га чини еквивалентним нули, док се циклови a и b не могу непрекидно трансформисати у тачку (наравно, без деформације торуса).
Генерално, сечење површи по циклу који је еквивалентан нули дели ту површ на два или више делова. Зато закључујемо да је $\beta_1 = 2$ јер можемо на два различита начина сећи торус по 1-цикловима a и b (замислимо те циклове као меридијане и паралеле) тако да он остане повезан. ∇



Слика 2.5: Четири начина лепљења квадрата да би се добиле четири различите затворене површи; лепе се појединачне стрелице заједно и дупле стрелице заједно у назначеним смеровима

Још један начин на који то можемо разумети јесте као када би квадрату лепили наспрамне странице, прво један па други пар, добили бисмо торус.

Или пак обрнуто, када бисмо секли торус по цикловима a и b са слике 2.4 добили бисмо квадрат (тачније правоугаоник) коме би наспрамне странице представљале нееквивалентне циклове. Илустрација овог објашњења дата је на слици 2.5 (в. [13]) за сферу, торус, Клајнову боцу и пројективну раван. На њој су представљена четири начина лепљења квадрата да би се добиле четири различите затворене површи. Лепе се појединачне стрелице заједно и дупле стрелице заједно у назначеним смеровима. На тај начин можемо да закључимо број сечења (или обрнуто лепљења) страница, односно β_1 број посматране фигуре. Треба имати у виду да је овакво интуитивно закључивање ограничено на човеку замислив простор и да је овде дато искључиво због лакшег разумевања прећашње теорије.

Глава 3

Перзистентна хомологија

3.1 Основни појмови

Перзистентна хомологија је метода за израчунавање тополошких карактеристика простора при различитим просторним скалама. Рачунањем хомолошких група за сваки комплекс на скали може се пратити када нека хомолошка класа настаје, а када нестаје. Отпорније класе откривају се помоћу скала широког опсега и сматра се да ће оне представљати суштинске карактеристике основног простора. Да би се одредила хомологија простора, исти прво мора бити представљен као симплицијални комплекс. Пресликавање дефинисано на основном простору одговара филтрацији симплицијалног комплекса. Сада ћемо ове појмове математички формализовати.

Дефиниција 3.1.1. Филтрација симплицијалног комплекса K је низ комплекса K_i, такав да:

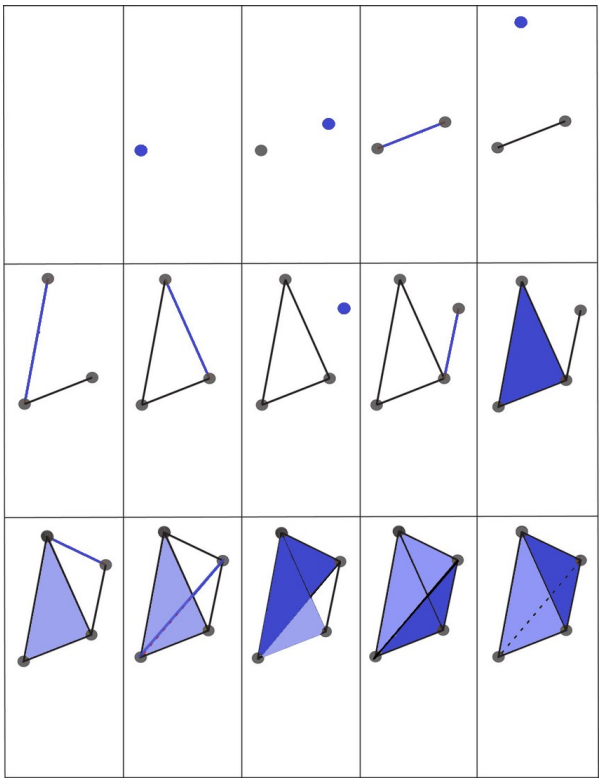
$$\varnothing = K_0 \subset K_1 \subset ... \subset K_n = K$$

Произвољна филтрација комплекса K дефинише парцијално уређење на скупу симплекса који чине K. За нас су од интереса само оне филтрације које воде до тоталног уређења јер нам је неопходна хијерархија комплекса.

Једна од предности оваквог начина представљања симплицијалног комплекса је што ће нам то омогућити да пратимо мењања тополошких инваријанти у складу са еволуцијом комплекса. Такође ћемо имати увид у то када се одређена хомолошка класа појављује, колико траје и када нестаје, а то је и основна идеја перзистентне хомологије.

Пример 3.1. На следећој слици 3.1 можемо видети еволуцију филтрације 3-симплекса од празног скупа па све до целог тетраедра. У сваком

кораку додато је једно ново теме или нова страна и то је на слици обележено тамно плавом бојом. ∇



Слика 3.1: Филтрација тетраедра

3.2 Перзистентне хомолошке групе

Посмтарамо симплицијални комплекс K димензије k са дефинисаном филтрацијом као у претходном поглављу. Међу комплексима $K_i = \{\sigma_1, ..., \sigma_n\}$ постоји инклузија па можемо дефинисати пресликавање:

$$\varnothing = K_0 \xrightarrow{i_0} K_1 \xrightarrow{i_1} ... \xrightarrow{i_{n-1}} K_n = K.$$

Сваком од комплекса K_i додељујемо ланчасти комплекс $C_*^{K_i}$, а пресликавања i_j линеарно продужимо на Абелове групе ланчастог комплекса (в. Главу 2).

На следећем дијаграму приказана је конструкције те структуре, где хоризонталне стрелице представљају граничне хомоморфизме међу групама у $C_*^{K_i}$, а вертикалне продужење инклузије међу одговарајућим комплексима филтрације:

$$\begin{array}{ccccc}
\downarrow & & \cdots & & \downarrow \\
C_k^{K_i} & \xrightarrow{\vartheta} & \cdots & \xrightarrow{\vartheta} & C_0^{K_i} \xrightarrow{\vartheta} 0 \\
\downarrow & & \cdots & & \downarrow \\
C_k^{K_{(i+1)}} & \xrightarrow{\vartheta} & \cdots & \xrightarrow{\vartheta} & C_0^{K_{(i+1)}} \xrightarrow{\vartheta} 0 \\
\downarrow & & \cdots & & \downarrow \\
\cdot & & & & \cdot \\
\cdot & & & & \cdot \\
\cdot & & & & \cdot \\
\downarrow & & \cdots & & \downarrow \\
C_k^{K_n} & \xrightarrow{\vartheta} & \cdots & \xrightarrow{\vartheta} & C_0^{K_n} \xrightarrow{\vartheta} 0
\end{array}$$

За свако $0 \leq i \leq k$ инклузија $i : K_{i-1} \to K_i$ индукује инклузију на ланчастим комплексима $i_* : C_*^{K_(i-1)} \to C_*^{K_i}$. И додатно важи:

$$i(Z_p^{K_{(i-1)}}) \subset i(Z_p^{K_i})$$
$$i(B_p^{K_{(i-1)}}) \subset i(B_p^{K_i})$$

Теорема 3.2.1. Пресликавање $f_p^{i-1} : H_p^{K_(i-1)} \to H_p^{K_i}$ индуковано инклузијом i_* је добро дефинисано и хомоморфизам је.

За произвољну класу $[\sigma] \in H_p^{K_(i-1)}$ важи:

$$f_p^{i-1}([\sigma]) = \begin{cases} [\sigma], & \text{ако је } [\sigma] \text{ граница неког ланца у } C_{p+1}^{K_i} \\ 0, & \text{иначе} \end{cases}$$

На овај начин можемо бројати појављивање и евентуално нестајање нових циклова, односно рупа, (детаљније у Глави 2 у поглављу 2.4.1,) кроз филтрацију у свим димензијама. Да бисмо, коначно, увели појам перзистентних хомолошких група дефинишимо још пресликавање:

$$f_p^{i,j} : H_p^{K_i} \to H_p^{K_j} \ ,$$

као композицију индукованих инклузија $f_p^{j-1} \circ \ldots \circ f_p^i$.

Дефиниција 3.2.1. За свако $0 \leq i \leq k$, p-перзистентна хомолошка група је:

$$H_p^{i,j} = Im f_p^{i,j} = Z_p^i / (B_p^j \cap Z_p^i),$$

где су Z_p^i и B_p^i групе, респективно, циклова и граница у комплексу K_i. Групе $H_p^{i,j}$ су коначно генерисане Абелове групе. Сада можемо и овде, на потпуно аналоган начин, као у сингуларној хомологији да дефинишемо Бетијеве бројеве.

3.2.1 Перзистентни Бетијеви бројеви

Дефиниција 3.2.1.1. p-перзистентни Бетијев број комплекса K са филтрацијом $\varnothing = K_0 \subset K_1 \subset ... \subset K_n = K$ је:

$$\beta_p^{i,j} = r(H_p^{i,j})$$

Дакле, $\beta_p^{i,j}$ је број независних хомолошких класа у $H_p^{i,j}$. Интуитивније, те класе можемо посматрати као хомолошке класе које су опстале бар до комплекса K_j, а у филтрацији су се појавиле у комплексу K_i или пре. Другим речима, перзистентни Бетијев број $\beta_p^{i,j}$ броји класе са животним веком бар $j - i$, као разликом умирања и рађања у одговарајућим симплицијалним комплексима у филтрацији. Односно, знајући вредности перзистентних Бетијевих бројева знамо и животне векове хомолошких класа.

3.3 Визуализација перзистенције

Типови скупова података који се могу проучавати помоћу перзистентне хомологије укључују дигиталне слике, ниво скупове реалних функција, мреже итд.. Као што смо споменули у уводу овог рада, коначни метрички простори се у литератури тополошке анализе података називају и облацима тачака. Са тополошког становишта, коначни метрички простори не садрже никакве занимљиве информације. Стога се облак тачака посматра на различитим скалама, а затим се анализира добијена еволуција облика. Тополошке инваријанте нам дају квалитативне карактеристике којима смо се бавили у овој и претходној глави. У овом поглављу говоримо о начинима визуализације тих карактеристика на различитим скалама да би се добио жељени сажетак структуре.

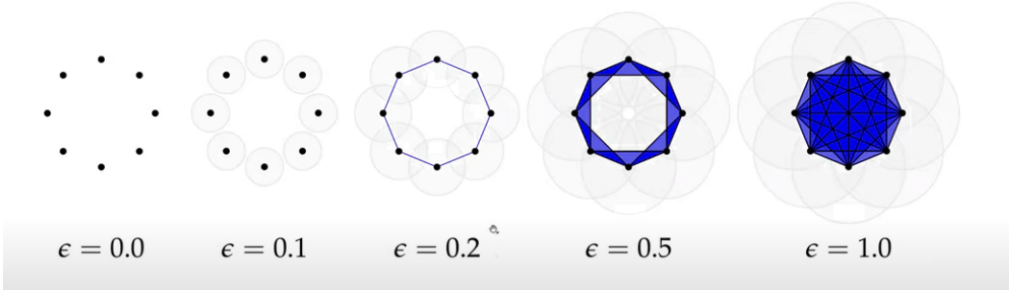
3.3.1 Виеторис-Рипс комплекс

Пре него што уведемо појмове перзистентних бар-кодова и дијаграма даћемо један веома интуитиван пример конструисања филтрације да бисмо имали основу за даљи рад.
Као илустрацију узмимо скуп тачака у $\mathbb{R}^2$ који приказујемо на слици 3.2.

Нека је ϵ параметар удаљености, ненегативни реални број. За различите вредности ϵ конструишемо простор S састављен од темена, ивица, троуглова и вишедимензионалних политопа према следећем правилу: За свака два темена конструишемо ивицу ако и само ако еуклидска удаљеност између њих није већа од ϵ; конструишемо троугао ако и само ако су све његове ивице већ у S; итд. за више димензије, аналогно. Приметимо, тада за $\epsilon' \leq \epsilon$ важи $S_{\epsilon'} \subseteq S_{\epsilon}$. Ово је груба конструкција филтрације Виеторис–Рипс комплекса. Сада ћемо дати и његову формалну дефиницију:

Дефиниција 3.3.1. За ненегативни реални број ϵ, Виеторис–Рипс комплекс у простору S са матриком d је:

$$VR_{\epsilon}(S) = \{\sigma \subseteq S | d(x,y) \leq 2\epsilon \text{ , за свако } x, y \in S\}$$



Слика 3.2: Виеторис-Рипс комплекс

Треба споменути још неке познате конструкције комплекса, као што су Чехов комплекс[1], Делоне комплекс[2] и алфа комплекс, али у овом раду се нећемо бавити свима њима појединачно (више о овим структурама се може пронаћи у литератури [14, 15]).

3.3.2 Перзистентни бар-кодови

Захваљујући тополошким алатима, можемо израчунати карактеристике простора као што су број и животни век хомолошких класа (компонената, циклова итд.). То можемо визуализовати помоћу коначне колекције полуотворених интервала познате као бар-код.

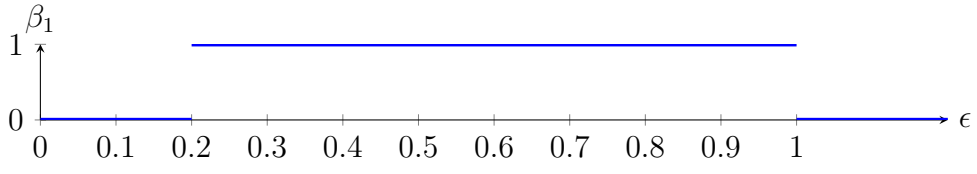
Дајемо и формално тумачење бар-кодова помоћу перзистентне хомолошке теорије. Кажемо да је $\sigma \in H_p(K_i)$ рођено у $H_p(K_i)$ ако не припада слици $f_p^{i-1,i}$ (тј. $f_p^{i-1,i}(\sigma) = 0$). Кажемо да $\sigma \in H_p(K_i)$ умире у $H_p(K_j)$ ако је $j > i$

[1]Eduard Čech - чешки математичар
[2]Boris Nikolaevich Delaunay - руски математичар

најмањи индекс за који је $f_p^{i,j}(\sigma) = 0$. Тада је животни век од класе σ представљен полуотвореним интервалом $[i, j)$. Ако је $f_p^{i,j}(\sigma) \neq 0$ за све j такве да је $i < j \leq k$, где је к димензија посматраног објекта, кажемо да је σ бесконачно перзистентна, а њен животни век представљен је интервалом $[i, \infty)$.

Грубо речено, лева крајња тачка интервала представља рођење класе, а десна крајња тачка представља смрт исте класе. Дакле, дужи интервали бар-кода представљају дуже животне векове. Бесконачно перзистентне класе представљамо бесконачним полуинтервалима које скалирамо на погодан начин зарад визуализације. На то указујемо стављањем врха стрелице на десну крајњу тачку интервала. Супротно томе, могу се појавити веома кратки интервали (значајно краћи од других) и такве класе можемо посматрати као шум у нашим подацима. Међутим, ту морамо бити опрезни како одређујемо ту минималну границу значајности животног века. Она се углавном одређује емпиријски у зависности од случаја употребе. У наредним главама ћемо видети како нам та граница заправо одређује тачан број кластера у ToMATo алгоритму кластеровања.

Као почетни пример погледајмо следећу слику 3.3. То је бар-код који одговара филтрацији са слике 3.2. Хомолошке класе се мењају са променом параметра растојања ϵ у Виеторис-Рипс комплексу. На x оси је вредност параметра ϵ, а на y оси је вредност β_1 броја јер посматрамо животни век цикличне рупе тј. једнодимензионалне хомолошке класе. Приметимо да се једна рупа појављује за вредност $\epsilon = 0.2$ и нестаје када се потпуно попуни њена унутрашњост на вредности параметра $\epsilon = 1$. Зато је $\beta_1 = 1$ на интервалу $(0.2, 1)$, а нула иначе.



Слика 3.3: Перзистентни бар-код за димензију 1 одговарајућих Виеторис-Рипс комплекса са слике 3.2

Један од најкреативнијих делова употребе перзистентне хомологије је статистичка интерпретација резултата. Са статистичке тачке гледишта, непозната је величина коју процењујемо, стога су потребне методе за квантитативну процену квалитета бар-кодова. Постоје препрека у томе што се

тополошке инваријанте, до недавно, нису ни сусреле са статистичким мерама, оценама и тестовима. Простору бар-кодова недостају геометријска својства која би олакшала дефинисање основних појмова као што су средња вредност, медијана итд. Та препрека је свакако и мотивација за тренутна истраживања која су усмерена на проучавање метода које тај простор мапирају у просторе који имају боље прилагођена геометријска својства статистичким алатима.
Дакле, потребан је начин за оцену статистичке значајности добијених резултата и упоређивање два различита излаза, као и начин за израчунавање одговарајућих статистичких сажетака (енг. summary statistics) . Неки од главних приступа проблему статистичке анализе бар-кодова су:

- У првом приступу се користе случајни симплицијални комплекси који инкорпорирају наше претпоставке о подацима. Они се у овом случају посматрају као нул-модели (енг. null model) за упоређивање са стварним подацима.
- У другом приступу се проучавају својства метричког простора чије су тачке перзистентни дијаграми који одговарају посматраним бар-кодовима.
- У трећем приступу се проучавају особине појединачних перзистентних дијаграма.

Сада је прави тренутак да формално уведемо појам перзистентних дијаграма.

3.3.3 Перзистентни дијаграми

Перзистентни дијаграм је презентација коначног броја тачака $(x, y) \in \bar{\mathbb{R}}^2$ које се могу понављати, односно бити мултиплициране, заједно са дијагоналом $\Delta = \{(x, y) \in \mathbb{R}^2 | x = y\}$. За сваку тачку дијаграма вредност на x- оси је тренутак рађања у одговарајућој филтрацији која се посматра, а вредност на y-оси тренутак умирања. Следи да је животни век сваке тачке сразмеран њеном растојању од дијагонале. Што је тачка удаљенија од дијагонале то је перзистентнија па на тај начин можемо раздвајати суштински битне особине од шума у подацима. Такође, и бесконачно перзистентне тачке обележавамо на y оси изнад свих осталих коначних вредности. Видимо да нам оваква визуализација даје исте информације као и бар-код.

Формално, перзистентни дијаграми визуализују информације добијене из свих перзистентних хомолошких група. Пре свега, помоћу перзистентних

Бетијевих бројева дефинишемо $\mu_p^{i,j}$ као број p-димензионалних хомолошких класа које су се родиле у комплексу K_i, а умрле у комплексу K_j у филтрацији:

$$\mu_p^{i,j} = (\mu_p^{i,j-1} - \mu_p^{i,j}) - (\mu_p^{i-1,j-1} - \mu_p^{i-1,j}),$$

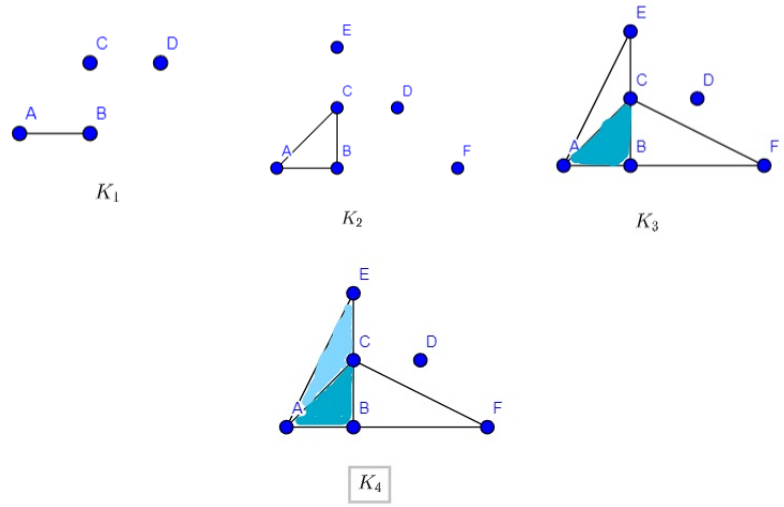
за свако $0 \leq i < j \leq k$. То је заправо разлика класа рођених пре K_{i+1}, а умрле су у K_j и класа рођених пре K_i, а умрле су у K_j.

Дефиниција 3.3.3.1. Унија тачака (i, j) мултиплицираних $\mu_p^{i,j}$ пута и свих тачака дијагонале назива се k-ти перзистентни дијаграм, у ознаци D_pK.

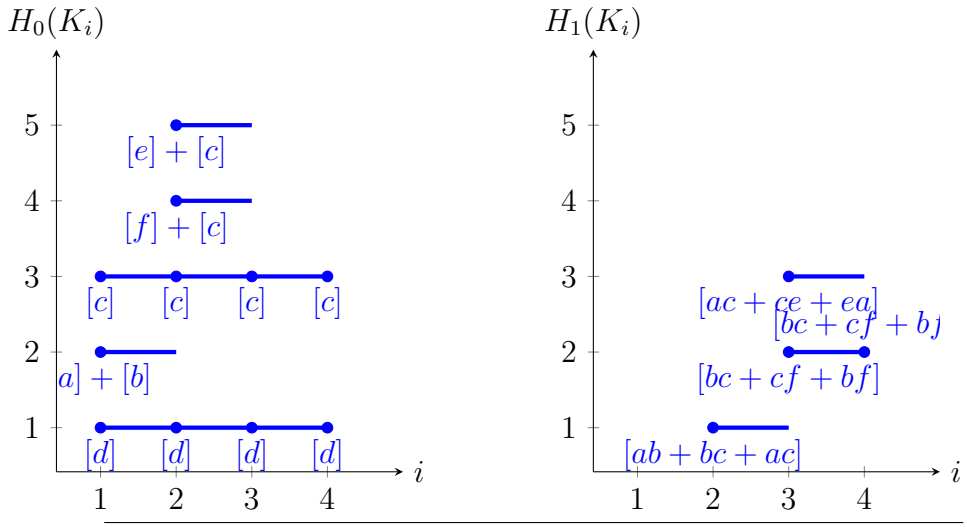
Следи пример у којем дајемо визуализацију горе уведених појмова:
Пример 3.2. Посматрајмо коначну филтрацију симплицијалног комплекса на слици 3.4 и њену еволуцију у корацима, који одговарају сваком комплексу у филтрацији, на слици 3.5. Интуитивно, на левом дијаграму у сваком кораку записујемо и бројимо повезане компоненте, а на десном рупе. Формално, користимо информације које нам дају векторски простори $H_p(K_i)$ заједно са пресликавањима $f_p^{i,j}$ између њих. Те информације можемо визуализовати цртањем дијаграма на следећи начин: у i-том кораку филтрације, цртамо онолико тачака колика је димензија векторског простора $H_p(K_i)$ тако да то буде представљено на y-оси. Затим повезујемо тачке тако што цртамо интервал између њих ако се хомолошка класа из i-тог корака филтрације налази у перзистентној хомолошкој групи $H_p^{i,i+1}$. Ако се не налази, односно није се одржала до $(i+1)$-ог корака, онда цртамо отворен интервал. Лево је представљена димензија 0 са бројем темена једнаким $H_0(K_i)$ и десно димензија 1 са $H_1(K_i)$ темена за свако $1 \leq i \leq 4$.

Након овога лако долазимо до перзистентног бар-кода на слици 3.6 и њему одговарајућег перзистентног дијаграма на слици 3.7. Бесконацно перзистентне класе су у бар-кодовима означене стрелицама на десној страни, а у дијаграму су означене празним кругом. Такође, величина тачака директно је пропорционална броју класа које они представљају.

Наравно, овако конструисани перзистентни дијаграми не садрже тачке испод дијагонале због тоталног уређења филтрације па увек важи $x \leq y$. Постоји и обрнути поредак који добијамо преусмеравањем филтрације. Избор поретка је искључиво формалност и не утиче на валидност теорије. Ми ћемо у даљем раду користити тај обрнути поредак, а мотивација за то ће бити ускоро дата у Глави 4. ∇



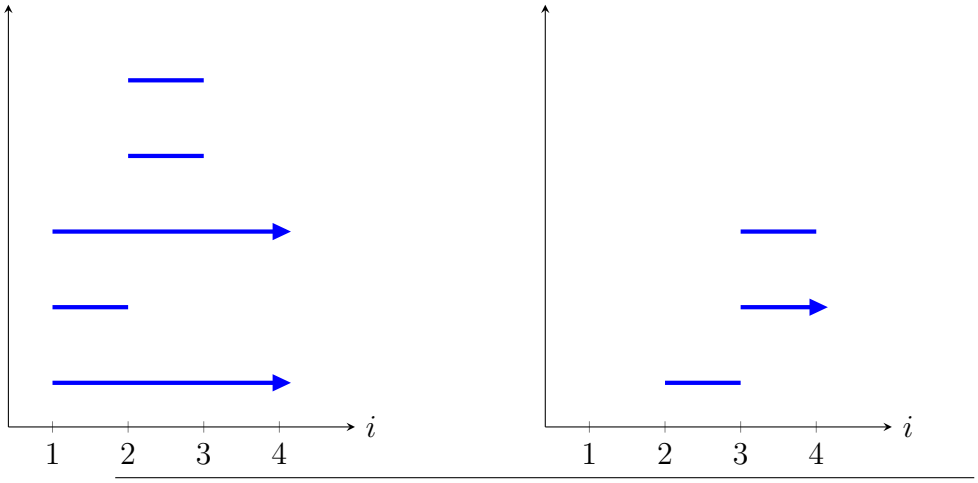
Слика 3.4: Филтрација датог симплицијалог комплекса



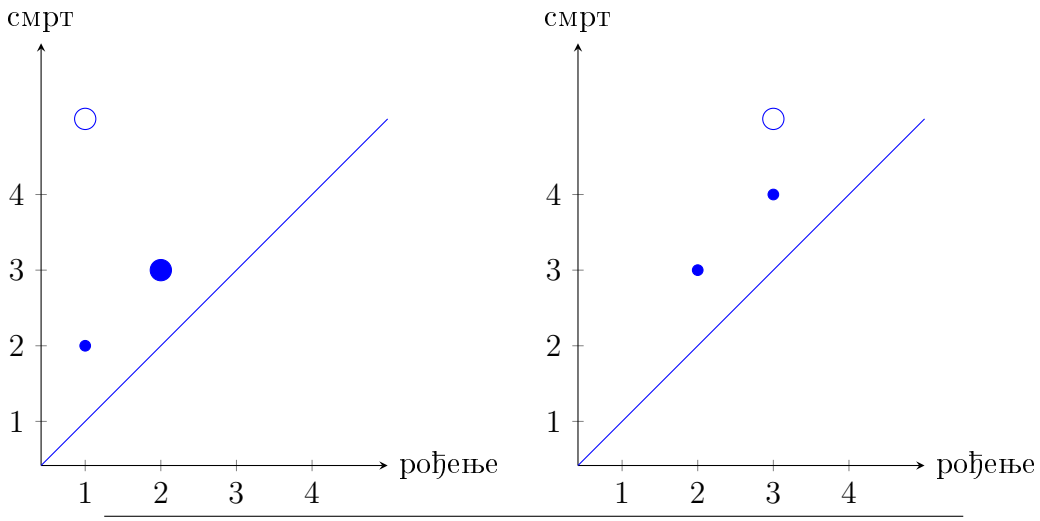
Слика 3.5: Еволуција филтрације у корацима

Видимо да ипак постоје одређене предности у тумачењу и употреби перзистентних дијаграма. Рекли смо да се у другом приступу посматра одговарајуће дефинисан простор дијаграма перзистенције, дефинише се метрика, проучавају се геометријска својства овог простора и врше се стандардне статистичке анализе. Односно, пожељно је имати могућност упоређивања дијаграма перзистенције проучавањем бијекција између њихових елемената. Зато дијаграми морају бити исте кардиналности, а то лако решавамо додавањем тачака на дијагонали. Један начин компарације је увођене Хаусдорфове метрике[3] [16] или bottleneck растојања на том простору.

[3]Felix Hausdorff - немачки математичар



Слика 3.6: Перзистентни бар-кодови за димензију 0 (лево) и димензију 1 (десно)



Слика 3.7: Перзистентни дијаграми за димензију 0 (лево) и димензију 1 (десно)

Дефиниција 3.3.3.2. Нека су X и Y два перзистентна дијаграма. Хаусдорфово растојање и bottleneck растојање су, редом, дати са:

$$d_H(X,Y) = max\{sup_x inf_y ||x-y||_\infty, sup_y inf_x ||y-x||_\infty\},$$
$$d_B(X,Y) = inf_\gamma sup_x ||x-\gamma(x)||_\infty,$$

за $x \in X$ и $y \in Y$ и све бијекције $\gamma : X \to Y$.

Дефиниција 3.3.3.2. Простор X триангуларизабилан простор ако постоји њему хомеоморфан симплицијални комплекс.

Дефиниција 3.3.3.3. Ниво скупови (енг. level sets) funkcije $f : X \to \mathbb{R}$, су скупови облика $f^{-1}((-\infty, \alpha])$ или $f^{-1}([\alpha, +\infty))$, где је ниво α неки реални број.

То нам даље даје теореме о стабилности перзистентних дијаграма под малим пертурбацијама тачака података што је фундаменталан резултат за примену перзистентне хомологије. Теорему о стабилности [6] наводимо без доказа.

Теорема 3.3.3.1. Нека је X триангуларизабилан простор и нека су $f, g : X \to \mathbb{R}$ непрекидне функције чији ниво скупови генеришу коначно димензионалне хомолошке групе и постоји коначно много хомолошких критичних вредности. Тада важи:

$$d_B(D(f) - D(g)) \leq ||f - g||_\infty.$$

Постоји још приступа проблему статистичке анализе, али сви су још увек у фази развијања и пуно је могућности за нове идеје. Неки од њих се заснивају на пресликавању простора перзистентних дијаграма у познате просторе, нпр.Банахов. То су перзистентни пејзажи (енг. landscapes) [17], Бетијеве криве, перзистентне слике и кернелизација. Иако нам нису од пресудне важности за наставак овог рада, овде их наводимо као погодне за манипулацију статистичким алатима, као и многим методама машинског учења.

Глава 4

Перзистенција и кластеровање

4.1 Идеја и мотивација

Веома је тешко дати прецизну уопштену дефиницију самог кластера, а топологија нам може и ту помоћи. Неформално, кластеровање се односи на процес поделе скупа података на више делова, односно кластера, који се препознатљиво разликују једни од других. У контексту коначних метричких простора, то отприлике значи да су тачке у кластерима ближе једна другој него што су тачке у различитим кластерима. Кластеровање можемо сматрати статистичким панданом геометријске конструкције повезаних компоненти простора, што је основа алгебарске топологије. Постоји много алгоритама који граде кластере на основу метричких информација, као што су к-средина, спектрално кластеровање итд. Иако је кластеровање, очигледно, веома важан део анализе података, начини на који се формулише и примењује препуни су вишезначности. Углавном су потешкоће са којима се суочавамо избор различитих параметара и прагова, као и недостатак робусности.

Ненадгледано учење (енг. unsupervised learning) је важан и користан алат за тумачење података у разним областима примене. Иако је природно груписање човеку у великом броју случајева очигледно, проблем кластеровања нема јединствено решење. Ипак, његова важност као алата за анализу података, порасла је због повећане доступности високодимензионалних скупова података јер је директна интерпретација таквих података тешка, ако не и немогућа.

Уобичајено је да се скуп података састоји од узорака извучених из неке непознате функције густине f и да је крајњи циљ анализе разумевање структуре те густине. С обзиром да функција густине обично није позната,

мора се оценити из доступних узорака. Методе кластовања се стога ослањају на оцене густине, које се сврставају у две основне категорије: параметарске оцене, које узимају параметарску фамилију функција као модел за густину; непараметарске оцене, које се изводе на основу локалног понашања функције густине. Методе засноване на параметарским оценама користе знање о густини да би постигле боље резултате, док су непараметарске методе општије јер нису везане за неки одређени модел густине.
Кластери узорака који долазе из функције густине f се могу идентификовати уз помоћ мода функције f. Интуитивно, кластер је скуп свих тачака које се уливају у исти локални максимум дуж тока дефинисаног векторским пољем градијента f. Још један начин на који можемо посматрати кластере јесте као басене атракције атрактора који одговарају модама функције густине. Следи и формална дефиниција:

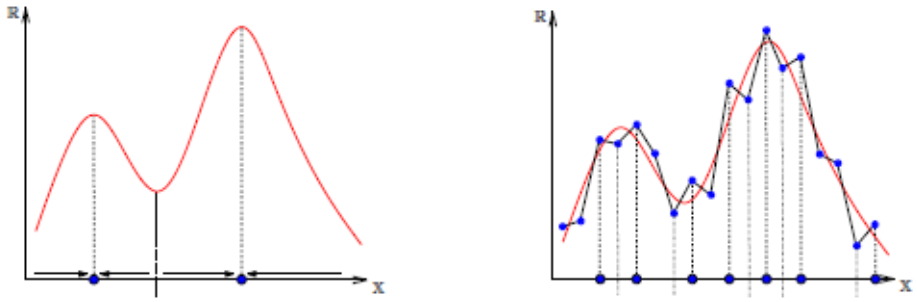
Дефиниција 4.1.1. Атрактор је стање (скуп) у који динамички систем (или његов подскуп) еволуира након довољно времена. Басен атракције B_p неког атрактора p је скуп тачака који ће еволуирати у њега самог.

Овај појам кластера није нов. Већ је предложен у алгоритму градијентног успона (енг. graph-based gradient ascent algorithm) [5] и користи се у бројним алгоритмима за детекцију мода, укључујући Mean-Shift [18] и његове наследнике. Уобичајени проблем са којим се суочавају ове методе је да су градијент и тачке екстрема функције густине f величине које су јако нестабилне под, чак и произвољно малим, пертурбацијама f. Будући да нам је f често непознато, зависимо од оцене густине f, чији се атрактори ретко поклапају са онима стварне густине. Погледајмо слику 4.1 (в. [2]) за илустрацију проблема. У неким методама да би се избегао овај проблем ради се тзв. заглађивање оцене густине (енг. smoothing) , што доводи до следећег проблема апрокцимације, а то је колико је заглађивања потребно.

Дефиниција 4.1.2. Надниво скупови (енг. superlevel sets) funkcije $f : X \to \mathbb{R}$, су скупови облика $f^{-1}([\alpha, +\infty))$, где је ниво α неки реални број. Означавамо их са F^α.

Уместо да директно проучава градијент функције f, тополошка перзистенција проучава развој топологије надниво скупова од f:

Дефиниција 4.1.3. Филтрација генерисана надниво скуповима F^α функције $f : X \to \mathbb{R}$ је низ комплекса у X добијених помоћу F^α, где се параметар α креће од $+\infty$ до $-\infty$.

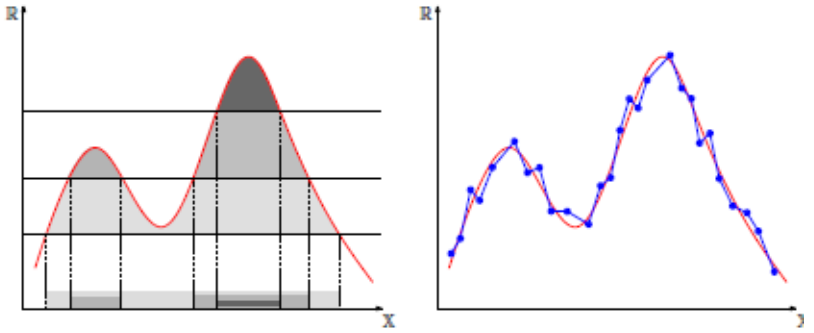


Слика 4.1: На левој страни приказана је функција густине са две екстремне вредности и подела на два њена басена атракције, а на десној одговарајуће вредности за линеарну интерполацију функције густине.

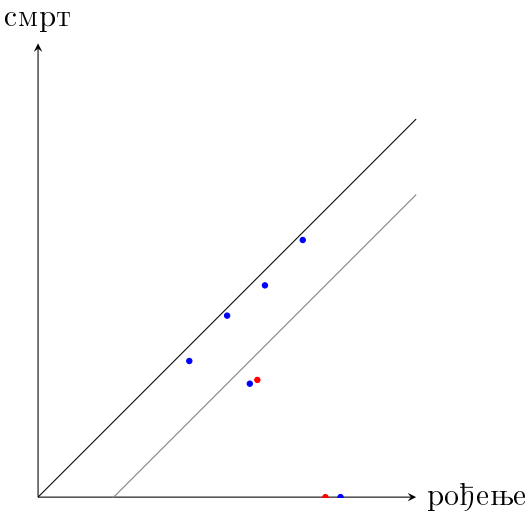
Дакле имамо дефинисано тотално уређење на X.

У контексту кластеровања, нас углавном занима перзистентна хомологија надниво скупова, тј. да ли су и како повезани надниво скупови. То је заправо посебан случај 0-димензионалне теорије перзистенције где као хомолошке класе посматрамо повезане компоненте простора. Слика 4.2 (в. [2]) приказује повезане компоненте три надниво скупа. Свака компонента се појављује када се достигне локални максимум посматране функције.

Штавише, перзистенција намеће строгу хијерархију компонентама: када се две од њих повежу једна са другом на неком надниво скупу од f, за компоненту генерисану нижим врхом каже се да је спојена у ону генерисану вишим врхом. Тада се свакој компоненти C може доделити животни век. Ако посматрамо као тачку p у равни: апциса p је време у којем се компонента C појављује у породици надниво скупова од f; ордината од p је време у коме се C спаја у другу компоненту генерисану неким вишим врхом функције f. Разлика $p_x - p_y$ је мера значајности компоненте C (у даљем тексту само значајност), или еквивалентно, важности њој одговарајуће моде. Колекција таквих тачака назива се 0-ти (нулти) перзистентни дијаграм функције f. Обратимо пажњу на то да је у овом случају вредност перзистентности компоненте једнака разлици тренутка рођења и тренутка смрти. То је због опадајуће скале параметра α помоћу које конструишемо филтрацију у том поретку. Наравно, све карактеристике остају на снази јер имамо само симетрију у односу на дијагоналу перзистентног дијаграма. Погледајмо слику 4.3 за илустрацију упоредних дијаграма. Уопштење перзистентних дијаграма за произвољно димензионалну теорију перзистентности се налази у Глави 3 овог рада где је детаљно објашњена њихова конструкција.



Слика 4.2: Еволуција надниво скупова функције густине и њене оцене



Слика 4.3: Упоредни перзистенти дијаграми за функцију густине (црвене тачке) и за њену оцену (плаве тачке)

У примеру са слике 4.3, упоређујући перзистентне дијаграме густине f и оцене од f, може се видети да су тачке удаљене од дијагонале, које одговарају веома истакнутим екстремима f, добро очуване под пертурбацијом, за разлику од тачака близу дијагонале, које одговарају незначајним екстремима и због тога се могу сматрати шумом. Ово својство стабилности је основни резултат теорије хомолошке перзистенције. Оно оправдава горе објашњену употребу перзистенције у контексту анализе података: када нам права густина $f : \mathbb{X} \to \mathbb{R}$ није дата, и даље је могуће добро апроксимирати понашање значајних екстрема функције f помоћу дијаграма перзистенције њене оцене.

4.2 Компарација са познатим методама кластеровања

Овде наводимо неколико стандардних метода кластеровања и предности ToMATo алгоритма у односу на њих:

- K-средина [19] је једна од метода која се најучесталије користи. У складу са унапред задатим фиксним бројем кластера, покушава да постави центре кластера и да одреди припадност одговарајућим кластерима тако што минимизује суму квадратних растојања до центра унутар сваког кластера. Познато је да је овај проблем минимизације НП-тежак проблем у рачунарској комплексности, па се неретко прибегава итеративном EM (енг. Expectation-Maximization) алгоритму који је оптималнији. За њега је загарантовано да ће конвергирати барем неком локалном минимуму, међутим овај минимум не мора бити глобални и у зависности од тога се могу добити скроз другачији резултати на истим подацима. Још један проблем алгоритма к-средина је да лоше ради на изразито неконвексним кластерима.

- Спектрално кластеровање [20] посебно је прилагођено за рад на неконвексним подацима. Ова метода користи спектар (сопствене вредности) и матрицу сличности података да би се редуковала димензионалност пре самог кластеровања. Матрица сличности је дата као улаз који се састоји од вредности мере сличности сваког пара тачака у скупу података. Мера сличности је функција са реалним вредностима која квантификује колико су нека два објекта међусобно слична. Израчунава се Лапласијан и његови сопствени вектори, а потом се примењује нека од основних метода (нпр. к-средина). Декомпозиција Лапласијана може правити велики проблем са порастом броја података. Спектар Лапласијана може бити показатељ тачног броја кластера, међутим овај приступ је веома осетљив на велика одступања и присуство шума у подацима и у тим случајевима се број кластера не може јасно одредити.

- Технике кластеровања засноване на густини претпостављају да су подаци узорковани из неке непознате функције густине f. Груписање тада постаје проблем апроксимације функције f. Популарни приступ је посматрање функције густине на неком фиксном нивоу и потом се повезане компоненте надниво скупа узимају као кластери, а остатак

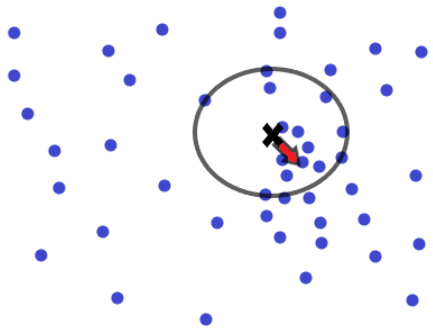
података као шум. Нажалост, због употребе фиксног прага вредности густине, ове технике не реагују добро на хијерархијске скупове података у којима имамо различите нивое груписања.

Дакле, мане горе споменутих метода су углавном неприлагођеност неконвексним подацима, осетљивост на велика одступања, комплексност итд. ToMATo алгоритам је у предности што се свих тих недостатака тиче, а ову тврдњу ћемо оправдати и математички, као и кроз примере у последње две главе овог рада.

Већ смо споменули у уводу да метода ToMATo комбинује две фазе: тражење мода и спајање кластера. Зато ћемо у наредна два потпоглавља посебно издвојити допринос перзистенције техникама попут тражења мода и хијерархијског кластеровања.

4.2.1 Тражење мода

Тражење мода је још један популаран приступ који се састоји у откривању локалних мода функције коју посматрамо. Даље се те моде користе као центри кластера, а подаци се деле у складу са њиховим басенима атракције које смо већ дефинисали. Прецизан појам басена атракције B_p атрактора p варира од случаја употребе, али увек важи да B_p одговара подскупу тачака које у неком тренутку достижу p неком грубом шемом градијентног успона. То илуструјемо на следећој слици 4.4:



Слика 4.4: Један корак у алгоритму детекције мода функције густине, помоћу шеме градијентног успона, на подацима представљеним графом суседа - померање посматране тачке (ознака **x**) ка центру масе свих посматраних суседа у кругу

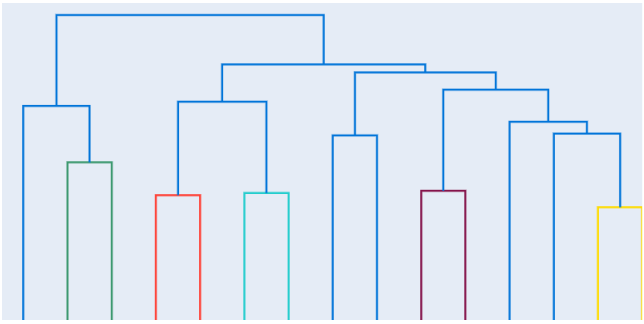
Као што смо већ нагласили, уобичајен проблем са којим се суочавају ове технике је да су градијент и екстремне тачке функције густине веома нестабилне, па њихова апроксимација добијена помоћу оцене густине

може довести до непоузданих резултата. Перзистенција нам помаже тако што детектује нестабилне кластере након њихове поделе. Спајањем таквих кластера добијамо на стабилности резултата. Детекцију нестабилних кластера вршимо посматрањем перзистентног дијаграма филтрације надниво скупова који нам даје значајност сваке моде функције као перзистенцију 0-димензионалне хомолошке класе. Поред тога, имамо и хијерархију мода која нам каже које компоненте су повезане, и у ком кораку филтрације надниво скупова. Ове две информације које добијамо из теорије перзистенције, значајност и хијерархију мода, користимо за унапређење технике тражења мода функције густине.

4.2.2 Хијерархијско кластеровање

Како нам сам назив каже, хијерархијско кластеровање даје хијерархију угњеждених кластера. То је група метода прилагођених за кластеровање података са вишеструким нивоима. Структура таквих података је, такође, хијерархијска. Заправо сваки кластер се може индивидуално даље делити на мање кластере који поседују заједничке особине нивоа родитељског кластера, али се разликују по неком новом нивоу особина. Најпознатија је агломеративна техника хијерархијског кластеровања која креће од тога да је свака тачка података засебан кластер и у сваком кораку спаја два кластера која су најсличнија под неким условом. Квалитет резултата највише зависи од дефинисања мере сличности између кластера. Највећа мана је што се грешке (углавном због утицаја веома далеких тачака података на мере растојања) направљене у раним корацима, тешко или никако, не могу исправити у даљој хијерархији. Хијерархију репрезентујемо као стабло које називамо дендрограм.

Једна од метода агломеративног приступа је једноструко повезивање [1]



Слика 4.5: Дендрограм агломеративног кластеровања над узорком величине 15

(енг. single-linkage) . Она у сваком кораку спаја два кластера који садрже најближи пар тачака. Треба напоменути да се перзистенција већ имплицитно користи у овим методама. Графови настали из алгоритма једноструког повеивања су исто што и симплицијални комплекси димензије један у Виеторис-Рипс филтрацији. Такође, дендрограми нису ништа друго до алтернативни прикази перзистентних дијаграма. Тада разматрана функција није густина већ удаљеност тачака. Хијерархија кластера индукована на овај начин даје сумарну анализу података из које се могу закључити најбољи прагови за даље кластеровање.
Дендрограми, као и перзистентни дијаграми, су стабилни под пертурбацијама података. Доказ за ту тврдњу се може пронаћи у литератури [21].
Нажалост, теоријске гаранције за број кластера и стабилност, доказане су само у случају ограниченог Хаусдорф модела шума [22]. Овај модел је веома рестриктиван јер ограничава појаву, горе поменутог, проблема ланчаног ефекта тј. утицај великих одступања у подацима (енг. outliers), а томе су подаци веома подложни у пракси. Ту нам опет помаже теорија перзистенције која гарантује стабилност перзистентних дијаграма и под слабијим условима.

4.3 Mapper алгоритам

Иако акценат овог рада није на Mapper алгоритму, уводимо га као додатну занимљиву идеју и још један начин примене тополошке анализе података. Алгоритам је први пут објављен у следећој литератури [8]. Већ је достигао успешну примену у финансијама, детекцији превара, а посебно биологији и медицини [23].
Mapper је комбинација редукције димензије, кластеровања и графова. Углавном се користи за:

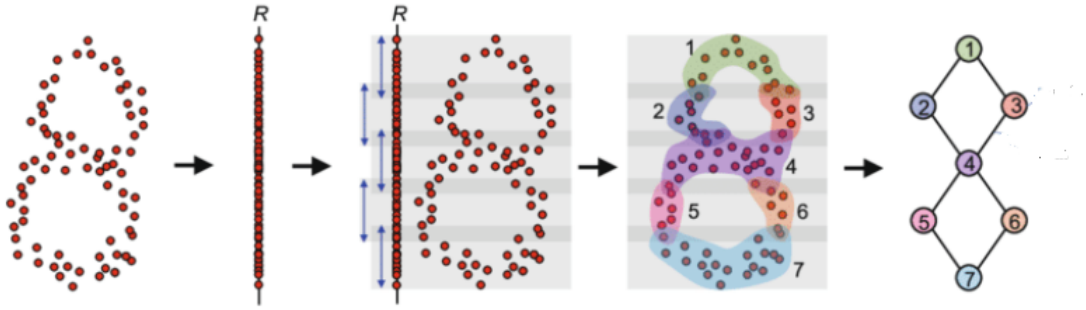
- визуализацију облика података
- детекцију кластера и занимљивих тополошких структура у подацима
- селекцију карактеристика (атрибута) које најбоље описују податке и повећавају интерпретабилност модела

Дефиниција 4.3.1. Отворени покривач тополошког простора X је колекција $U = (U_\alpha)_{\alpha \in A}$ отворених скупова $U_\alpha \subseteq X$, $\alpha \in A$ где је A скуп такав да важи $\bigcup_{\alpha \in A} U_\alpha = X$.

Опишимо сада сваки корак алгоритма:

1. Дефинишемо филтер функцију (енг. lens) на простору података које посматрамо $f : X \to Z$ која пресликава податке у мању димензију
2. Конструишемо покривач $(U_\alpha)_{\alpha \in A}$ скупа Z тј. важи $\bigcup_{\alpha \in A} U_\alpha = Z$
3. Одредимо $X_\alpha = f^{-1}(U_\alpha)$ и на сваки од њих посебно применимо кластеровање
4. Конструишимо граф коме су чворови добијени кластери, а гране правимо између кластера који имају заједничке тачке

Илустрацију ових корака дајемо на слици 4.6 (в. [24]).



Слика 4.6: Mapper алгоритам примењен на дводимензионалне податке (тачке скроз лево); филтер функција је пројекција на y осу и покривач се састоји од 5 интервала; скроз десно је добијени граф као резултат алгоритма

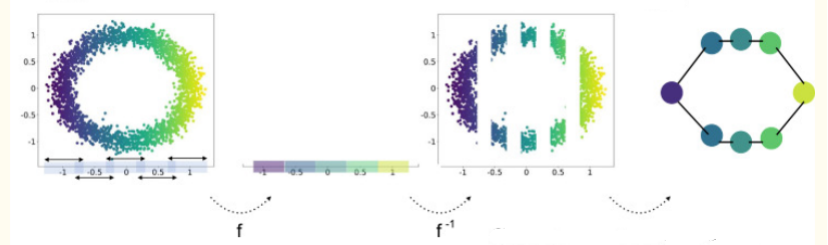
Дефиниција 4.3.2. За дати покривач U тополошког простора X, $U = (U_\alpha)_{\alpha \in A}$, његов нерв дефинишемо као апстрактни симплицијални комплекс $N(U)$ коме је скуп темена U и за кога важи:

$$\sigma = [U_{i_0}, ..., U_{i_k}] \in N(U) \text{ акко } \bigcap_{j=0}^{k} U_{i_k} \neq \emptyset.$$

Приметимо да чворове добијеног графа можемо посматрати као 0-димензионалне симплексе, а гране као 1-димензионалне симплексе. Нека је $f^\star(U)$ покривач од X_α. Тада је резултат Mapper алгоритма нерв тог покривача тј. $M(U, f) = N(f^\star(U))$. И то је наша веза Mapper методе и хомолошке теорије.

Избор филтер функције је очигледно веома битан. Ако желимо једнодимензионални случај, обично се узима нека пројекција (као на слици 4.6), ексцентрицитет или растојање од неке конкретне тачке или фиксног броја тачака. За вишедимензионалне случајеве типичан избор су методе редукције димензије (нпр. PCA, t-SNE). Суптилнији резултати могу се постићи

пројектовањем на неки латентни простор добијен помоћу аутоенкодера. Све у свему, избор филтера се врши у зависности од типа података које анализирамо. На пример, на следећој слици 4.7 (в. [25]) филтер је растојање од тачке скроз лево на кругу, односно најтамније плаве тачке са слике. На тај начин нам резултујући граф открива облик скупа података.



Слика 4.7: Mapper алгоритам примењен на круг, а филтер функција је растојање од најтамније плаве тачке, скроз лево на кругу

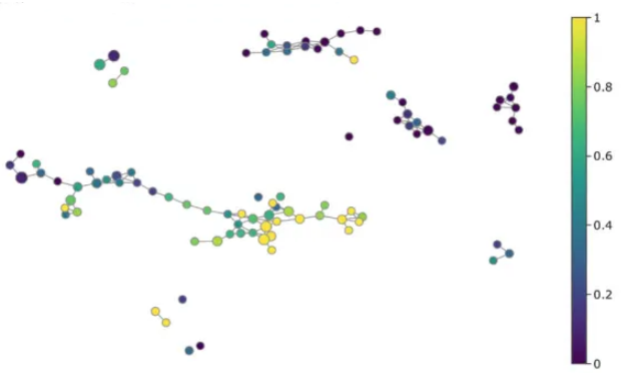
Што се тиче покривача, стандардни избор је колекција d-димензионалних интервала исте дужине који се међусобно преклапају са задатим процентом преклапања.
Методу кластеровања опет бирамо сами у зависности од случаја употребе. Важно је да се не морамо уско ограничавати на једну методу или на исте параметре за сваки интервал у покривачу.

Закључујемо да је велика предност Mapper алгоритма у произвољности избора методе кластеровања, избора филтер функције и покривача. Такође, граф који добијамо нам открива тополошку структуру података коју много лакше интерпретирамо на тај начин. Главна предност Mapper алгоритма је у томе што се, након избора филтер функције, покривач враћа инверзном функцијом у првобитни простор високе димензије и кластеровање се ту дешава. Тако не долази до губитка информација као код класичних метода редукције димензије.

Добијени граф нам омогућава да видимо зашто су неке групе међусобно повезане. Тако можемо да видимо које карактеристике података су значајне и извршимо селекцију атрибута. Бојењем графа вредношћу атрибута можемо видети које вредности одређених атрибута праве групе у подацима или се пак издвајају. Коначно, визуализација графом нам помаже и у интерпретацији модела тако што видимо на којим местима модел највише греши или најбоље предвиђа податке.

Пример 4.1. Овим примером илуструјемо на који начин уз помоћ Mapper алгоритма вршимо селекцију атрибута. Посматрамо скуп који се користи за предвиђање болести срца код пацијената на основу атрибута као што су старост, холестерол, крвни притисак итд. На слици 4.8 (в. [23]) Mapper граф, коме су чворови обојени степеном срчане болести, је раздвојио различите групе у подацима. Тако можемо обојити граф и вредностима било ког атрибута и видети који то атрибути корелишу са циљном променљивом, које вредности им се издвајају у посебне групе и како објашњавају податке. Употребљена филтер функција је дводимензионална UMAP пројекција [26], покривач има 10 дводимензионалних интервала са 20 процената преклапања и DBSCAN [27] је алгоритам кластеровања. ∇

Видимо да резултујући граф нуди флексибилност у анализи података. Међутим, имамо само приказ који је резултат фиксног конструисаног покривача. Већ смо рекли да је веза Mapper алгоритма и хомолошке теорије у томе што резултат алгоритма можемо посматрати као симплицијални комплекс. То је и била инспирација за Multiscale mapper алгоритам. У том алгоритму угњеждени низ покривача индукује филтрацију симплицијалних комплекса. Даље се добијена структура проучава помоћу перзистентне хомологије и њених метода визуализације попут перзистентног дијаграма.



Слика 4.8: Mapper граф коме су чворови обојени степеном срчане болести

Глава 5

ToMATo алгоритам

5.1 О алгоритму

У трећој глави овог рада смо навели фундаменталну теорему о стабилности перзистентних дијаграма (Теорема 3.3.3.1.). Њен доказ се може пронаћи у недавном раду [6] који нам говори више о томе. Нама је довољан њен следећи специјалан случај:
Нека је f непозната реална функција дефинисана над непознатим простором X и дат је облак тачака L из простора X заједно са апроксимацијом $\tilde{f}$ од f над тачкама од L. Резултат теореме омогућава веродостојну апроксимацију перзистентног дијаграма помоћу конструкције графа суседа над L и оцене густине $\tilde{f}$. Следи:

$$d_B(D(f) - D(\tilde{f})) \leq ||f - \tilde{f}||_\infty,$$

где је d_B ознака за bottleneck растојање (в. дефиницију 3.3.3.2.).
У претходном резултату [28], ово својство апроксимације важи ако се у простору X постигне нека минимална густина узорковања, што није погодно у контексту кластеровања, где је облак тачака L узоркован у складу са функцијом густине f. Први циљ јесте да покажемо да ова стабилност остаје на снази и под слабијом верзијом апроксимације. Ти слабији услови су погодни за кластеровање јер је под тим условима узорковање директно пропорционално регионима функције густине.

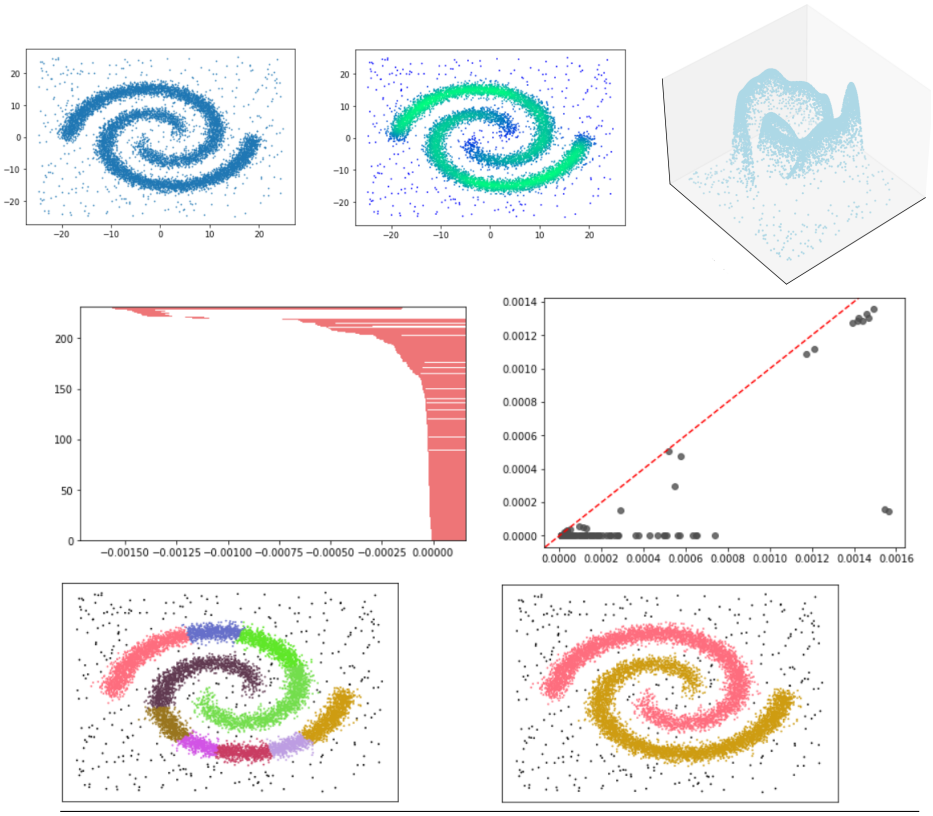
Захваљујући, горепоменутом, новом резултату о стабилности, предложена је нова метода кластеровања која комбинује тражење мода (енг. mode-seeking) засновано на графу, помоћу стандардне шеме успона (енг. graph-based hill-climbing scheme), и спајање (енг. merging) кластера засновано на теорији перзистентности. У фази детекције мода: за задати параметар $\sigma \geq 0$, конструишемо граф суседа G, повезивањем сваког пара тачака под

неким условом (нпр. међусобно растојање највише σ или k најближих суседа); затим, за оцену функције густине f, конструишемо разапињућу шуму графа G повезивањем сваког чвора v са својим суседом у G коме је вредност оцене $\tilde{f}$ највећа; ако сви суседи од чвора v имају мање вредности $\tilde{f}$ од њега, тада је v проглашавамо модом од $\tilde{f}$ и оно постаје корен неког стабла у шуми. Као што је поменуто у Глави 4 и илустровано на слици 4.2, овај алгоритам је веома осетљив и на мале пертурбације, а то не желимо у пракси због апроксимације функције густине.
Новост приступа је у начину спајања кластера у другој фази алгоритма где користимо перзистентну хомологију и на тај начин повећавамо стабилност резултата. Већ смо споменули да значајност кластера, односно повезане компоненте, одређујемо као растојање од дијагонале тачке на перзистентном дијаграму која представља ту компоненту. За параметар τ, спајамо сваки кластер значајности мање од τ са његовим родитељем у хијерархији кластера дефинисаној помоћу перзистенције. Перзистенција и хијерархија се могу израчунати у ходу током прве фазе, под условом да су чворови сортирани по опадајућој вредности $\tilde{f}$. Заправо, видећемо да прва и друга фаза могу бити извршене симултано.
Излаз алгоритма су кластери значајности бар τ. Додатно, као визуелизацију имамо и перзистентни дијаграм $D_0\tilde{f}$ када је параметар τ једнак $+\infty$ (то јест, када је сваки кластер спојен са својим родитељем у хијерархији) јер тако добијамо информације о целој хијерархији.
Параметри σ, односно k, и τ су врло различите природе. Параметром σ задајемо скалу на којој ће подаци бити посматрани у зависности од тога какву структуру у њима тражимо. У пракси није лако то одредити без претходног знања о подацима, али тада се можемо ослонити на информације које нам дају дендрограми. Релевантне вредности за параметар τ читамо са перзистентног дијаграма $D_0\tilde{f}$ који нам визуелно издваја значајне компоненте и тако добијамо минималну вредност значајности τ коју даље користимо у спајању кластера.

Валидност овог алгоритма је теоријски загарантована. Под условом да имамо довољно велики облак тачака и уз одговарајући избор параметра σ гарантујемо да је добијени перзистентни дијаграм оцене густине близу перзистентном дијаграму стварне густине у њиховом простору. Под претпоставком да постоји јасан јаз између значајних и незначајних класа у D_0f, можемо доказати да се тај јаз уочава и на перзистентном дијаграму који добијамо из прве фазе алгоритма. На тај начин бирамо вредност параметра τ тако да у другој фази алгоритма добијемо значајне кластере. То даје

теоријски смисао појму тачног броја кластера. Поред ове гаранције о броју кластера, показаћемо и корелацију просторне локације кластера са басенима атракције локалних максимума функције густине f. Ови резултати су детаљно описани у наредним поглављима овог рада (Теорема 5.4.1. и Теорема 5.4.2.).

Пример 5.1.



Слика 5.1: ToMATo алгоритам примењен на синтетички дводимензионални облак тачака са додатним шумом

Посматрајмо дводимезионални синтетички конструисан скуп података у облику две учешљане спирале са додатним шумом. Облак се састоји од 10 хиљада тачака од чега 5% додатног шума. На слици 5.1 илуструјемо кораке припреме података и примене ToMATo алгоритма на те податке:

- у првом реду имамо иницијалне податке, податке обојене вредностима оцене функције густине, као и оцену функцију густине представљену на z-оси
- у другом реду се налазе перзистентни бар-код и дијаграм - прво приметимо две издвојене тачке које су рано рођене у филтрацији и које

су значајно удаљене од дијагонале; видимо да имамо доста тачака веома блиских дијагонали које проглашавамо незначајним и спајамо их накнадно, али имамо, такође, доста бесконачно перзистентних тачака које су рођене касније, што нам говори да су то тачке са малим вредностима оцене густине, али су веома удаљење у нашим подацима па су због тога као издвојене опстале у филтрацији; користили смо Рипсов граф са параметром $\sigma = 1$ (в. дефиницију 5.2.2.4.)

- коначно, у трећем реду добијамо резултат након примене само стандардне шеме успона засноване на графу (лево) и резултат након спајања кластера помоћу перзистенције (десно) где смо за вредност параметра спајања узели $\tau = 0.0012$ (в. Главу 6 за поступак избора) и на тај начин издвојили два значајно перзистентна кластера рођена раније и удаљене бесконачно перзистентне тачке рођене касније у филтрацији

5.2 Теоријске основе

У овом поглављу се упознајемо са појмовима неопходним за пролазак кроз све кораке алгоритма у наредном поглављу. Даћемо математичке дефиниције и објашњења појмова редом којим их касније користимо.

5.2.1 Риманова многострукост

Дефиниција 5.2.1.1. Нека је T_pX тангентни простор глатке (диференцијабилне) многострукости X у тачки $p \in X$. Риманова метрика свакој тачки $p \in X$ додељује скаларни производ $g_p : T_pX \times T_pX \to \mathbb{R}$. Глатка многострукост X на којој је дефинисана ова метрика g назива се Риманова многострукост и обележавамо је са (X, g).

Дефиниција 5.2.1.2. Геодезијска удаљеност две тачке на многострукости је дужина локално најкраће путање на тој многострукости која повезује те две тачке.
За $x \in X$ и реални радијус $r \geq 0$, геодезијска лопта је подскуп $B_X(x, r)$ од X за који важи $B_X(x, r) = \{y \in X | d_X(x, y) \leq r\}$.

Надаље у раду, ако није другачије назначено, X означава Риманову многострукост, а d_X означава геодезијску удаљеност на њој. За све довољно мале вредности $r \geq 0$, познато је да је лопта $B_X(x, r)$ јако конвексна,

односно важи: за сваки пар тачака y, y' у $B_X(x,r)$ постоји јединствена најкраћа путања у X између њих која припада $B_X(x,r)$. Својство да за сваку тачку на X постоји позитиван радијус за који је лопта јако конвексна, важи на пример када је X компактан или када је $X = \mathbb{R}^m$.

За m-димензионалну Риманову многострукост, са H^m обележавамо m-димензионалну Хаусдорфову меру индуковану Римановом метриком ([16]).

Дефиниција 5.2.1.3. Под функцијом густине вероватноће над X заједно са H^m подразумевамо ненегативну функцију $f : X \to \mathbb{R}$ која је интеграбилна у односу на меру H^m и важи $\int_X f \, dH^m = 1$.

У остатку рада сматрамо да су све функције густине вероватноће овако дефинисане на X.

5.2.2 Надниво филтрација

До сада су нам већ познати појмови филтрације и перзистентне хомологије који су прецизно уведени у Глави 3 овог рада. Сада ћемо покрити и специјалан случај филтрације који користимо у имплементацији ToMATo алгоритма.

Посматрамо надниво скупове F^α функције густине $f : X \to \mathbb{R}$ као затворене подскупове од X. Приметимо да овде имамо непрекидно индексирање скупова у $\mathbb{R}$. И поред тога, у већ поменутом раду о стабилности перзистентних дијаграма [6], постоји тврђење које нам говори да се скуп свих перзистентних хомолошких група генерисаних надниво скуповима функције под благим условима (коначно димензионалне хомолошке групе и коначан број хомолошких критичних вредности функције) може представити перзистентним дијаграмом, као што то радимо и у дискретном случају.

Ми се у алгоритму првенствено бавимо 0-димензионалном перзистентном хомологијом, која представља повезаност простора. У том случају се онда услови горњег тврђења своде на претпоставку да f има само коначан број локалних максимума.

Да бисмо имплементирали алгоритам, морамо конструисати дискретне структуре на подацима. Како се бавимо 0-димензионалном перзистентном хомологијом, конструишемо једноставне неоријентисане графове који су заправо специјални случајеви апстрактних симплицијалних комплекса. Сада дефинишемо граф који ћемо користити је Рипсов граф (Елијаху Рипс[1]), познат и као σ-граф суседа:

[1]Eliyahu Rips - израелски математичар

Дефиниција 5.2.2.4. Нека је дат коначан облак тачака L, метрички простор (X, d_X) и реални параметар $\sigma > 0$. Тада је Рипсов граф $R_\sigma(L, d_X)$ граф са скупом чворова L, а ивице одговарају паровима тачака $x, y \in L$ које задовољавају услов $d_X(x, y) \leq \sigma$.

Убудуће можемо изостављати d_X јер знамо са којом метриком радимо па пишемо само $R_\sigma(L)$. За функцију $f : X \to \mathbb{R}$ и ниво $\alpha \in \mathbb{R}$ са L^α означавамо надниво скуп F^α функције f над облаком тачака L тј. $L^\alpha = L \cap F^\alpha$.

Дефиниција 5.2.2.5. Рипсова надниво филтрација функције f, за фиксирано σ, у ознаци $R^f_\sigma(L)$, је угњеждени низ подскупова од $R_\sigma(L)$:

$$R^f_\sigma(L) = \{R_\sigma(L^\alpha)\}_{\alpha \in \mathbb{R}},$$

где се параметар α креће од $+\infty$ до $-\infty$.

Такође, убудуће можемо изостављати да се ради о Рипсовој надниво филтрацији јер се то подразумева, ако не нагласимо другачије.
Приметимо да иако α пролази кроз цео скуп $\mathbb{R}$ до промене у $R_\sigma(L^\alpha)$ долази само ако имамо нови чвор v за који важи $f(v) = \alpha$. Дакле, филтрација $R^f_\sigma(L)$ садржи коначан број различитих графова.

5.2.3 Морсеова теорија

У математици, посебно у диференцијалној топологији, Морсеова (Марстон Морзе[2]) теорија омогућава анализирање многострукости посматрањем диференцибилних функција на њој. Морсеова теорија се бави декомпозицијама на многострукостима и добијањем информација о њиховој хомологији. Ми ћемо се у наредном поглављу, у непрекидној поставци алгоритма, упознати са једним таквим партиционисањем.

Дефиниција 5.2.3.1. Критична тачка реалне функције $f : X \to \mathbb{R}$ је тачка домена X у којој функција није диференцијабилна или јој је извод једнак нула у тој тачки.
Вредност функције у критичној тачки назива се критична вредност.

Дефиниција 5.2.3.2. Ако је функција $f : X \to \mathbb{R}$ бар C^2 глатка онда је извод функције једнак нула у критичним тачкама. Ако је у критичној тачки m, матрица парцијалних других извода (Хесијан) несинуларна онда је m недегенерисана критична тачка, а у супротном је дегенерисана критична тачка.

[2]Marston Morse - амерички математичар

Дефиниција 5.2.3.2. Бар C^2 глатка функција $f : X \to \mathbb{R}$ на многострукости X је Морсеова функција ако нема дегенерисаних критичних тачака.

Основни резултат Морсеове теорије нам каже да су скоро све функције Морсеове функције. Како нас занима топологија надниво скупова F^α, сада ћемо без доказа дати две важне теореме. Оне говоре о томе да се топологија надниво скупова мења само након проласка α кроз критичну вредност и колика је промена након проласка.

Теорема 5.2.3.1. Нека је дата глатка функција $f : X \to \mathbb{R}$ на многострукости X, нека је за сваки затворени интервал $[a, b] \subset \mathbb{R}$ инверзна слика $f^{-1}([a, b])$ компактан подскуп од X и нека нема критичних тачака у интервалу $[a, b] \subset \mathbb{R}$. Тада је F^a дифеоморфно F^b, односно то је тополошко увлачење F^b у потпростор F^a.

Теорема 5.2.3.2. Нека је дата глатка функција $f : X \to \mathbb{R}$ на n-димензионалној многострукости X и нека је m њена недегенерисана критична тачка и $f(p) = q$. Ако постоји ϵ тако да је инверзна слика $f^{-1}[q-\epsilon, q+\epsilon]$ компактна и не садржи друге критичне тачке осим p, онда важи да је $F^{q+\epsilon}$ хомотопски еквивалентно са $F^{q-\epsilon}$ и додатим Декартовим производом n затворених интервала.

5.3 Алгоритам

Алгоритам можемо посматрати у непрекидној поставци, односно када као улаз имамо непрекидне вредности параметара, или у дискретној поставци, када улазни параметри могу имати коначно много вредности. Прво дајемо интуитивну идеју алгоритма у непрекидној поставци. Затим, у следећем делу дајемо детаље алгоритма у дискретној поставци, као и псеудокод, који се користе у пракси.

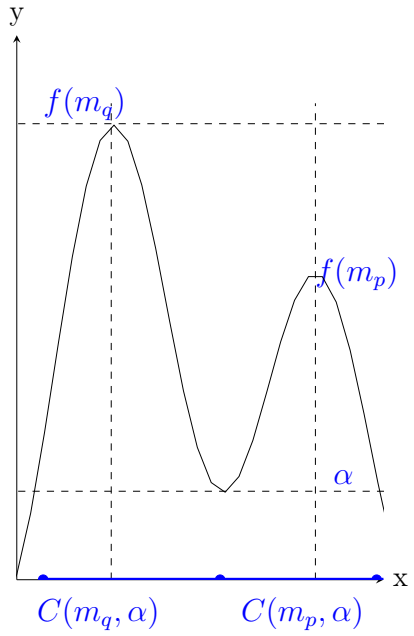
5.3.1 Непрекидна поставка

Дефиниција 5.3.1.1. Узлазна регија критичне тачке m, ознака $A(m)$, је подскуп тачака од X које након довољно времена достижу тачку p крећући се у правцу индукованом пољем вектора градијента од f.
Посматрајмо n-димензионалну Риманову многострукост X и на њој дефинисану Морсеову функцију $f : X \to \mathbb{R}$. Претпоставимо да f има коначан

број критичних тачака.
За све тачке $k \in A(m)$ кажемо да је m њихов корен.

Додатно, под претпоставком да X нема границу и да је функција f одозго ограничена и да је за све $a < b$, $f^{-1}([a, b])$ компактан подскуп од X, онда узлазне регије локалних екстремних вредности функције f покривају X до на подскуп Хаусдорфове мере нула. Зато је природно користити те регије за партиционисање многострукости X.

Нека је за дато $x \in X$ и $\alpha \in \mathbb{R}$, $C(X, \alpha)$ компонентна наднивоа F^α која садржи x. Морсеова теорија нам говори да када се локални максимум m_p од f појави у филтрацији надниво скупова у тренутку $\alpha = f(m_p)$, нова повезана компонента $C(m_p, \alpha)$ се појави у надниво скупу F^α. У терминима хомологије, m_p се назива генератор компоненте рођене у тренутку $f(m_p)$. Тиме смо описали рађање компоненте, односно њој одговоарајуће моде.
Илуструјмо сада тренутке рађања и смрти једне компоненте на слици 5.2. Повезана компонента $C(m_p, \alpha)$ престаје да буде независна компонента од F^α у тренутку када се повеже са другом компонентом генерисаном неким вишим локалним максимумом m_q. У теорији перзистенције кажемо да се компонента $C(m_p, \alpha)$ спојила са компонентом $C(m_q, \alpha)$ у тренутку α. Тада m_p престаје да буде генератор, док m_q остаје генератор и по аналогији постаје корен новог стабла, ознака $m_q = r(m_p)$.



Слика 5.2: Хијерархија мода функције густине одређена перзистенцијом

Дакле, ако у овом случају посматрамо 0-ти перзистентни дијаграм $D_0 f$, рођење генератора m_p одређено је апцисом тачке p на дијаграму са $p_x = f(m_p)$, а смрт њеном ординатом $p_y = \alpha \leq p_x$. Разлика $p_x - p_y$ између времена рођења и смрти зовемо значајност локалног максимума m_p.

За дати праг $\tau \geq 0$, посматрамо само локалне максимуме којима је значајност бар τ. Интуитивно, тачке на X које се налазе у басену атракције B_{m_p} су оне тачке које припадају узлазним регијама од m_p које се евентуално споје помоћу теорије перзистенције са генератором m_p, пре него што се у неком тренутку $\alpha_\tau(m_p)$ споје у компоненту било ког другог генератора значајности бар τ.
Формално, за сваки локални максимум m_q од f, произвољне значајности, итеративно пређимо кроз њихове корене $r(m_q)$ све док не стигнемо до неког значајности бар τ. Означимо добијено мапирање са r_τ^*. Тада басен атракције од m_p можемо дефинисати као унију узлазних регија свих локалних максимума мапираних у m_p са r_τ^*:

$$\forall m_p \; p_x - p_y \geq \tau, B_\tau(m_p) = \bigcup_{r_\tau^*(m_q)=m_p} A(m_q).$$

Приметимо да $B_\tau(m_p)$ садржи $A(m_p)$ јер је m_p фиксна тачка мапирања.

5.3.2 Псеудокод

Пре свих корака алгоритма кластеровања неопходно је да оценимо функцију густине помоћу посматраних података. Када имамо ту оцену, алгоритам узима као улаз n-димензионални вектор вредности $\tilde{f}$ са реалним координатама, $n \times n$ симетричну матрицу D са ненегативним реалним коефицијентима и два реална параметра $\sigma, \tau \geq 0$. Димензија n представља n тачака у облаку L; вектор представља функцију $\tilde{f} : L \to \mathbb{R}$, а поље $D_{i,j} = D_{j,i}$ матрице D даје растојање између i-те и j-те тачке облака L. Како је избор координата произвољан, алгоритам се може применити у било ком метричком простору. На самом почетку, алгоритам израчунава Рипсов граф $R_\sigma(L)$ из улаза D и σ. Затим следе две главне фазе алгоритма у којима поступамо на следећи начин:

1. Прво, сортирамо тачке облака L по опадајућим вредностима функције $\tilde{f}$: за сваки чвор i апроксимирамо градијент функције густине повезивањем i са суседом на графу $R_\sigma(L)$ коме је највећа вредност функције $\tilde{f}$. Ако такав сусед не постоји тј. сви суседи имају ниже вредности, тада i проглашавамо модом функције $\tilde{f}$. Након проласка

кроз све тачке добијамо разапињућу шуму графа $R_\sigma(L)$: свако дрво у овој шуми се може посматрати као аналогон басена атракције у непрекидној поставци алгоритма.

2. Друго, користимо дисјунктни-сет (енг. union-find) структуру [30] за спајање стабала у шуми. Један подскуп e такве структуре представља унију одговарајућих стабала у шуми. Корен подскупа e, ознака $r(e)$, је његов елемент који има највећу вредност функције $\tilde{f}$. По конструкцији, $r(e)$ мора бити корен једног од стабала садржаних у e, па према томе и мода функције $\tilde{f}$ у графу $R_\sigma(L)$. Спајање два подскупа дисјунктни-сет структуре у дискретној поставци аналогно је спајању два басена атракције у непрекидној поставци алгоритма. Подскупове спајамо у складу са хијерархијом одређеном перзистенцијом. Прецизније, још једном пролазимо кроз тачке графа $R_\sigma(L)$ сортиране по опадајућим вредностима функције $\tilde{f}$. За сваки чвор i посматрамо тачке наднивоа од i у $R_\sigma(L)$. Нека је e_i подскуп структуре који садржи i. Ако било која тачка наднивоа повезује e_i са неким другим подскупом e_j, чији корен $r(e_j)$ има мању вредност функције $\tilde{f}$ од корена $r(e_i)$, тада нам хијерархија одређена перзистенцијом говори да спојимо подскупове e_i и e_j. Међутим, овде уводимо модификацију помоћу параметра τ и вршимо спајање само ако је перзистенција корена $r(e_i)$ мања од задатог прага τ. Овај услов се своди на проверу да ли је $\tilde{f}_{r(e_j)} - f_i < \tau$. Када спојимо све незначајне суседне подскупове са e_i, онда проверавамо да ли треба и сам e_i спојити са суседним подскупом $\bar{e}$ који има највећу вредност корена, наравно ако такав сусед постоји и различит је од e_i. Дакле, имамо још један услов који се своди на проверу да ли је $\tilde{f}_{r(e_i)} - f_i < \tau$.

Испод су дати детаљни псеудокови за описане кораке алгоритма. Као што можемо приметити, кораци се могу имплементирати симултано, са само једним проласком кроз тачке на графу $R_\sigma(L)$. Видимо и да граф $R_\sigma(L)$ не мора бити претходно израчунат већ у свакој итерацији рачунамо надниво филтрацију чвора i.

На финалном излазу, алгоритам нам даје колекцију подскупова дисјунктни-сет структуре U, која партиционише облак тачака L у кластере. Заправо, он даје само оне подскупове e структуре U чији корен $r(e)$ задовољава $\tilde{f}_{r(e)} \geq \tau$. Овај услов мотивисан је случајевима када неке тачке које веома одступају од остатка података формирају независне повезане компоненте које се не могу спојити. Такве повезане компоненте имају врло

Алгоритам 1 Кластеровање

Улаз: n-димензионални вектор вредности $\tilde{f}$, $n \times n$ симетрична матрица D, реални параметри $\sigma, \tau \geq 0$.

Сортирај индексе од L тако да $\tilde{f}_1 \leq \tilde{f}_2 \leq ...\tilde{f}_n$;
Иницијализуј дисјунктни-сет структуру U;
for $i = n$ to 1 **do**
    Израчунај надниво скуп $S_i = \{(i, j_1), ..., (i, j_k)\}$ од i у $R_\sigma(L)$;
    if $S_i = \varnothing$ **then** ▷ чвор i је локални максимум од $\tilde{f}$ над $R_\sigma(L)$
        $g(i) \leftarrow$ *нулл*; ▷ g додељује оцену градијента у свакој тачки
        Креирај нови подскуп у структури U који садржи стабло $\{i\}$;
    else ▷ чвор i је локални максимум од $\tilde{f}$ над $R_\sigma(L)$
        $g(i) \leftarrow argmax_{j \in \{j_1, ..., j_k\}} f(j)$;
        Додај чвор i стаблу које садржи $g(i)$;
        $U \leftarrow$ Спајање$(\tilde{f}, U, i, S_i, \tau)$;

Излаз: Колекција подскупова e од U који задовољавају $\tilde{f}_{r(e)} \geq \tau$.

Алгоритам 2 Спајање

Улаз: n-димензионални вектор вредности $\tilde{f}$, структура U, i, $S = \{j_1, ..., j_k\}$, $\tau \geq 0$.

e_i подскуп у U који садржи i;
▷ пронађи подскупове у U који имају непразан пресек са S и чији корени су мање од τ значајни и споји их са e_i
for $j \in \{j_1, ..., j_k\}$ **do**
    e_j подскуп у U који садржи j;
    if $e_j \neq e_i$ i $\tilde{f}_{r(e_j)} - f_i < \tau$ **then**
        Избаци e_j из U и споји га са e_i;
▷ пронађи подскуп $\bar{e}$ у U који има непразан пресек са e_i и чији корен има највећу вредност
$\bar{e} \leftarrow null$
for $j \in \{j_1, ..., j_k\}$ **do**
    e_j подскуп у U који садржи j;
    if $\bar{e} = null$ ili $\tilde{f}_{r(e_j)} > \tilde{f}_{r(\bar{e})}$ **then**
        $\bar{e} \leftarrow e_j$;
▷ споји e_i са $\bar{e}$ ако је и значајност од e_i мања од τ
if $e_j \neq \bar{e}$ i $\tilde{f}_{r(e_i)} - f_i < \tau$ **then**
    Избаци e_i из U и споји га са $\bar{e}$;

Излаз: Ажурирана структура U.

ниске вредности функције густине па их проналазимо горњим условом. Такође, алгоритам нам даје додатне повратне информације о животном веку компонената у структури U у облику колекције интервала, као и хијерархију тих компонената. Када је параметар τ једнак $+\infty$, излазна колекција интервала није ништа друго до 0-ти перзистентни дијаграм филтрације $R_\sigma^{\tilde{f}}(L)$.

5.3.3 Избор параметара

ToMATo алгоритам на самом почетку захтева три улаза: оцену густине $\tilde{f}$, граф суседа и параметар спајања кластера τ. Иако произвољност која је остављена кориснику у избору ових улаза даје алгоритму велику флексибилност, ипак не смемо дозволити да значајно утиче на цену и време извршавања алгоритма. Зато у овом одељку дајемо неке смернице за избор параметара.

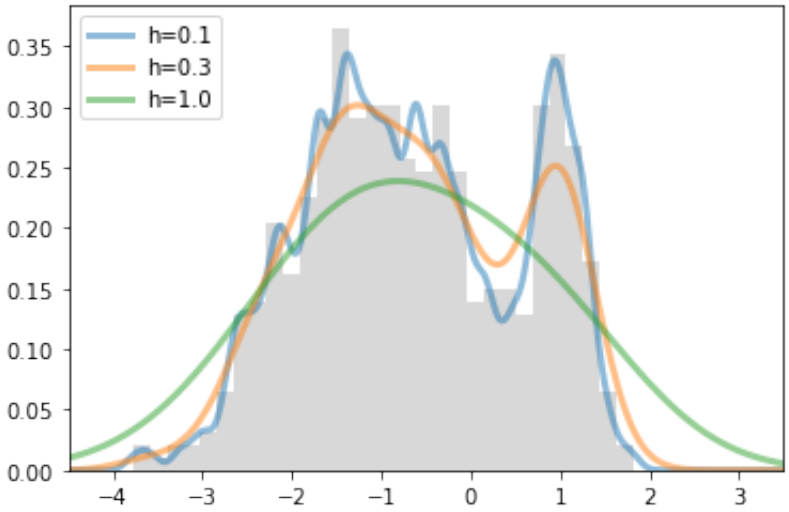
- Оцена густине: ToMATo алгоритам у потпуности даје слободу избора оцењивача функције густине. И поред тога, често се користе методе оцене густине засноване на кернелима.

  $$\tilde{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)$$

  Интуитивно, кернел (као функција језгра) представља функцију сличности. Што је вредност кернела за неке две тачке података већа, то се оне могу сматрати сличнијим. У различитим контекситма, праве се врло различите претпоставке везане за кернеле, па се и сами концепти који се под тим изразом подразумевају врло разликују. Кернели углавном имају одређене параметре, којима се финије подешава њихово понашање. Један од најпознатијих кернела, који смо користили и у примерима у овом раду, јесте Гаусов кернел:

  $$K_h(x) = \frac{1}{(\sqrt{2\pi}h)^n} exp(-\frac{|x|^2}{2h^2})$$

  Параметар h је параметар размере (односно ширина кернела, енг. bandwidth). Генерално, мала вредност параметра h води преприлагођавању, док велика води губитку информација у подацима. Ширина кернела независна је од тачке x па се зато мора водити рачуна о њеној оптималној вредности за све области различите густине у подацима. То се види и на слици 5.3.



Слика 5.3: Оцена густине Гаусовим кернелом са различитим вредностима параметра размере

- Граф суседа: ToMATo алгоритам се у великој мери ослања на информације добијене од суседа одређених графом. Одабир релевантног графа и метрике је проблем са којим се суочавају и многе друге технике кластеровања. У овом раду смо првенствено користили Рипсов граф суседа $R_\sigma(L)$ који зависи само од параметра σ и растојања између тачака података. Његова чисто метричка дефиниција омогућава употребу у произвољним метричким просторима и интерпретацију добијених перзистентних дијаграма захваљујући теоријским гаранцијама добијених за Рипсов граф. Вредност параметра σ директно је пропорционална скали на којој ћемо посматрати податке. Наравно, у зависности и од посматраних података, различити избори σ могу открити и различите структуре. Због тога препоручујемо покретање ToMATo алгоритма на неколико скала. Захваљујући ефикасности алгоритма то неће бити проблем чак ни за велике скупове података. За превелике вредности параметра σ нећемо ухватити стварну структуру, док ће премале вредности дати много бесконачно перзистентних кластера због недовољно повезаних компонената. Посматрајући више скала, компромис тражимо између тих вредности. Постоје још неки популарни избори графова суседа као што су k-најближих суседа (енг. k-NN). Његова главна предност је што остаје проређен без обзира на растојање између тачака података. У последњој глави показаћемо и примену алгоритма са овим графом и добијене резултате. У пракси је највећи изазов пронаћи оптималан избор k суседа. Међутим, ови

емпиријски резултати нису теоријски поткрепљени. Такође, у неким радовима, употребљен је и Делоне граф [14, 15] коме је главна предност јер не зависи од параметра, али то доноси и одређене проблеме чије решавање отвара врата за нове идеје алгоритма.

- Параметар спајања: Током фазе спајања кластера, ToMATo на крају спаја све кластере значајности мање од прага τ. Дакле, избор параметра τ одређује који локални максимуми од $\tilde{f}$ се сматрају битним, а који се третирају као шум. Да бисмо изабрали релевантну вредност за параметар τ, пролазимо кроз ToMATo алгоритам два пута. У првом проласку τ узима вредност $+\infty$, што као резултат даје стандардни 0-ти перзистентни дијаграм скаларног поља $\tilde{f}$ над графом суседа. На тај начин добијамо вредност и значајност сваке моде оцене функције густине $\tilde{f}$, као и њихову хијерархију. Стога тај резултат користимо за избор параметра τ који користимо у другом проласку кроз алгоритам да бисмо добили финално кластеровање. У случајевима када на перзистентном дијаграму имамо велики јаз који одваја мали скуп од m веома значајних мода од остатка структуре, закључујемо да је тачан број кластера вероватно m, па τ постављамо на било коју вредност значајности која се налази у том јазу. Тада финално кластеровање садржи тачно m кластера. Велики јаз можемо једноставно хеуристички открити. У последњој глави овог рада ћемо илустровати ово закључивање помоћу хистограма значајности. Чак и у случајевима када не постоји јасно уочљив јаз на перзистентном дијаграму, он и даље пружа исту везу између избора параметра τ и финалног броја кластера. Избор одређене вредности некада се додатно може утврдити и у зависности од специфичног случаја употребе података које посматрамо.

У пракси се може јавити проблем да нам нису позната геодезијска растојања тачака. Зато је неопходно апроксимирати растојања. На пример, када су тачке података у еуклидском простору $\mathbb{R}^m$ са познатим координатама, а X је непозната подмногострукост, геодезијска растојања у X могу се апроксимирати помоћу удаљености тачака графа на неком погодно изабраном графу суседа. Више о избору графа суседа за апроксимацију геодезијског растојања се може пронаћи у литератури [28].

5.3.4 Комплексност

Као што смо већ рекли, граф суседа $R_\sigma(L)$ не морамо унапред израчунати, јер само рачунамо надниво филтрацију чвора i у свакој итерацији. То значи да је основна употреба меморије $O(n)$, где је n број тачака у облаку L. Сваки чвор у $R_\sigma(L)$ представља нови унос у структури U, док свака ивица у $R_\sigma(L)$ захтева два тражења подскупова у U и потенцијално једно спајање у алгоритму. Како имамо n темена и $m = O(n^2)$ ивица, не може бити више од $n-1$ спајања и $2m$ тражења, па следи да је укупно време извршавања алгоритма $O(n+m\alpha(n))$, где је α инверзна Акерманова (Вилхелм Фридрих Акерман[3]) функција. У пракси, параметар σ се бира довољно мали тако да је $m = O(n)$. Тада време извршавања алгоритма постаје скоро линеарно зависно величини облака тачака што је разумно оптимално.

5.4 Теоријске гаранције

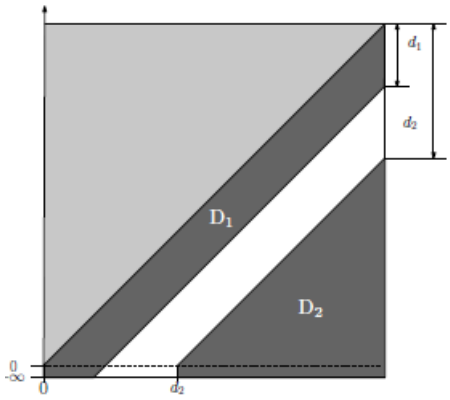
У овом поглављу наводимо две главне теореме које нам дају теоријске гаранције о броју кластера и њиховој просторној локацији, као и скице доказа тих теорема. Због обима математичке теорије која поткрепљује ове теореме, сматраћемо да је читалац већ упознат са том теоријом. Детаљна уводна теорија и докази се могу пронаћи у следећој литератури [2]. Ми ћемо пре навођења теорема, ради лакшег разумевања без неопходног читања додатне литературе, дефинисати само круцијалне појмове које користимо.
Подразумевамо да је X m-димензионална Риманова многострукост, а $f : X \to \mathbb{R}$ Липшиц непрекидна функција густине вероватноће у односу на m-димензионалну Хаусдорфову меру. Подразумевамо и да су тачке облака L узорковане из X као независне и једнако расподељене случајне величине (енг. i.i.d.) са функцијом густине f. Надаље ћемо сматрати да већ имамо узоркован облак тачака, као и њихова геодезијска растојања. Уводимо сада нове круцијалне појмове:

Дефиниција 5.4.1. Нека је $Y \subseteq X$ и нека је реалан параметар $\epsilon > 0$. Тада кажемо да је L геодезијски ϵ-узорак од Y ако је свака тачка на Y највише на растојању ϵ од L, односно: $\forall y \in Y, min_{x \in L} d(x, y) \leq \epsilon$.

Дефиниција 5.4.2. За $d_1, d_2 \geq 0$, кажемо да је перзистентни дијаграм $D_0 f$ (d_1, d_2)-раздвојен ако свака тачка лежи у области D_1 изнад праве $y = x - d_1$ или у области D_2 испод праве $y = x - d_2$ и десно од вертикале $x = d_2$.

[3]Wilhelm Friedrich Ackermann - немачки математичар

Интуитивно, то схватамо као раздвојеност значајних кластера D_2 и тополошког шума D_1. Услов за тачке десно од вертикале биће објашњен кроз примере у последњој глави. Следећа слика 5.4 (в. [2]) перзистентног дијаграма то илуструје:



Слика 5.4: Подела перзистентног дијаграма на области значајних кластера D_2 и шума D_1

Први главни резултат нам даје корелацију између броја кластера добијених ToMATo алгоритмом и броја значајних мода функције f. Користећи стабилност перзистентних дијаграма можемо доказати да области D_1 и D_2 остају раздвојене чак и под пертубацијама узрокованим апроксимацијом функције густине f. Дакле, могу се одвојити помоћу неког прага τ. Ако је $D_0 f$ добро раздвојен и број тачака у облаку L је довољно велики, тада постоје вредности за Рипсов параметар σ и параметар τ тако да је број кластера добијених ToMATo алгоритмом, са великом вероватноћом, једнак броју мода функције густине f значајности барем τ.

Други главни резултат нам даје корелацију између кластера добијених ToMATo алгоритмом и басена атракције значајних мода функције густине f. Међутим, то можемо гарантовати само у случају стабилних делова басена атракције јер басени атакције, генерално, нису стабилни чак ни у глатком случају. Исто, ако је $D_0 f$ добро раздвојен и број тачака у облаку L је довољно велики, тада постоје вредности за Рипсов параметар σ и параметар τ тако да за сваку барем τ значајну моду p функције густине f, са великом вероватноћом, алгоритам даје кластер који се поклапа са басеном атракције $B_p(\tau)$ све до тренутка α када се $B_p(\tau)$ споји са неким другим значајним кластером. Приметимо да за басконачно перзистентне кластере имамо потпуно поклапање.

Приметимо да су главни резултати пробабилистички. То је због тога што захтевамо да улазни облак тачака L мора да буде геодезијски ϵ-узорак неког надниво скупа од f да би алгоритам имао шансу да тачно апроксимира $D_0 f$, што се дешава са великом вероватноћом јер L узоркујемо насумично као i.i.d. случајне величине.

Дефиниција 5.4.3. Нека је $\rho(x)$ супремум радијуса r за које је геодезијска лопта $B_X(x,r)$ јако конвексна. Инфимум вредности $\rho(x)$ за сваку тачку $x \in X$ зовемо радијусом јаке конвексности, у ознаци $\rho(X)$.

Дефиниција 5.4.4. $N_r(A) \in \mathbb{N} \cup \{+\infty\}$ је минимални број геодезијских лопти $B_X(x,r)$ радијуса r неопходан да се покрије цео скуп A.
$V_r(A)$ је инфимум Хаусдорфове мере на свим геодезијским лоптама радијуса r у A тј. $V_r(A) = inf_{x \in A} H^m(B_X(x,r))$.

Теорема 5.4.1. (Тачан број кластера) Нека је X Риманова многострукост са позитивним радијусом јаке конвексности и нека је $f : X \to \mathbb{R}$ Липсхиц непрекидна, са константом c, функција густине вероватноће са коначаним бројем локалних максимума. Ако је $D_0 f$ (d_1, d_2)-раздвојен, $d_2 > d_1 \geq 0$, тада за сваки позитивни параметар $\sigma < min\{\rho(X), \frac{d_2 - d_1}{5c}\}$ и сваки праг $\tau \in \{d_1 + 2c\sigma, d_2 - 3c\sigma\}$ и сваки n-димензионални скуп тачака узоркован у складу са густином, тада је број кластера добијених алгоритмом једнак броју локалних максимума функције f значајности бар d_2, са вероватноћом већом од $1 - N_{\sigma/8}(F^{c\sigma}) e^{-n\frac{3}{4}c\sigma V_{\sigma/8}(F^{c\sigma})}$.

Теорема 5.4.2. (Кластери као басени атракције) Нека је X Риманова многострукост са позитивним радијусом јаке конвексности и нека је $f : X \to \mathbb{R}$ Липшиц непрекидна, са константом c, функција густине вероватноће са коначаним бројем локалних максимума. Ако је $D_0 f$ (d_1, d_2)-раздвојен, $d_2 > d_1 \geq 0$, тада за сваки позитивни параметар $\sigma < min\{\rho(X), \frac{d_2 - d_1}{5c}\}$ и сваки праг $\tau \in \{d_1 + 2c\sigma, d_2 - 3c\sigma\}$ и сваки n-димензионални скуп тачака узоркован у складу са густином, тада са вероватноћом већом од $1 - N_{\sigma/8}(F^{c\sigma}) e^{-n\frac{3}{4}c\sigma V_{\sigma/8}(F^{c\sigma})}$ тврдимо: за сваку тачку $p \in D_2$ и њен кластер $B^R_\tau(p)$ добијен алгоритмом, важи $B^R_\tau(p) \cap F^\alpha = B_\tau(p) \cap L \cap F^\alpha$ за све $\alpha \in (\alpha_\tau(p) + d_1 + \frac{5}{2}c\sigma, p_x]$.

Глава 6

Примена ToMATo алгоритма

Сада ћемо дати експерименталну примену идеја и резултата из претходних глава. ToMATo алгоритам примењујемо на разне скупове података у разним димензијама. То су следеће три групе скупова података:

1. Синтетички скуп генерисан у две (поглавље 6.1.1) и три (поглавље 6.1.2) димензије
2. Компоненте боје слика у рачунару (поглавље 6.2.1) и додатне просторне информације на сликама (поглавље 6.2.2)
3. Деветодимензионални скуп тачака генерисан из видео записа простирије

Оцена густине у свим експериментима јесте оцена помоћу Гаусовог кернела. Параметар размере изабран је, наравно независно за сваки узорак, методом k-слојне унакрсне валидације.

Као што смо већ рекли у Глави 5, параметри σ и τ су веома различите природе и зато их бирамо различитим методама у експериментима. Прво, да бисмо изабрали параметар σ конструишемо дендрограм (в. слику 4.5 и објашњење). Он нам пружа релевантну скалу за посматрање растојања тачака података и њихово груписање. Избор σ даље потврђујемо правилом лакта (енг. elbow method) са графика зависности просечног броја суседа од избора параметра σ у Рипсовом графу суседа R_σ. На тај начин добијамо конкретну вредност или сужен скуп вредности параметра σ за које проверамо добијене резултате. Илустрацију овог одређивања ћемо изоставити. Затим покрећемо алгоритам кластеровања са изабраним параметром σ и

са $\tau = +\infty$. На основу излаза алгоритма конструишемо перзистентни дијаграм функције густине из којег одређујемо животне векове кластера. Напоменимо да су бесконачно перзистентни кластери на дијаграму, због сразмерног приказа, илустровани као кластери који су последњи умрли у тренутку 0. Потом тражимо вредност параметра τ тако што сортирамо тачке са перзистентног дијаграма по значајности опадајуће, а затим тражимо највећи јаз у том низу значајности. Такође, јаз можемо тражити помоћу хистограма значајности. Са њега читамо број кластера одређене коначне значајности. Перзистентне дијаграме и њихово тумачење илустроваћемо кроз примере. Коначно, покрећемо алгоритам за оба изабрана параметра и добијамо финални излаз.

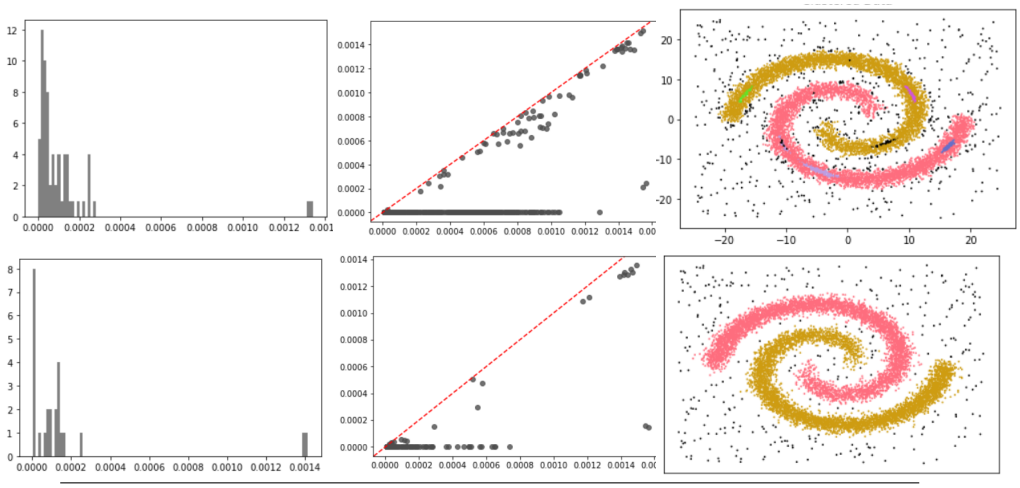
Дајемо имплементацију псеудокода из поглавља 5.3.2 у *Python* програмском језику. Сви потребни кодови могу се пронаћи на следећем Github репозиторијуму [31].

```
def define_clusters(vec_sorted, radius, tau):
unf = UnionFind()
n=len(vec_sorted)
births, deaths = {}, {}
for i in reversed(range(n)):
    idx = index_sorted[i]
    nei = kdt.query_radius([x[idx]], radius, return_distance=
   False)[0]
    S = [elem for elem in nei if elem in index_sorted[(i+1):]]
    if not S:
        unf.insert_objects([idx])
        births[idx] = -vec_sorted[idx]
    else:
        parent = S[np.asarray([vec_sorted[j] for j in S]).
   argmax()]
        unf.union(parent, idx)
        roots = [unf.find(ele) for ele in S]
        highest = roots[np.asarray([vec_sorted[elem] for elem in
    roots]).argmax()]
        for root in roots:
            if (root != parent) & (vec_sorted[root] - vec_sorted
   [idx] < tau):
                unf.union(parent, root)
                deaths[root] = -vec_sorted[idx]
        if (highest != parent) & (vec_sorted[parent] -
   vec_sorted[idx] < tau):
            unf.union(highest, parent)
            deaths[parent] = -vec_sorted[idx]
return unf, births, deaths
```

6.1 Синтетички скуп података

6.1.1 Дводимензионални случај

Већ смо илустровали на слици 5.1 примену ToMATo алгоритма на дводимензионални синтетички скуп података. Ту смо видели разлику у резултату алгоритма и резултату тражења мода шемом успона. Сада ћемо на истом скупу података приказати како избор параметра σ мења финални резултат ToMATo алгоритма. Такође, илустроваћемо симултано и избор параметра τ.

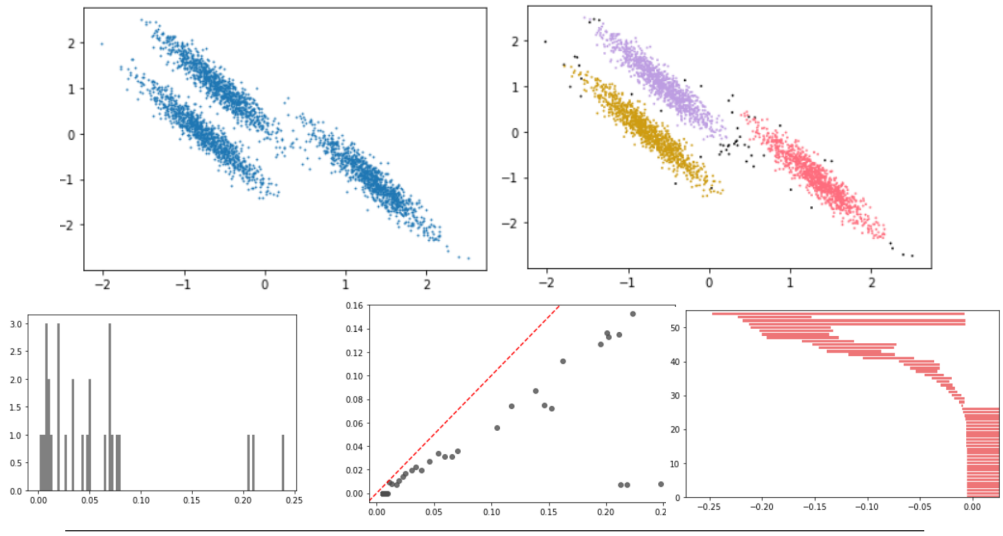


Слика 6.1: Утицај параметра σ на ToMATo алгоритам: Дати су хистограм значајности, перзистентни дијаграм и финални резултат за вредности $\sigma = 0.5$ и $\sigma = 1$, тим редом

Прокоментаришимо прво избор параметра τ на горњој слици. Приметимо да се на оба хистограма значајности види јасан јаз између значајности два издвојена кластера и осталих. Дакле, за другу итерацију кроз алгоритам бирамо праг значајности као било коју вредност у том јазу. Конкретно, у оба случаја, узета је вредност $\tau = 0.0012$. Издвојеност та два кластера примећујемо и на перзистентним дијаграмима где можемо видети да су они рођени први у филтрацији за разлику од бесконачно перзистентних кластера који су рођени знатно касније. То нам говори да су то тачке са великим одступањима у нашим подацима. Такође, оба кластера су веома значајна за разлику од кластера близу дијагонале који су кратког животног века па њих проглашавамо незначајним и спајамо у другој фази алгоритма. За мању вредност параметра σ имамо више незначајних и више малих бесконачно перзистентих кластера јер податке посматрамо на мањој скали па

немамо довољно повезаних компонената. Зато у финалном резултату у првом случају уочавамо неке ситне неправилности у кластеровању. Све то нам сугерише да податке треба да посматрамо на већој скали да бисмо добили прецизан финални резултат, као што смо и урадили у другом случају. Напоменимо да су сви кластери величине мање од неке вредности (у овом примеру 50) обојени црном бојом на финалном излазу због прегледности резултата.

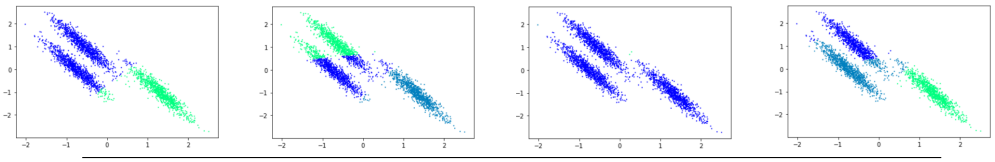
Посматрајмо сада други синтетички скуп података који ћемо користити за поређење ToMATo алгоритма са стандардним техникама кластеровања. Скуп података састоји се од 3000 тачака вештачки груписаних у три групе неправилног облика. Прво на слици 6.2 дајемо резултат ToMATo алгоритма заједно са перзистентним дијаграмом и бар-кодом.



Слика 6.2: Примена ToMATo алгоритма на дводимензионални синтетички скуп података; на слици су дати иницијални подаци и финални резултат алгоритма након друге фазе, у првом реду, и хистограм значајности, перзистентни дијаграм и бар-код након прве фазе алгоритма, у другом реду

За овај скуп података користили смо вредност параметра $\sigma = 0.2$. Вредност прага $\tau = 0.2$ читамо са хистограма значајности са слике 6.2 на ком се види јаз који одваја три најзначајнија кластера. Коначно, након друге фазе добијамо финални резултат. Приметимо како је ToMATo алгоритам, за задате вредности улазних параметара, посебно издвојио изоловане тачке (на слици обојене црно). То је његова важна предност у односу на остале технике кластеровања и то илуструјемо на следећој слици 6.3. Поредимо

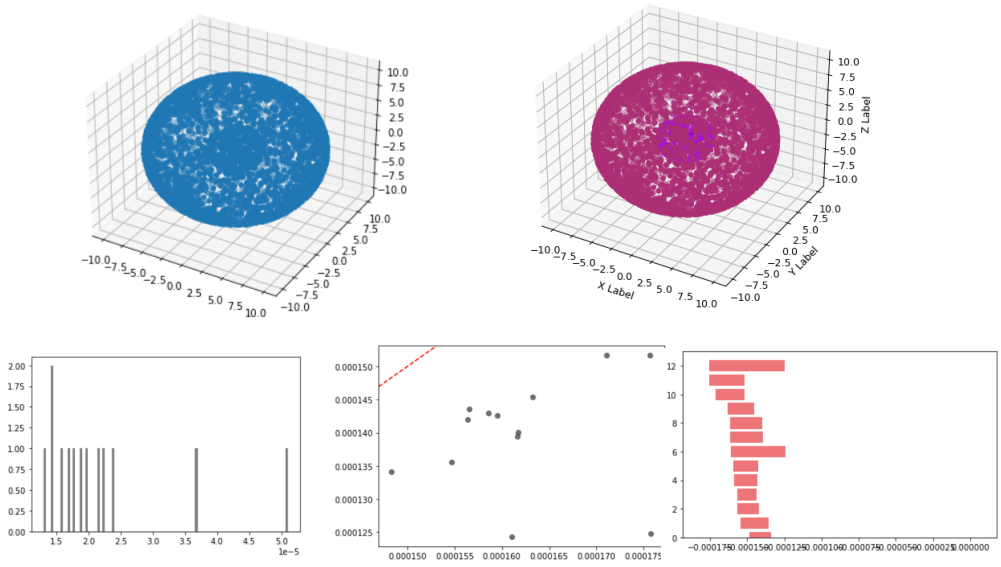
ToMATo алгоритам са, редом: Mean-Shift алгоритмом [18], алгоритмом k-средина [19], агломеративним кластеровањем са једноструким повезивањем [1] и спектралним кластеровањем [20] примењеним на исти скуп података.



Слика 6.3: Резултати кластеровања добијени, редом, Mean-Shift алгоритмом, алгоритмом k-средина, агломеративним кластеровањем са једноструким повезивањем и спектралним кластеровањем

6.1.2 Тродимензионални случај

Посматрајмо синтетички генерисан скуп података у три димензије. Скуп се састоји од укупно 5000 тачака које су груписане у две концентричне сфере. Тачке на обе сфере су узорковане тако да је функција густине иста. Поента оваквог вештачког скупа података је примена ToMATo алгоритма на податке са одређеном хијерархијом и детекција вишеструких нивоа у тим подацима. На слици 6.4 илуструјемо добијене резултате:



Слика 6.4: Примена ToMATo алгоритма на тродимензионални синтетички скуп података; на слици су дати иницијални подаци и финални резултат алгоритма након друге фазе, у првом реду, и хистограм значајности, перзистентни дијаграм и бар-код након прве фазе алгоритма, у другом реду

За овај скуп података користили смо вредност параметра $\sigma = 3$. Вредност прага $\tau = 3.5 \times 10^{-5}$ читамо са хистограма значајности са слике 6.4

на ком се види јаз након два издвојена значајна кластера. Коначно, након друге фазе добијамо финални резултат.

6.2 Сегментација слика

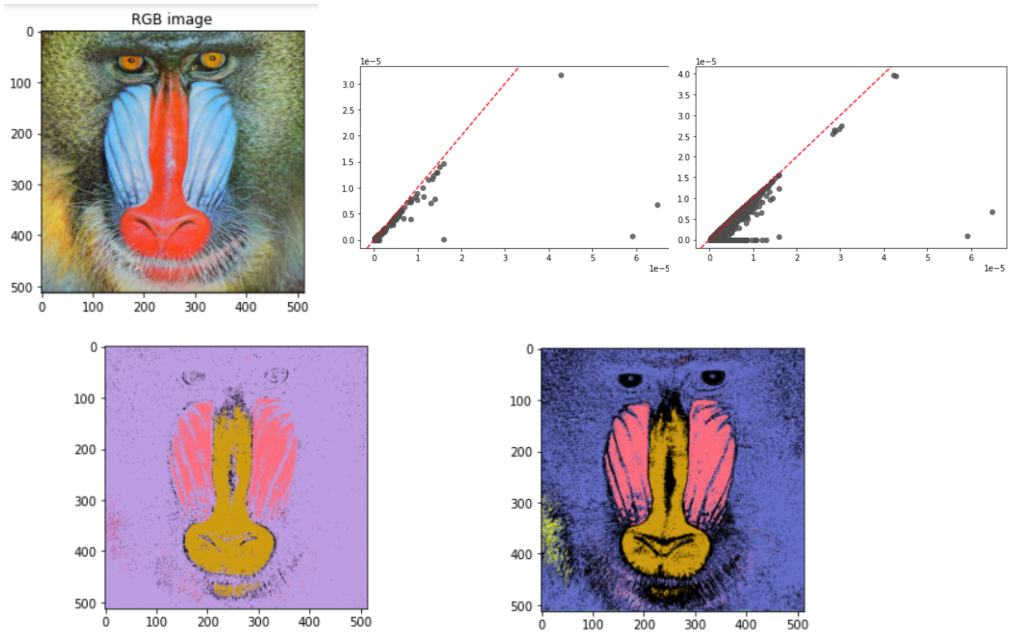
6.2.1 Компоненте боје

У овом случају посматрамо дату слику и сваки њен пиксел мапирамо у тачку у тродимензионалном простору боја Лув [32] (енг. CIELUV). Разлог зашто користимо Лув компоненте уместо, стандардног избора, РГБ модела боја, је што је еуклидска удаљеност у Лув простору природна метрика. Већ смо спомињали у уводу рада у Глави 1 да нема смисла користити еуклидски простор јер нам не даје смислен осећај за растојање тачака тј. боја.

У овом делу експеримената посматрамо само простор боја и занемарујемо просторне информације на слици. Проблем са којим се суочавамо у имплементацији алгоритма, код неких слика, је велики број ивица у Рипсовом графу код слика на којима постоји велики број тачака које се налазе у непосредној близини у Лув простору. Да бисмо убрзали извршавање, смањујемо број тачака у узорку који улази у ToMATo алгоритам на следећи начин: Почнимо од тога да су све тачке необележене. За сваку тачку p која није обележена узимамо тачке на удаљености σ од ње, као и до сада, али међу њима уклањамо тачке на растојању $\frac{\sigma}{m}$ и тако добијамо улазни скуп тачака. Уклоњене тачке обележавамо да бисмо их на крају приписали кластеру тачке p. Обично користимо вредност $m \in \{10, 20\}$ у зависности од скале података.
У пракси, тежимо томе да узимамо мање околине да бисмо видели целу структуру и боље препознали мале независне делове слика. Црном бојом су обојени кластери кардиналности мање од задатог прага (углавном 100 тачака).

Сада дајемо примере сегментације слика:

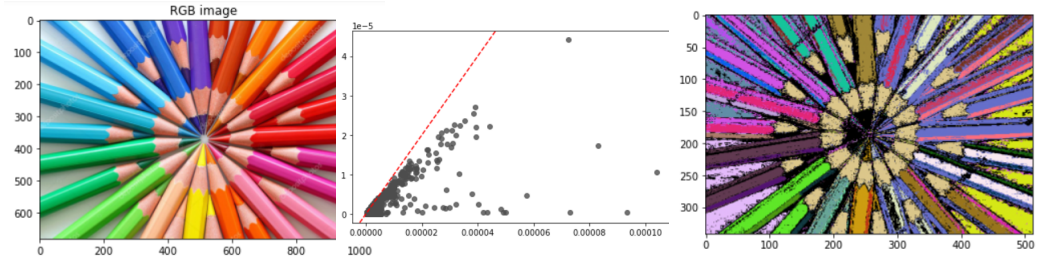
Прво посматрамо слику мајмуна димензија 512×512 на слици 6.5. Помоћу дендрограма и правила лакта одређујемо скалу на којој посматрамо податке. Поредимо резултате за вредности $\sigma = 2$ и $\sigma = 1$ редом. У оба случаја, након прве фазе, одређујемо $\tau = 1 \times 10^{-5}$. Приметимо како у финалним резултатима, у зависности од параметра σ, имамо компромис између јасније структуре података и већег броја бесконачно перзистентних



Слика 6.5: Иницијални подаци и резултати примене ToMATo алгоритма на компоненте боје дате слике за вредности $\sigma = 2$ и $\sigma = 1$ редом

кластера који су веома касно рођени и кардиналност им је испод задатог прага.

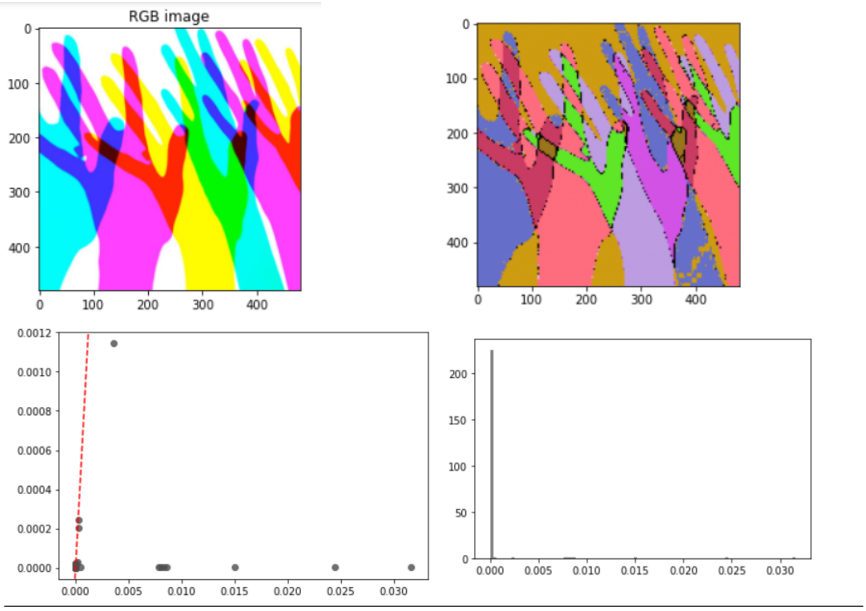
Други пример је слика разнобојних оловака, величине 341×512. Изабрали смо вредност параметра $\sigma = 1.5$, а потом вредност прага $\tau = 1 \times 10^{-5}$. Овде смо користили вредност $m = 20$, док је у првом примеру била $m = 10$. То бирамо експериментално на основу скале растојања тачака на слици и величине узорка који, после тог уклањања, улази у алгоритам. Резултат ToMATo алгоритма је дат на слици 6.6 и приметимо како су црном бојом обојене ивице оловака, чак и сенке на њиховим контурама.



Слика 6.6: Иницијални подаци и резултат примене ToMATo алгоритма на компоненте боје дате слике

Трећи пример, дат на слици 6.7, је 480×480 слика разнобојних шака које се преклапају. Користили смо вредност параметра $\sigma = 3$ и као што

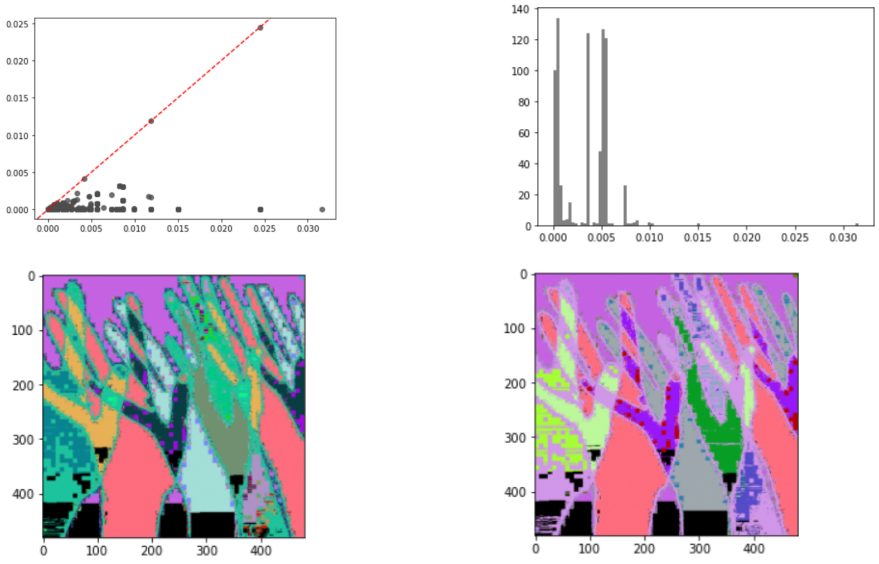
видимо са хистограма значајности смислено је узети праг $\tau = 0.002$. Овај пример има за циљ да покаже сегментацију вишеструког преклапања независних делова на слици. Још важније, ову слику користимо и као пример примене графа k-најближих суседа (енг. k-NN). Његову главну предност (в. поглавље 5.3.3 Избор параметара у Глави 5), што не дивергира без обзира на растојање између тачака, у овом случају користимо због доминантне беле позадине и веома ограниченог спектра боја на слици. Из тог разлога, Рипсов граф је у овом примеру мање оптималан за израчунавање. На слици 6.8 дајемо финалне излазе ToMATo алгоритма за број суседа $k = 200$ и $k = 400$, тим редом, као и хистограм значајности и перзистентни дијаграм у случају $k = 200$. Због занемарљиве разлике у истим, изостављамо их за случај $k = 400$. Можемо приметити да је резултат на десној страни само боље заглађен од резултата на левој, што је и очекивано понашање. Другачији резултат бисмо очекивали за много веће k, али тада се губи главна идеја о оптимизацији. Такође, приметимо да и на овај начин препознајемо делове који се преклапају. Међутим, највећа мана је управо константан број суседа за сваку тачку без обзира на њене атрибуте, па нам тачност финалног излаза варира за различите групе пиксела на слици.



Слика 6.7: Иницијални подаци и резултат примене ToMATo алгоритма на компоненте боје дате слике

6.2.2 Компоненте боје и просторне информације

Као што смо видели у претходном поглављу, кластеровањем у Лув простору удаљени пиксели на слици могу да заврше у истом кластеру. Зато у

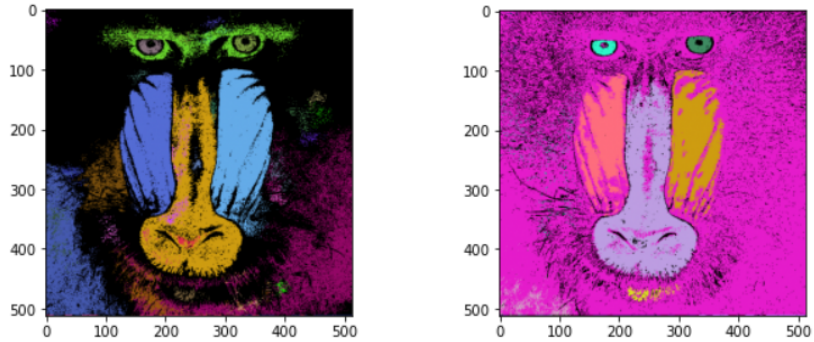


Слика 6.8: Иницијални подаци и резултат примене ToMATo алгоритма на компоненте боје дате слике употребом графа k-најближих суседа за $k = 200$ и $k = 400$ редом

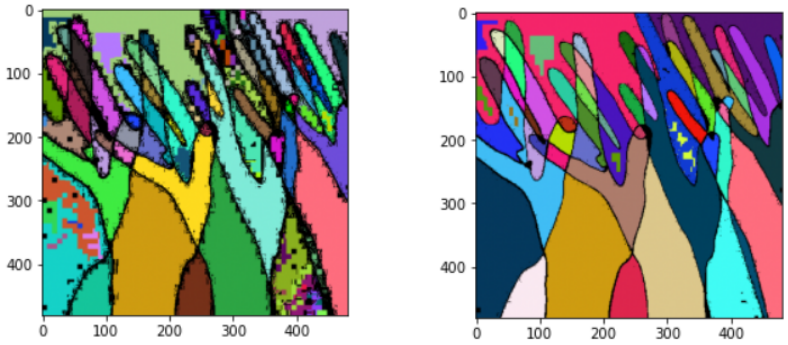
овом делу експеримената користимо и додатну просторну информацију, у смислу близине пиксела. Ову додатну информацију користимо у зависности од контекста и жељеног резултата.

Проблем са којим се овде суочавамо је израчунавање удаљености у новом петодимензионалном простору који се састоји од две координатте пиксела и три компоненте боје. Зато поступамо на следећи начин: Још увек посматрамо тачке у тродимензионалном Лув простору и оцењујемо густину као у претходном поглављу. Међутим, да би две тачке биле повезане у Рипсовом графу, сада захтевамо да буду близу и просторно. Радимо обрнутим редоследом, односно прво повезујемо суседне пикселе, а затим бришемо (енг. pruning) ивице које су удаљене у Лув просору. Разлог за то је константан број суседних пиксела па алгоритам ради брже јер има мање брисања ивица. У експериментима користимо 5×5 суседне околине пиксела. И овде су, такође, црном бојом обојени кластери кардиналности мање од задатог прага.

На следећим сликама 6.9 и 6.10 дајемо резултате примене ToMATo алгоритма на слике из првог и трећег примера у претходном поглављу. За обе слике имамо два резултата у зависности од вредности параметра σ којим задајемо колику околину боје желимо да посматрамо. На левој страни је дат резултат за мању вредност σ, а на десној за већу, па се јасно виде и разлике у финалним излазима алгоритма.



Слика 6.9: Резултат примене ToMATo алгоритма на компоненте боје и просторне информације слике 6.5 за различите вредности параметара

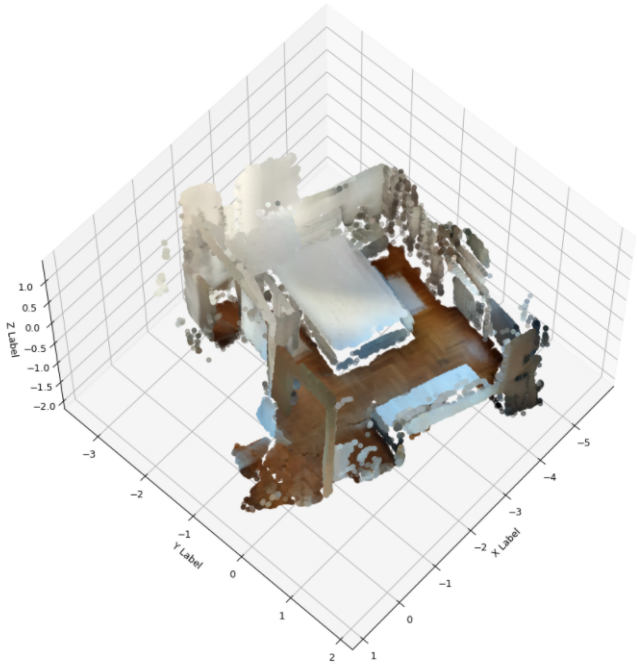


Слика 6.10: Резултат примене ToMATo алгоритма на компоненте боје и просторне информације слике 6.7 за различите вредности параметара

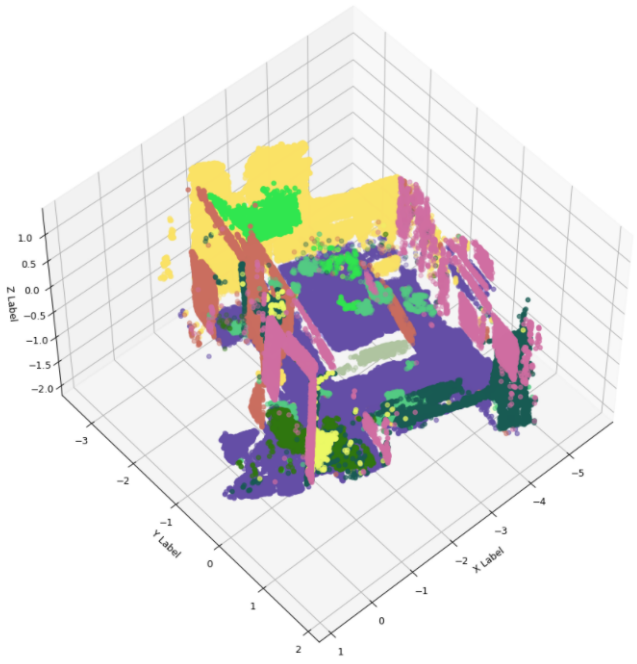
6.3 Сегментација просторије

Последња примена ToMATo алгоритма у овом раду односи се на кластеровање скупа од 300 хиљада тачака, у 9 димензија, генерисаног из видео записа просторије. Димензије представљају три компоненте боје, три координате положаја тачке у простору и три координате правца вектора нормале на ту тачку. Иницијални подаци су дати на слици 6.11.

И за сегментацију просторија користимо исту идеју конструкције Рипсовог графа као за сегментацију слика у поглављу 6.2.2. Резултат сегментације просторије добијен тако што захтевамо да су две тачке близу у Рипсовом графу ако су близу у векторском простору нормала, а затим и у Лув простору боје, приказујемо на слици 6.12. Дакле, у овом примеру смо користили шест димензија скупа података.

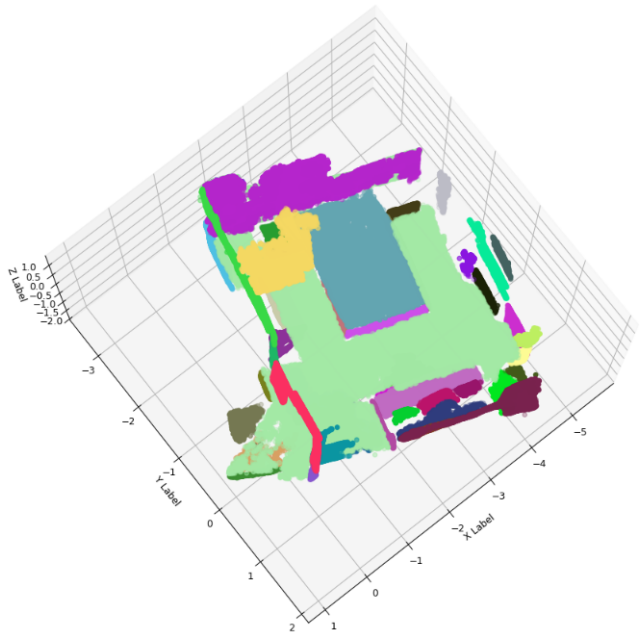


Слика 6.11: Облак тачака обима 300 хиљада генерисан из видео записа просторије



Слика 6.12: Резултат сегментације просторије применом ToMATo алгоритма на компоненте боје и правце нормале у генерисаном облаку тачака

Нови резултат на слици 6.13 је добијен тако што, додатно, користимо и положаје тачака у простору и захтевамо близину у Рипсовом графу и по том атрибуту. Дакле, у овом примеру смо користили свих девет димензија скупа података.



Слика 6.13: Резултат сегментације просторије применом ToMATo алгоритма на компоненте боје, правце нормале и положаје у генерисаном облаку тачака

Глава 7

Закључак

У овом раду представљена је метода кластеровања, под називом ToMATo (од Topological Mode Analysis Tool) . Алгоритам комбинује фазу детекције мода са фазом спајања кластера. Најважнији допринос је употреба тополошке перзистенције у другој фази алгоритма. Додатно, на излазу алгоритма добијамо и додатне визуелизације резултата, попут перзистентног бар-кода и дијаграма. Такође, дали смо и теоријске гаранције о броју кластера и њиховој просторној локацији. На самом почетку рада, упознали смо се са појмовима хомологије и Бетијевих бројева, а затим и перзистентне хомологије. Поред теоријских оквира алгоритма, дали смо и интуитивни приступ и мотивацију употребе перзистенције у кластеровању. На крају, све смо поткрепили експерименталним примерима - синтетичким и реалним скуповима различитих димензија и области примене. Како нам сам ToMATo алгоритам пружа велику слободу избора, овај рад је отворио много питања за будуће истраживање. Пре свега, не морамо се ограничити на један метод оцене густине и требало би видети како неке друге оцене утичу на резултат алгоритма. Занимљиво је размишљати, шта би било када се не бисмо ограничили на 0-димензионалну хомологију? Свакако бисмо открили и неке суптилније структуре, али би комплексност алгоритма порасла са димензијом коју посматрамо. У том случају остаје отворено питање оптимизације или можда неког другог приступа. Још једно занимљиво питање је аутоматизовање избора графа суседа и његових параметара на улазу у алгоритам. Хеуристички начин избора употребљен у овом раду би требало теоријски оправдати.

Библиографија

[1] F.Murtagh, P.Contreras. *Algorithms for hierarchical clustering: an overview.* DOI: 10.1002/widm.53. 2011

[2] F.Chazal, L.J.Guibas, S.Oudot, P.Skraba. *Persistence-Based Clustering in Riemannian Manifolds.* [Research Report] RR-6968, INRIA. 2009, 47 p. inria-00389390

[3] S.Oudot. *Persistence Theory: From Quiver Representations to Data Analysis.* 2010 Mathematics Subject Classification, ISBN-10: 1-4704-2545-9, 2015

[4] G.Carlsson. *Topology and Data.* Bulletin of the American Mathematical Society, vol. 46, no. 2, April 2009, 255–308 p.

[5] W.L.Koontz, P.M.Narendra, K.Fukunaga.*A graph-theoretic approach to nonparametric cluster analysis.* IEEE Trans. on Computers, 24:936–944, September 1976.

[6] F.Chazal, D.Cohen-Steiner, L.J.Guibas, M.Glisse, S.Oudot. *Proximity of persistence modules and their diagrams.* In Proc. 25th ACM Sympos. Comput. Geom., 2009.

[7] D.Cohen-Steiner, H.Edelsbrunner, J.Harer. *Stability of persistence diagrams.* Discrete Computational Geometry 37:103–120, 2007. DOI: 10.1007/s00454-006-1276-5

[8] G.Singh, F.Mémoli, G.Carlsson. *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.* Eurographics Symposium on Point-Based Graphics, 2007

[9] М.Марјановић. *Топологија.* Математички факултет, Београд, 1990.

[10] Triangulations of the sphere, URL: https://fenicsproject.discourse.group/t/triangulations-of-the-sphere/1420. 16.08.2021.

[11] A.Piro. *The fundamental theorem for finite Abelian groups: A brief history and proof.* Georgia College & State University, Department of Mathematics

[12] Torus cycles, URL: https://en.wikipedia.org/wiki/Homology_(mathematics)#/media/File:Toruscycles1.svg. 16.08.2021.

[13] Flat surfaces, URL: https://en.wikipedia.org/wiki/Homology_(mathematics)#/media/File:Flatsurfaces.svg. 16.08.2021.

[14] H.Edelsbrunner, E.P.Mücke. *Three-dimensional Alpha Shapes.* Department of Computer Science, University of Illinois, 1994

[15] H.Edelsbrunner, D.Letscher, A.Zomorodian. *Topological Persistence and Simplification.* Discrete Computational Geometry 28:511–533, 2002. DOI: 10.1007/s00454-002-2885-2

[16] URL:https://en.wikipedia.org/wiki/Hausdorff_measure, 16.08.2021. *Hausdorff measure*

[17] P.Bubenik. *Statistical Topological Data Analysis Using Persistence Landscapes.* arXiv:1207.6437v4, 2015

[18] Y.Cheng. *Mean Shift, Mode Seeking, and Clustering.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no.8, 1995

[19] М.Николић, А.Зечевић. *Машинско учење.* Београд, 2019

[20] A.Y.Ng, M.Jordan, Y.Weiss. *On spectral clustering: analysis and an algorithm.* Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, 2001

[21] G.Carlsson, F.Mémoli. *Characterization, Stability and Convergence of Hierarchical Clustering Methods.* Journal of Machine Learning Research 11 (2010) 1425-1470

[22] B.J.Maiseli. *Hausdorff Distance with Outliers and Noise Resilience Capabilities.* SN Computer Science, 2021. DOI: 10.1007/s42979-021-00737-y

[23] M.Minervino. *Topological data analysis with Mapper.* Computer vision, Quantmetry.

[24] T.Liao, Y.Wei, M.Luo, G.Zhao, H.Zhou. *tmap: an integrative framework based on topological data analysis for populationscale microbiome stratification and association studies.* Genome Biology (2019) 20:293. DOI: 10.1186/s13059-019-1871-4

[25] F.Chazal, B.Michel. *Covers and nerves: union of balls, geometric inference and Mapper.* Barcelona, 2016

[26] L.McInnes, J.Healy, J.Melville. *UMAP:Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv:1802.03426v3 [stat.ML], 2020

[27] M.Ester, H.Kriegel, J.Sander, X.Xu. *A density-based algorithm for discovering clusters in large spatial databases with noise.* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. ISBN 1-57735-004-9.

[28] F.Chazal, L.J.Guibas, S.Oudot, P.Skraba. *Analysis of scalar fields over point cloud data.* In Proc. 20th ACM-SIAM Sympos. Discrete Algorithms, 2009.

[29] N.Otter, M.Porter, U.Tillmann, P.Grindrod, H.Harrington. *A roadmap for the computation.* EPJ Data Science, 2017. DOI: 10.1140/epjds/s13688-017-0109-5 of persistent homology

[30] URL:https://en.wikipedia.org/wiki/Disjoint-set_data_structure, 16.08.2021. *Disjoint-set data structure*

[31] *Github, ToMATo clustering.* URL: https://github.com/Vildana96/ToMATo-clustering

[32] M.Fairchild. *Color Appearance Models.* Reading, MA: Addison-Wesley, 1998

Биографија

Вилдана Бакаревић је рођена 25.јануара 1996.године у Пријепољу. Похађала је основну школу "Владимир Перић Валтер"у Пријепољу. Након тога 2011.године долази у Београд где завршава Математичку гимназију као вуковац и потом уписује Математички факултет Универзитета у Београду, на смеру статистика, актуарска и финансијска математика. Основне студије завршава 2019.године са просечном оценом 9.98. Љубав према математици гаји одмалена, а главне области интересовања су јој вероватноћа и статистика, као и геометрија и машинско учење.