Difuzijska preslikavanja s primjenom

Ramljak, Tatjana

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:018050

Difuzijska preslikavanja s primjenom

Ramljak, Tatjana

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:018050

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tatjana Ramljak

DIFUZIJSKA PRESLIKAVANJA S PRIMJENOM

Diplomski rad

Mentor:
prof. dr. sc. Zlatko Drmač

Zagreb, srpanj, 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem svima koji su svojim prijedlozima, savjetima i podrškom pridonijeli izradi ovog rada. Zahvaljujem svom mentoru, prof. dr. sc. Zlatku Drmaču, na pomoći, strpljenju i vodstvu tijekom izrade rada. Zahvaljujem svojim prijateljima i kolegama koji su mi olakšali ovaj put i pridonijeli uspješnom završetku studija. Posebna zahvala mojim roditeljima, braći i sestrama koji su mi pružali bezuvjetnu podršku i ljubav tijekom mog školovanja.

Sadržaj

Sadržaj **iv**

1 Uvod **1**
1.1 Redukcija dimenzije 2
1.2 Neke metode za redukciju dimenzije 4

2 Markovljevi lanci **9**
2.1 Definicija i osnovna svojstva 10
2.2 Granična i stacionarna distribucija 12
2.3 Markovljevi lanci unatrag 15
2.4 Spektralna svojstva tranzicijske matrice reverzibilnog lanca 16
2.5 Uvod u vrijeme miješanja Markovljeva lanca 19
2.6 Vrijeme relaksacije 20

3 Teorija difuzijskih preslikavanja **23**
3.1 Modeliranje preko grafova 23
3.2 Povezanost točaka 25
3.3 Difuzijski proces 27
3.4 Spektralna analiza Markovljeva lanca 31
3.5 Difuzijska udaljenost 34
3.6 Konstrukcija preslikavanja 35

4 Povezanost sa spektralnim grupiranjem **45**
4.1 Normalizirani rez 46
4.2 Veza između Markovljeve šetnje i normaliziranog reza 49
4.3 Stohastičke matrice s po dijelovima konstantnim svojstvenim vektorima . 51

5 Primjene **53**
5.1 Spektralno ulaganje u niždimenzionalni potprostor i grupiranje podataka 53
5.2 Odabir parametra ε 54

5.3 Primjeri 55
5.4 Ograničenja algoritma 63

A Familija difuzija **68**

B Dijelovi koda korišteni za vizualizaciju metode **70**

Bibliografija **74**

Poglavlje 1

Uvod

Količina elektroničkih podataka koja je danas dostupna u mnogim znanstvenim područjima brzo raste, samim time povećava se njihova dimenzija i složenost. Stoga grupiranje podataka i njihova niždimenzionalna reprezentacija predstavljaju bitne probleme u mnogim područjima primjene, poput robotike, bioinformatike, biomedicine i multimedijske tehnologije. Mnogi skupovi podataka vrlo su složeni, ali imaju i jednostavnu unutarnju strukturu koja omogućava modeliranje takvih podataka bez gubitka informativnosti. Kako bismo otkrili takve strukture, podatkovne točke možemo analizirati s pomoću grafa, koji omogućuje modeliranje relacija između elemenata skupova. Geometrijska organizacija grafova i skupova podataka u $\mathbb{R}^n$ središnji je zadatak u statističkoj analizi podataka.

U ovom radu predstavljamo metodu zasnovanu na difuzijskom procesu za pronalaženje smislenih struktura u skupovima podataka. Odgovarajuće svojstvene funkcije matrice prijelaza Markovljeva lanca mogu se koristiti za izgradnju koordinata nazvanih difuzijskim preslikavanjima, koje pronalaze efikasne prikaze složenih geometrija. U praksi, gornje spomenute svojstvene funkcije omogućuju nižedimenzionalno ulaganje podataka u $\mathbb{R}^k$, $k \ll n$, tako da obična euklidska udaljenost u novom prostoru bude dobra mjera sličnosti podataka.

Mnoge od ovih ideja pojavljuju se u različitim kontekstima analize podataka, kao što su spektralna teorija grafova, mnogostruko učenje, nelinearne glavne komponente i jezgrene metode (engl. *kernel methods*). Te pristupe upotpunjavamo pokazujući da je difuzijska udaljenost ključna svojstvena geometrijska veličina koja povezuje spektralnu teoriju Markovljeva procesa, Laplaceove operatore te jezgre s odgovarajućom geometrijom i gustoćom podataka. Ovo otvara vrata primjeni metoda numeričke analize i obrade signala do analize funkcija i transformacija podataka.

1.1 Redukcija dimenzije

Smanjenje dimenzije prostora u kojem su smješteni podatci zauzima središnje mjesto u mnogim područjima kao što je teorija informacija, strojno učenje i teorija uzorkovanja. Naime, cilj je promjena reprezentacije skupa podataka koji uključuje velik broj varijabli, tako da se podatci opišu koristeći se samo malim brojem slobodnih parametara. Nova reprezentacija podataka trebala bi vjerno opisati podatke čuvajući određena svojstva koja su nam od interesa, poput lokalne međusobne udaljenosti. Problem smanjenja dimenzije analogan je pronalaženju smislenih struktura u skupovima podataka. Ideja je ovdje malo drugačija, cilj je izvući relevantne značajke iz podataka kako bi se dobio uvid i razumijevanje fenomena koji je generirao podatke.

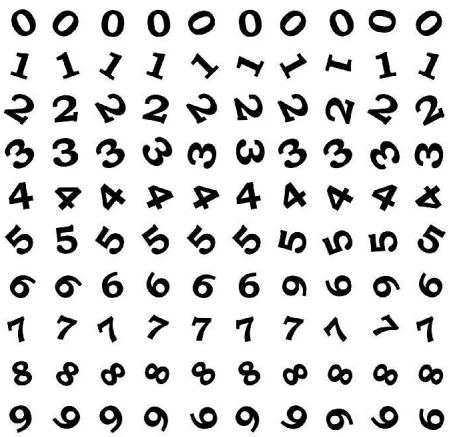
U današnjem društvu usmjerenom na informacije prisutni su mnogi podatci koji su često velike dimenzije, broj varijabli izmjerenih po uzorku lako može iznositi tisuće. Na primjer, ako imamo sliku koja ima 1000×1000 piksela, pri čemu svaki piksel predstavlja varijablu, ukupno dolazimo do dimenzije od 1000000. Pretpostavimo da imamo kolekciju takvih slika u vektorskom prostoru dimenzije 1000000. Tada u tom vektorskom prostoru možemo pronaći sustav izvodnica takav da s pomoću njega možemo prikazati bilo koju novu sliku tog prostora. U tako velikom prostoru značajki, reprezentacija je novih slika rijetka, tj. novu sliku možemo zapisati kao linearnu kombinaciju tako da koristimo malo koeficijenata različitih od nula. Takva reprezentacija uzrokuje brojne probleme, neki se algoritmi usporavaju ili u potpunosti ne rade, opisano u [16]. Nadalje, neka imamo kolekciju slika s 128×128 piksela, tako da svaka kodira broj između 0 i 9, vidi sliku 1.1. Slike se razlikuju po svojoj orijentaciji. Čovjek, suočen sa zadaćom organiziranja takvih slika, vjerojatno bi prvo primijetio različite znamenke, a nakon toga njihove različite orijentacije. Promatrač intuitivno pridaje veću vrijednost parametrima koji kodiraju veća odstupanja u opservacijama, stoga najprije grupira podatke u 10 skupina, po jednu za svaku znamenku. Unutar svake od 10 grupa znamenke su nadalje raspoređene prema kutu rotacije. Ova organizacija vodi jednostavnoj dvodimenzionalnoj parametrizaciji koja značajno smanjuje dimenziju skupa podataka uz očuvanje svih važnih atributa.

S druge strane, računalo svaku sliku vidi kao podatkovnu točku u $\mathbb{R}^{128\times128}$. Podatci su po prirodi organizirani prema svom položaju u koordinatnom prostoru, gdje je najčešća mjera sličnosti euklidska udaljenost. Mala euklidska udaljenost između vektora gotovo sigurno ukazuje na to da su oni vrlo slični, dok velika udaljenost daje vrlo malo informacija. Stoga euklidska udaljenost pruža dobru mjeru samo lokalne sličnosti. Dodatno, u višedimenzionalnim prostorima udaljenosti su često velike, s obzirom na rijedak prostor obilježja.

Na slici 1.2 vidimo tri slike bora pod tri različita kuta rotacije. Analogno opisu primjera sa znamenkama, želimo analizirati kolekciju slika prikazanih borova pod različitim kutem rotacija te prikazati s pomoću manje dimenzija. Ključno je za nelinearno smanje-

nje dimenzije spoznaja da su podatci često ugrađeni u nižedimenzionalni potprostor. U tom kontekstu bilo bi moguće okarakterizirati podatke i odnos između pojedinih točaka koristeći se manjim brojem dimenzija, računajući udaljenosti na samom ugrađenom prostoru umjesto u originalnom. Na primjer, uzimajući u obzir njegovu globalnu strukturu, mogli bismo predstaviti podatke u našem primjeru s borovima koristeći se samo dvjema varijablama. Takav prikaz s pomoću difuzijskih preslikavanja pokazan je u poglavlju 3.6.2.

Slika 1.1: Deset različitih znamenki pod deset različitih kutova rotacije.



Slika 1.2: Tri slike istog bora različitih rotacija

Izazov je odrediti strukturu podataka u nižoj dimenziji koja određuje podatke, što dovodi do smislene parametrizacije. Takav prikaz postiže smanjenje dimenzije uz očuvanje važnih odnosa između točaka podataka. Jedna su realizacija rješenja difuzijska preslikavanja koja su obrađena u ovom radu.

1.2 Neke metode za redukciju dimenzije

Postoji niz tehnika redukcije dimenzija. One se mogu široko kategorizirati u one koje mogu otkriti nelinearne strukture i one koje ne mogu. Intuitivno, možemo razmišljati kako linearna transformacija mijenja i rasteže podatke, a nelinearna transformacija donosi dramatičnije promjene u podatcima. Svaka metoda ima za cilj očuvanje određenog svojstva od interesa. Jedne su od najpoznatijih metoda glavnih komponenti (PCA), višedimenzionalno skaliranje (MDS) i izometrijsko preslikavanje značajki (isomap).

MDS ulaže podatke u prostor niže dimenzije tako da čuva uparene udaljenosti između podatkovnih točaka, $x_1, \ldots, x_n$. Najprije definiramo matricu udaljenosti D_x čiji su elementi udaljenosti između točaka u karakterističnom prostoru. Cilj je tada pronaći novi, nižedimenzionalni skup vektora $y_1, \ldots, y_n$, za koje matrica udaljenosti, $D_y[i, j] = d(y_i, y_j)$, minimizira funkciju troška, $\rho(D_x, D_y)$. Postoji mnogo različitih raspoloživih funkcija troška od kojih je najpopularnija tzv. *strain* funkcija:

$$\rho_{\text{strain}}(D_x, D_y) = \|J^T(D_x^2 - D_y^2)J\|_F^2.$$

Ovdje je J matrica centriranja, definirana s $J = I_n - 1/n\mathbf{1}\mathbf{1}^T$ [1], tako da $J^T XJ$ oduzme vektorsku sredinu svake komponente u X. *Frobeniusova* norma realne matrice $A \in M^{m\times n}$ definirana je s $\|A\|_F^2 = \sum_{i=1}^{n}\sum_{i=j}^{m} a_{ij}^2$. Intuitivno, ova funkcija troška zadržava varijacije u daljinama, a ne same vrijednosti pa iz toga slijedi da skaliranje konstantnim faktorom nema utjecaja. Naime, d najvećih svojstvenih vektora kovarijacijske matrice od X, koja je dana sa $\Sigma = XJJ^T X^T$, zahvaćaju najveće varijacije u $J^T D^2 J$ i time daju koordinate preslikavanja. Nedostatak MDS-a pri korištenju euklidskom udaljenosti davanje je jednakih težina velikim i malim udaljenostima, pa ne uspijeva dobro kada su pitanju nelinearne strukture.

Isomap, opisan u [5], tehnika je nelinearne redukcije dimenzija koja se temelji na MDS-u. Za razliku od MDS-a, među točkama se ne gleda euklidska udaljenost podataka, nego geodetska udaljenost, koja predstavlja ravnu liniju u zakrivljenom prostoru ili najkraću krivulju duž geometrijske strukture definirane našim podatkovnim točkama. Isomap traži preslikavanje tako da se geodetska udaljenost između podatkovnih točaka podudara s odgovarajućom euklidskom udaljenosti u transformiranom prostoru. Time se zadržava istinska geometrijska struktura podataka. Nadalje, bez poznavanja geometrijske strukture naših podataka, geodetsku udaljenost između točaka možemo aproksimirati. Pretpostavljamo da je u malom susjedstvu (npr. točke unutar određenog radijusa) euklidska udaljenost dobra aproksimacija za geodetsku udaljenost. Za daljnje točke geodetska udaljenost aproksimira se kao zbroj euklidskih udaljenosti duž najkraćeg povezanog puta. Postoji nekoliko algoritama zasnovanih na grafovima za izračunavanje ove aproksimacije. Jednom kada se postigne geodetska udaljenost, MDS se izvodi kao što je gore objašnjeno. Aproksimacija geodetske udaljenosti nije robusna na šum, što je nedostatak ovog algoritma.

[1] Ovdje I_n označava $n \times n$ jediničnu matricu, a $\mathbf{1}$ k-dimenzionalni vektor jedinica

1.2.1 Analiza glavnih komponenti

Ponekad među opažanim varijablama (svojstvima) nije naizgled očito koje od njih jesu, a koje nisu korelirane. To, međutim, ne znači da među njima ne postoje nekorelirane varijable koje ih opisuju (npr. neka varijabla je linearna kombinacije nekih nekoreliranih varijabli). Pronalazak takvih varijabli bio bi od velike važnosti upravo na već spomenutim skupovima podataka gdje je promatranih varijabli jako puno. Jedan je od načina kako iz originalnih varijabli pronaći linearno nekorelirane varijable koje ih opisuju analiza glavnih komponenti (engl. *principal component analysis* ili skraćeno PCA). PCA je tehnika linearne redukcije dimenzija čiji je cilj pronaći linearno preslikavanje između prostora visoke dimenzije (npr. n) i potprostora dimenzije d ($d < n$) koji bilježi većinu varijabilnosti u podatcima. Podtprostor je određen s d ortogonalnih vektora koje zovemo glavnim komponentama. Formalno, neka su $x_1, \ldots, x_n$ opažene varijable, matrica $C \in \mathbb{R}^{n\times n}$ td. je $C_{ij} = \mathrm{Cov}(x_i, x_j)$ (na mjestu (i, j) je kovarijanca[2] varijabli x_i i x_j). Glavne komponente su određene dominantinim svojstvenim vektorima kovarijacijske matrice podataka C. Prva glavna komponenta zahvaća najveću varijancu originalnih opažanja, a $(i + 1)$-ta glavna komponenta zahvaća najveću varijancu originalnih opažanja projiciranih na ortogonalni komplement potprostora prvih i glavnih komponenti.

Za svaku linearnu kombinaciju varijabli (u slučaju dvodimenzionalnog prostora radi se o pravcima u ravnini) promotrimo ortogonalnu projekciju točaka na taj pravac, što je prikazano na slici 1.3 . Vidimo da, ovisno o nagibu pravca, projekcije su raspoređene rjeđe ili gušće.[3] Po gornjem opisu glavnih komponenti, (prva) glavna komponenta bila bi ona linearna kombinacija koja zahvaća najveću varijancu među podatcima, to jest, ona u čijoj su ortogonalnoj projekciji podatci najraspršeniji. Dakle, potrebno je maksimizirati raspršenost projekcije na pravac, što je prikazano na slici 1.3 (prva slika u drugom redu).

Za $x_1, \ldots, x_n$ opažene varijable i kovarijacijsku matricu $C \in \mathbb{R}^{n\times n}$ tražimo matrice $W, \Lambda \in \mathbb{R}^{n\times n}$ takve da je Λ dijagonalna matrica i da vrijedi

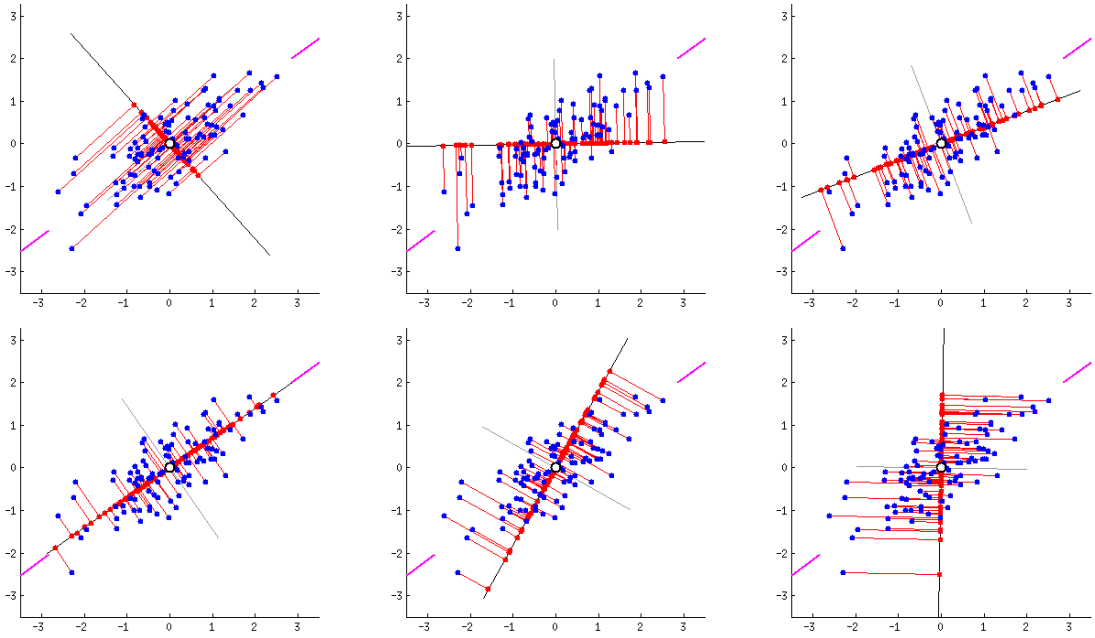
$$C = W\Lambda W^{\mathrm{T}}.$$

Osim toga, zahtijevat ćemo i da su svojstvene vrijednosti (od kojih su sve nenegativne) u Λ poredane silazno od mjesta $(1, 1)$ do mjesta (n, n) i da su stupci matrice W normirani. Ako su svojstvene vrijednosti jednostruke, takva je dekompozicija jedinstvena do na orijentaciju i poredak vektora (stupaca) u W. S druge strane, ako postoji neka višestruka svojstvena vrijednost, npr. kratnosti k, onda postoji k-dimenzionalni potprostor u tom prostoru u kojem je svaki vektor norme 1 svojstveni vektor toj svojstvenoj vrijednosti.

[2] Kovarijanca dviju varijabli X, Y u nekom vjerojatnosnom prostoru jednaka je njihovoj zajedničkoj varijabilnosti, tj. $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. Dodatno vrijedi $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ i $\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var(X)}$.

[3] Slike su preuzete sa stranice `https://giphy.com/gifs/pca-Lyejb62QjQepG`.

Slika 1.3: Prikaz ortogonalnih projekcija točaka na različite pravce u dvodimenzionalnom prostoru. Pronalazak prve glavne komponente postiže se na prvoj slici u drugom redu, u slučaju kada pravac zahvaća najveću varijancu među podatcima.



Formalno, vektori u W koji pripadaju svojstvenoj vrijednosti 0 nisu glavne komponente, oni ne zahvaćaju nikakvu varijancu originalnih varijabli. No, ponekad se zanemaruju vektori koji pripadaju "vrlo malim" svojstvenim vrijednostima jer oni vjerojatno imaju vrlo malu informativnost te njihovo zanemarivanje neće utjecati na točnost modela.

Za neku matricu Y tako da stupci matrice Y odgovaraju redom varijablama $x_1, \ldots, x_n$, a redci opažanjima projekcija Y' dobije se množenjem

$$Y' = YW.$$

Ako nas zanima projekcija samo na prvih k glavnih komponenti, za matricu W_k, čiji su stupci redom prvih k stupaca matrice W, možemo promatrati i projekciju

$$Y'_k = YW_k.$$

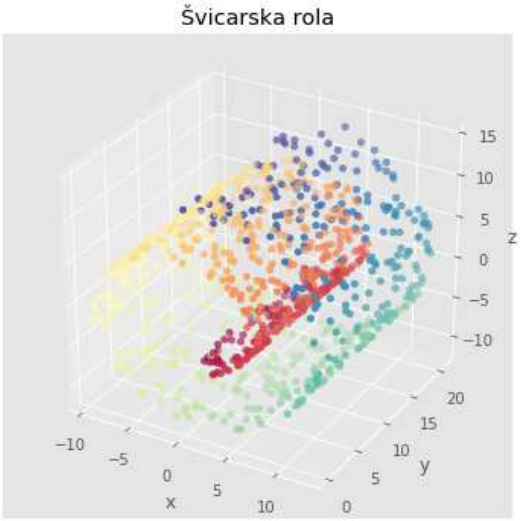
Ova metoda jednostavna je za provedbu, ali mnogi skupovi podataka u stvarnom svijetu imaju nelinearne karakteristike koje PCA preslikavanje ne uspijeva otkriti.U području multivarijatne statistike, postoje neke metode koje su proširenje klasične analize glavnih komponenata:

i) Jezgrena analiza glavnih komponenti (*kernel* PCA) koristi se jezgrenom funkcijom za projekciju podataka u prostor koji je linearno separabilan. Za razliku od PCA, može se uspješno primijeniti za nelinearne strukture.

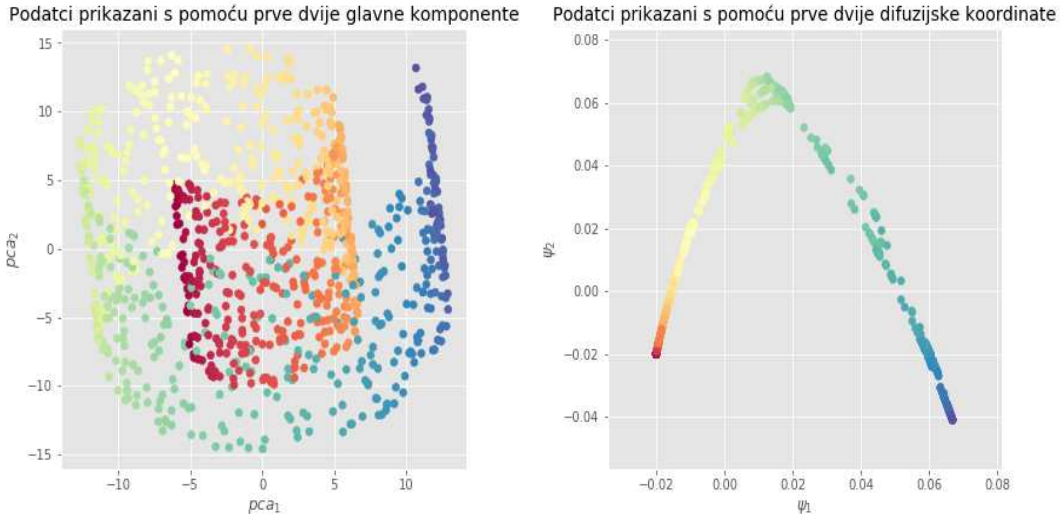
ii) Multilinearna analiza glavnih komponenti (MPCA) multilinearno je proširenje PCA koje se koristi za analiziranje i modeliranje tenzora.

Primjer 1.2.1. Jedan vrlo jasan primjer početnog skupa podataka na kojem PCA "pada" trodimenzionalna je tzv. švicarska rola. PCA pretpostavlja da podatci imaju linearnu projekciju na novom skupu osi. Jedna od nelinearnih alternativa PCA također su difuzijska preslikavanja, koja uzimaju u obzir temeljni oblik podataka. Primjer je ilustriran na slikama. Na slici 1.4 vidimo originalni skup podataka generiran u programskom jeziku *Python*. Slika 1.5 prikazuje projekciju švicarske role u $\mathbb{R}^2$ redom na prve dvije glavne komponente i na prve dvije difuzijske koordinate (definirane u 3. poglavlju). Vidimo kako se analizom glavnih komponenti podatci ne mogu smisleno projicirati na nove osi, ne prepoznaje se struktura originalnih podataka. S druge strane, difuzijsko preslikavanje omogućava prikaz "odmotane" švicarske role, gdje je druga slika samo jedna takva realizacija, što će biti objašnjeno dalje u radu.

Slika 1.4: Prikaz originalnih podataka, 1000 točaka koje generiraju švicarsku rolu.



Slika 1.5: Prikaz razlike ulaganja podatkovnih točaka švicarske role koristeći se prvim dvjema PCA koordinatama i prvim dvjema difuzijskim koordinatama. Primijetimo kako PCA ne uspijeva otkriti strukturu originalnih podataka za razliku od difuzijskog preslikavanja koje ćemo obraditi u ovom radu.



Poglavlje 2

Markovljevi lanci

U svijetu postoji aktivna i raznolika zajednica istraživača koji koriste Markovljeve lance u računalnim znanostima, fizici, statistici, bioinformatici, inženjerstvu i mnogim drugim područjima. Cilj je klasične teorije Markovljevih lanaca procijeniti stopu konvergencije do stacionarnosti vjerojatnosne distribucije u vremenu t, kako $t \to \infty$. U posljednja dva desetljeća, kako se povećao interes za ovu temu, pojavila se drugačija asimptotska analiza. Naime, zanima nas ciljna udaljenost do stacionarne distribucije; broj koraka potrebnih za postizanje ovog cilja, tzv. vrijeme miješanja (engl. *mixing time*) Markovljeva lanca, u oznaci t_{mix}. Za statističare i fizičare Markovljevi lanci korisni su u Monte Carlo simulacijama, posebno za konačne modele, vidi [17]. Vrijeme t_{mix} može odrediti vrijeme rada za simulaciju. Dodatno, koriste se i kao modeli dinamičkih procesa. U isto su vrijeme matematičari intenzivno proučavali miješanje karata i slučajne šetnje po grupama. Spektralne metode i vjerojatnosne tehnike, poput tzv. spajanja dvije vjerojatnosne mjere (engl. *coupling*[1]), igrale su važnu ulogu. Postoje mnoge metode za određivanje asimptotske konvergencije u stacionarnost, koje koriste funkcije veličine i geometrije prostora stanja. Prije svega, proučavanje konvergencije Markovljeva lanca središnji je dio moderne teorije vjerojatnosti, no postoje veze i s nekoliko drugih matematičkih područja. Ponašanje slučajne šetnje na grafu otkriva značajke geometrije grafa. Mnogi fenomeni, koji se mogu primijetiti kod konačnih grafova, također se javljaju u diferencijalnoj geometriji [1].

U ovom poglavlju navodimo neke osnovne definicije i svojstva Markovljevih lanaca koji su nam potrebna za razumijevanje metode opisane u radu.

[1] Spajanje dvije vjerojatnosne mjere μ i υ na skupu S je par slučajnih varijabli (X, Y) koje imaju zajedničku distribuciju q na $S \times S$, tako da su marginalne distribucije dane sa $\mathbb{P}(X = x) = \sum_{y \in S} q(x, y) = \mu(x)$ te $\mathbb{P}(Y = y) = \sum_{x \in S} q(x, y) = \upsilon(y)$, za svaki $x, y \in S$.

2.1 Definicija i osnovna svojstva

Definicija 2.1.1. *Neka je S skup.* ***Slučajan proces*** *s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli (ili elemenata) definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u S. Dakle, za svaki $n \geq 0$, je $X_n : \Omega \to S$ slučajna varijabla.*

Definicija 2.1.2. *Neka je S prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$, s vrijednostima u skupu S je* ***Markovljev lanac*** *ako vrijedi*

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i), \tag{2.1}$$

za svaki $n \geq 0$ i za sve $i_0, \ldots i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane. Svojstvo u relaciji (2.1) naziva se Markovljevim svojstvom.

Slika 2.1: Markovljevo svojstvo

$$\mathbb{P}(\text{budućnost} \mid \text{sadašnjost, prošlost}) = \mathbb{P}(\text{budućnost} \mid \text{sadašnjost})$$

Pretpostavimo da se nalazimo u vremenskom trenutku n. Tada vrijeme $n+1$ predstavlja neposrednu budućnost, dok vremena $0, 1, \ldots, (n-1)$ predstavljaju prošlost. Markovljevo svojstvo nam govori da je ponašanje Markovljeva lanca u neposrednoj budućnosti, uvjetno na sadašnjost i prošlost, jednako ponašanju Markovljeva lanca u neposrednoj budućnosti, uvjetno na samo sadašnjost. Dodatno, Markovljev lanac zovemo i Markovljev proces.

Definicija 2.1.3. *Neka je $\mu = (\mu_i : i \in S)$ vjerojatnosna distribucija na S, te neka je $P = (p_{ij} : i, j \in S)$ stohastička matrica. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s prostorom stanja S je homogen Markovljev lanac s početnom distribucijom μ i prijelaznom matricom P ako vrijedi*

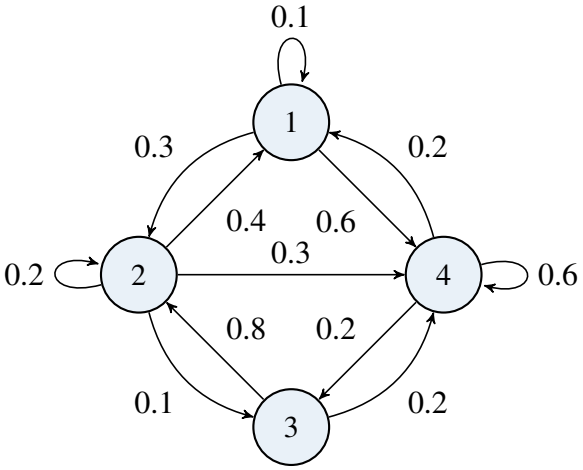
(i) $\mathbb{P}(X_0 = i) = \mu_i, \quad$ *za sve* $i \in S$,

(ii) $\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p_{ij}, \quad$ *za svaki* $n \geq 0$ *i za sve* $i_0, \ldots, i, j \in S$.

Radi jednostavnosti, Markovljev lanac iz definicije 2.1.3 zovemo (μ, P)-Markovljevim lancem. U radu ćemo pretpostaviti da je skup stanja S konačan. Na slici 2.2 vidimo primjer jednog Markovljeva lanca s četiri skupa stanja $(1, 2, 3, 4)$ i matricom prijelaza P. Strelice iz pojedinih stanja označavaju moguće tranzicije, a brojevi pored vjerojatnosti tih tranzicija.

Za razumijevanje evolucije Markovljeva lanca važno je znati koji su putevi kroz prostor stanja mogući. Postavlja se pitanje koja stanja lanac uopće može posjetiti krenuvši iz nekog zadanog stanja.

Slika 2.2: Prikaz primjera Markovljeva lanca i pripadna matrica prijelaza iz mogućih stanja. Primijetimo kako je suma svakog reda matrice P jednaka 1 (jer je matrica P stohastička).



$$P = \begin{bmatrix} 0.1 & 0.3 & 0 & 0.6 \\ 0.4 & 0.2 & 0.1 & 0.3 \\ 0 & 0.8 & 0 & 0.2 \\ 0.2 & 0 & 0.2 & 0.6 \end{bmatrix}$$

Definicija 2.1.4. *Neka je $X = (X_n : n \geq 0)$ Markovljev lanac s prostorom stanja S i prijelaznom matricom P. Za $B \subset S$ definiramo **prvo vrijeme pogađanja** tog skupa kao*

$$T_B = \min\{n \geq 0 : X_n \in B\}, \tag{2.2}$$

uz konvenciju da je $\min \varnothing = \infty$. U slučaju $B = \{j\}$ za $j \in S$ zbog jednostavnosti pišemo T_j umjesto preciznijeg $T_{\{j\}}$.

Ako je μ početna distribucija Markovljeva lanca $X = (X_n : n \geq 0)$ definiranog na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ sa skupom stanja S takva da $\mathbb{P}(X_0 = i) > 0$, $i \in S$, tada možemo definirati uvjetnu vjerojatnost $\mathbb{P}_i$ na sljedeći način

$$\mathbb{P}_i(A) = \mathbb{P}(A|X_0 = i), \quad \text{za bilo koji događaj } A \in \mathcal{F}.$$

Definicija 2.1.5. *Za stanja $i, j \in S$ kažemo da je j dostižno iz i u oznaci $i \longrightarrow j$ ako vrijedi*

$$\mathbb{P}_i(T_j \leq \infty) > 0. \tag{2.3}$$

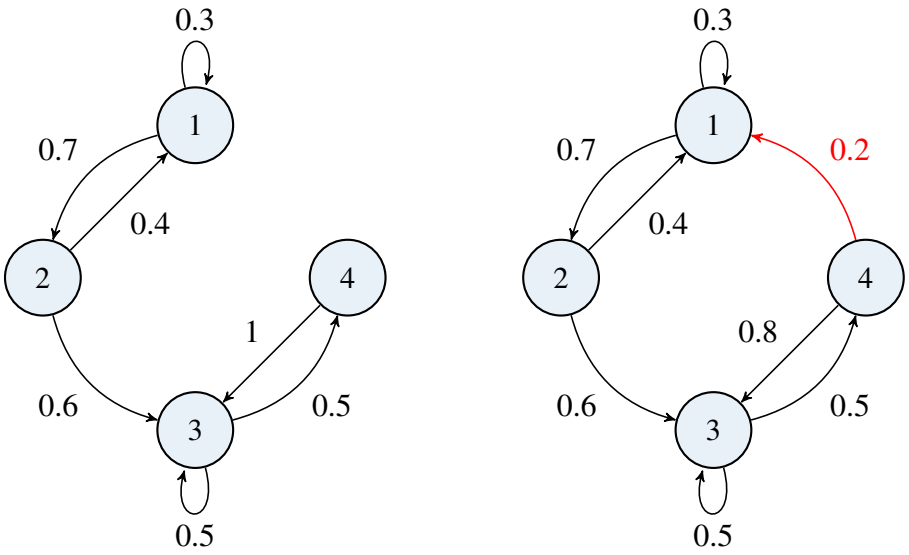
Dodatno, stanja komuniciraju, u oznaci $i \longleftrightarrow j$, ako vrijedi $i \longrightarrow j$ i $j \longrightarrow i$.

Relacija komuniciranja relacija je ekvivalencije na $S \times S$, stoga inducira particiju prostora stanja S na klase. Sva stanja iz jedne klase međusobno komuniciraju.

Definicija 2.1.6. *Markovljev lanac X je **ireducibilan** ako se prostor stanja S sastoji samo od jedne klase komuniciranja, t.j., za sve $i, j \in S$ vrijedi $i \longleftrightarrow j$.*

Na slici 2.3 prikazana su dva Markovljeva lanca. Prvi nije ireducibilan, jer npr. jednom kada dođemo u stanje 3 ili 4, ne možemo doći više u stanje 1 ili 2. Druga slika prikazuje kako, nakon što na lijevi lanac dodamo mogućnost prelaska iz stanja 4 u stanje 1, lanac postaje ireducibilan.

Slika 2.3: Ilustracija svojstva ireducibilnosti Markovljeva lanaca. Lijeva slika prikazuje lanac koji nije ireducibilan zbog činjenice da kada jednom dođemo u stanje 3 ili 4 ne možemo doći u stanje 1 ili 2. Desni je lanac ireducibilan, svako stanje je dostižno iz svih ostalih.



2.2 Granična i stacionarna distribucija

Nadalje, definiramo graničnu i stacionarnu distribuciju, pokazujemo njihovu vezu te rješavamo pitanje egzistencije granične distribucije.

Definicija 2.2.1. *Neka je $X = (X_n : n \geq 0)$ Markovljev lanac s prebrojivim skupom stanja S i prijelaznom matricom P. Vjerojatnosna distribucija $\pi = (\pi_i : i \in S)$ na S je **stacionarna distribucija** (ili invarijantna distribucija) Markovljeva lanca X (odnosno prijelazne matrice P) ako vrijedi*

$$\pi = \pi P, \tag{2.4}$$

odnosno po komponentama

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, \qquad za\ sve\ j \in S. \tag{2.5}$$

Teorem 2.2.2. *Neka je S konačan skup stanja, te pretpostavimo da za neki $i \in S$,*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \qquad za\ sve\ \ j \in S,$$

gdje je $p_{ij}^{(n)} = \mathbb{P}_i(X_n = j)$. Tada je $\pi = (\pi_j : j \in S)$ stacionarna distribucija.

Dokaz. Vrijedi

$$\sum_{j\in S} \pi_j = \sum_{j\in S} \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{j\in S} p_{ij}^{(n)} = 1,$$

pa je π vjerojatnosna distribucija. Zamjena limesa i sume opravdana je zbog konačnosti skupa S. Nadalje,

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = lim_{n\to\infty} p_{ij}^{(n+1)} = \lim_{n\to\infty} \sum_{k\in S} p_{ik}^{(n)} p_{kj} = \sum_{k\in S} \pi_k p_{kj}.$$

Dakle, π je stacionarna distribucija. □

Definicija 2.2.3. *Neka je $X = (X_n : n \geq 0)$ Markovljev lanac na skupu stanja S s prijelaznom matricom P. Vjerojatnosna distribucija $\pi = (\pi_i : i \in S)$ na S naziva se **graničnom distribucijom** Markovljeva lanca X (odnosno prijelazne matrice P) ako za sve $i, j \in S$ vrijedi*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j.$$

Definicija 2.2.4. *Niz $\lambda = (\lambda_i : i \in S)$ naziva se mjera ako je $\lambda_i \in [0, \infty)$ za sve $i \in S$. Mjera λ je netrivijalna ako postoji $i \in S$ takav da je $\lambda_i > 0$. Neka je $X = (X_n : n \geq 0)$ Markovljev lanac s prijelaznom matricom P. Netrivijalna mjera λ na S je **invarijantna mjera** Markovljeva lanca X(odnosno prijelazne matrice P) ako vrijedi*

$$\lambda = \lambda P,$$

odnosno po komponentama

$$\lambda_j = \sum_{k\in S} \lambda_k p_{kj}, \qquad za\ sve\ j \in S.$$

Definicija 2.2.5. *Neka je $X = (X_n : n \geq 0)$ Markovljev lanac na skupu stanja S s prijelaznom matricom P. Za stanje $i \in S$ označimo s $d(i)$ najveći zajednički djelitelj (nzd) skupa $\{n \geq 1 : p_{ii}^n > 0\}$, gdje je $d(i) = 1$ ako je taj skup prazan. Kažemo da je stanje i **aperiodično** ako je $d(i) = 1$. U suprotnom je i periodičko stanje, a $d(i)$ njegov period.*

Teorem 2.2.6. *Neka je μ proizvoljna vjerojatnosna distribucija na skupu stanja S. Pretpostavimo da je $X = (X_n : n \geq 0)$ (μ, P)-Markovljev lanac koji je ireducibilan i aperiodičan te ima stacionarnu distribuciju π. Tada je*

$$\lim_{n\to\infty} \mathbb{P}(X_n = j) = \pi_j, \qquad za\ sve\ j \in S. \tag{2.6}$$

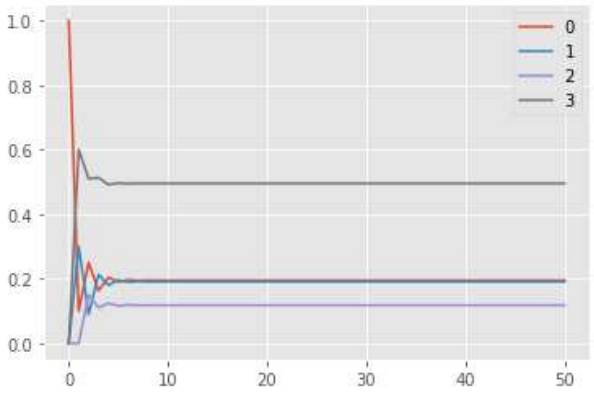
Specijalno,

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \qquad za\ sve\ i, j \in S, \tag{2.7}$$

t.j. stacionarna distribucija ujedno je i granična.

Dokaz. Vidi [18], 53.-54. stranica. □

Slika 2.4: Stacionarna distribucija Markovljeva lanca iz primjera prikazanog na slici 2.2



Još jedno zanimljivo svojstvo vezano za ponašanje Markovljeva svojstva je ergodičnost. Ako je lanac ireducibilan, tada je i ergodičan. Pretpostavimo da imamo funkciju f na skupu stanja S koja preslikava u $\mathbb{R}$. Tada možemo definirati srednju vrijednost ove funkcije duž određene putanje i prostornu sredinu, srednju vrijednost funkcije po skupu stanja S otežanu stacionarnom distribucijom.

Ergodski teorem govori nam da je vremenska srednja vrijednost, kada putanja postane beskrajno duga, jednaka prostornoj sredini, tj. za $n \in N$, imamo

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \sum_{e\in E} \pi(e) f(e).$$

2.3 Markovljevi lanci unatrag

Markovljevo svojstvo kaže da su budućnost i prošlost uvjetno nezavisni uz danu sadašnjost. Prema tome, prošlost i budućnost simetrične su, što sugerira da Markovljev lanac promatramo u vremenu koje teče unatrag. S druge strane, po teoremu o graničnoj distribuciji jasno je da distribucija koncentrirana u danom stanju teži prema stacionarnoj distribuciji kada vrijeme teče unaprijed. To pokazuje da potpunu simetriju u vremenu nećemo moći dobiti osim u slučaju da je i početna distribucija lanca stacionarna.

Definicija 2.3.1. *Za stohastičku matricu P i mjeru λ kažemo da su u **detaljnoj ravnoteži** ako vrijedi*

$$\lambda_i p_{ij} = \lambda_j p_{ji}, \qquad \text{za sve } i, j \in S.$$

Pretpostavimo da u stanju i imamo masu veličine λ_i, a u stanju j masu λ_j. Redistribuiramo li te mase s pomoću prijelazne matrice P, iz stanja i u stanje j odlazi masa $\lambda_i p_{ij}$, dok iz stanja j u stanje i dolazi jednaka masa $\lambda_j p_{ji}$. To pokazuje da je λ invarijantna mjera za P.

Lema 2.3.2. *Ako su P i λ u detaljnoj ravnoteži, tada je λ invarijantna mjera za P.*

Definicija 2.3.3. *Neka je λ distribucija na S. Za ireducibilan (λ, P)-Markovljev lanac $X = (X_n : n \geq 0)$ kažemo da je **reverzibilan** ako je za sve $N \geq 1$ $(X_{N-n} : 0 \leq n \leq N)$ ponovno (λ, P)-Markovljev lanac.*

Tipičan su primjer reverzibilnih Markovljevih lanaca slučajne šetnje na grafovima.

Primjer 2.3.4. Neka je G povezan, lokalno konačan graf, $G = (V, E)$, gdje je V skup vrhova, a E skup bridova. Na skup bridova gledamo kao na podskup Kartezijeva produkta $V \times V$ sa svojstvom $(i, j) \in E$ ako i samo ako je $(j, i) \in E$ (ako brid povezuje vrhove i i j, onda povezuje i vrhove j i i). Povezanost grafa znači da za svaka dva vrha $i, j \in V$ postoji konačan niz bridova $(i, i_1), (i_1, i_2), ..., (i_n, j)$ koji povezuju ta dva vrha, što se postiže kada je lanac ireducibilan.[2] Lokalna konačnost znači da iz svakog vrha izlazi najviše konačno mnogo bridova. Pretpostavimo da je svakom bridu $e \in E$ pridružen strogo pozitivan broj c_e koji zovemo *provodljivost* brida. Dakle, $c : E \to (0, \infty)$. Ako je $e = (i, j) = (j, i)$, pišemo $c_e = c_{ij} = c_{ji}$. Ako $(i, j) \notin E$ imamo $c_{ij} = 0$. Za vrh $i \in V$ definiramo *kapacitet* tog vrha kao $c_i = \sum_{k \in V} c_{ik}$. Slučajna šetnja na G Markovljev je lanac $X = (X_n : n \geq 0)$ sa skupom stanja V i prijelaznom matricom $P = (p_{ij} : i, j \in V)$ gdje je

$$p_{ij} = \frac{c_{ij}}{c_i}.$$

[2]Prema definiciji 2.1.6, ireducibilnog lanca, znamo da sva stanja međusobno komuniciraju, što znači da postoji put između njih. Dakle, ireducibilan graf je povezan. Dodatno, vrijedi i obrat.

Drugim riječima, šetnja iz vrha i prelazi u jedan od susjednih vrhova s vjerojatnošću proporcionalnom provodljivosti odgovarajućeg brida. Zbog pretpostavke o povezanosti grafa slijedi da je šetnja X ireducibilna. Stavimo $c = \sum_{i \in V} c_i$, te definiramo $\pi = (\pi_i : i \in V)$ sa

$$\pi_i = \frac{c_i}{c}.$$

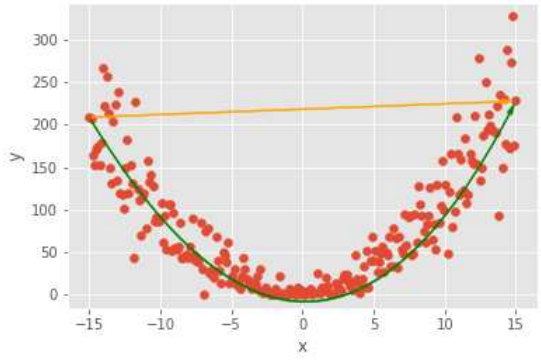
Tada je

$$\pi_i p_{ij} = \frac{c_i}{c}\frac{c_{ij}}{c_i} = \frac{c_{ij}}{c} = \frac{c_j}{c}\frac{c_{ij}}{c_j} = \pi_j p_{ji},$$

odnosno P i π su u detaljnoj ravnoteži. To pokazuje da je π stacionarna distribucija od X te da je X reverzibilan.

Na slici 2.5 vidimo dvije moguće putanje između slučajno odabranih točaka. Putanja koju napravi slučajna šetnja, označena zelenom bojom, kreće duž geometrijske strukture. S druge strane, narančasti put je najkraći put između odabranih točaka, ali on nam ne daje nikakve informacije o strukturi podataka.

Slika 2.5: Dvije moguće putanje slučajne šetnje između dvije slučajno odabrane krajnje točke. U analizi nelinearnih struktura, poželjni su putevi koje slučajna šetnja postiže duž geometrijske strukture (put označen zelenom bojom).



2.4 Spektralna svojstva tranzicijske matrice reverzibilnog lanca

Sljedećom lemom pokazujemo osnovne činjenice svojstvenih vrijednosti matrice prijelaza reverzibilnog Markovljeva lanca.

Lema 2.4.1. *Neka je P tranzicijska matrica konačnog Markovljeva lanca.*

(i) Ako je λ svojstvena vrijednost od P, onda je $|\lambda| \leq 1$.

(ii) Ako je P ireducibilan Markovljev lanac, svojstveni vektor koji odgovara svojstvenoj vrijednosti 1 je jednodimenzionalni prostor generiran vektorom $\mathbf{1} := (1, \ldots, 1)^T$.

Dokaz. Za bilo koju funkciju f vrijedi

$$\|Pf\|_\infty = \max \left| \sum_{y \in \Omega} P(x,y) f(y) \right| \leq \|f\|_\infty. \tag{2.8}$$

Neka je φ svojstveni vektor koji pripada svojstvenoj vrijednosti λ. Tada imamo $P\varphi = \lambda\varphi$ pa vrijedi $\|P\varphi\|_\infty = |\lambda| \|\varphi\|_\infty$. Kako (2.8) vrijedi za bilo koju funkciju f, posebno vrijedi i za φ pa imamo

$$|P\varphi\|_\infty = |\lambda| \|\varphi\|_\infty \leq \|\varphi\|_\infty,$$

iz čega slijedi tvrdnja pod (i).

Nadalje, uvrštavanjem $\lambda_0 = 1$ u $P\varphi = \lambda\varphi$ i korištenjem činjenice da je P stohastička matrica slijedi tvrdnja pod (ii). □

Označimo sa $\langle \cdot, \cdot \rangle$ standardni skalarni produkt na $\mathbb{R}^S$, dan sa $\langle f, g \rangle = \sum_{x \in S} f(x) g(x)$. Nadalje, definiramo još jedan skalarni produkt, u oznaci $\langle \cdot, \cdot \rangle_\pi$:

$$\langle f, g \rangle_\pi := \sum_{x \in S} f(x) g(x) \pi(x).$$

$L^2(\pi)$ označava vektorski prostor $\mathbb{R}^S$ s upravo definiranim skalarnim produktom. Budući da elemente $\mathbb{R}^S$ smatramo funkcijama, sa S u $\mathbb{R}$, svojstvene vektore matrice P nazivamo još svojstvenim funkcijama. Kako smo već spomenuli, tranzicijska matrica P reverzibilna je u odnosu na stacionarnu distribuciju π ako vrijedi $\pi(x)p(x,y) = \pi(y)p(y,x)$ za svaki $x, y \in S$, gdje je S skup stanja, a $p(x,y) = \mathbb{P}(X_1 = x \mid X_0 = y)$. Posebno, za $t \geq 0$, vjerojatnost prijelaza iz x u y u t vremenskih koraka dana je $p_t(x,y)$. Preciznije, $p_t(x_i, x_j) = \mathbb{P}(X_t = x_j \mid X_0 = x_i) = P^t_{ij}$. Izraz $P^t(x,y)$ upotrebljavamo pod pretpostavkom da postoje indeksi i, j takvi da $x = x_i$, a $y = x_j$ i tada je ta vjerojatnost jednaka P^t_{ij}. Vrijedi sljedeća lema koja nam daje svojstvenu dekopomoziciju potrebnu za uspješnost difuzijskih preslikavanja.

Lema 2.4.2. *Neka imamo reverzibilan Markovljev lanac s matricom prijelaza P i stacionarnom distribucijom π, skupom stanja $S = \{x_0, \ldots, x_{n-1}\}$ i $n = |S|$.*

(i) Prostor $(\mathbb{R}^S, \langle \cdot, \cdot \rangle_\pi)$ ima ortonormiranu bazu koju formiraju svojstvene funkcije $\{f_j\}_{j=0}^{n-1}$ matrice P koji odgovaraju svojstvenim vrijednostima $\{\lambda_j\}_{j=0}^{n-1}$.

(ii) Za $x = x_i$, $y = x_k$, $i,k \in \{0,\dots,n-1\}$ i bilo koje vrijeme t, matrica P ima sljedeću dekompoziciju

$$\frac{P^t(x,y)}{\pi(y)} = \sum_{j=0}^{n-1} f_j(x) f_j(y) \lambda_j^t.$$

(iii) Svojstvena funkcija f_1 koja odgovara svojstvenoj vrijednosti 1*, može biti konstantan vektor* $\mathbf{1}$*, što implicira sljedeće*

$$\frac{P^t(x,y)}{\pi(y)} = 1 + \sum_{j=1}^{n-1} f_j(x) f_j(y) \lambda_j^t.$$

Dokaz. Definiramo $A(x,y) := \frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}} P(x,y)$, gdje iz reverzibilnosti od P slijedi simetričnost od A. Spektralni teorem[3] za simetrične matrice nam pokazuje kako prostor $(\mathbb{R}^S, \langle\cdot,\cdot\rangle)$ ima ortonormiranu bazu $\{\varphi_j\}_{j=0}^{n-1}$ tako da je φ_j svojstvena funkcija s realnom svojstvenom vrijednosti λ_j

Direktnim uvrštavanjem vrijedi da je $\sqrt{\pi}$ (korijen iz stacionarne distribucije) svojstveni vektor od A za svojstvenu vrijednost 1. Označimo s D_π dijagonalnu matricu takvu da za $x = x_i$ vrijedi $D_\pi(x,x) = \pi(x)$, tada vrijedi $A = D_\pi^{1/2} P D_\pi^{-1/2}$. Dodatno, $f_j := D_\pi^{-1/2}\varphi_j$ je svojstveni vektor od P tj. vrijedi

$$Pf_j = PD_\pi^{-1/2}\varphi_j = D_\pi^{-1/2}(D_\pi^{1/2} P D_\pi^{-1/2})\varphi_j = D_\pi^{-1/2} A\varphi_j = D_\pi^{-1/2}\lambda_j\varphi_j = \lambda_j f_j.$$

Skup $\{f_j\}$ je ortonormiran u odnosu na skalarni produkt $\langle\cdot,\cdot\rangle_\pi$:

$$\delta_{ij} = \langle\varphi_i,\varphi_j\rangle = \langle D_\pi^{1/2} f_j, D_\pi^{1/2} f_j\rangle = \langle f_i, f_j\rangle_\pi.$$

Prva jednakost slijedi iz činjenice da je skup $\{\varphi_j\}$ u odnosu na standardni skalarni produkt, pa je dokazana tvrdnja pod (i).

Neka je δ_y funkcija definirana s

$$\delta_y(x) = \begin{cases} 1 & \text{ako je } y = x, \\ 0 & \text{ako je } y \neq x. \end{cases}$$

Unitarni prostor $(\mathbb{R}^S, \langle\cdot,\cdot\rangle_\pi)$ ima ortonormiranu bazu[4] $\{f_0,\dots,f_{n-1}\}$, pa možemo svaki element v tog prostora jedinstveno prikazati kao $v = \sum_{i=0}^{n-1}\beta_i f_i$. Skalarnim množenjem

[3] Realna simetrična matrica $A \in M_{nn}$ ima n ortogonalnih svojstvenih vektora s realnim svojstvenim vrijednostima.

[4] Ortonormiran skup u unitarnom prostoru je ortonormirana baza ako je taj skup ujedno i baza prostora.

prethodne jednakosti s f_k, zbog ortogonalnosti elementa, dobivamo $\beta_k = \langle v, f_k\rangle_\pi$ za sve $k = 0, \ldots, n-1$. Dakle, funkciju $\delta_y(x)$ definiranu gore možemo prikazati kao

$$\delta_y = \sum_{j=0}^{n-1} \langle \delta_y, f_j\rangle_\pi f_j = \sum_{j=0}^{n-1} f_j(y)\pi(y) f_j.$$

Nadalje, povratom na notaciju s indeksima, za $x = x_i$ i $y = x_k$ vrijedi sljedeće

$$\begin{aligned}
P^t\delta_y(x) &= e_i^t P^t e_k \\
&= \begin{bmatrix} 0 & \cdots 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} p_t(x_1,x_1) & \cdots & p_t(x_1,x_k) & \cdots & p_t(x_1,x_n) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ p_t(x_i,x_1) & \cdots & p_t(x_i,x_k) & \cdots & p_t(x_i,x_n) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ p_t(x_n,x_1) & \cdots & p_t(x_n,x_k) & \cdots & p_t(x_n,x_n) \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \cdots 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} p_t(x_1,x_k) \\ \vdots \\ p_t(x_i,x_k) \\ \vdots \\ p_t(x_n,x_k) \end{bmatrix} \\
&= p_t(x_i,x_k) = P^t(x,y),
\end{aligned}$$

gdje je e_k kanonski vektor (stupac) s jedinicom na k-tom mjestu.

Iz $P^t f_j = \lambda_j^t f_j$ i $P^t(x,y) = (P^t\delta_y)(x)$ imamo

$$P^t(x,y) = \sum_{j=0}^{n-1} f_j(y)\pi(y)\lambda_j^t f_j(x),$$

pa dijeljenjem gornje jednadžbe s $\pi(y)$ slijede tvrdnje pod (ii) i (iii). □

Stoga za funkciju $f : S \to \mathbb{R}$ vrijedi

$$P^t f = \sum_{j=1}^{n-1} \langle f, f_j\rangle_\pi f_j \lambda_j^t. \tag{2.9}$$

2.5 Uvod u vrijeme miješanja Markovljeva lanca

Kako bismo analizirali asimptotsko ponašanje konačnog Markovljeva lanca i odredili brzinu konvergencije, moramo odabrati odgovarajuću metriku na prostoru distribucija.

Definicija 2.5.1. *Ukupna udaljenost varijacija (engl. total variation distance) između dvije vjerojatnosne distribucije π i υ na skupu stanja S definirana je s*

$$\|\mu - \upsilon\|_{TV} = \max_{A \subset S} |\mu(A) - \upsilon(A)|.$$

Dakle, udaljenost između μ i υ najveća je razlika između pripadnih vjerojatnosti ako idemo po svim događajima.

Nadalje, predstavljamo sljedeći teorem koji je od ključne važnosti za daljnju provjeru brzine konvergencije prema stacionarnoj distribuciji π.

Teorem 2.5.2 (Teorem o konvergenciji)**.** *Neka je X ireducibilan i aperiodičan Markovljev lanac, s matricom prilaza P, skupom stanja S i stacionarnom distribucijom π. Tada postoji $\alpha \in \langle 0, 1\rangle$ i $C > 0$ takvi da*

$$\max_{x \in S} \|P^t(x, \cdot) - \pi\|_{TV} \leq C\alpha^t. \tag{2.10}$$

Kako bismo ograničili maksimalnu udaljenosti između $P^t(x, \cdot)$ i π, prikladno je definirati

$$d(t) := \max_{x \in S} \|P^t(x, \cdot) - \pi\|_{TV}. \tag{2.11}$$

Nadalje, kao što je već spomenuto na početku ovog poglavlja, korisno je uvesti pojam vremena miješanja (engl. *mixing time*), koji mjeri vrijeme potrebno Markovljevu lancu do stacionarnosti u odnosu na parametar greške ϵ. Definiramo,

$$t_{\text{mix}}(\epsilon) := \min\{t \ : \ d(t) \leq \epsilon\}.$$

2.6 Vrijeme relaksacije

Za svaku tranzicijsku matricu P reverzibilnog Markovljeva lanca (sa skupom stanja S), koristeći 2.4.1 možemo svojstvene vrijednosti označiti u padajućem poretku:

$$1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{n-1} \geq -1.$$

Definiramo

$$\lambda_* := \max\{|\lambda| \ : \ \lambda \text{ je svojstvena vrijednost od } P, \lambda \neq 1\}. \tag{2.12}$$

Razliku $\gamma_* = 1 - \lambda_*$ nazivamo *apsolutna spektralna praznina* (engl. *spectral gap*). Prema lemi (2.4.1) slijedi da je P aperiodičan i ireducibilan, tada $\gamma_* > 0$. *Spektralna praznina* reverzibilnog lanca definiran je s $\gamma := 1 - \lambda_2$.

Definicija 2.6.1. *Vrijeme relaksacije (engl. relaxation time), u oznaci t_{rel}, reverzibilnog Markovljeva lanca s apsolutnom spektralnom prazninom γ_* je*

$$t_{rel} := \frac{1}{\gamma_*}. \tag{2.13}$$

U knjizi [4] detaljno je obrađena veza između vremena relaksacije, *mixing* vremena i stacionarne distribucije Markovljeva lanca.

Nadalje, kako vrijedi $\mathbb{E}_\pi(f) = \sum_{y\in S} f(y)\pi(y)$, slijedi da

$$\begin{aligned}
\mathbb{E}_\pi(P^t f) &= \sum_{x\in S}(P^t f)(x)\pi(x) \\
&= \sum_{x\in S}\sum_{y\in S} f(y)P^t(x,y)\pi(x) \\
&= \sum_{y\in S} f(y)\sum_{x\in S} P^t(x,y)\pi(x) \quad \text{(kako je } \pi \text{ stacionarna distribucija, iz (2.5) slijedi)} \\
&= \sum_{y\in S} f(y)\pi(y) = \mathbb{E}_\pi(f).
\end{aligned}$$

Prvi svojstveni vektor od P^t je vektor jedinica $f_1 = \mathbf{1}$ pa možemo pisati $\mathbb{E}_\pi(P^t f) = \mathbb{E}_\pi(f) = \langle P^t f, f_1\rangle_\pi$ i prema (2.9) slijedi da

$$P^t f - \mathbb{E}_\pi(P^t f) = \sum_{j=2}^{|S|}\langle f, f_j\rangle_\pi f_j \lambda_j^t.$$

Kako je skup $\{f_j\}$ ortonormiran u odnosu na skalarni produkt $\langle\cdot,\cdot\rangle_\pi$ slijedi

$$\begin{aligned}
\mathrm{Var}_\pi(P^t f) &= \langle P^t f - \mathbb{E}_\pi(P^t f), P^t f - \mathbb{E}_\pi(P^t f)\rangle_\pi \\
&= \sum_{j=2}^{|S|}\langle f, f_j\rangle_\pi^2 \lambda_j^{2t} \le (1-\gamma_*)^{2t}\sum_{j=2}^{|S|}\langle f, f_j\rangle_\pi^2.
\end{aligned}$$

Primijetimo da

$$\sum_{j=2}^{|S|}\langle f, f_j\rangle_\pi^2 = \sum_{j=1}^{|S|}\langle f, f_j\rangle_\pi^2 - \mathbb{E}_\pi^2(f) = \mathbb{E}_\pi(f^2) - \mathbb{E}_\pi^2(f) = \mathrm{Var}_\pi(f),$$

pa dobivamo jedno operativno značenje vremena opuštanja pokazano u ([4]) iz sljedeće nejednakosti:

$$\mathrm{Var}_\pi(P^t f) \le (1-\gamma_*)^{2t} Var_\pi(f). \tag{2.14}$$

Prema teoremu konvergencije (([4], 52. str) $P^t f(x) \to \mathbb{E}_\pi(f)$, za bilo koji $x \in S$, $P^t f$ se približava konstantnoj funkciji. S pomoću prethodne nejednakosti može se donijeti kvantitativni iskaz: ako je $t \geq t_{rel}$, tada je standardna devijacija od $P^t f$ ograničena s umnoškom $1/e$ i standardne devijacije od f. Sljedeća dva teorema pokazuju veze između t_{rel} i t_{mix}.

Teorem 2.6.2. *Neka je P tranzicijska matrica reverzibilnog, ireducibilnog Markovljeva lanca sa skupom stanja S te $\pi_{\min} = \min_{x \in S} \pi(x)$. Tada vrijedi*

$$t_{mix}(\epsilon) \leq \log\left(\frac{1}{\epsilon \pi_{\min}} t_{rel}\right).$$

Teorem 2.6.3. *Za reverzibilan, ireducibilan i aperiodičan Markovljev lanac*

$$t_{mix}(\epsilon) \geq (t_{rel} - 1)\log\left(\frac{1}{2\epsilon}\right). \tag{2.15}$$

Dokazi prethodna dva teorema i nejednakosti (2.14) nalaze se u ([4], 159. - 160.str).

Poglavlje 3

Teorija difuzijskih preslikavanja

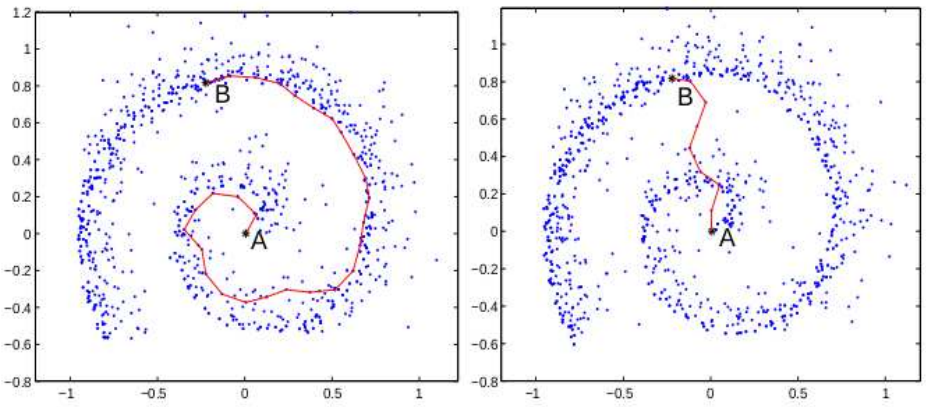
3.1 Modeliranje preko grafova

Brojne tehnike rudarenja podataka (engl. *data mining*) i strojnog učenja oslanjaju se na algoritme bazirane na grafovima. U pogledu strukture podataka, grafovi su jednostavni, interpretabilni i pogodni za predstavljanja složenih odnosa između podataka. Težinski grafovi obično se koriste kako bismo predstavili pojam geometrije temeljene na lokalnoj sličnosti ili interakciji između podataka. U mnogim je situacijama svaki uzorak podataka predstavljen skupom numeričkih atributa pa se tada uvjet za povezanost između dva čvora temelji na blizini odgovarajućih podataka u prostoru značajki. U kontekstu mreža (npr. društvene, računalne, komunikacijske ili transportne mreže) podatci se prirodno modeliraju s pomoću grafova. Primjerice, metode s grafovima igraju veliku ulogu u kvantitativnom modeliranju društvenih mreža. U kombinaciji s tehnikama na Markovljevim lancima, metode temeljene na grafovima mogu biti vrlo uspješne. Kod klasifikacije i grupiranja, slučajna šetnja na grafovima pokazala se vrlo učinkovitom u pronalaženju odgovarajućih struktura u složenoj geometriji. Na primjer, u [12], L^1 udaljenost [1] između vjerojatnosti prijelaza koristi se kao metrika između podatkovnih točaka u svrhu određivanje klasa. Pokazalo se da je ova tehnika prilično uspješna kada podatci imaju nelinearne oblike. Za spektralno grupiranje na grafu gradimo Markovljev lanac, a predznak vrijednosti nekonstantnog svojstvenog vektora odgovarajuće prijelazne matrice koristi se za pronalaženje grupa i računanje rezova [7].

Na slici 3.1 imamo dvije spirale sa šumom, s istaknutim točkama A i B (slika preuzeta iz [9]). U idealnom slučaju najkraći put između A i B trebao bi slijediti granu spirale (lijevo). Međutim, neke realizacije mogu stvoriti prečace, što drastično smanjuje duljinu najkraće staze (desno).

[1] Za $x = (x_1, \ldots, x_n)$ i $y = (y_1, \ldots, y_n)$, $d_{L^1}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$.

Slika 3.1: Prikaz dviju putanja slučajne šetnje na spirali. Slika lijevo prikazuje putanju koja se kreće duž geometrijske strukture podataka, dok je desna putanja kraća i ne daje informativnost o strukturi.



3.1.1 Različiti grafovi sličnosti

Postoji nekoliko popularnih konstrukcija transformacije danog skupa podataka $\{x_1, \ldots, x_n\}$ sa sličnostima parova točaka ili udaljenosti točaka na grafu. Pri konstruiranju grafova sličnosti cilj je modelirati odnose lokalnog susjedstva između podataka.

- Graf ϵ-susjedstva

  Ovdje se povezuje sve točke čija je međusobna kvadratna udaljenost manja od ϵ. Ako uzmemo u obzir euklidsku udaljenost, ovaj uvjet znači $\|x_i - x_j\|^2 \leq \epsilon$. Kako su udaljenosti između svih povezanih točaka približno iste (najviše ϵ), težine na rubovima ne bi povećavale informativnost podataka. Geometrijska motivacija i činjenica da je ovakva veza prirodno simetrična prednost je ovog pristupa. Njegov je nedostatak što vodi do nekoliko povezanih komponenti te je ponekad teško izabrati parametar ϵ.

- Graf k najbližih susjeda

  Čvorovi i i j povezani su ako je j među k najbližih susjeda od i. Ova je relacija simetrična. Prednost ovog pristupa je njegova jednostavnost te činjenica da ne teži nepovezanom grafu.

- Potpuno povezan graf

  Ovdje jednostavno povezujemo sve točke s pozitivnom funkcijom sličnosti k. Kako bi graf trebao predstavljati odnose lokalnog susjedstva, ova je konstrukcija korisna samo ako sama funkcija sličnosti modelira lokalno okruženje. Primjer je takve funkcije sličnosti je Gaussova jezgra definirana u idućem poglavlju.

3.2 Povezanost točaka

Neka je $(\mathbf{X}, \mathcal{A}, \mu)$ izmjeriv prostor, $\mathbf{X}$ je metrički prostor, a μ mjera na $(\mathbf{X}, \mathcal{A})$. Pretpostavimo da na skupu $\mathbf{X}$ definiramo slučajnu šetnju $(X_t,\, t \geq 0)$, skakanje između točaka tog prostora, (vidi sliku 2.2). Za $\mathbf{X} = \{x_1, \ldots, x_n\}$, gdje je svaki $x_i \in \mathbb{R}^p$, $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ predstavlja početnu distribuciju slučajne šetnje, tj. $\mu_i = \mathbb{P}(X_0 = x_i)$. Između točaka prostora definiramo tzv. difuzijsku jezgru $k : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$. Takva nenormirana vjerojatnosna funkcija k definira lokalu mjeru sličnosti u određenom području izvan kojeg funkcija brzo teži u 0. Popularna Gaussovska jezgra jedan je primjer takve funkcije, za $\varepsilon > 0$:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\varepsilon}\right).$$

Susjedstvo ili okruženje oko neke točke x u prostoru možemo definirati kao skup svih elemenata y za koje vrijedi $k(x, y) \geq \epsilon$ gdje je $0 \leq \epsilon \leq 1$. Unutar tog područja vjerujemo u točnost naše mjere sličnosti (npr. euklidska udaljenost). Promjenom vrijednosti ε u Gaussovoj jezgri biramo veličinu susjedstva na temelju prethodnog znanja o strukturi i gustoći podataka. Nadalje, u ovom je radu pokazano koliko je ključan odabir ovog parametra i koliko utječe na dobivene rezultate. Osim Gaussove, moguće su i druge jezgre te je često izbor ispravne jezgre presudan za uspješnost metode.

Jezgra k zadovoljava sljedeća svojstva:

1. k je simetrična: $k(x, y) = k(y, x)$,
2. k je pozitivna: $k(x, y) \geq 0$.

Ona predstavlja neki pojam sklonosti ili sličnosti između točaka skupa $\mathbf{X}$ jer opisuje odnos između parova točaka, pa možemo promatrati podatkovne točke kao čvorove simetričnog grafa čija je težinska funkcija k. Dodatno, karakterizira definiciju lokalne geometrije skupa $\mathbf{X}$, s obzirom na to da će izvući specifične značajke skupa podataka, za razliku od globalnih metoda poput analize glavnih komponenti ili višedimenzionalnog skaliranja gdje se uzimaju u obzir sve korelacije između podataka. Ovdje polazimo od ideje da su u mnogim aplikacijama visoke korelacijske vrijednosti jedina značajna informacija u skupu podataka. Ova pojava ilustrirana je u dodatku A definiranjem jednoparametarske familije jezgrenih funkcija. Nakon dobro odabrane jezgre, odgovarajuće se difuzije mogu koristiti za analizu geometrije, statistike ili neke dinamike podataka.

Bilo kojem reverzibilnom Markovljevu procesu može se pridružiti simetrični graf te obratno, iz grafa definiranog s $(\mathbf{X}, k)$ može se konstruirati reverzibilni Markovljev lanac $(X_t : t \geq 0)$ na $\mathbf{X}$. U različitim područjima primjene ova je tehnika klasična i poznata kao konstrukcija normaliziranog *Laplaciana* grafa. Neka je

$$d(x) = \int_{\mathbf{X}} k(x, y) d\mu(y)$$

lokalna mjera volumena (ili stupnja u grafu)[2].

Vidjet ćemo kasnije u radu da nam prvo svojstvo simetričnosti funkcije k omogućava korisna spektralna svojstva difuzijske matrice definirane u 3.2.1. Drugo svojstvo omogućuje nam tumačenje difuzijske jezgre kao skalirane vjerojatnosti, tako da vrijedi

$$\frac{1}{d(x)} \int_{\mathbf{X}} k(x,y) d\mu(y) = 1.$$

Promotrimo slučajnu šetnju na $\mathbf{X}$. Intuitivno, znamo da je skok na obližnju točku u podatcima vjerojatniji od skoka na drugu koja je daleko. Ovo promatranje daje odnos između udaljenosti među točkama u karakterističnom prostoru i vjerojatnosti prijelaza slučajne šetnje. Stoga možemo definirati povezanost između podatkovnih točaka x i y kao vjerojatnost prelaska s x na y u jednom koraku slučajne šetnje

$$\text{povezanost(x,y)} = p(x,y) = \frac{k(x,y)}{d(x)}, \tag{3.1}$$

gdje je $d(x)$ normalizacijska konstanta, a $p(x,y) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$.

Također vrijedi

$$\int_{\mathbf{X}} p(x,y) d\mu(y) = 1.$$

Definicija 3.2.1. *Neka je $\mathbf{X} = \{x_1, \ldots, x_n\}$ skup točaka. Definiramo **matricu povezanosti** K tako da vrijedi $K_{ij} = k(x_i, x_j)$ i dijagonalnu matricu D tako da vrijedi*

$$D_{ii} = \sum_{j=1}^{n} K_{ij}.$$

Posebno, $d(x) = D_{ii}$, za $x = x_i$, $i \in \{1, \ldots, n\}$.

*Dodatno, definiramo **difuzijsku matricu** P (matricu prijelaza) tako da vrijedi*

$$P_{ij} = p(x_i, x_j) = \mathbb{P}(X_{t+1} = x_j \mid X_t = x_i).$$

Svaka vrijednost upravo definirane matrice P osigurava povezanost između dvije točke x_i i x_j te obuhvaća ono što se zna lokalno. Kao što smo gore definirali, svaka vrijednost ova matrica predstavlja vjerojatnost kretnje jednog koraka Markovljeva lanca. Točnije, P_{ij} je vjerojatnost prijelaza u točku x_j počevši iz x_i u jednom koraku. Potenciranjem difuzijske matrice P povećava se broj pređenih koraka između točaka. Na primjer, uzmimo difuzijsku matricu 2×2

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}.$$

[2]Općenito, stupanj vrha u grafu je broj vrhova koji su incidentni s njime. U našem slučaju težinskog grafa taj broj vrhova je otežan pa se definicija svodi na zbroj težina bridova.

Kao što je definirano gore, svaki element $P_{ij} = p(x_i, x_j)$ predstavlja vjerojatnost prijelaza u jednom vremenskom koraku od čvora x_i do čvora x_j. Kvadriranjem matrice P dobivamo

$$P^2 = \begin{bmatrix} P_{11}P_{11} + P_{12}P_{21} & P_{12}P_{22} + P_{11}P_{12} \\ P_{21}P_{12} + P_{22}P_{21} & P_{22}P_{22} + P_{21}P_{12} \end{bmatrix}.$$

Primijetimo da je, na primjer, prvi element ove matrice $P_{11}P_{11} + P_{12}P_{21}$, što je zbroj dvije vjerojatnosti: vjerojatnost ostanka u točki x_1 u 2 koraka i vjerojatnost prelaska u točku x_2 pa ponovno vraćanje u x_1.

Na slici 3.2 vidimo prikazane matrice prijelaza P i njenih potencija (za $t = 8$, $t = 16$, $t = 32$) za podatke iz primjera 3.3.1. Podatci su sortirani po grupama, tako da je na matrici vidljiva grupacija podataka. Za manje potencije grupe su prepoznatljive, ali već za $t = 32$ podatkovne točke imaju velike vjerojatnosti prelaska iz jedne u drugu grupu.

S gledišta analize podataka, razlog je za proučavanje ovog Markovljeva lanca taj što matrica P sadrži geometrijske informacije o skupu podataka $\mathbf{X}$. Doista, tranzicije koje definira izravno odražavaju lokalnu geometriju koju definiraju neposredni susjedi svakog čvora u grafu. Za $t \geq 0$, vjerojatnost prijelaza iz x u y u t vremenskih koraka dana je s $p_t(x, y)$, dobivene iz P^t, tj. vrijedi $p_t(x_i, x_j) = \mathbb{P}(X_t = x_j \mid X_0 = x_i) = P^t_{ij}$. Jedna je od glavnih ideja da će nam kretanje lanca naprijed ili, ekvivalentno, uzimanje većih potencija od P, omogućiti otkrivanje relevantne geometrijske strukture skupa $\mathbf{X}$ u različitim vremenima [6].

3.3 Difuzijski proces

Potenciranjem matrice P za rastuće vrijednosti od t dobivamo matrice P^t koje nam omogućuju analizu skupa podataka u različitom vremenu t. Tako stvaramo difuzijski proces, neprekidan stohastički proces koji nam osigurava globalnu povezanost podataka.

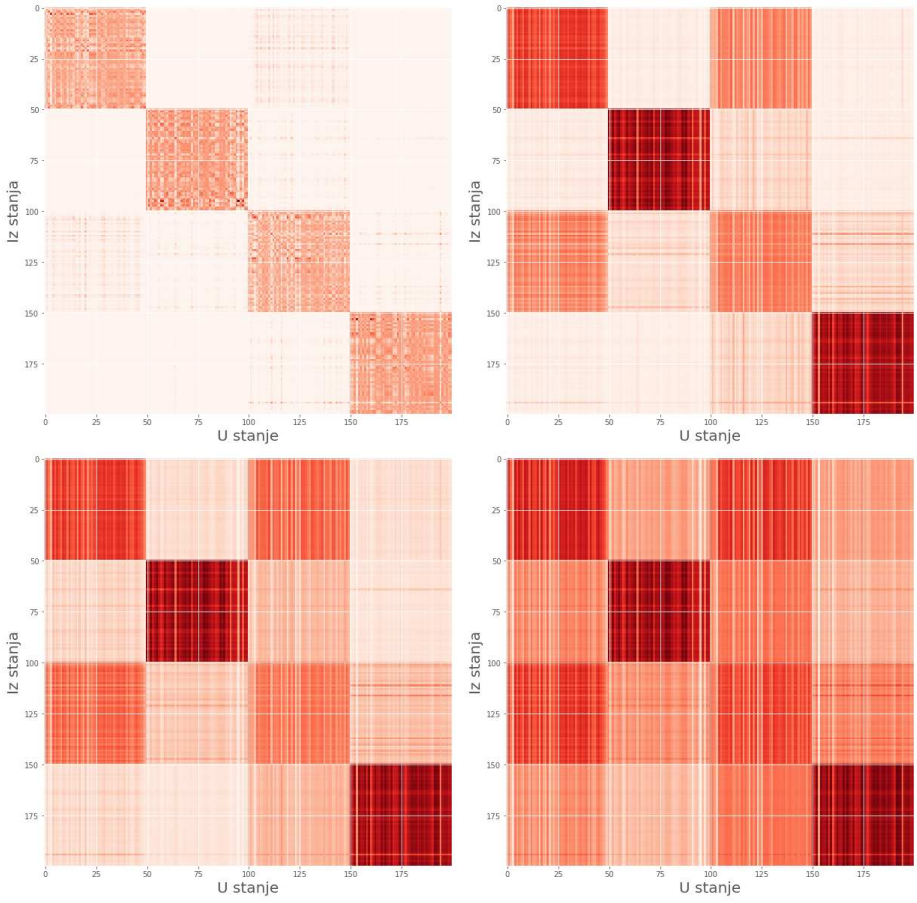
S povećanjem vrijednosti od t difuzijski se proces kreće naprijed te se povećava vjerojatnost da slijedimo put duž osnovne geometrijske strukture skupa podataka. To se događa zato što su duž geometrijske strukture točke guste i stoga visoko povezane (povezanost je funkcija euklidske udaljenosti između dviju točaka)[6]. Staze se formiraju uz kratke skokove velike vjerojatnosti. S druge strane, staze koje ne slijede ovu strukturu uključuju jedan ili više dugih skokova male vjerojatnosti koji smanjuju ukupnu vjerojatnost puta. Vjerojatnost $p_t(x, y)$ teži u stacionarnu distribuciju $\pi(y)$ pa time udaljenosti između tih vjerojatnosti teže u 0.

Primjer 3.3.1. Slučajno generiramo dvodimenzionalni skup podataka od 200 točaka s pomoću funkcije `make_blobs` iz `sklearn.datasets` u programskom jeziku *Python*. Na prvoj slici 3.3 prikazani su podatci, a na drugoj isti podatci obojeni po grupama, tj. prikazani

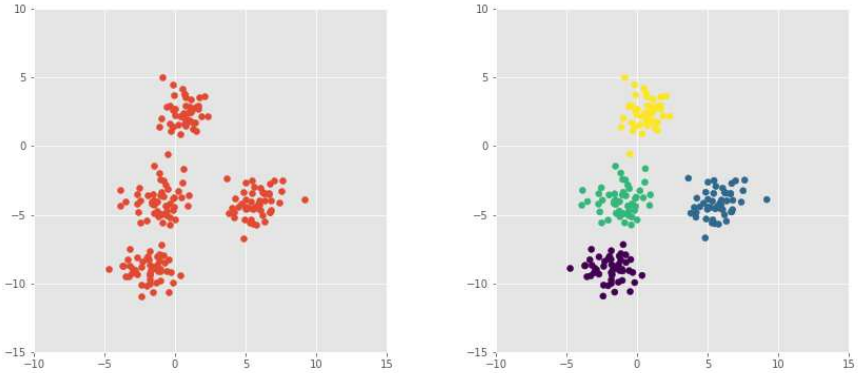
kako bi ih algoritam trebao grupirati. Takvu ćemo grupaciju postići koristeći difuzijske koordinate definirane u 3.20. Koristimo implementirane algoritme, što je opisano u dodatku B gdje su također detaljnije opisani parametri algoritma.

Na slici 3.5 prikazani su podatci s pomoću difuzijskih koordinata. U prvom redu dani su rezultati algoritma za parametre $\varepsilon = 0.0625$, $k = 20$ i $t = 1$, koordinate prepoznaju jednu grupu (označenu smeđom bojom), dok preostale tri grupira zajedno. U drugom su redu rezultati za parametre $\varepsilon = 0.03125$, $k = 7$ i $t = 1$, gdje se prepoznaju dvije različite grupe. Treći red prikazuje rezultate za parametre $\varepsilon = 1.46658$, $k = 4$ i $t = 1$ uz koje je algoritam dobro prepoznao sve četiri grupe. Željeni rezultat možemo postići promatrajući difuzijsku matricu u vremenu, bez odabira parametara, što je prikazano na slici 3.4.

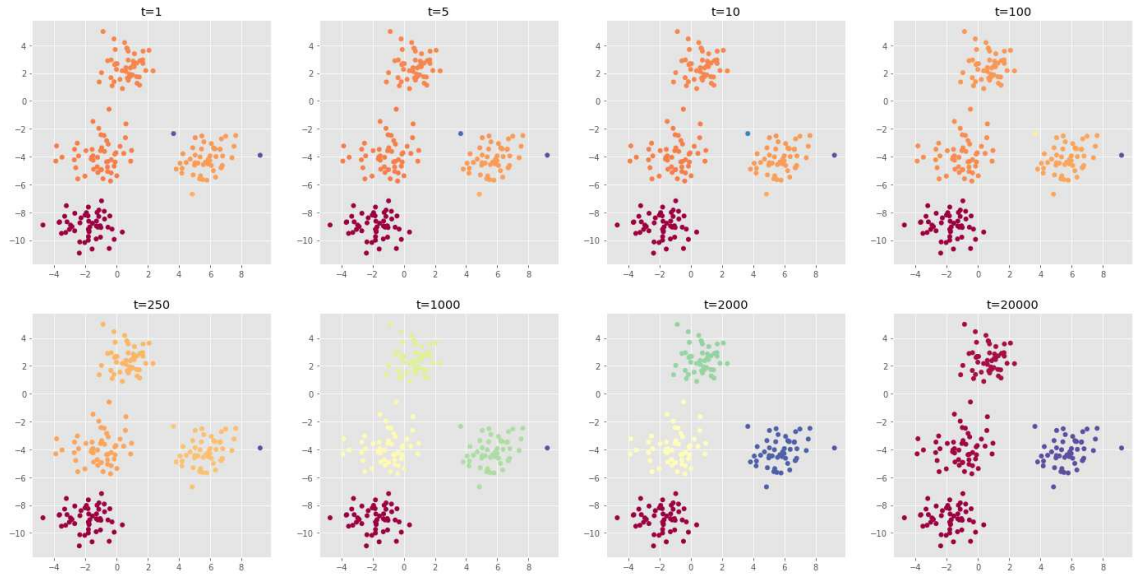
Slika 3.2: Vizualizacija tranzicijskih matrica P, P^8, P^{16} i P^{32} podatkovnih točaka iz primjera 3.3.1. Na prvoj slici vidljivo je postojanje četiriju različitih grupa, ali su već u drugom slučaju grupe manje prepoznatljive.



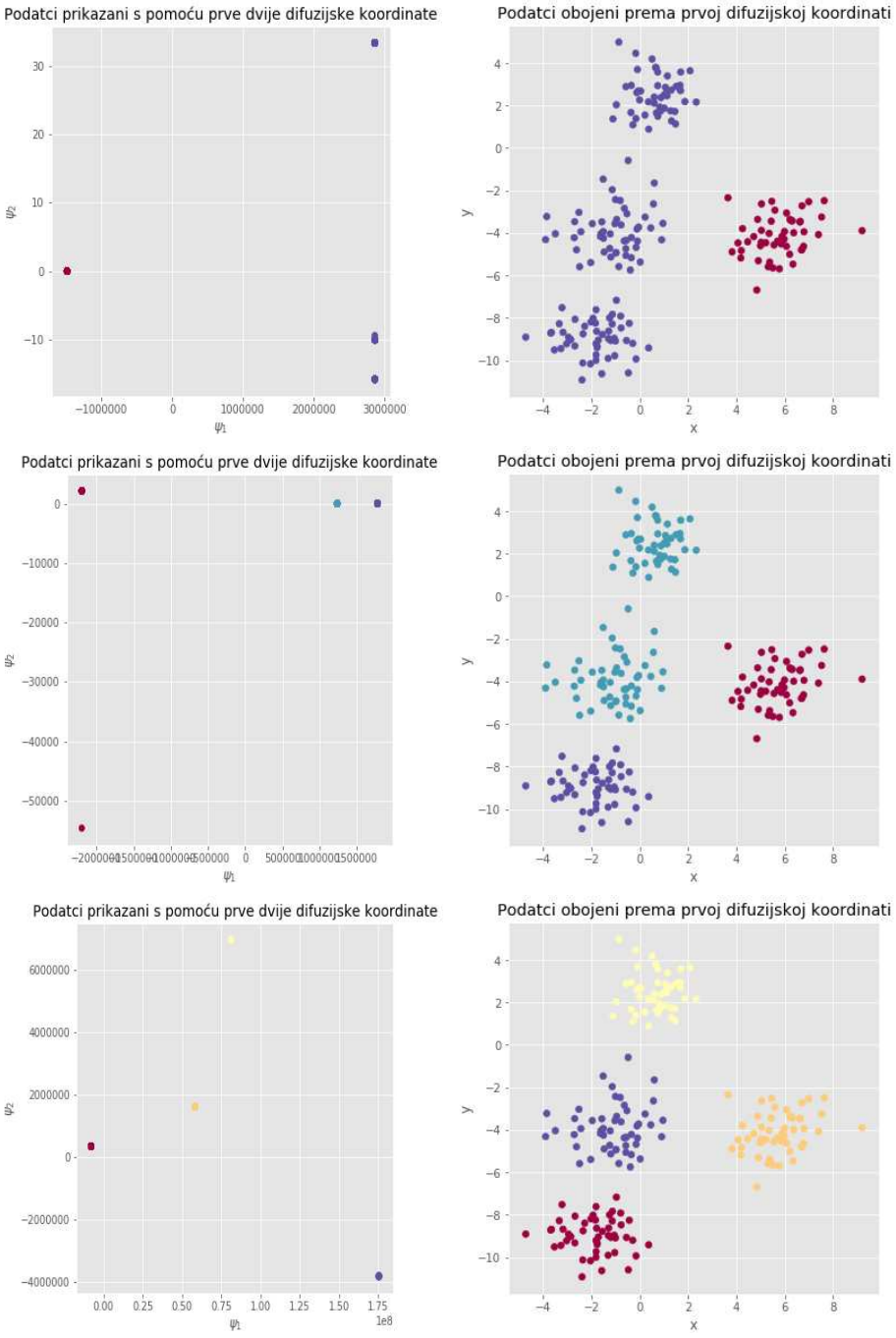
Slika 3.3: Prikaz 200 podatkovnih točaka iz primjera s četiri različite grupe. Lijevo su prikazani originalni podatci, a desno isti, ali obojeni po grupama.



Slika 3.4: Rezultati dobiveni korištenjem zadanih parametara implementiranog algoritma u 5.1 za podatke iz primjera 3.3.1. Primijetimo kako se grupacija podataka otkriva kako vrijeme raste, ali se u nekom trenutku gubi. Za $t = 20000$ vidimo kako algoritam prepoznaje samo dvije različite grupe.



Slika 3.5: Rezultati difuzijskih preslikavanja za podatke iz primjera 3.3.1. Primjećujemo da su rezultati drugačiji za različite parametre algoritma. Izbor optimalnih parametara ključan je za uspješnost ove metode.



3.4 Spektralna analiza Markovljeva lanca

Iz prethodnog odjeljka zaključujemo da potenciranje P predstavlja objekt od interesa za proučavanje geometrijskih struktura skupa $\mathbf{X}$. Potencije nekog operatora možemo opisati i analizirati koristeći se sa spektralnom teorijom, tj. svojstvenim vektorima i vrijednostima. Iako općenito za prijelazne matrice Markovljevih lanaca postojanje spektralne teorije nije zajamčeno, slučajna šetnja koju smo izgradili pokazuje vrlo posebna matematička svojstva:

- Ako je graf povezan, Markovljev lanac je ireducibilan i aperiodičan, a P ima svojstvene vrijednosti $\{\lambda_k\}_{k\geq 0}$ takve da vrijedi $1 = \lambda_0 > |\lambda_1| \geq |\lambda_2| \geq \cdots$. Tada postoji jedinstvena stacionarna distribucija dana s

  $$\pi(y) = \lim_{t\to\infty} p_t(x,y) = \phi_0(y),$$

  gdje je ϕ_0 lijevi svojstveni vektor matrice P, pokazano u teoremu 3.4.1. On odgovara svojstvenoj vrijednosti $\lambda = 1$, a prema primjeru 2.3.4 eksplicitno je dan sljedećom formulom

  $$\pi(x_i) = \phi_0(x_i) = \frac{D_{ii}}{\sum_j D_{jj}}.$$

- Lanac je reverzibilan tj. u detaljnoj ravnoteži (prema primjeru 2.3.4):

  $$\pi(x)p(x,y) = \pi(y)p(y,x). \tag{3.2}$$

- Ako je $\mathbf{X}$ konačan i graf povezan, onda je lanac ergodski.

- Za konačno vrijeme t, prema teoremu 3.4.2 dobivamo sljedeću dekompoziciju

  $$p_t(x,y) = \phi_0 + \sum_{k\geq 1} \lambda_k^t \psi_k(x)\phi_k(y), \tag{3.3}$$

  gdje su ϕ, ψ lijevi i desni svojstveni vektori matrice P.

Sljedeća dva teorema pokazuju koje su svojstvene vrijednosti od matrice P te kako vrijedi dekompozicija (3.3). Pretpostavljamo da imamo skup $\{x_i, \ldots, x_n\}$, gdje je svaki $x_i \in \mathbb{R}^p$.

Teorem 3.4.1. *Neka je K simetrična jezgrena matrica takva da $K_{ij} = k(i,j)$ te D dijagonalna matrica koja normalizira redove od K tako da vrijedi*

$$P = D^{-1}K. \tag{3.4}$$

Tada za matricu S definiranu s

$$S = D^{\frac{1}{2}} P D^{-\frac{1}{2}} \tag{3.5}$$

vrijede sljedeća svojstva:

- *simetrična je,*
- *ima iste svojstvene vrijednosti kao matrica P,*
- *lijevi svojstveni vektori od S skalirani su s $D^{-\frac{1}{2}}$, dok su desni skalirani s $D^{\frac{1}{2}}$.*

Dokaz. Uvrštavanjem (3.4) u (3.5) dobivamo

$$S = D^{-\frac{1}{2}} K D^{-\frac{1}{2}}. \tag{3.6}$$

Kako je K simetrična slijedi da je S također simetrična. Iz (3.5) imamo

$$P = D^{-\frac{1}{2}} S D^{\frac{1}{2}} \tag{3.7}$$

Kako je S simetrična, ona je dijagonalizabilna, pa ima n realnih svojstvenih vrijednosti $\{\lambda_k\}_{k=0}^{n-1}$ čiji odgovarajući svojstveni vektori $\{\upsilon_k\}$ čine ortonormirani skup (koji je baza za $\mathbb{R}^n$). Znamo da tada vrijedi:

$$S = W \Lambda W^T, \tag{3.8}$$

gdje je Λ dijagonalna matrica koja sadrži svojstvene vrijednosti od S i W matrica čiji su stupci ortonormirani svojstveni vektori od S. Uvrštavanjem (3.8) u (3.7) imamo

$$P = D^{-\frac{1}{2}} W \Lambda W^T D^{\frac{1}{2}} = (D^{-\frac{1}{2}} W) \Lambda (D^{-\frac{1}{2}} W)^{-1}$$

Iz posljednje jednakosti slijedi da su desni svojstveni vektori matrice P stupci matrice Q takve da

$$Q = D^{-\frac{1}{2}} W, \tag{3.9}$$

dok su joj lijevi svojstveni vektori redovi matrice

$$Q^{-1} = D^{\frac{1}{2}} W^T, \tag{3.10}$$

Dobivamo jednadžbu za svojstvene vektore P u preko svojstvenih vektora υ_l od S. Desni svojstveni vektor od P je

$$\psi_k = \upsilon_k D^{-\frac{1}{2}}, \tag{3.11}$$

dok je lijevi

$$\phi_k = \upsilon_k D^{\frac{1}{2}}. \tag{3.12}$$

□

Kako su svojstveni vektori $\{\upsilon_k\}$ ortonormirani u standardnom skalarnom produktu u $\mathbb{R}^n$, slijedi da su ϕ_i i ψ_j biortogonalni tj. vrijedi

$$\langle \phi_i, \psi_j \rangle = \delta_{ij}. \tag{3.13}$$

Teorem 3.4.2. *Difuzijska matrica P ima svojstvenu dekompoziciju danu s*

$$P_{ij} = \sum_{k\geq 0} \lambda_k \psi_k[i]\phi_k[j], \tag{3.14}$$

gdje je

$$\psi_k[i] = \upsilon_k[i]D^{-\frac{1}{2}}[i,i], \tag{3.15}$$

$$\phi_k[j] = \upsilon_k[j]D^{\frac{1}{2}}[j,j], \tag{3.16}$$

D dijagonalna normalizacijska matrica, a $\{\upsilon_k\}$ *skup ortonormiranih svojstvenih vektora simetrične matrice S. Dodatno, svojstveni vektori matrice S formiraju ortonormiranu bazu za* $L^2(\mathbb{R}^n, d\mu)$ *te lijevi svojstveni vektori od P formiraju ortonormiranu bazu za* $L^2(\mathbb{R}^n, \frac{d\mu}{\pi})$.

Dokaz. Ovaj je teorem analogan lemi 2.4.2. Kako je S simetrična matrica, znamo da se može dekomponirati na sljedeći način

$$S_{ij} = \sum_{k\geq 0} \lambda_k \upsilon_k[i]\upsilon_k[j], \tag{3.17}$$

gdje je $\{\upsilon_k\}$ ortonormirani skup svojstvenih vektora od S. Koristeći se (3.15) i (3.16), dobivamo sljedeće

$$\upsilon_k[i] = D^{\frac{1}{2}}[i,i]\psi_k[i],$$
$$\upsilon_k[j] = D^{-\frac{1}{2}}[j,j]\phi_k[j],$$

iz čega uvrštavanjem u (3.17) slijedi

$$S_{ij} = \sum_{k\geq 0} \lambda_k D^{\frac{1}{2}}[i,i]\psi_k[i]D^{\frac{-1}{2}}[j,j]\phi_k[j].$$

Iz relacije (3.7) zaključujemo da množenjem slijeva s $D^{\frac{-1}{2}}[i,i]$ i množenjem zdesna s $D^{\frac{1}{2}}[j,j]$ dobivamo dekompoziciju matrice P kao u (3.14).

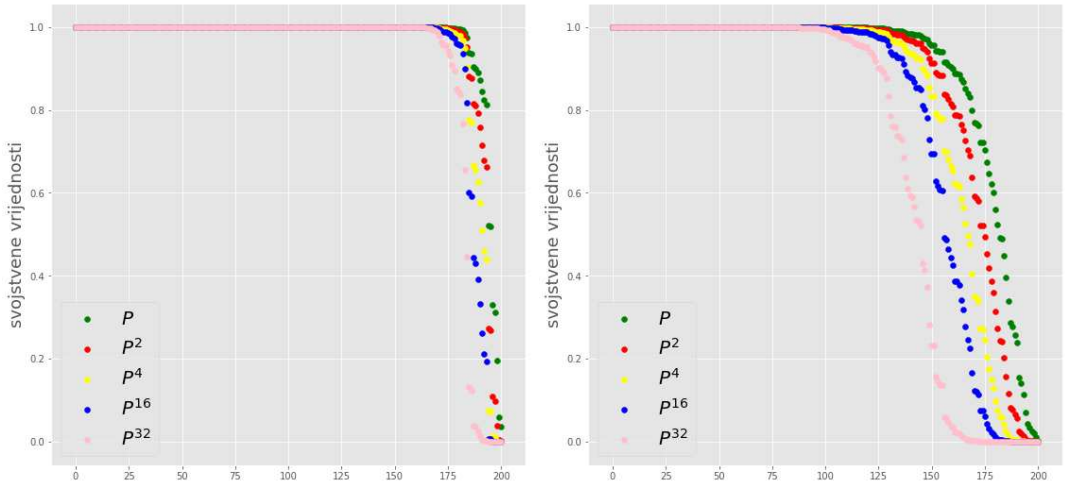
Prema lemi 2.4.2 svojstveni vektori $\{\upsilon_k\}$ formiraju ortonormiranu bazu za $L^2(\mathbb{R}^n, d\mu)$ te uz korištenje (3.11) i (3.12), možemo lako vidjeti da lijevi svojstveni vektori formiraju bazu za $L^2(\mathbb{R}^n, \frac{d\mu}{\pi})$. Koeficijenti $\{\lambda_k\psi_k[i]\}$ u (3.14) osiguravaju koordinate vektora difuzijske matrice u novom koordinatnom sustavu definiranom preko lijevih svojstvenih vektora.

□

Na slici 3.6 prikazane su svojstvene vrijednosti difuzijske matrice P i P^t (iz primjera 3.3.1), za $t = 1$, $t = 2$, $t = 4$, $t = 16$, $t = 32$. Kao što je pokazano u 2.4.1, vidimo da je raspon

svojstvenih vrijednosti između 0 i 1 te kako se t povećava, smanjuje se padaju relativni omjeri između svojstvenih vrijednosti.

Slika 3.6: Svojstvene vrijednosti tranzicijske matrice za podatke iz primjera 3.6.4 i 3.3.1 za različite potencije. Povećavanjem potencije matrice P padaju relativni omjeri između svojstvenih vrijednosti. Uočimo da na prvoj slici krivulja počinje ubrzano padati od vrijednosti λ_{175}, dok na drugoj od λ_{100}. Ovdje smo uzeli u obzir samo 200 točaka švicarske role radi usporedbe.



3.5 Difuzijska udaljenost

Spektralna svojstva Markovljeva lanca povezujemo s geometrijom skupa podataka $\mathbf{X}$. Kao što je već spomenuto, ideja definiranja slučajne šetnje na podatcima temelji se na činjenicu da jezgra k određuje lokalnu geometriju podataka i bilježi neka geometrijska obilježja koja su od interesa.

Definicija 3.5.1. *Metrika koja mjeri sličnosti dviju točaka u opservacijskom prostoru kao povezanost između njih naziva se difuzijska udaljenost i dana je sljedećom formulom:*

$$D_t(x_i, x_j)^2 = \|p_t(x_i, \cdot) - p_t(x_j, \cdot)\|^2_{L^2(\mathbf{X}, d\mu/\pi)} = \int_{\mathbf{X}} \Big(p_t(x_i, u) - p_t(x_j, u)\Big)^2 \frac{d\mu(u)}{\pi(u)}. \tag{3.18}$$

Kako se difuzijski proces kreće naprijed, otkrivajući geometrijsku strukturu podataka, glavni su doprinos difuzijske udaljenosti staze duž te strukture. Drugim riječima, $D_t(x,y)$ je funkcionalna težinska L^2 udaljenost između dvije *posterior* distribucije $u \to p_t(x,u)$ i $u \to p_t(y,u)$. Po definiciji, blizina koju definira odražava povezanost u grafu podataka.

Doista, $D_t(x,y)$ će biti mala ako postoji veliki broj kratkih staza koje povezuju x i y, tj. ako su te staze velike vjerojatnosti. Pored toga, kao što je ranije napomenuto, t ima ulogu vremenskog parametra. Stoga podvlačimo tri glavna svojstva difuzijske udaljenosti:

- Odražava povezanost podataka, točke su bliže ako su visoko povezane u grafu, stoga naglašava pojam grupiranja u podatcima.
- Ona je robusna na smetnje jer zbraja sve moguće staze duljine t između točaka.
- S gledišta strojnog učenja, isto promatranje omogućava nam zaključak da je ova udaljenost prikladna za dizajn algoritama zaključivanja na temelju većine jer uzima u obzir sve veze između točaka.

Razmotrimo pojam $p_t(x,u)$ u definiciji 3.5.1. To je vjerojatnost skoka s točke x na u (za bilo koju točku skupa podataka) u t koraka, a zbraja vjerojatnost svih mogućih staza duljine t između x i u. Kao što je objašnjeno u prethodnom odjeljku, ovaj izraz ima velike vrijednosti za staze duž osnovne geometrijske strukture podataka. Da bi difuzijska udaljenost $D_t(x,y)$ ostala mala, vjerojatnosti putanje između x, u i u, y moraju biti približno jednake. To se događa kada su x i y dobro povezani preko u.

Dakle, difuzijska udaljenost uspijeva uhvatiti sličnost dviju točaka s obzirom na istinske parametre promjene temeljne geometrijske strukture određenog skupa podataka. Kako je pokazano u teoremu 3.6.1, ova udaljenost može biti izražena preko svojstvenih vrijednosti i vektora tranzicijske matrice kao:

$$D_t^2(x,y) = \sum_{k\geq 1} \lambda_k^{2t}(\psi_k(x) - \psi_k(y))^2. \tag{3.19}$$

Primijetimo da je ψ_0 konstanta pa izostavljamo $k = 0$.

3.6 Konstrukcija preslikavanja

U prethodnom smo odjeljku pronašli metriku, difuzijsku udaljenost koja je sposobna približiti udaljenosti duž ove strukture. Računanje difuzijskih udaljenosti je računalno skupo. Stoga je prikladno preslikati podatkovne točke u euklidski prostor prema difuzijskoj metrici. Difuzijska udaljenost u podatkovnom prostoru postaje euklidska udaljenost u ovom novom difuzijskom prostoru. Preslikavanjem koordinata podataka u difuzijski prostor reorganiziramo podatke prema difuzijskoj udaljenosti te tako postižemo smanjenje dimenzionalnost i grupiranje podataka. Difuzijsko preslikavanje čuva unutrašnju geometriju skupa podataka, a budući da mjeri udaljenosti na strukturi nižih dimenzija, očekujemo da će biti potrebno manje koordinata za predstavljanje točaka podataka u novom prostoru. Nastaje

pitanje koje dimenzije treba zanemariti kako bi se optimalno sačuvala difuzijska udaljenost. S time na umu ispitujemo sljedeće preslikavanje:

$$Y_i := \begin{bmatrix} p_t(x_i, x_1) \\ p_t(x_i, x_2) \\ \vdots \\ p_t(x_i, x_n) \end{bmatrix}.$$

Za ovo preslikavanje, euklidska udaljenost između dviju preslikanih točaka Y_i i Y_j jednaka je difuzijskoj udaljenosti između x_i i x_j u opservacijskom prostoru, što pokazuje teorem 3.6.1.

Ovo osigurava reorganizaciju koju smo tražili s obzirom na difuzijsku udaljenost. No, još uvijek nismo postigli redukciju dimenzije, ona je i dalje n. Preostaje nam zanemariti određene dimenzije difuzijskog prostora koristeći teorem 3.4.2. Uzmemo li normaliziranu difuzijsku matricu $P = D^{-1}K$, gdje je D dijagonalna matrica čije su vrijednosti dobivene zbrajanjem redova matrice K. Tada za proizvoljni t možemo difuzijsko preslikavanje zapisati kao

$$Y_i^t := \begin{bmatrix} \lambda_1^t \psi_1[i] \\ \lambda_2^t \psi_2[i] \\ \vdots \\ \lambda_n^t \psi_n[i] \end{bmatrix}, \tag{3.20}$$

gdje je $\psi_1[i]$ i-ti element prvog svojstvenog vektora matrice P. Opet je euklidska udaljenost između preslikanih točaka $Y_i^{'}$ i $Y_j^{'}$ difuzijska udaljenost. Skup lijevih ortogonalnih svojstvenih vektora matrice P čini bazu za difuzijski prostor, a pripadajuće svojstvene vrijednosti λ_k određuju važnost svake dimenzije. Svojstvene vrijednosti $\lambda_1, \lambda_2 \ldots$ teže u 0 i strogo su manje od 1 pa difuzijsku udaljenost možemo izračunati s unaprijed zadanom točnošću $\delta > 0$. Ako definiramo

$$s(\delta, t) = \max\{k \in \mathbb{N} : |\lambda_k|^t > \delta |\lambda_1|^t\},$$

tada, do relativne preciznosti δ, imamo

$$D_t^2(x, y) = \sum_{k=1}^{s(\delta,t)} \lambda_k^{2t} (\psi_k(x) - \psi_k(y))^2.$$

Smanjenje dimenzionalnosti postiže se zadržavanjem $s(\delta, t)$ dimenzija povezanih s dominantnim svojstvenim vektorima, što osigurava najbolju aproksimaciju difuzijske udaljenosti $D_t(x, y)$. Stoga je difuzijsko preslikavanje koje optimalno čuva unutarnju geometriju

podataka za neki $x_i = x$ i $\psi_k[i] = \psi_k(x)$ dano s

$$\mathbf{\Psi_t(x)} := \begin{bmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \vdots \\ \lambda_{s(\delta,t)}^t \psi_{s(\delta,t)}(x) \end{bmatrix}. \tag{3.21}$$

Svaka koordinata od $\Psi_t(x)$ zove se difuzijska koordinata. Preslikavanje $\Psi_t : \mathbf{X} \to \mathbb{R}^{s(\delta,t)}$ ugrađuje podatke u euklidski prostor dimenzije $s(\delta, t)$.

Teorem 3.6.1. *Ako izaberemo difuzijske koordinate kao u (3.20), difuzijska udaljenost između točaka u originalnom prostoru jednaka je euklidskoj udaljenosti u difuzijskom prostoru.*

Dokaz. Tvrdimo da vrijedi

$$\begin{aligned} D_t^2(x_i, x_j) &= \|p_t(x_i, \cdot) - p_t(x_j, \cdot)\|^2_{L^2(\mathbb{R}^n, \frac{d\mu}{\pi})} \\ &= \|Y_i^t - Y_j^t\|^2_{L^2(\mathbb{R}^n, \frac{d\mu}{\pi})} = \sum_{k\geq 0} \lambda_k^{2t} (\psi_k[x_i] - \psi_k[x_j])^2. \end{aligned} \tag{3.22}$$

Neka je $\mathbf{X}$ skup podataka, pretpostavimo $t = 1$. Tada vrijedi

$$\begin{aligned} D^2(x_,x_j) &= \|p(x_i, \cdot) - p(x_j, \cdot)\|^2_{L^2(\mathbb{R}^n, d\mu/\pi)} \\ &= \sum_{u\in\mathbf{X}} \left| \sum_{k\geq 0} \lambda_k \psi_k[x_i]\phi_k[u] - \sum_{k\geq 0} \lambda_k \psi_k[x_j]\phi_k[u] \right|^2 \\ &= \sum_{u\in\mathbf{X}} \left| \sum_{k\geq 0} \lambda_k \phi_k[u](\psi_k[x_i] - \psi_k[x_j]) \right|^2 \\ &= \sum_{u\in\mathbf{X}} \left| \sum_{k\geq 0} \lambda_k \upsilon_k[u] D^{\frac{1}{2}}[u,u](\psi_k[x_i] - \psi_k[x_j]) \right|^2 \\ &= \sum_{u\in\mathbf{X}} \left| D^{\frac{1}{2}}[u,u] \sum_{k\geq 0} \lambda_k \upsilon_k[u](\psi_k[x_i] - \psi_k[x_j]) \right|^2. \end{aligned}$$

Druga je jednakost dobivena korištenjem (3.3), a četvrta slijedi uvrštavanjem (3.16). Kako je $\{\upsilon_k\}$ ortonormirani skup, znamo da vrijedi

$$\sum_{u\in\mathbf{X}} \sum_{i\neq j} \upsilon_i[u]\upsilon_j[u] = 0.$$

Stoga slijedi

$$D^2(x,z) = \sum_{u\in\mathbf{X}} D[u,u] \sum_{k\geq 0} \lambda_k^2 v_k^2[u](\psi_k[x_i] - \psi_k[x_j])^2 \tag{3.23}$$

$$= \sum_{k\geq 0} \lambda_k^2 (\psi_k[x_i] - \psi_k[x_j])^2 \left(\sum_{u\in\mathbf{X}} \phi_k^2[u] \right). \tag{3.24}$$

Prema teoremu 3.4.2 difuzijske koordinate $\{\phi_k\}$ formiraju ortogonalnu bazu za difuzijski prostor $L^2(\mathbb{R}^n, \frac{d\mu}{\pi})$. U difuzijskom prostoru znamo da vrijedi

$$\sum_{u\in\mathbf{X}} \phi_k^2[u] = 1,$$

pa slijedi

$$D^2(x,z) = \sum_{k\geq 0} \lambda_k^2 (\psi_k[x] - \psi_k[y])^2.$$

□

Sljedeća lema daje ogradu pogreške aproksimacije difuzijske udaljenosti.

Lema 3.6.2. *Za svaki $t \geq 0$ greška aproksimacije difuzijske udaljenosti ograničena je s*

$$|D_t^2(x_i,x_j) - \sum_{l=1}^{k} \lambda_l^{2t} (\psi_l[i] - \psi_l[j])^2| \leq \lambda_{k+1}^{2t} \left(\frac{1}{\pi(x_i)} + \frac{1}{\pi(x_j)} \right). \tag{3.25}$$

Dokaz. Koristeći se spektralnom dekompozicijom (3.19), imamo

$$\left| D_t^2(x_0,x_1) - \sum_{j=1}^{k} \lambda_j^{2t} (\psi_j(x_0) - \psi_j(x_1))^2 \right| = \sum_{j=k+1}^{n-1} \lambda_j^{2t} (\psi_j(x_0) - \psi_j(x_1))^2$$

$$\leq \lambda_{k+1}^{2t} \sum_{j=k+1}^{n-1} (\psi_j(x_0) - \psi_j(x_1))^2.$$

U trenutku $t = 0$ vrijedi

$$D_0^2(x_0,x_1) = \sum_{j=k+1}^{n-1} (\psi_j(x_0) - \psi_j(x_1))^2.$$

Međutim, prema definiciji difuzijske udaljenosti 3.5.1 slijedi

$$D_0^2(x_0,x_1) = \|p_0(x_0,y) - p_0(x_1,y)\|^2_{L^2(\mathbf{X},d\mu/\pi)} = \left(\frac{1}{\pi(x_0)} + \frac{1}{\pi(x_1)} \right).$$

Kombiniranjem zadnje dvije jednakosti imamo tvrdnju. □

3.6.1 Vjerojatnosna interpretacija

Posljednji teorem pruža vjerojatnosnu interpretaciju nelinearne ugradnje točaka skupa $\mathbf{X}$ iz originalnog prostora (npr. $\mathbb{R}^p$) u difuzijski prostor $\mathbb{R}^m$. Geometrija je u difuzijskom prostoru značajna i može se interpretirati s pomoću Markovljevih lanaca te nam dati informacije o strukturi podataka. Prednost difuzijske udaljenosti u odnosu na standardnu udaljenost u originalnom prostoru očita je. Dok je u originalnom prostoru udaljenost između bilo kojeg para točaka nezavisna od lokacije točaka, difuzijska udaljenost ovisi o svim mogućim putovima koji povezuju neke dvije točke. Tako difuzijska udaljenost mjeri dinamičku blizinu između točaka na grafu u skladu s njihovom povezanošću.

Difuzijska udaljenost i difuzijska preslikavanja ovise o vremenskom parametru t. Za vrlo kratka vremena sve točke u difuzijskom prostoru su daleko, dok se vrijeme povećava u beskonačnost, sve međusobne udaljenosti konvergiraju u nulu jer $p_t(x,y)$ konvergira u stacionarnu distribuciju.

Teorem 3.6.1 pokazuje kako svojstvene vrijednosti i svojstveni vektori matrice P obuhvaćaju karakteristično vrijeme relaksacije (t_{rel} definiran u 2. poglavlju) i procesa slučajne šetnje na grafu. Većina svojstvenih vrijednosti bilježi strukturu detalja, a samo prvih nekoliko najvećih bilježe grube globalne strukture grafa. U slučajevima kada matrica P ima spektralni razmak (engl. *spectral gap*), razmak između susjednih svojstvenih vrijednosti, tako da ima samo nekoliko svojstvenih vrijednosti blizu 1, a sve preostale značajno manje od 1, difuzijska udaljenost za dovoljno veliki t može se dobro aproksimirati sa samo prvih nekoliko k svojstvenih vektora $\psi_1(x),\ldots,\psi_k(x)$, sa zanemarivom pogreškom. Nadalje, kao što je pokazano u [14], kretanje ovog difuzijskog procesa kroz vrijeme, ekvivalentno je skupljanju slučajne šetnje zadržavajući samo njegove najsporije procese opuštanja (one za koje je t_{rel} velik). Specijalno, kako je opisano u [3], za uspješnost spektralnog grupiranja potrebno je da prosječno vrijeme izlaska iz svake grupe bude značajno veće nego najveće t_{rel} unutar individualnih grupa. Grupacija se podataka s pomoću difuzijska preslikavanja postiže ako odaberemo takve parametre algoritma da slučajna šetnja ima male vjerojatnosti izlaska iz pojedinih grupa.

Sljedeći teorem pokazuje kako je k-dimenzionalna aproksimacija optimalna prema kriteriju srednje kvadratne pogreške.

Teorem 3.6.3. *Od svih k-dimenzionalnih aproksimacija vjerojatnosne distribucije $p_t(x,y)$ oblika*

$$\hat{p}_t(x,y) = \pi(y) + \sum_{j=1}^{k} a_j(t,x) w_j(y),$$

ona koja minimizira srednje kvadratnu grešku

$$\mathbb{E}_x\left(\|p_t(x,y) - \hat{p}_t(x,y)\|^2_{L^2(\mathbf{X}, d\mu/\pi)} \right),$$

sa stacionarnom distribucijom π, dana je sljedećim parametrima $w_j(y) = \phi_j(y)$ i $a_j(t,x) = \lambda_j^t \psi_j(x)$. Dakle, optimalna k-dimenzionalna aproksimacija dana je s

$$\hat{p}_t(x,y) = \pi(y) + \sum_{j=1}^{k} \lambda_j^t \psi_j(x)\phi_j(y).$$

Dokaz. Dokaz proizlazi iz metode težinskih glavnih komponenti primijenjene na matricu P, koristeći biortogonalnost lijevih i desnih svojstvenih vektora [3]. □

Princip proučavanja spektralnih svojstava Markovljeva lanca definiran na skupu $\mathbf{X}$ zapravo kombinira ideje iz teorije potencijala, spektralne geometrije i proučavanja parcijalnih diferencijalnih operatora. Iz teorije potencijala znamo da se posebnosti neke domene (šiljci, uglovi) odražavaju na ponašanje rješenja *Dirichletovih* i *Neumannovih* problema (vidi u [13] str. 11-12.). Spektralna geometrija postavlja pitanje je li geometrija *Riemannove* mnogostrukosti određena spektrom *Laplaceova* operatora. Općenitije, studij spektralne asimptotike za parcijalne diferencijalne operatore povezuje geometrijske karakteristike domene $\mathbf{X}$ s rastom svojstvenih vrijednosti takvih operatora. Zajednički je nazivnik ovih ideja da se geometrija skupa $\mathbf{X}$ može proučavati analizom prostora funkcija definiranih na $\mathbf{X}$ i linearnih operatora na tim prostorima. Spektralna analiza difuzijskog operatora P služi kao alat za geometrijsku analizu $\mathbf{X}$.

Prilikom pronalaska difuzijskih preslikavanja (opisano u dodatku B) potrebno je odabrati broj difuzijskih koordinata, širinu jezgre ε, broj susjeda k u grafu sličnosti te α parametar opisan u dodatku A. Za uspješnost metode potrebno je dobro odabrati parametre algoritma. U primjeru 3.3.1 vidljivo je kako se grupacija podataka postiže samo za neki odabir parametara.

Primjer 3.6.4. Na prvoj slici 3.7 prikazano je 10000 točaka u $\mathbb{R}^3$ koji generiraju švicarsku rolu, a na trećoj ulaganje podataka u $\mathbb{R}^2$ s pomoću difuzijskih koordinata. Primijetimo kako je ovakvo ulaganje grafički vrlo drugačije od onog iz primjera 1.2.1 gdje nisu korišteni optimalni parametri. Podatci su generirani na sljedeći način:

$$\begin{aligned}
\phi &= l_\phi \cdot \texttt{random.rand(m)} \\
t &= \texttt{random.rand(m)} \\
X &= 1/6 \cdot (\phi + \sigma \cdot t)\sin(\phi) \\
Y &= 1/6 \cdot (\phi + \sigma \cdot t)\cos(\phi) \\
Z &= l_Z \cdot \texttt{random.rand(m)},
\end{aligned}$$

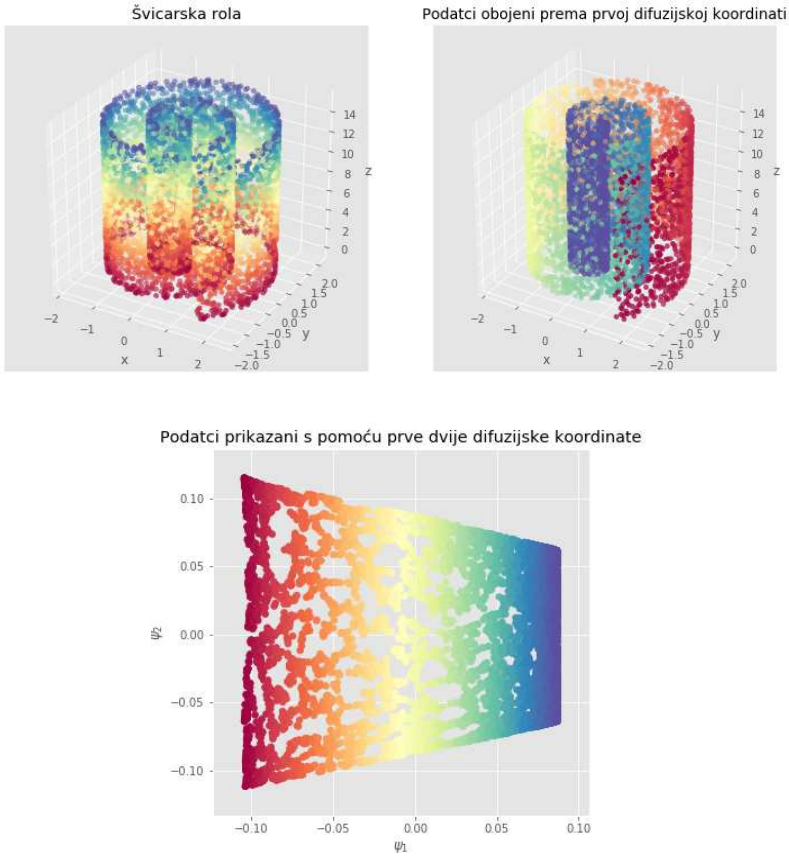
gdje je l_ϕ duljina role u kutnom smjeru, l_Z duljina role u z smjeru, σ snaga šuma, m broj točaka, a `random.rand` `numpy` funkcija koja generira niz danog oblika iz uniformne $(0,1)$ distribucije. Podatci su generirani za sljedeće parametre: $l_\phi = 15$, $l_Z = 15$, $\sigma = 0.1$, $m =$

10000. Na slici 3.8 vidimo zašto difuzijske koordinate daju željenu ugradnju podataka u $\mathbb{R}^2$. Prva difuzijska koordinata visoko je korelirana s kutem ϕ, a druga sa Z-koordinatom.

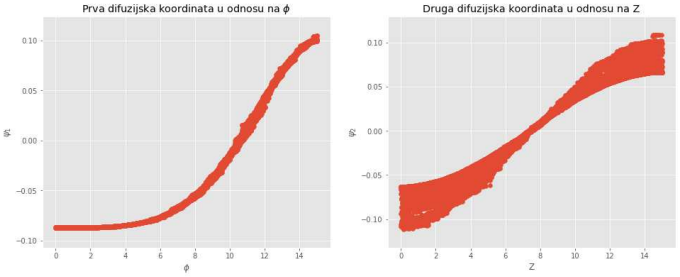
Promotrimo sliku 3.9 na kojoj je prikazo ulaganje švicarske role s pomoću različitih koordinata difuzijskog preslikavanja, u prvom slučaju za parametre $t = 1$, $\varepsilon = 2$, $k = 100$, $\alpha = 0$, dok u drugom $t = 1$, $\varepsilon = 25$, $k = 10$, $\alpha = 0$. Različite difuzijske koordinate daju nam drugačije realizacije odmotane švicarske role, "hvataju" različite geometrijske karakteristike skupa podataka.

Nadalje, na slici 3.10 prikazan je utjecaj parametra k, broja susjednih točaka koje gledamo u algoritmu te kako se difuzijske koordinate ponašaju u vremenu. U ovom primjeru, parametar k nema utjecaja. Za različite vrijednosti tog parametra dobivamo gotovo iste difuzijske koordinate. No, u slučaju grupiranja podataka ponekad je ključno odabrati dobru vrijednost ovog parametra.

Slika 3.7: Vizualizacija švicarske role s pomoću prve dvije difuzijske koordinate za parametre $\varepsilon = 0.015625$, $k = 200$ i $t = 1$. Za ovaj odabir parametara postižemo željeni prikaz "odmotane" role u $\mathbb{R}^2$.



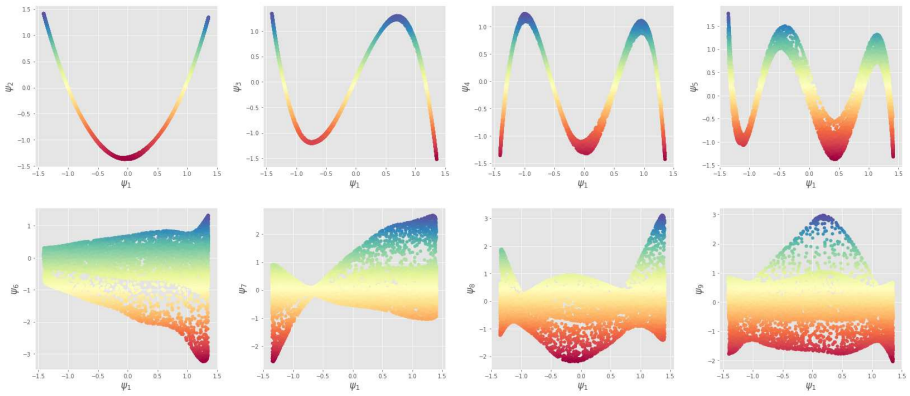
Slika 3.8: Grafički prikaz korelacija između difuzijskih koordinata i varijabli koje generiraju švicarsku rolu. Prva difuzijska koordinata visoko je korelirana s kutem ϕ, a druga sa Z-koordinatom.



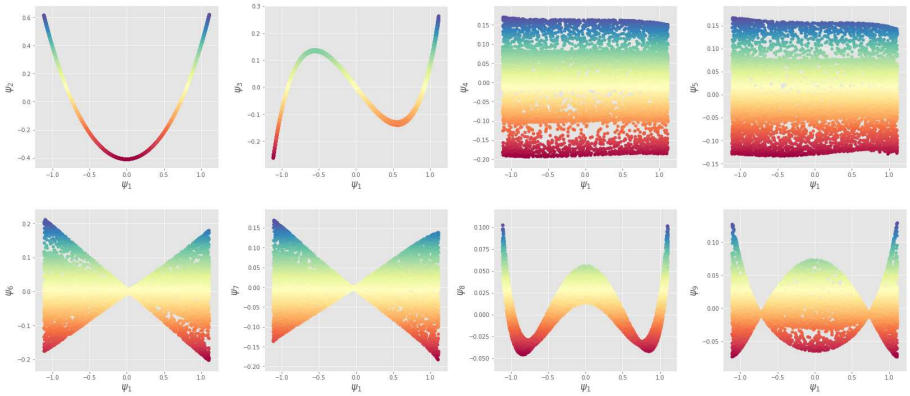

```
Korelacija između varijabli na prvom grafu:
[[1.         0.92760337]
 [0.92760337 1.        ]]
Korelacija između varijabli na drugom grafu:
[[1.         0.97796051]
 [0.97796051 1.        ]]
```

Slika 3.9: Prikaz švicarske role s pomoću različitih difuzijskih koordinata koje su označene na koordinatnim osima. U prvom slučaju za parametre $t = 1$, $\varepsilon = 2$, $k = 100$, $\alpha = 0$, a u drugom za $t = 1$, $\varepsilon = 25$, $k = 10$, $\alpha = 0$. Primijetimo kako za prvu kombinaciju parametara npr. šesta difuzijska koordinata daje najbolju realizaciju odmotane švicarske role. Za drugu kombinaciju to se postiže koristeći četvrtu ili petu difuzijsku koordinatu.

(a)

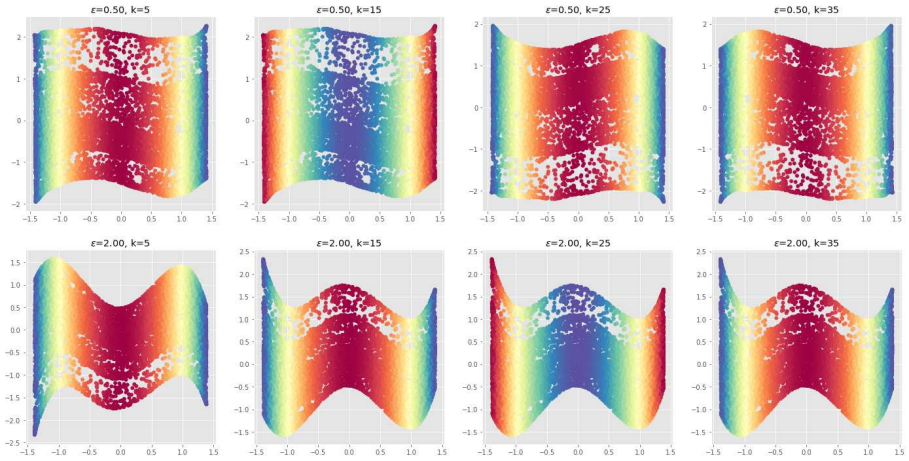


(b)

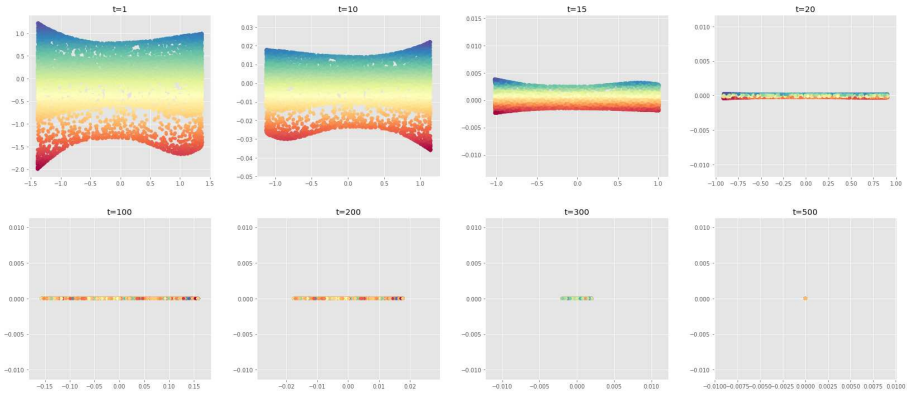


Slika 3.10: Ilustracija utjecaja parametara k na promjenu difuzijskih koordinata ((a) dio) za različite vrijednosti parametra ε, gdje su vrijednosti označene u naslovu grafova. Ponašanje difuzijskih koordinata kroz vrijeme prikazano je pod b), gdje vidimo gubljenje strukture podataka kako t raste.

(a)



(b)

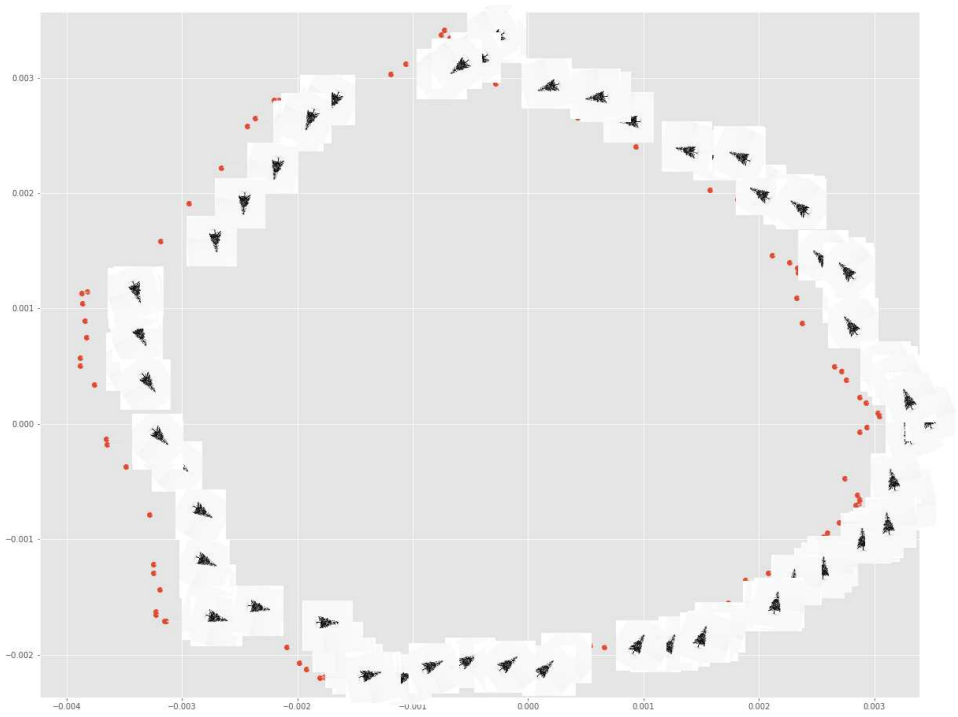


3.6.2 Parametrizacija podataka

Prethodni odjeljak kaže da difuzijska preslikavanja nude prikaz podataka u euklidskom prostoru. Ovaj prikaz karakterizira činjenica da je udaljenost između dviju točaka jednaka difuzijskoj udaljenosti u originalnom skupu podataka. Stoga, difuzijsko preslikavanje reorganizira podatkovne točke prema međusobnim difuzijskim udaljenostima. Prikaz organizacijske snage difuzijskog preslikavanja dan je slikom 3.11. Generirali smo kolekciju slika istog bora (slika 1.2 iz uvodnog poglavlja), ali pod različitim kutevima rotacija. Prve dvije difuzijske koordinate reorganiziraju slike tako da prepoznaju drugačije kutove rota-

cije, dimenzija kolekcije 60 slika od 256×256 piksela prikazana samo dvama parametrima (prvim dvjema difuzijskim koordinatama).

Slika 3.11: Organizacijska snaga difuzijskog preslikavanja



Poglavlje 4

Povezanost sa spektralnim grupiranjem

U ovom dijelu rada predstavljamo pogled grupiranja i segmentacije podataka preko sličnosti parova točaka, opisan u [7]. Kao i dosad u radu, tumačimo sličnosti kao bridove slučajne Markovljeve šetnje na podatcima i proučavamo svojstvene vrijednosti i vektore tranzicijske matrice. Pokazujemo da ovaj pogled omogućava dobivanje vjerojatnosnih temelja spektralnih metoda za grupiranje i segmentaciju podataka. Uvodimo pojam metode normalizacijskog presjeka ili reza (engl. *normalized cut*) koji prirodno proizlazi iz ovog rada te pokazujemo vezu s *Laplacianom* grafa.

Pristup redukcije dimenzionalanosti opisan ranije, koristeći se difuzijskim preslikavanjima, služi se svojstvenim vektorima *Laplaciana* grafa te pridruženih operatora koji ga aproksimiraju[1]. Ovo rješenje može se primijeniti i kao tehnika spektralnog grupiranja podataka i ima blisku povezanost sa segmentacijom slika ([7]), uravnoteženosti opterećenja i konstrukcijom strujnog kruga. Blisko povezan algoritam nedavno predložen u [2] ima pristup koji koristi globalnu povezanost s eksponencijalno opadajućim težinama. Parametar opadanja je iznimno važan. U mnogim visoko dimenzionalnim problemima, minimalna i maksimalna udaljenost između točaka sasvim su blizu pa za matricu povezanosti ne očekujemo da ima puno nula za bilo koji stupanj opadanja. Za razliku od statističkog grupiranja podataka, gdje pretpostavljamo vjerojatnosni model koji je generirao podatke, grupiranje u parovima (engl. *pairwise clustering*) definira funkciju sličnosti između parova točaka te iz toga formulira kriterij koji grupiranje mora optimizirati. Intuitivno, optimizacijski kriterij govori kako su točke u istoj grupi slične, dok su točke u različitim grupama različite.

[1] Opisano u [13], 13 - 17. str gdje su definirani takvi operatori te dokazana njihova svojstva.

4.1 Normalizirani rez

Pretpostavimo da imamo skup od n objekata koji želimo podijeliti u konačan broj grupa. Koristimo već definiranu matricu povezanosti K (definicija 3.2.1). Tada modeliramo podatke pomoću težinskog grafa $G = (V, E)$, gdje su V vrhovi određeni točkama skupa proizvoljno numerirani, E bridovi između njih određeni težinama K. Težina K_{ij} povezana s bridom e_{ij} sličnost je između vrhova v_i i v_j. Matrica K je simetrična i pretpostavljamo da je odgovarajući neusmjeren graf povezan.[2]

Graf $G = (V, E)$ možemo particionirati u dva razdvojena podskupa A i B, tako da vrijedi $A \cup B = V$, $A \cap B = \emptyset$, jednostavno micanjem bridova koji ih povezuju. Rez (engl. *cut*) mjera je različitosti između dvaju dijelova grafa i može biti jednostavno definirana kao ukupna težina bridova koji su maknuti kako bi skupovi A i B bili razdvojeni:

$$\mathrm{cut}(A, B) = \sum_{u \in A, v \in B} K(u, v).$$

Optimalna biparticija grafa ona je za koju je rez minimalan. No, minimiziranjem gornje funkcije rješavamo se slabo povezanih *outliera*, što vodi do loše kvalitete dijeljenja. Ovo nije iznenađujuće s obzirom na to da ovako definirana funkcija povećava svoju vrijednost s povećanjem broja rubova koji idu preko dva podijeljena dijela. Kako bismo izbjegli taj problem, predstavljena je nova mjera razdvajanja.

Za neki podskup A, prvo definiramo značaj ili važnost u odnosu na ostale vrhove.

$$\mathrm{vol}(A) = \sum_{u \in A, v \in V} K(u, v).$$

Nadalje, definiramo normalizirani rez (engl. *normalized cut*):

$$\mathrm{ncut}(A, B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right). \tag{4.1}$$

Ovako definirano razdvajanje skupa koji particioniraju mali broj izoliranih točaka neće imati male vrijednosti normaliziranog reza. No, kako je opisano u [7], minimiziranje *ncut*-a po svim particijama skupa V je np težine, ali ako se ograničimo na realnu domenu, možemo riješiti problem minimizacije u polinomijalnom vremenu s proizvoljnom preciznošću. Efikasno rješenje pronalazimo koristeći svojstvene vrijednosti i vektore *Laplaceove* matrice $L = D - K$, gdje su matrice D i K definirane u definiciji (3.2.1). L je simetrična i pozitivno definitna matrica.

Algoritam se svodi na generalizirani problem određivanja svojstvenih vrijednosti i vektora

$$Lx = \lambda Dx. \tag{4.2}$$

[2] Ako graf nije povezan, postoji mnogo algoritama koji pronalaze njegove povezane komponente.

Fokus je na drugoj najmanjoj svojstvenoj vrijednosti, označimo je s λ^L, te pripadajućim svojstvenim vektorom x^L. Kada postoji particija skupa vrhova V takva da vrijedi

$$x_i^L = \begin{cases} \alpha, & v_i \in A \\ \beta, & v_i \in B, \end{cases}$$

tada su A i B optimalni i vrijedi $\text{ncut}(A, B) = \lambda^L$. Ovaj rezultat predstavlja bazu spektralne segmentacije i grupiranja s pomoću normalizacijskih rezova. Nakon rješavanja generaliziranog problema (4.2), pronalazimo particiju tako da podijelimo x^L u dva skupa koji sadrže približno jednake vrijednosti. Takva particija inducira particiju skupa vrhova V. U praksi, particioniranje grafa na k dijelova postižemo primjenjujući rekurzivno ovu metodu, tako da pojedine dijelove grafa ponovno podijelimo. Ovu proceduru nazivamo *ncut* algoritam.

Kako je pokazano u [10], vrijedi sljedeća jednakost

$$x^T Lx = \frac{1}{2} \sum_{i,j} (x_i - x_j)^2 K_{ij}.$$

Ako stavimo $a = \text{vol}(A)$ i $b = \text{vol}(B)$, te

$$x_i = \begin{cases} \frac{1}{\text{vol}(A)}, & \text{ako je } v_i \in A \\ -\frac{1}{\text{vol}(B)}, & \text{ako je } v_i \in B, \end{cases}$$

tada imamo

$$x^T Lx = \frac{1}{2} \sum_{i,j} (x_i - x_j)^2 K_{ij} = \frac{1}{2} \sum_{v_i \in A, v_j \in B} \left(\frac{1}{a} + \frac{1}{b} \right)^2 K_{ij} = \frac{1}{2} \left(\frac{1}{a} + \frac{1}{b} \right)^2 \text{cut}(A, B).$$

Također imamo

$$x^T Dx = \sum_i x_i^2 D_{ii} = \sum_{v_i \in A} \frac{1}{a^2} d_{ii} + \sum_{v_i \in B} \frac{1}{b^2} d_{ii} = \frac{1}{a^2} \text{vol}(A) + \frac{1}{b^2} \text{vol}(B) = \frac{1}{a} + \frac{1}{b}.$$

Iz toga slijedi

$$\frac{x^T Lx}{x^T Dx} = \frac{1}{2} \text{cut}(A, B) \left(\frac{1}{a} + \frac{1}{b} \right) = \frac{1}{2} \text{ncut}(A, B).$$

Primijetimo da $x^T D\mathbf{1} = 0$, gdje je $\mathbf{1}$ vektor stupac jedinica. Dakle, imamo problem minimizacije pod uvjetom $x^T D\mathbf{1} = 0$. Ako stavimo $y = D^{1/2} x$, tada

$$\frac{x^T Lx}{x^T Dx} = \frac{y^T D^{-1/2} L D^{-1/2} y}{y^T y},$$

gdje je x okomit na $D^{1/2}\mathbf{1}$. Matricu $L' = D^{-1/2}LD^{-1/2}$ zovemo normaliziran *Laplacian* grafa. Ona je simetrična, pozitivno semidefinitna te najmanjoj svojstvenoj vrijednosti 0 pripada svojstveni vektor $D^{1/2}1$. Stoga se $\min_{y \perp D^{1/2}\mathbf{1}} \frac{y^T L' y}{y^T y}$ postiže za y koji odgovara svojstvenom vektoru druge najmanje svojstvene vrijednosti. Naravno, 0 može biti višestruka svojstvena vrijednost ako graf nije povezan, tj. ima više zasebno povezanih komponenti. Ovaj se pristup razlikuje od onog opisanog u našoj teoriji u činjenici da su svojstvene vrijednosti pomaknute za 1, pokazano u 4.2.1.

Ovo je promatranje u [15] prošireno tako da graf $G = (V, E)$ s težinama K particioniramo u k različitih dijelova. Preciznije, želimo podijeliti skup vrhova $V = \{x_1, \ldots, x_n\}$ u k disjunktnih skupova, tj. prikazati V tako da $V = \cup_{l=1}^{k} V_l$ i $V_l \cap V_j = \emptyset$, za svaki $l \neq j$. Ovu particiju označavamo s $\Gamma_V^k = \{V_1, \ldots, V_k\}$.

Neka su $A, B \subset V$. Definiramo novi normalizirani rez s

$$\text{ncuts}(A, B) = \frac{\text{cut}(A, B)}{\text{vol}(A)}. \tag{4.3}$$

Dobro grupiranje podataka zahtijeva čvrste veze unutar grupa, a labave veze između grupa. Stoga, definiramo sljedeće

$$\text{knassoc}(\Gamma_V^k) = \sum_{l=1}^{k} \text{ncuts}(V_l, V_l), \tag{4.4}$$

$$\text{kncuts}(\Gamma_V^k) = \sum_{l=1}^{k} \text{ncuts}(V_l, V \setminus V_l), \tag{4.5}$$

gdje $\text{ncuts}(V_l, V_l)$ mjeri koliko veza ostaje u V_l, a $\text{ncuts}(V_l, V \setminus V_l)$ koliko ne ostaje. Kako je $\text{knassoc}(\Gamma_V^k) + \text{kncuts}(\Gamma_V^k) = 1$, maksimiziranje jedne vrijednosti minimizira drugu. Dakle, dovoljno je da maksimiziramo sljedeću funkciju

$$\xi(\Gamma_V^k) = \text{knassoc}(\Gamma_V^k).$$

Koristimo $n \times k$ particijsku matricu $X = [X_1, \ldots, X_k]$ za reprezentaciju Γ_V^k, gdje je

$$X_{il} = \begin{cases} 1 & \text{ako } i \in V_l, \\ 0 & \text{ako } i \notin V_l, \end{cases} \tag{4.6}$$

za sve $i \in V$ i $l \in \{1, \ldots, k\}$. Ovdje je X_l binarni indikator za V_l. Kako je svaki vrh dodijeljen jednoj particiji od V, mora vrijediti $X\mathbf{1}_K = \mathbf{1}_n$ ($\mathbf{1}_n$ je vektor od n jedinica). Matricu stupnjeva D tada možemo izračunati kao $D = \text{Diag}(K\mathbf{1}_n)$, a cut i vol na sljedeći način

$$\text{cut}(V_l, V_l) = X_l^T K X_l, \tag{4.7}$$

$$\text{vol}(V_l) = X_l^T D X_l. \tag{4.8}$$

Dakle, želimo maksimalnu vrijednost

$$\xi(X) = \frac{1}{k} \sum_{l=1}^{k} \frac{X_l^T K X_l}{X_l^T D X_l},$$

uz diskretni uvjet $X \in \{0,1\}^{n \times k}$, $X\mathbf{1}_k = \mathbf{1}_n$. U [15] dano je rješenje ovog problema, sprovedeno u dva dijela. Optimizacijski diskretni uvjet zapravo kaže da želimo particijsku matricu koja ima u svakom retku samo jednu jedinicu. Ovaj uvjet mijenjamo tako da gledamo matrice X koje imaju ortonormirane stupce, ali ne moraju biti diskretni. Tako se svodimo na rješavanje problema svojstvenih vrijednosti Ako stavim takve uvjete minimum se postiže na matrici koja ima k ortonormiranih stupaca koji su svojstveni vektori jednog svojstvenog problema. Ti svojstveni vektori nam daju za n točaka koordinate u prostoru dimenzije k. Tako dobivamo skup svih globalnih optimuma te među njima pronalazimo onoga koji je najviše diskretan.

4.2 Veza između Markovljeve šetnje i normaliziranog reza

Kako bismo dobili intuiciju, želimo znati zašto je svojstveni vektor x^L po dijelovima konstantan. Interpretacija slučajne šetnje daje nam odgovor. Normaliziranjem matrice sličnosti K imamo stohastičku matricu $P = D^{-1}K$, suma redova je 1. Prema teoriji Markovljeve slučajne šetnje, opisanoj u ovom radu, P_{ij} predstavlja vjerojatnost prelaska s čvora i na j u jednom koraku. Svojstvene vrijednosti matrice P su $\lambda_1 = 1 \geq \lambda_2 \geq \cdots \lambda_n$, a $\psi_1, \ldots, \psi_n$ svojstveni vektori. Prvi svojstveni vektor $\psi_1 = \mathbf{1}$ očito je vektor jedinica.

Propozicija 4.2.1. *Ako su λ, x rješenja spektralnog problem $Px = \lambda x$, tada su $(1-\lambda)$ i x rješenja od (4.2).*

Dokaz. Jednostavnim raspisom za $L = D - K$ ($K = D - L$) i $P = D^{-1}K$ imamo

$$\begin{aligned} Px &= \lambda x, \\ D^{-1}Kx &= \lambda x, \\ Kx &= \lambda Dx, \\ (D-L)x &= \lambda Dx, \\ Dx - Lx &= \lambda Dx, \\ Lx &= Dx - \lambda Dx, \\ Lx &= (1-\lambda)Dx, \end{aligned}$$

gdje smo u trećem redu pomnožili jednakost slijeva s D. □

Nadalje definiramo $\mathbb{P}(B \mid A)$, vjerojatnost tranzicije slučajne šetnje sa skupa $A \subset V$ na $B \subset V$ u jednom koraku, s početnim stanjem iz A.

Propozicija 4.2.2. *Neka je G povezan i nebipartitan graf te $(X_t)_{t\in\mathbb{N}}$ slučajna šetnja koja kreće iz X_0 prema stacionarnoj distribuciju π. Tada za $\mathbb{P}(B \mid A)$=$\mathbb{P}(X_1 \in B \mid X_0 \in A)$ vrijedi*

$$ncut(A, A^c) = \mathbb{P}(A^c \mid A) + \mathbb{P}(A \mid A^c).$$

Dokaz.

$$\begin{aligned} \mathbb{P}(X_0 \in A, X_1 \in A^c) &= \sum_{i\in A, j\in A^c} \mathbb{P}(X_0 = i, X_1 = j) \\ &= \sum_{i\in A, j\in A^c} \pi_i p_{ij} = \sum_{i\in A, j\in A^c} \frac{d_i}{\mathrm{vol}(V)} \frac{k_{ij}}{d_i} \\ &= \frac{1}{\mathrm{vol}(V)} \sum_{i\in A, j\in A^c} k_{ij}. \end{aligned}$$

Koristeći se formulom uvjetne vjerojatnosti dobiva se

$$\begin{aligned} \mathbb{P}(X_1 \in A^c \mid X_0 \in A)) &= \frac{\mathbb{P}(X_0 \in A, X_1 \in B)}{\mathbb{P}(X_0 \in A)} \\ &= \left(\frac{1}{\mathrm{vol}(V)} \sum_{i\in A, j\in A^c} k_{ij} \right) \left(\frac{\mathrm{vol}(A)}{\mathrm{vol}(V)} \right)^{-1} \\ &= \frac{\sum_{i\in A, j\in A^c} k_{ij}}{\mathrm{vol}(A)}. \end{aligned}$$

Analogno pokažemo da vrijedi

$$\mathbb{P}(X_0 \in A \mid X_1 \in A^c) = \frac{\sum_{i \in A, j \in A^c} k_{ij}}{\mathrm{vol}(A^c)},$$

pa vrijedi tvrdnja. □

Ova propozicija dovodi do dobre interpretacije normaliziranog reza, a time i do normaliziranog spektralnog grupiranja. Otkriva nam da prilikom minimiziranja normaliziranog reza zapravo tražimo rez grafa tako da slučajna šetnja rijetko prijeđe iz A u A^c i obrnuto. Naime, ako je normalizirani rez male vrijednosti za neku particiju A i A^c skupa V, znači da je mala vjerojatnost izbjegavanja skupa A kada je šetnja na njemu, te obratno. Intuitivno, particionirali smo skup V u dva dijela tako da slučajna šetnja koja se kreće na jednom dijelu ima sklonost ostati u njemu. Nadalje, ovaj rez jako je povezan s konceptom niske provodljivosti skupova u Markovljevoj slučajnoj šetnji. Nisko provodljiv skup $A \subset V$ ima malu vrijednost funkcije h dane sljedećom formulom

$$h(A) = \max(\mathbb{P}(A, A^c \mid A), \mathbb{P}(A^c, A \mid A^c)).$$

4.3 Stohastičke matrice s po dijelovima konstantnim svojstvenim vektorima

Koristimo tranzicijsku maticu P kako bismo postigli bolje razumijevanje ncut algoritma. Kao što smo već rekli, algoritam normaliziranog reza traži drugu najveću svojstvenu vrijednost matrice P kako bi particionirao neki skup. Nadalje, definiramo vektor w koji je po dijelovima konstanta u odnosu na particiju $\Delta = (A_1, A_2, \ldots, A_k)$ nekog skupa I ako i samo ako $w_i = w_j$ za sve točke i, j iz skupa A_s, $s = 1, \ldots k$. Primijetimo kako je prvi svojstveni vektor ($\mathbf{1}$) matrice P uvijek po dijelovima konstanta. S obzirom na to da je ovakvo svojstvo ključno za spektralno grupiranje, važno je razumjeti kada matrica P ima željeno obilježje. Prema [11], za matricu P indeksiranu prema I s nezavisnim redcima i stupcima vrijede sljedeće tvrdnje:

- Matrica P ima k svojstvenih vektora koji su po dijelovima konstante u odnosu na particiju Δ i odgovaraju nenegativnim svojstvenim vrijednostima ako i samo ako je suma $P_{is} = \sum_{j \in A_s} P_{ij}$ konstanta za sve $i \in A_{s'}$, $s, s' = 1, \ldots, k$ te je matrica $R = [P_{ss'}]_{s,s'=1\ldots,k}$ (gdje je $P_{ss'} = \sum_{j \in A_{s'}, i \in A_s} P_{ij}$) nesingularna.

- Ako matrica P dimenzije n i oblika $P = D^{-1}S$, S simetrična, D nesingularna, tada P ima nezavisne svojstvene vektore.

Stohastičku matricu koja zadovoljava prvu tvrdnju zovemo blok-matricom. Intuitivno, stohastička matrica ima po dijelovima konstantne svojstvene vektore ako je vjerojatnost tranzicije iz neke točke x_i u segment A_s jednaka za sve točke iz segmenta kojem pripada x_i. Dakle, spektralno grupiranje postižemo koristeći se sličnostima tranzicijskih vjerojatnosti na segmentima te ako odaberemo takve parametre algoritma da slučajna šetnja ima male vjerojatnosti izlaska iz pojedinih grupa.

U primjerima (3.3.1), (5.3.2), (5.3.4) i (5.3.5) difuzijske su koordinate prepoznale grupaciju podataka. Ilustrirano je kako su podaci prikazani s pomoću prve dvije difuzijske koordinate konstante.

Poglavlje 5

Primjene

5.1 Spektralno ulaganje u niždimenzionalni potprostor i grupiranje podataka

Neka su $\{x_i\}$ točke uzorkovane uniformno iz niždimenzionalnog prostora ugrađene u višedimenzionalni prostor. Lema 3.6.2 pokazuje da je zadržavanjem prvih k koordinata difuzijskog preslikavanja greška zanemariva. No, ovo nije nužno točan kriterij za provedbu algoritma. Cilj je niždimenzionalne reprezentacije sačuvati informacije o globalnoj strukturi prostora, a zanemariti detalje. Pitanje je pod kojim je uvjetima ovo zadovoljeno. Problem možemo formulirati na sljedeći način. Neka je $\mathcal{Y} = \{y_i\}_{i=1}^n \subset \mathbb{R}^m$ slučajan uzorak iz neke glatke vjerojatnosne distribucije definiran na kompaktnoj domeni $\mathbb{R}^m$, tako da je $\mathcal{X} = f(\mathcal{Y})$, gdje je $f : \mathbb{R}^m \to \mathbb{R}^q$ glatko preslikavanje takvo da $q \geq m$. Želimo procijeniti m i odrediti koordinate točaka $\{y_i\}$, što je problem s puno stupnjeva slobode. Dok se općenita teorija učenja potprostora nije u potpunosti razvila, pružamo uvid u spektralno ulaganje baziran na vjerojatnosnoj interpretaciji opisanoj u radu. Pokazujemo da u određenim slučajevima algoritam radi, ako odaberemo ispravne parametre. Počinjemo s najjednostavnijim primjerom jednodimenzionalne krivulje uložene u višedimenzionalni prostor. Uspješno se pokazuje kako difuzijska preslikavanja otkrivaju dimenzionalnost podataka i daju reprezentaciju dužine luka krivulje. Pokazujemo to sljedećim teoremom.

Teorem 5.1.1. *Pretpostavimo da su podatci uniformno distribuirani iz glatke 1-D krivulje koja se ne presijeca i uložena je u višedimenzionalni prostor. Tada za velik broj uzoraka i malu širinu u jezgre prva difuzijska koordinata daje bijektivnu parametrizaciju krivulje. Nadalje, u slučaju da je krivulja zatvorena, prve dvije difuzijske koordinate preslikavaju krivulju u kružnicu.*

Dokaz. Neka je $\Gamma : [0,1] \to \mathbb{R}^p$ označava konstantu parametrizaciju krivulje tj. vrijedi $\|d\Gamma(s)/d(s)\|) = \text{const}$. Kada $n \to \infty$ i $\varepsilon \to 0$, difuzijske koordinate konvergiraju u svoj-

stvene vektore pripadnog *Fokker-Planck* operatora (za detalje i dokaz vidi [13]). U slučaju nepresijecajuće krivulje, taj operator jednak je

$$\mathcal{H}\psi = \frac{d^2\psi}{ds^2}, \tag{5.1}$$

gdje je s duljina luka duž Γ, s *Neumannovim* graničnim uvjetima u rubovima $s = 0, 1$. Prva su dva netrivijalna svojstvena vektora $\psi_1 = \cos(\pi s)$ i $\psi_2 = \cos(2\pi s)$. Prvi svojstveni vektor daje jednoznačnu parametrizaciju krivulje, i stoga može biti iskorišten za ulaganje u $\mathbb{R}$. Druga svojstvena funkcija može se prikazati kao kvadratna funkcija prve, $\psi_2 = 2\psi_1^2 - 1$. Ova povezanost i procjena lokalne gustoće točaka potvrđuju da dani podatci stvarno leže u 1-D prostoru. Pretpostavimo sada da imamo zatvorenu krivulju u $\mathbb{R}^p$. Tada nemamo rubne uvjete operatora, nego tražimo periodične svojstvene funkcije. Prva su dva nekonstantna svojstvena vektora $\sin(\pi s + \theta)$ i $\cos(\pi s + \theta)$, gdje je θ proizvoljan kut rotacije. Oni preslikavaju podatke u kružnicu u $\mathbb{R}^2$. □

Rezultati sljedećih primjera dobiveni su korištenjem biblioteke `pydiffmap` u programskom jeziku `Python`. Nadalje, za promatranje kroz vrijeme se koristi implementacija preuzetu s *git://github.com/satra/mapalign*. Ova implementacija nam omogućuje analizu podataka u direktno zadanom vremenu t. Dijelovi koda nalaze se u dodatku B, gdje su dodatno objašnjeni parametri algoritma.

5.2 Odabir parametra ε

Kao što smo već spomenuli, u algoritmu je odabir parametra ε presudan kako bismo odabrali dobre težine K_{ij}. Izbor ε ovisi o originalnom skupu podataka. Tijekom pokretanja algoritma često se može pronaći ε za koji nema konvergencije.

U ovom je radu za odabir parametra ε korišten način predstavljen u [1] koji ima dobru fizičku interpretaciju. Opravdanost pristupa leži u tome što kada je ε prevelik u usporedbi s $\|x_i - x_j\|^2$, tada su vrijednosti težina vrlo blizu 1, dok kada je premali, težine su blizu 0. Stoga se vrijednost od interesa nalazi između ove dvije krajnosti. Na temelju ove ideje, predstavljena je sljedeća procedura:

i) Konstruiraj težinsku matricu $\mathbf{K} := \mathbf{K}(\varepsilon)$ za različite vrijednosti ε.

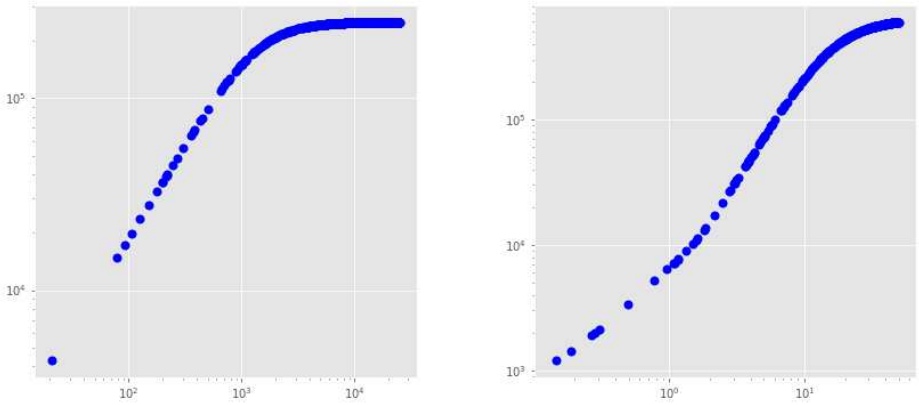
ii) Izračunaj $L(\varepsilon) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{K}_{ij}(\varepsilon)$.

iii) Nacrtaj $L(\varepsilon)$ koristeći logaritamski graf.

iv) Odaberi ε tamo gdje graf izgleda linearno.

Na slici 5.1 prikazan je logaritamski graf $L(\varepsilon)$ za spiralu u $\mathbb{R}^3$ iz primjera 5.3.6 i spiralu u $\mathbb{R}^2$ iz primjera 5.3.1. Na grafovima su vidljive različite vrijednosti na x-osi, čime pokazujemo ovisnost odabranog parametra o skupu podataka. Za svaki od skupova podataka koji su korišteni u radu na ovaj je način procijenjen raspon mogućih vrijednosti ε . Kod korišten za dobivanje ovakvog grafa nalazi se u dodatku B.

Slika 5.1: Prikaz logaritamskih grafova koji nam služe za pronalazak optimalnog parametra ε. Lijevo je prikaz takvog grafa za podatke iz primjera 5.3.6, a desno iz primjera 5.3.1.



5.3 Primjeri

Primjer 5.3.1. Skup od 400 točaka u $\mathbb{R}^2$ prikazan na prvom grafu (slika 5.2) generiran je na sljedeći način:

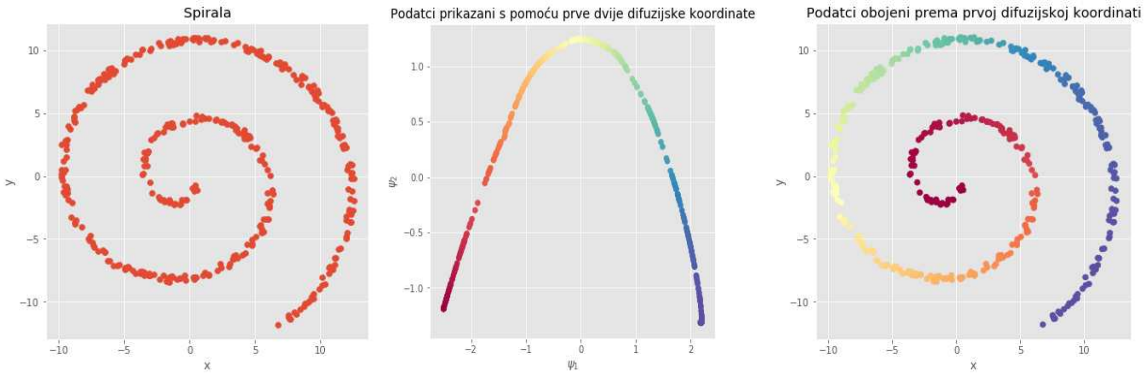
$$t = \text{random.rand}(n,1) \cdot 780 \cdot (2 \cdot \pi)/360$$
$$x = \cos(t) * t + \text{random.rand}(n,1) * \sigma$$
$$y = -\sin(t) * t + \text{random.rand}(n,1) * \sigma$$

gdje je `random.rand` `numpy` funkcija koja generira niz danog oblika iz uniformne $(0,1)$ distribucije, a $\sigma = 0.5$ šum. Drugi graf slike 5.2 prikazuje prve dvije difuzijske koordinate, a treći originalnu spiralu obojenu prema prvoj difuzijskoj koordinati.

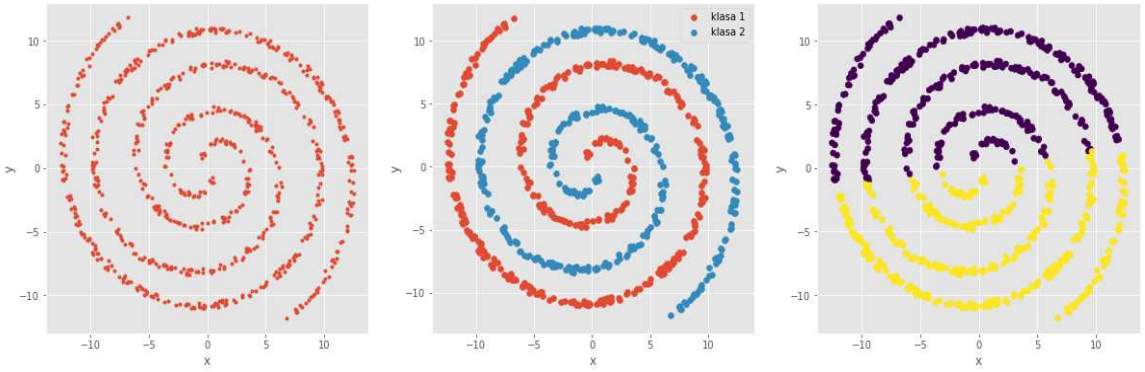
Primjer 5.3.2. Generiramo skup od 800 točaka u $\mathbb{R}^2$. 400 točaka određuju spiralu kao u primjeru 5.3.1, a preostalih 400 drugu spiralu tako da za x, y (iz spomenutog primjera) vrijedi $x = -x$, a $y = -y$. Podatci su prikazani na prvom grafu slike 5.3. Želimo s pomoću

difuzijskog preslikavanja prepoznati dvije različite spirale, tj. postići grafički prikaz kao na drugoj slici. Na slici 5.4 grafički je prikazan skup podataka s pomoću difuzijskih koordinata pronađenih koristeći `DiffusionMap.from_sklearn` za različite parametre preslikavanja. Vidimo kako se za različite parametre dobivaju potpuno drugačije koordinate. U prvom slučaju, nemamo željene rezultate, dok su u drugom prepoznate dvije krivulje, ali nisu grupirane kao dvije različite klase. No, treći slučaj pokazuje moć grupiranja podataka s pomoću difuzijskog preslikavanja, difuzijske koordinate prepoznaju dvije različite spirale.

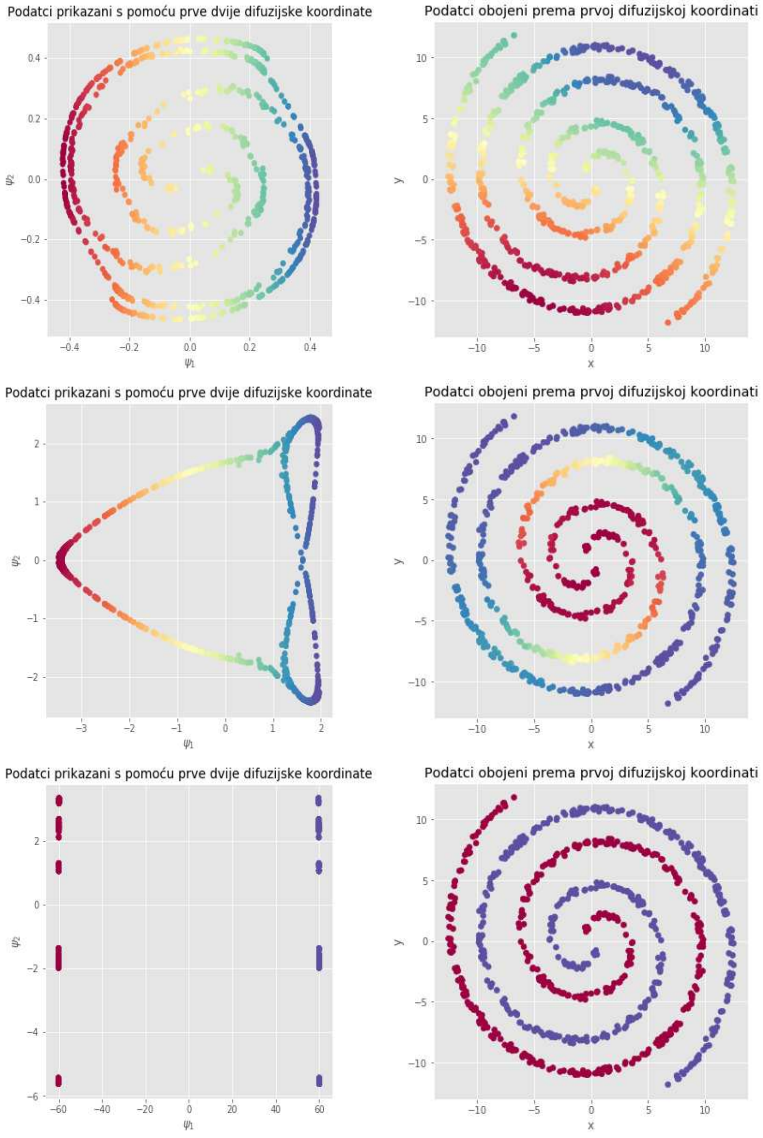
Slika 5.2: Spirala prikazana s pomoću difuzijskih koordinata, za parametre $\varepsilon = 2$, $\alpha = 0$ i $k = 30$. Vidimo kako druga difuzijska koordinata prepoznaje strukturu spirale.



Slika 5.3: Prikaz dviju spirala koje želimo prepoznati s pomoću difuzijskog preslikavanja. Na trećoj slici vidimo rezultate klasičnog algoritma k-sredina koji ne uspijeva dobro grupirati nelinearne podatke.



Slika 5.4: Dvije različite spirale prikazane s pomoću prve dvije difuzijske koordinate u trenutku $t = 1$ za različite parametre algoritma. U prvom redu za $\varepsilon = 2, k = 200$, drugom za $\varepsilon = 8.5, k = 20$, a zadnjem $\varepsilon = 0.0625, k = 20$. Treći slučaj pokazuje nam sposobnost grupiranja podataka s pomoću difuzijskih preslikavanja.



Primjer 5.3.3. U ovom primjeru generiramo 10000 točaka iz sfere te pokazujemo ulaganje u $\mathbb{R}^2$ koristeći se difuzijskim koordinatama, ilustrirano na slici 5.5, za parametre algoritma $t = 1$, $\varepsilon = 0.00048828$, $k = 400$, $\alpha = 1$. U ovom skupu podataka postoji rotacijska simetrija. Da bismo ju uklonili, definiramo 'sjeverni pol' kao točku u kojoj prva difuzijska koordinata

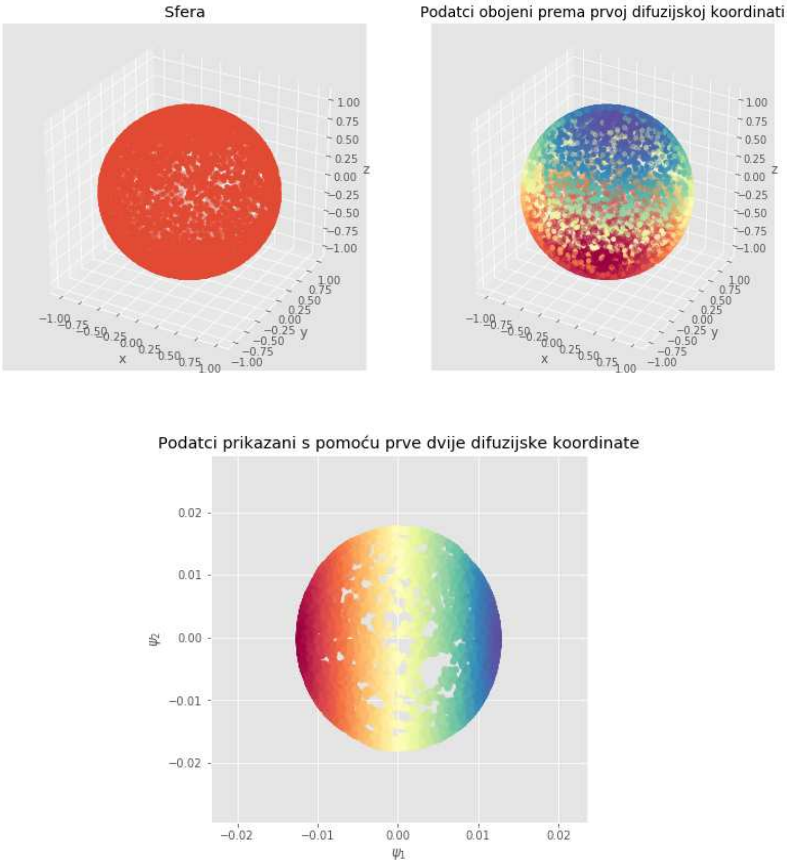
postiže svoju maksimalnu vrijednost. Kada je skup podataka rotiran, možemo provjeriti koliko dobro prva difuzijska koordinata aproksimira

$$Z = Y_1(\theta, \phi) = \sin(\theta),$$

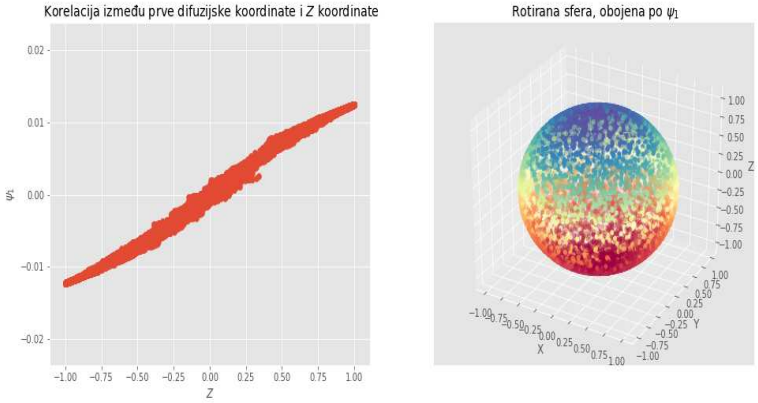
što je prikazano na slici 5.6, gdje su ϕ i θ kutevi u sferi generirani na sljedeći način

$$\phi = 2 \cdot \pi \cdot \texttt{random.rand}(m) - \pi$$
$$\theta = \pi \cdot \texttt{random.rand}(m) - 0.5 \cdot \pi.$$

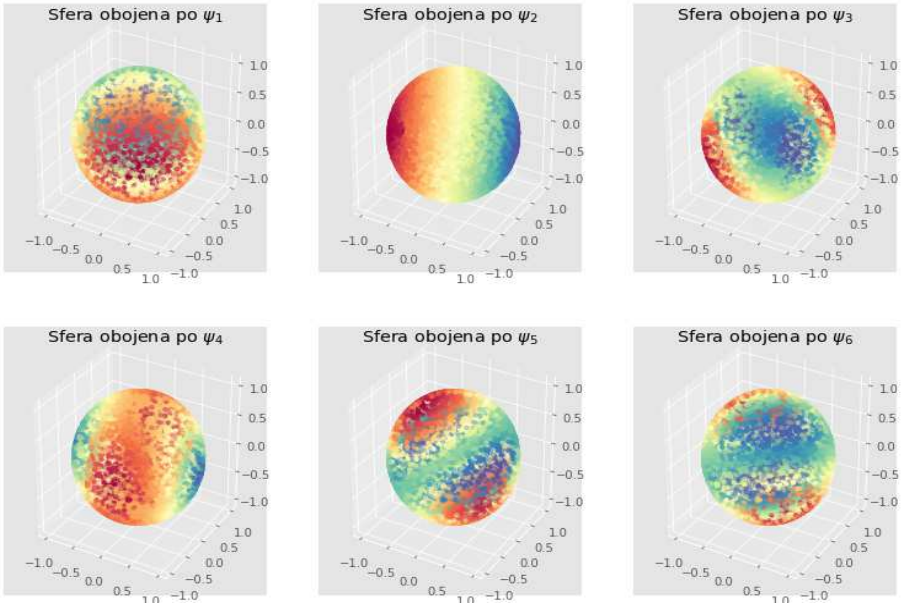
Slika 5.5: Vizualizacija generiranih točaka sfere (prva slika), podatci obojeni prvom difuzijskom koordinatom (druga slika) i ulaganje podataka u $\mathbb{R}^2$ s pomoću prve dvije difuzijske koordinate (treća slika).



Slika 5.6: Prikaz visoke korelacija između prve difuzijske koordinate i Z-osi sfere.



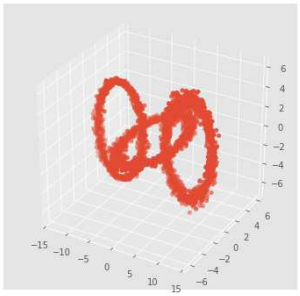
Slika 5.7: Grafički prikaz sfere obojene različitim difuzijskim koordinatama. Primijetimo kako svaka difuzijska koordinata zahvaća drugačiji smjer gibanja točaka.



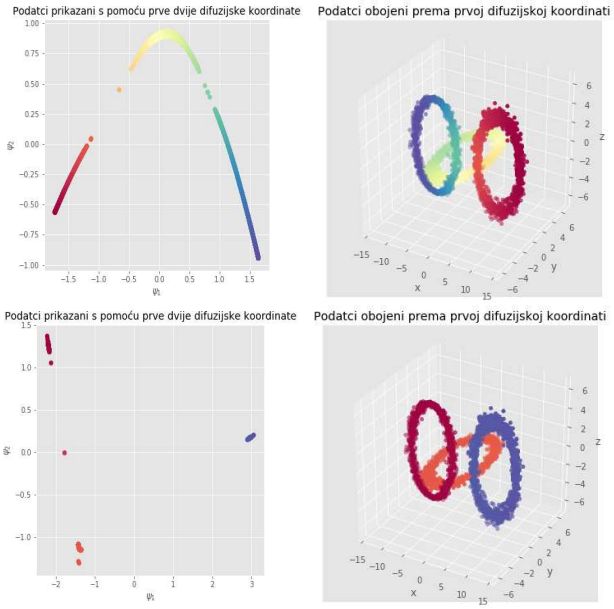
Primjer 5.3.4. Slika 5.15a prikazuje 1000 točaka koje generiraju tri kružnice sa šumom u R^3. Želimo pronaći takve difuzijske koordinate koje prepoznaju tri različite kružnice. Na slici 5.8b prikazani su neki rezultati. U prvom redu vidimo ugradnju u dvodimenzionalni prostor sa sljedećim parametrima $t = 1$, $\varepsilon = 1$, $k = 30$, $\alpha = 1$, dok je u drugom redu ε promijenjen na 0.015625. Nadalje, drugi način kako bismo mogli prepoznati grupe je koristeći se zadanim parametrima drugog algoritma (koristeći `mapalign`) te algoritam *k-sredina* za grupiranje difuzijskih koordinata. Rezultati su prikazani na slici 5.9.

Slika 5.8: Originalne podatkovne točke koje generiraju tri kružnice u $\mathbb{R}^3$ prikazane su pod a). Ulaganje podataka u $\mathbb{R}^2$ s pomoću difuzijskih preslikavanja prikazano je na slici pod b). Primijetimo koliko promjena parametra ε s 1 (u prvom slučaju) na 0.016525 (u drugom slučaju) utječe na grupaciju točaka. U drugom slučaju difuzijske su koordinate konstante i savršeno prepoznaju tri različite kružnice.

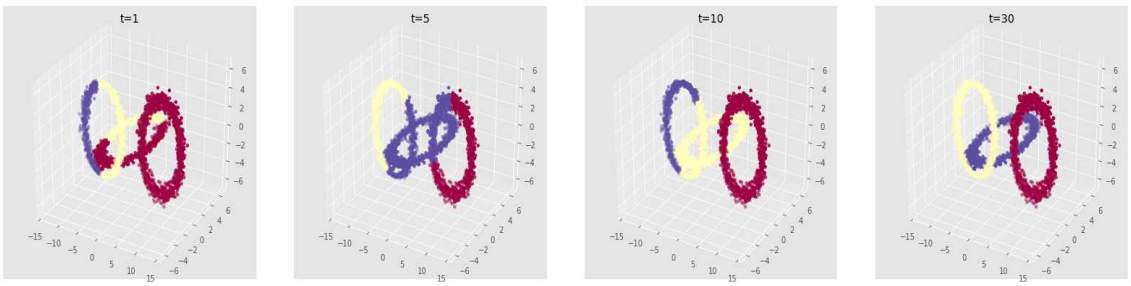
(a) Originalni podatci



(b) Difuzijsko preslikavanje



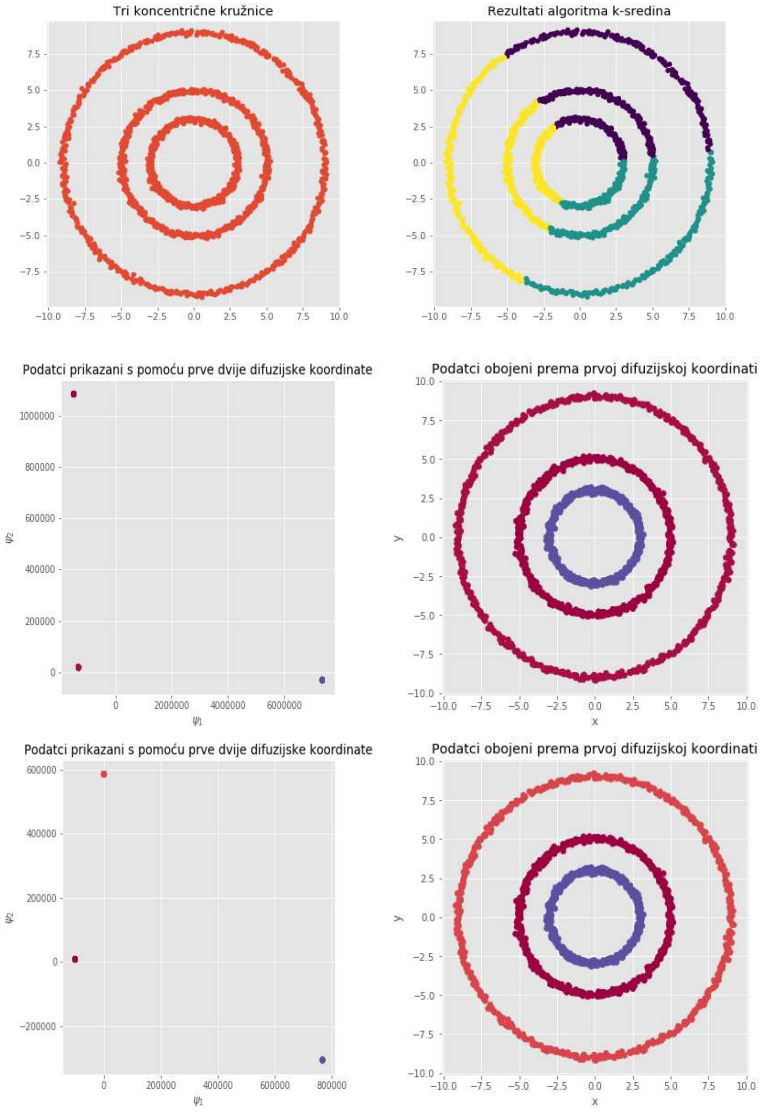
Slika 5.9: Grupiranja podataka primjenom algoritma k-sredina na difuzijske koordinate. Rezultati su prikazani za zadane parametre implementiranog algoritma u 5.1 za različite parametre vremena t. Za sve $t \geq 30$ postiže se grupiranje kao na posljednjoj slici.



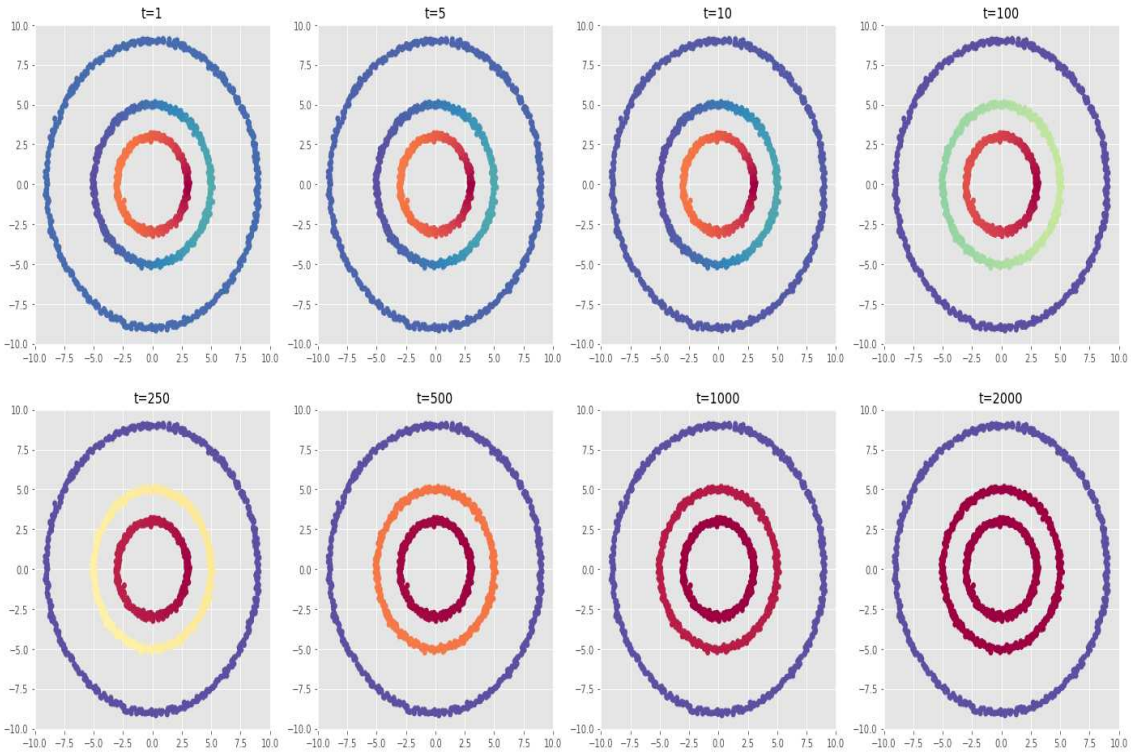
Primjer 5.3.5. U ovom primjeru prepoznajemo tri različite koncentrične kružnice, ilustrirano na slici 5.10. U prvom redu (druga slika) vidimo kako difuzijske koordinate, za $t = 1$, $k = 25$, $\varepsilon = 0.0078$, $\alpha = 1$, prepoznaju dvije grupe, vanjsku kružnicu kao jednu i unu-

tarnje dvije kao drugu. U drugom redu vidimo kako kroz vrijeme difuzijske koordinate prepoznaju tri različite kružnice za $t = 1$, $k = 20$, $\varepsilon = 0.01$, $\alpha = 1$,

Slika 5.10: U prvom su redu prikazani originalni podatci i rezultati algoritma k-sredina koje ne uspijeva prepoznati tri različite koncentrične kružnice. Drugi i treći red prikazuju podatke u difuzijskom prostoru za različite parametre i koncentrične kružnice obojene s pomoću prve difuzijske koordinate. Primijetimo kako u zadnjem redu algoritam prepoznaje tri različite kružnice, dok u prethodnom slučaju prepoznaje dvije grupe.

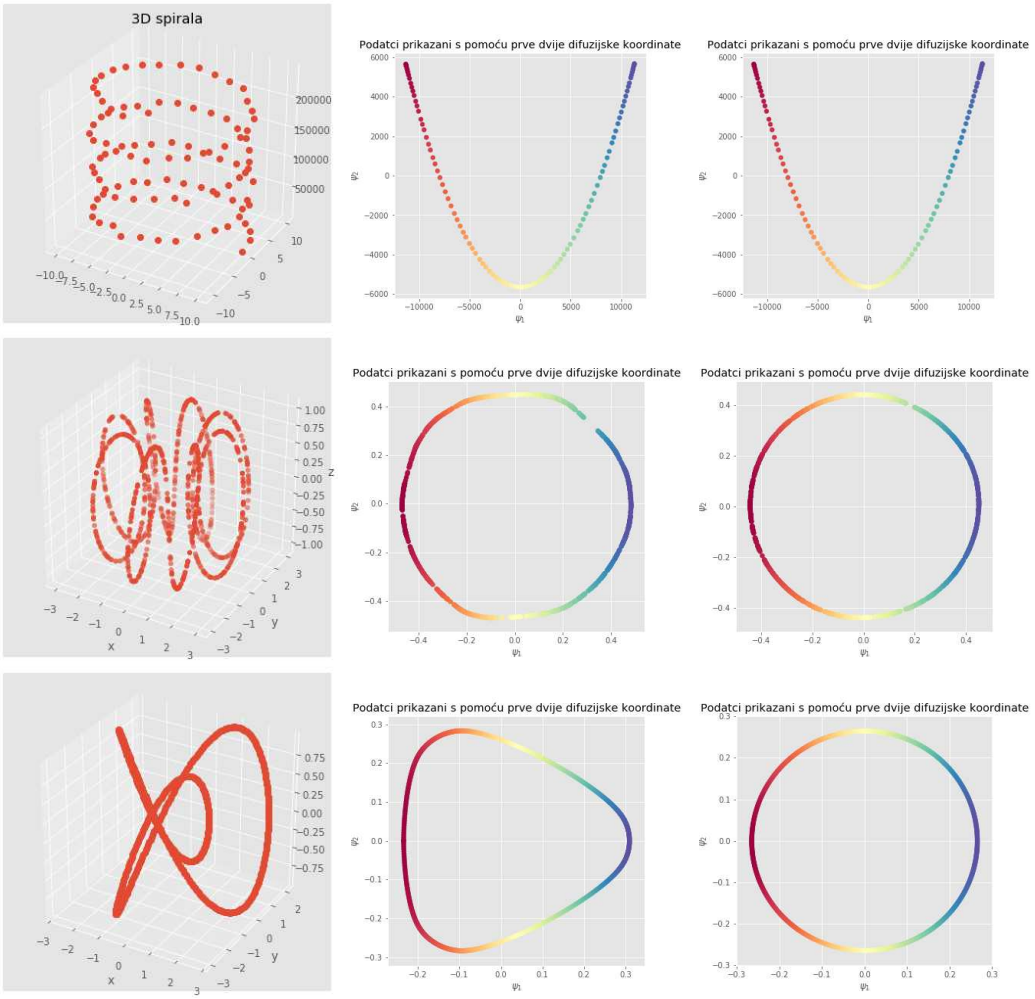


Slika 5.11: Koncentrične kružnice obojene prvom difuzijskom koordinatom za različite vrijednosti parametra t čija je vrijednost označena na grafu.



Primjer 5.3.6. Skup od 100 točaka u $\mathbb{R}^3$, uniformno distribuiranih iz spiralne krivulje, prikazan je na prvoj slici 5.12. Na drugoj su slici prikazane prve dvije difuzijske koordinate. Kao što smo očekivali, prva daje parametrizaciju krivulje, dok je druga njena kvadratna funkcija. Koristimo se procijenjenim $\varepsilon = 2097152.0$, $\alpha = 0$, $k = 10$. Na trećoj je slici prikazan slučaj kada je $\alpha = 1$ koji je isti kao i prethodni jer su točke uniformno distribuirane. U drugom i trećem redu slike 5.12 prikazani su rezultati za zatvorenu spiralu i zatvorenu krivulje u $\mathbb{R}^3$ za različite slučaje parametra α. Očekivano prema teoremu 5.1.1, u zadnjem slučaju krivulja se ugrađuje u $\mathbb{R}^2$ kao kružnica.

Slika 5.12: Slijeva na desno: originalni skup podataka, ulaganje podataka s pomoću difuzijskih preslikavanja za $\alpha = 0$ te u zadnjem stupcu u slučaju $\alpha = 1$.



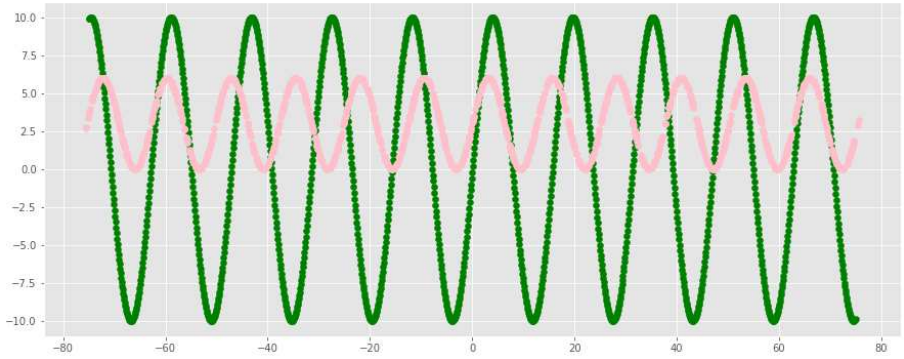
5.4 Ograničenja algoritma

Generiramo dvije različite presijecajuće krivulje u $\mathbb{R}^2$ koje su grafički prikazane na slici 5.13. Želimo da algoritam difuzijskih preslikavanja prepozna te dvije krivulje kao različite grupe. Pokušaj pronalaska takvih difuzijskih koordinata nije doveo do dobrih rezultata, što je prikazano na slici 5.14. No, analizom kretnji različitih difuzijskih koordinata, primjećujemo kako su difuzijske koordinate točaka koje generiraju jednu krivulju potpuno jednake onima koje generiraju drugu krivulju, tj. njihove se koordinate podudaraju. Ova opservacija je prikazana na slici 5.15a, a na 5.15b vidimo kako izgledaju difuzijske koor-

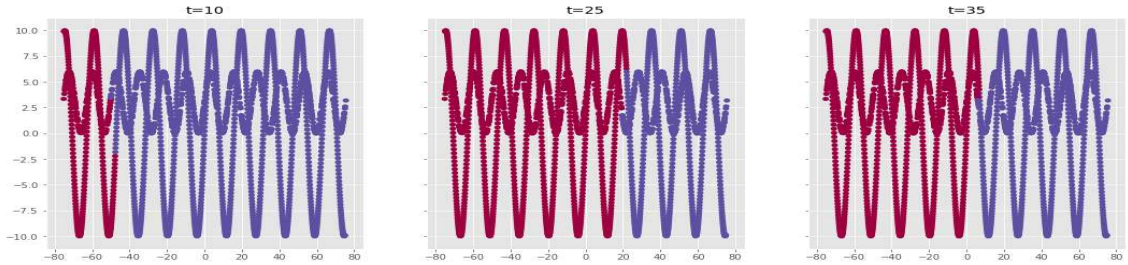
dinate nakon što prvih 1500 točaka translatiramo. Primjenom algoritma k-sredina na tako transformiranim koordinatama dobivamo željenu grupaciju, što je prikazano na slici 5.16.

To nas je motiviralo da uložimo podatke u $\mathbb{R}^3$ (prva slika 5.17) te izračunamo nove difuzijske koordinate. Na drugoj i trećoj slici 5.17 grafički je prikazan jedan takav rezultat, gdje prva difuzijska koordinata prepoznaje manju sinusoidu kao posebnu grupu. Promatrajući ostale difuzijske koordinate kroz vrijeme, dolazimo da željenog cilja. Za odabrane parametre algoritma $t = 100$, $k = 15$, $\varepsilon = 0.17$, peta nam difuzijska koordinata, grafički prikazana na trećoj slici 5.18, omogućava grupaciju podataka (druga slika 5.18).

Slika 5.13: Prikaz dvije različite krivulje u $\mathbb{R}^2$ koje želimo prepoznati pomoću difuzijskih koordinata.

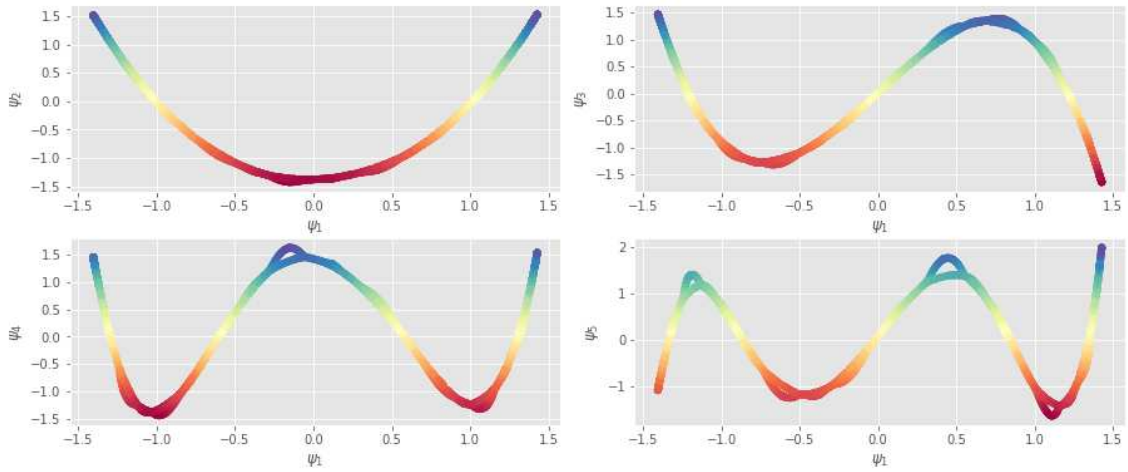


Slika 5.14: Za različite odabire parametara algoritma difuzijska preslikavanja ne uspijevaju prepoznati grupaciju podataka. Jedan je takav neuspjeh grafički prikazan na ovoj slici gdje su rezultati dobiveni primjenom algoritma k-sredina na difuzijske koordinate za različite vremenske parametar.

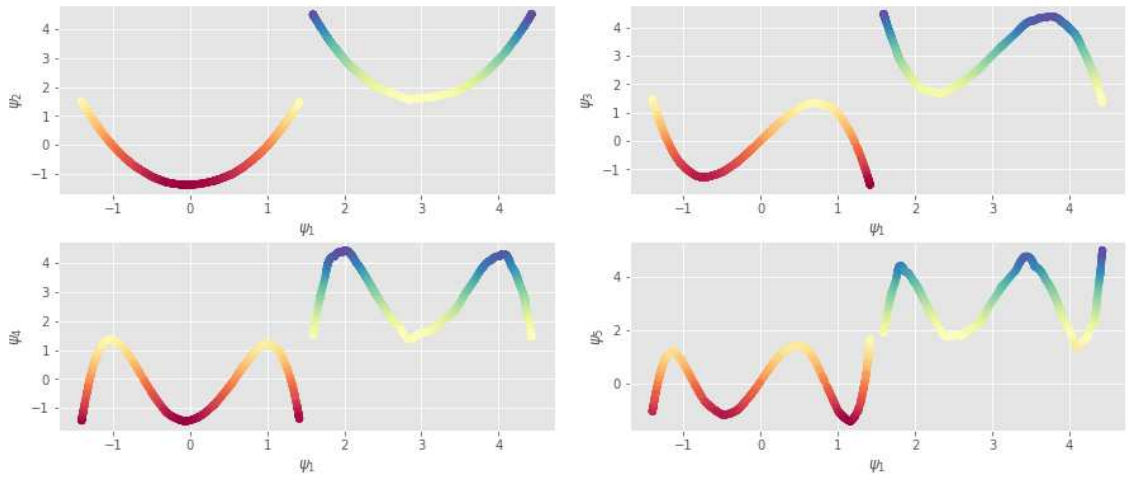


Slika 5.15: Slike pokazuju preklapanje difuzijskih koordinata skupa podataka iz 5.4. Difuzijske su koordinate prvih 1500 točaka koje generiraju manju krivulju i drugih 1500 točaka koje generiraju veću krivulju gotovo potpuno jednake.

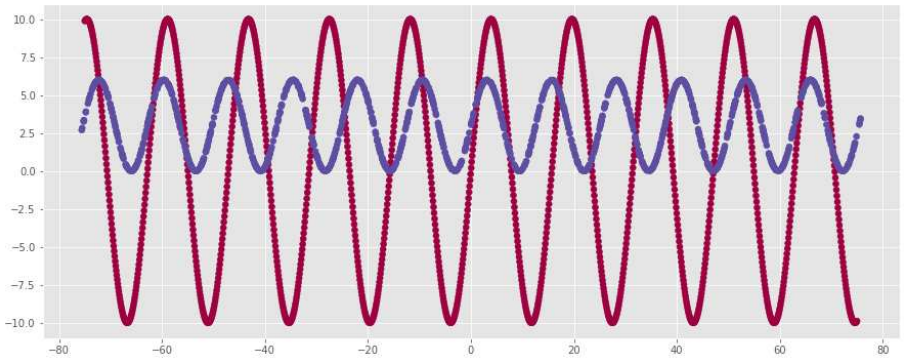
(a)



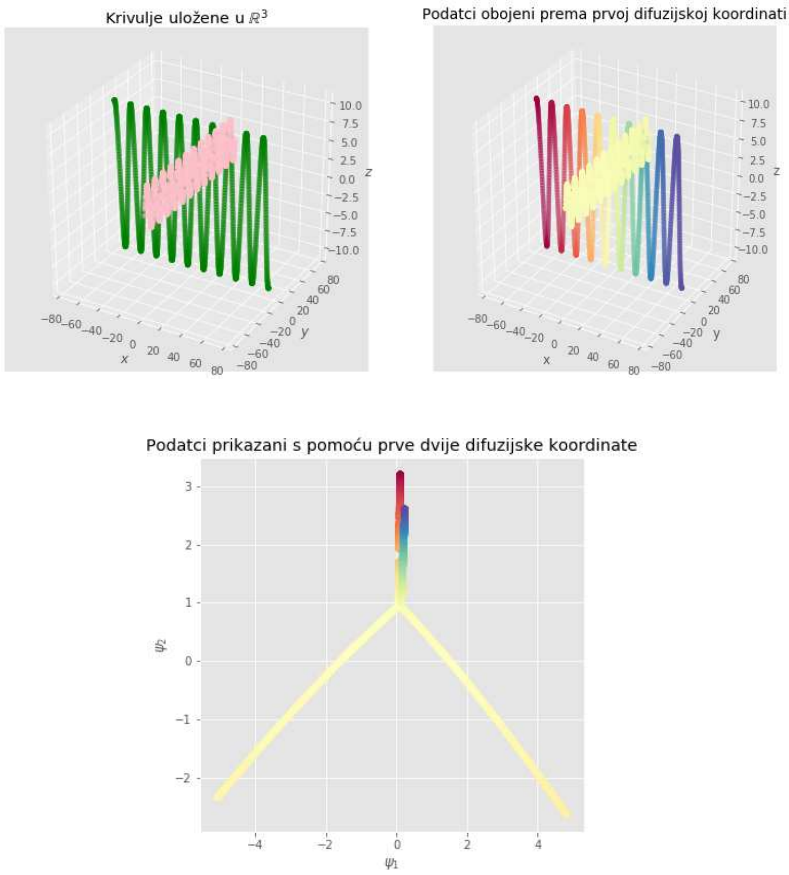
(b)



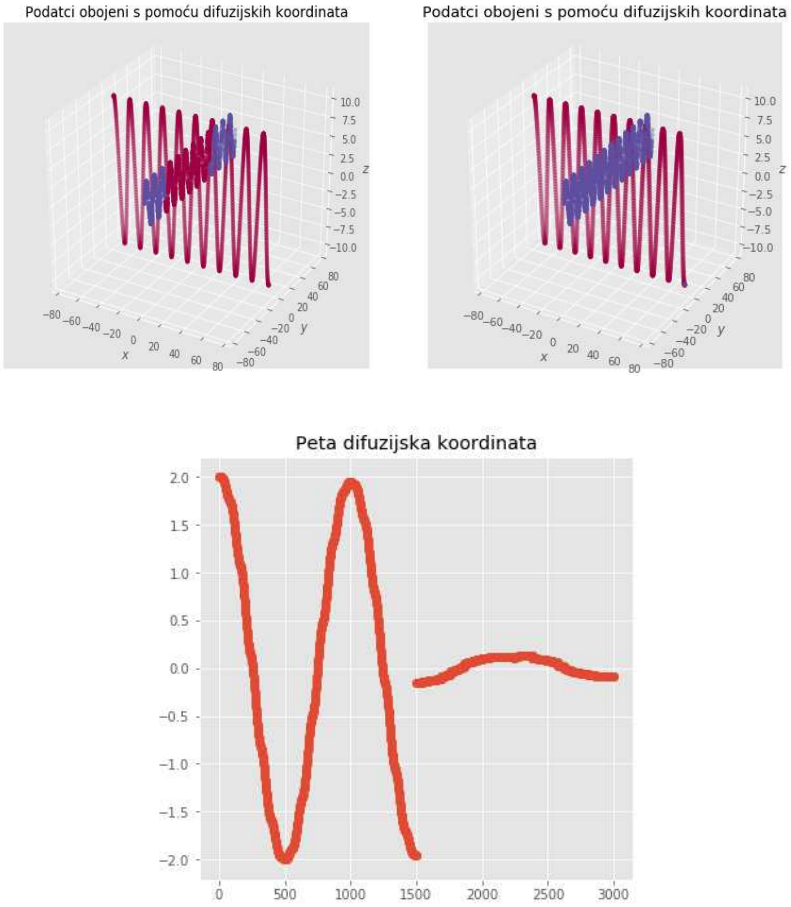
Slika 5.16: Skup podataka obojen tako da smo primijenili algoritam k-sredina na translatirane difuzijske koordinate.



Slika 5.17: Na prvoj slici vidimo dvije različite krivulje, iz primjera 5.4, uložene su u $\mathbb{R}^3$ iz prostora $\mathbb{R}^2$. Nadalje druga slika daje prikaz podataka obojenih prema prvoj difuzijskoj koordinati, dok su na trećoj podaci prikazani s pomoću prve dvije difuzijske koordinate, za parametre algoritma $t = 1$, $\varepsilon = 0.17$, $k = 15$. Vidimo kako je manja krivulja prepoznata kao posebna grupa (obojena žutom bojom), ali druga nije.



Slika 5.18: Rezultati grupacije podataka iz primjera 5.4, dobiveni korištenjem difuzijskih preslikavanja. Odabrani parametri su: $t = 100, \varepsilon = 0.17$, i $k = 15$. Prva slika pokazuje grupaciju korištenjem algoritma k-sredina na difuzijskim koordinatama, dok druga samo na petoj difuzijskoj koordinati.



Dodatak A

Familija difuzija

Promatramo skup podatkovnih točaka u euklidskom prostoru $\mathbb{R}^n$. Prostorni su modeli važni u mnogim primjenama, poput analize slike i računalnog vida [8], a posebno su zanimljivi oni za dobivanje niždimenzionalne reprezentacije skupa podataka. Budući da uzorkovanje podataka općenito nije povezano s geometrijom prostora, željeli bismo otkriti strukturu prostora bez obzira na distribuciju točaka podataka. U slučaju kada se podatkovne točke uzorkuju iz ravnotežne distribucije stohastičkog dinamičkog sustava, situacija je posve drugačija jer je gustoća točaka vrijednost od interesa. Doista, za neke dinamičke fizičke sustave, područja velike gustoće odgovaraju minimumima slobodne energije sustava. Kao posljedica toga, dugotrajno ponašanje dinamike ovog sustava rezultira interakcijom između statistike (gustoće) i geometrije skupa podataka.

Vrlo je primamljivo obrađivati skupove podataka u $\mathbb{R}^n$ uzimajući u obzir graf formiran podatkovnim točkama čije su težine određene nekom izotropnom jezgrenom funkcijom, npr. $k_\varepsilon(x,y) = e^{-\|x-y\|^2/\varepsilon}$, za neke pažljivo odabran parametar ε. U [10], Belkin i Niyogi predlažu da se izračuna *Laplacian* grafa te jezgre i iskoriste spektralna svojstva odgovarajuće difuzije za grupiranje i organiziranje podataka, kao što je opisano u ovom radu. Iako su vrline ovog vrste pristupa dobro poznate za općenite grafove, više se može reći za poseban slučaj točaka u euklidskom prostoru. Želimo znati koliki je utjecaj gustoće točaka i geometrije mogućih osnova skupa podataka na svojstvene funkcije i spektar difuzija.

Uvodimo familiju anizotropnih difuzijskih procesa koji su svi dobiveni kao limesi Laplaciana. Ova je familija parametrizirana s $\alpha \in \mathbb{R}$ koji se može prilagoditi u svrhu određivanja količine utjecaja gustoće u infinitezimalne (beskrajno mala količina) tranzicije difuzije [13]. Ključno je što se normalizacija *Laplaciana* grafa ne primjenjuje na grafovima s izotropnim težinama, već na renormaliziranom grafu. Tri vrijednosti parametra $\alpha \in \mathbb{R}$ su posebno zanimljive:

- Kada je $\alpha = 0$, difuzija se svodi na klasičan normalizirani *Laplacian* grafa te normalizaciju koja se primjenjuje na grafu s izotropnim težinama. Utjecaj gustoće je u

ovom slučaju maksimalan.

- Za srednji slučaj $\alpha = \frac{1}{2}$, Markovljev lanac je aproksimacija difuzije *Fokker-Planckove* jednadžbe koja omogućava približavanje dugotrajnog ponašanja ili točkovne distribucije sustava opisanog određenom stohastičkom diferencijalnom jednadžbom.
- Kada je $\alpha = 1$ i ako točke približno leže na potprostoru od $\mathbb{R}^n$, dobiva se aproksimacija *Laplace-Beltrami* operatora. U ovom slučaju, moguće je vratiti Riemannianovu geometriju skupa podataka, bez obzira na raspodjelu točaka. Ovaj je slučaj osobito važan u mnogim primjenama.

U nastavku ćemo objasniti konstrukcije ove familije difuzija, slučajevi detaljnije opisani i objašnjeni u [13].

Neka je $q(x)$ gustoća točaka na $\mathcal{M}$. Najprije renormaliziramo rotacijski invarijantne težine u anizotropnoj jezgra te iz novog grafa određujemo difuziju normaliziranog *Laplaciana* grafa.

1. Fiksirajmo $\alpha \in \mathbb{R}$ i rotacijski invarijantan $k_\epsilon(x,y) = h(\frac{\|x-y\|^2}{\epsilon})$.
2. Neka je
$$q_\epsilon(x) = \int_{\mathbf{X}} k_\epsilon(x,y)q(y)dy$$
te izgradimo novu jezgru
$$k_\epsilon^{(\alpha)}(x,y) = \frac{k_\epsilon(x,y)}{q_\epsilon^\alpha(x)q_\epsilon^\alpha(y)}.$$
3. Primjenom normalizacije težinskog *Laplacian* grafa na ovu jezgru imamo
$$d_\epsilon^{(\alpha)}(x) = \int_{\mathbf{X}} k_\epsilon^{(\alpha)}(x,y)q(y)dy,$$
što dovodi do definicije anizotropnog tranzicijske jezgre
$$p_{\epsilon,\alpha}(x,y) = \frac{k_\epsilon^{(\alpha)}(x,y)}{d_\epsilon^{(\alpha)}(x)}.$$

Definira se operator induciran tranzicijskom jezgrom:

$$P_{\epsilon,\alpha}f(x) = \int_{\mathbf{X}} p_{\epsilon,\alpha}(x,y)f(y)q(y)dy.$$

Markovljev lanac definira na podacima brze i spore smjerove širenja, temeljene na vrijednostima jezgrene funkcije, a kako se kreće naprijed, informacije o lokalnoj geometriji šire se i akumuliraju na isti način kao što se lokalni prijelazi sustava (dani diferencijalim jednadžbama) mogu integrirati kako bi se dobila globalna karakterizacija podataka.

Dodatak B

Dijelovi koda korišteni za vizualizaciju metode

Grafički su prikazi rezultata u radu dobiveni korištenjem implementiranih metoda u *Pythonu*. Difuzijske koordinate dobivene su korištenjem funkcije `diffusion_map` iz biblioteke `pydiffmap`. Ona u sebi ima klasu `pydiffmap.diffusion_map.DiffusionMap` koja za dane parametre izračunava difuzijske koordinate. U toj klasi postoji metoda `from_sklearn` koja izvodi difuzijska preslikavanja s pomoću jezgre konstruirane koristeći `sklearn` objekt najbližeg susjeda. Također, postoji klasa `pydiffmap.kernel.Kernel` koja može izračunati optimalni parametar ε[1].

Sljedećim je kodom dan jedan od primjera izračunavanja difuzijskih koordinata i njihova vizualizacija za podatke iz primjera 5.3.1, gdje je `n_evecs` broj difuzijskih koordinata, `k` broj susjeda u grafu sličnosti (u kojem konstruiramo jezgru), `epsilon` parametar jezgre, a `alpha` parametar opisan u dodatku A. U primjerima 3.3.1, 5.3.4 i 5.3.5 se može primijetiti kako je ključno za grupiranje podataka (osim optimalne vrijednosti ε) odabrati što manju vrijednost parametra k. Međutim, odabir 'premale' vrijednosti može dovesti do nepovezanog grafa.

```
from pydiffmap import diffusion_map as dm
from pydiffmap.visualization import embedding_plot,
data_plot

# Inicijalizacija Diffusion map objekta
mydmap = dm.DiffusionMap.from_sklearn(n_evecs=2, k=30,
```

[1]Trenutno postoji samo jedna ugrađena metoda za procjenjivanje tzv. 'bgh' metoda (Berry, Giannakis and Harlim).

```
epsilon=2, alpha=0.5)

# Fitanje podataka i pronalazak difuzijskog preslikavanja
dmap = mydmap.fit_transform(dataset2a)

# Vizualizacija
embedding_plot(mydmap, scatter_kwargs =
{'c': mydmap.dmap[:,0], 'cmap': 'Spectral'})
data_plot(mydmap, dim=2, scatter_kwargs =
{'cmap': 'Spectral'})
plt.show()
```

Na sljedeći način koristimo implementiranu metodu za procjenjivanje parametra ε.

```
# Inicijalizacija  Diffusion map objekta
# neighbor_params = {'algorithm': 'ball_tree'}
mydmap = dm.DiffusionMap.from_sklearn(n_evecs=5, k=200,
epsilon='bgh', alpha=1.0)

# Fitanje podataka i pronalazak difuzijskog preslikavanja
dmap = mydmap.fit_transform(swiss_roll)

# Ispis procjenjenog epsilona
mydmap.epsilon_fitted
```

Nadalje, rezultati dobiveni promatranje kroz vrijeme koriste implementaciju preuzetu s *git://github.com/satra/mapalign*. Ova implementacija ima slične parametre, ali nam omogućuje direktno zadati vremenski parametar t. Slijedi primjer koda koji pronalazi prvih deset za podatke iz primjera 5.3.4. Parametar `gamma` je ε, a `n_neighbors` je već spomenuti k.

```
from mapalign.embed import DiffusionMapEmbedding

de = DiffusionMapEmbedding(alpha=0.5, diffusion_time=t,
affinity='markov', n_components=10, gamma = 0.074393,
n_neighbors = 70).fit_transform(dataset3.copy())
```

Sljdećim je kodom dobiven grafički prikaz 5.9, kako difuzijske koordinate izgledaju u različitim vremenima (omogućava prikaz za 8 različitih vremena).

```
fig = plt.figure(figsize=(13,13))
fig.subplots_adjust(wspace=0.25, hspace=0.3,left=0.05,
right=0.95)
t = [1, 5, 10, 30, 50, 100, 200, 300]

for idx in range(8):
    de = DiffusionMapEmbedding(alpha=0.5,
    diffusion_time=t[idx], affinity='markov',
    n_components=10).fit_transform(dataset3.copy())
    ed = KMeans(n_clusters=3).fit(de).labels_

    ax = fig.add_subplot(2,4,idx+1,projection='3d')
    ax.scatter(dataset3[:, 0], dataset3[:, 1],
    dataset3[:,2], c=ed, cmap=plt.cm.Spectral, linewidths=0)
    plt.axis('tight')
    plt.title('t=%d'%(t[idx]))
```

Ovdje pronalazimo difuzijske koordinate pa ih grupiramo pomoću algoritma *k-sredina* te tako odredimo grupaciju originalnih podataka. To možemo postići i bez korištenja algoritma *k-sredina* (bez unaprijed danog broja grupa), transformacijom difuzijskih koordinata, pokazano sljedećim kodom.

```
cluster = None
f, axarr = plt.subplots(1, 7, sharex=True, sharey=True,
figsize=(25, 5))
for idx, t in enumerate([1, 5, 10, 100, 250, 500,1000]):
    de = DiffusionMapEmbedding(alpha=1.0,
    diffusion_time=t, affinity='markov',
    n_components=6).fit_transform(X.copy())
    ed = (de - de[0, :])
    ed = np.sqrt(np.sum(ed * ed , axis=1))
    ed = ed/max(ed)
    if cluster is not None:
        ed = KMeans(n_clusters=cluster).fit(de).labels_
    plt.axes(axarr[idx])
    if cluster is not None:
        plt.scatter(X[:, 0], X[:, 1], c=ed,
        cmap=plt.cm.Set1, linewidths=0)
    else:
        plt.scatter(X[:, 0], X[:, 1], c=ed,
```

```
        cmap=plt.cm.Spectral, linewidths=0)
    plt.axis('tight')
    plt.title('t={:g}'.format(t))
```

Naime, ovdje je *ed* vektor udaljenosti redaka matrice *de* (difuzijskih koordinata) od njezinog prvog retka normiran po normi ∞. Ovaj kod omogućava prikaz grupacije tri koncentrične kružnice, vidi sliku 5.11.

Pronalazak raspona parametra ε, koji je opisan u radu, nam daje sljedeći kod. Funkcija `l` daje nam vrijednosti $L(\varepsilon)$, a `epsilon` nam prikazuje logaritamski graf te funkcije.

```
def epsilon(X,n,rang):
    start_time = time.time()
    fig = plt.figure(figsize=(9,8))
    eps = np.random.uniform(0,rang,n)
    ax = fig.add_subplot(1, 1, 1)
    y = np.zeros(n)
    for i in range (n):
        y[i]=l(eps[i],X)
    ax.scatter(eps,y, color='blue', lw=2)
    ax.set_yscale('log')
    ax.set_xscale('log')
    plt.axis('equal')

    plt.autoscale(enable=True, axis='x', tight=True)
    plt.show()
    print("--- %s seconds ---" % (time.time() - start_time))

def l(eps,X):
    dists = euclidean_distances(X, X)
    K = np.exp(-dists**2 / (eps**2))
    return K.sum()
```

Bibliografija

[1] I. G. Kevrekidis, R. R. Coifman, A. Singer, R. Erban, *Detecting the slow manifold by anisotropic diffusion maps*, 2007, `https://people.maths.ox.ac.uk/erban/papers/PNASdiffusionmaps.pdf`.

[2] Y. Weiss A. Y. Ng, M. I. Jordan, *On Spectral Clustering: Analysis and an Algorithm*, Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (Cambridge, MA, USA), NIPS'01, MIT Press, 2001, str. 849–856.

[3] R. Coifman, I.G. Kevrekidis, B. Nadler, S. Lafon, *Diffusion Maps - a Probabilistic Interpretation for Spectral Embedding and Clustering Algorithms*, sv. 58, str. 239–259, Springer, Berlin, Heidelberg, Cham, 2008, ISBN 978-3-540-73750-6.

[4] E. L. Wilmer, D. A. Levin, Y. Peres, *Markov chains and mixing times*, American Mathematical Society, 2006, `https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf`.

[5] J. C. Langford, J. B. Tenenbaum, V. de Silva, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, sv. 290, str. 2319–2323, Science, 2000, `https://web.mit.edu/cocosci/Papers/sci_reprint.pdf`.

[6] W. Hereman S. J. van der Walt J. de Porte, B.M. Herbst, *An Introduction to Diffusion Maps*, Applied Mthematics Division, Department of Mathematical Sciences **79** (2008), br. 8, 10.

[7] J. Malik, J. Shi, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000), br. 3, 1373 – 1396, ISSN 1939-3539.

[8] A. B. Lee, K. S. Pedersen, *Toward a Full Probability Model of Edges in Natural Images. In: Heyden A., Sparr G., Nielsen M., Johansen P. (eds) Computer Vision*, Applied Mthematics Division, Department of Mathematical Sciences **2350** (2002), 328–342.

[9] Stéphane Lafon i Ann B. Lee, *Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization*, IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006), 1393–1403.

[10] P. Niyogi, M. Belkin, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Massachusetts Institute of Technology **15** (2003), br. 3, 1373 – 1396, https://www2.imm.dtu.dk/projects/manifold/Papers/Laplacian.pdf.

[11] J. Shi, M. Meila, *A Random Walks View of Spectral Segmentation*, 2001, https://sites.cs.ucsb.edu/~veronika/MAE/arandomwalksviewofimgsegmt_meila_shi_nips00.pdf.

[12] T. Jaakkola, M. Szummer, *Partially labeled classification with Markov random walks*, Advances in Neural Information Processing Systems 14 (T. G. Dietterich, S. Becker i Z. Ghahramani, ur.), MIT Press, 2002, str. 945–952.

[13] S. Lafon, R. Coifman, *Diffusion maps*, Applied and Computational Harmonic Analysis **21** (2006), 5–30, http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/Lafon06.pdf.

[14] A. B. Lee, S. S. Lafon, *Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization*, **28** (2006), br. 9, 1393–1403.

[15] J. Shi, S. X. Yu, *Multiclass spectral clustering*, Proceedings Ninth IEEE International Conference on Computer Vision, 2003, str. 313–319 vol.1.

[16] S.S.Lafon, *Diffusion maps and Gemetric Harmonics*, PhD thesis, Yele University (2004), https://sites.google.com/site/stefansresearchpapers/.

[17] S. D. Brown, Van Ravenzwaaij, P. Cassey, *A simple introduction to Markov Chain Monte–Carlo sampling.*, sv. 25, 2018, str. 1433–154, https://doi.org/10.3758/s13423-016-1015-8.

[18] Z. Vondraček, *Markovljevi lanci*, 2009, https://web.math.pmf.unizg.hr/~vondra/ml12-predavanja.html.

Sažetak

Spektralno ulaganje i grupiranje podataka metode su nelinearne redukcije dimenzije i grupiranja kompleksnih visoko dimenzionalnih podataka. Kako bismo uspješno proveli ove metode, ponekad je ključno prikazati skup podatka s pomoću odgovarajućeg skupa koordinata. Jedan od načina kojim to možemo postići jesu difuzijska preslikavanja.

Povezanosti između točaka mjerimo s pomoću jezgrene funkcije koja definira lokalnu mjeru sličnosti točaka. Za danu matricu susjedstva svih točaka u određenom skupu podataka definiramo slučajnu šetnju po grafu u kojem su vrhovi točke podataka. Nadalje, pronalazimo tranzicijsku matricu pripadnog Markovljeva lanca, čije svojstvene vrijednosti i vektore koristimo za konstrukciju difuzijskih koordinata. One preslikavaju podatke u difuzijski prostor, gdje je definirana difuzijska udaljenost jednaka euklidskoj. Odbacivanjem nekih difuzijskih koordinata, tako da ne gubimo informacije o strukturi podataka, pronalazimo parametre koji opisuju strukturu u nekom nižedimenzionalnom prostoru. Promatranjem većih potencija matrice prijelaza Markovljeva lanca možemo dobiti bolji uvid u globalnu strukturu skupa podataka te se upravo u nekim većim vremenskim skalama otkriva njihova dinamika.

Ovaj rad pruža vjerojatnosnu analizu difuzijskih preslikavanja. Dobivamo povezanost spektralnih svojstava Markovljevih procesa s njihovom geometrijskom prirodom. Za razliku od drugih popularnih metoda redukcije dimenzije, kao što je analiza glavnih komponenti difuzijska preslikavanja mogu otkriti nelinearne strukture. Pokazano je da je ova tehnika otporna na šumove i računski brza. U radu su obrađeni mnogi primjeri koji ilustriraju glavne teorijske principe.

Summary

Spectral embedding and data clustering are methods for nonlinear dimensional reduction and grouping of complex high - dimensional data. For a successful application of these methods,, sometimes it is crucial to display set of data using an appropriate set of coordinates. One way to achieve this is by using diffusion mapping.

The connection between points is measured using a kernel function that defines a local measure of point similarity. For a given neighborhood matrix of all points in certain data set, we define random walk on graph in which vertices are data points. Furthermore, in that way we find the transition matrix of the Markov chain, whose eigenvalues and eigenvectors are used for a construction of the diffusion coordinates. Those coordinates map the data in the diffusion space, where the defined diffusion distance is equal to Euclidean distance. By discarding some diffusion coordinates, information about data structure does not get lost and therefore, we can find parameters that describe the structure in some lower-dimensional space. By observing the greater potentials of the Markov chain transition matrix we can gain a better insight into the global structure of the data set whose dynamics can be revealed in some greater time scales.

This thesis provides a probabilistic analysis of diffusion maps. We obtain the connection of the spectral properties of Markov processes with their geometric nature. Unlike other popular dimensional reduction methods, such as the principal component analysis, diffusion maps can detect nonlinear structures. It has been proven that this techniques is noise resistant and that is computationally fast. The thesis provides many case study examples that illustrate the main theoretical principles.

Životopis

Moje ime je Tatjana Ramljak. Rođena sam 8. kolovoza 1995. u Sisku. Oduvijek sam imala iznimno veliki interes prema matematici i informatici te bila sposobna rješavati različite matematičke probleme. Završila sam prirodoslovno-matematičku gimnaziju u Kutini te 2014. upisala prirodoslovno-matematički fakultet u Zagrebu. Tri godine nakon, 2017. godine, stekla sam titulu prvostupnice matematike (*bacc. math*) nakon koje sam upisala diplomski sveučilišni studij Matematička statistika.