



UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I INFORMATIKU



Srđan Lazendić

MATEMATIČKI ALATI ZA REDUKCIJU DIMENZIONALNOSTI SIGNALA

- master rad -

Mentor:
dr Sanja Konjik

Novi Sad, 2016.

Predgovor

U današnje vrijeme velika je potreba za računarima i računarskim naukama. U posljednjih nekoliko godina radi se na olakšavanju određenih poslova i predviđanju određenih rezultata koristeći računare. Veliku primjenu u tome našlo je *mašinsko učenje*. Mašinsko učenje je podoblast vještačke inteligencije i koristi statistiku u cilju učenja na osnovu prethodno dostupnih podataka. Mašinsko učenje ima za cilj da stvori softvere koji mogu da uče na osnovu prethodnog znanja i iskustva, dok vještačka inteligencija ima za cilj ne samo da imitira ljudsko razmišljanje kroz učenje nego da ta mašina ima i apstraktno mišljenje kao i mogućnost rasuđivanja.

Veliku primjenu mašinsko učenje našlo je u analizi podataka (eng. data mining). U većini problema sa kojima se susrećemo imamo veoma velike baze podataka, pri čemu se javlja problem kako da iz velike količine podataka izdvojimo potrebne i važne podatke. Data mining je proces pretraživanja informacija sa određenim podacima i osobinama koristeći računare. Prilikom pripreme baze podataka potrebno je organizovati sirove informacije, transformisati i grupisati te podatke i često smanjiti broj promjenljivih tako da možemo dobiti dobre rezultate nakon same analize informacija. Cilj je da dobijemo rezultate u vidu dijagrama ili tabela u dvije ili tri dimenzije kako bismo mogli da vizualizujemo dobijene rezultate.

U ovom radu naš cilj je, kao što i sam naslov rada kaže, da predstavimo tehnike pomoću kojih možemo smanjiti dimenziju određenog skupa podataka. Napomenimo da se signal definiše kao skup podataka koji nosi informacije o određenom fenomenu, tako da u ovom radu nećemo eksplicitno navoditi da posmatramo signal, nego ćemo posmatrati skupove podataka. Predstavili smo neke od najpoznatijih tehnika za redukciju dimenzionalnosti, koje se dijele na linearne i nelinearne tehnike. Upravo smo Glavu 3 i Glavu 4 posvetili linearnim, odnosno nelinearnim tehnikama, redom. Ovaj rad se sastoji od četiri glave.

Prvi dio je Uvod i u njemu smo naveli motivaciju za bavljenje ovom temom. Tačnije naveli smo tri primjera u kojima se javlja potreba za tehnikama za redukciju dimenzionalnosti. Ovakvih primjera je zaista mnogo, a lijep pregled može se naći u [25].

Glava 1 je rezervisana za osnovne pojmove iz linearne i matrice algebre, vjerovatnoće i statistike, kao i multivarijacione analize, na kojima počivaju tehnike za redukciju dimenzionalnosti. Naveli smo osnovne definicije i tvrđenja u vezi sa vektorima i matricama, kao i različite dekompozicije matrice (spektralna, SVD i Cholesky). U drugom poglavlju ove glave najviše smo se bazirali na višedimenzionalne podatke i njihovu matricnu reprezentaciju. Vidjećemo kako možemo da posmatramo višedimenzionalne podatke i kako da odredimo njihove numeričke karakteristike, koje će nam biti ključne u Glavi 3.

U drugoj glavi naveli smo najvažnije pojmove koji su potrebni za definisanje nelinearnih tehnika za redukciju dimenzionalnosti. To su pojmovi iz diferencijalne geometrije i teorije grafova. U većini slučajeva višedimenzionalni podaci leže na nelinearnim strukturama takozvanim mnogostrukostima. Na početku smo uveli pojam diferencijabilnih mnogostrukosti i tangentnog prostora, kao i pojam Rimanovih mnogostrukosti koje su bitne da opišemo zajedničke osobine između tačaka skupa podataka. Cilj je da mjerimo rastojanja (geodezijska rastojanja) između tačaka na mnogostrukosti, ali s obzirom da ne znamo eksplicitno jednačinu mnogostrukosti, to nije izvodljivo u većini slučajeva. Olakšavajuća okolnost je što su skupovi podataka uglavnom konačni, tako da možemo iskoristiti predstavljanje skupova podataka pomoću grafova, tako da smo u drugom poglavlju naveli osno-

vne pojmove vezane za orijentisane i neorijentisane grafove, kao i pojmove matrice povezanosti i težinskih grafova, koji su za nas od posebnog značaja.

Na samom početku treće glave upoznali smo se sa pojmom i problemom dimenzionalnosti kao i sa klasifikacijom tehnika i problema redukcije dimenzionalnosti. Zatim smo dali detaljan pregled linearnih tehnika, nakon čega smo razmotrili dvije najpoznatije linearne tehnike, a to su Analiza glavnih komponenti - PCA i Klasično multidimenzionalno skaliranje - CMDS. Detaljno smo razmotrili dva pristupa PCA, geometrijski i statistički pristup. Kao što smo i kasnije za svaku tehniku, i ovdje smo naveli Matlab kod za PCA kao i primjere primjene PCA, da bismo na kraju naveli i Eigenface algoritam (metod za prepoznavanje lica) koji se zasniva na PCA metodi. Drugo poglavlje je posvećeno multidimenzionalnom skaliranju MDS. Definisali smo Euklidsku matricu rastojanja kao i Gramovu matricu pomoću kojih smo definisali pojam sličnosti proizvoljnog skupa podataka. Na kraju smo detaljnije razmotrili CMDS, a naveli smo i njegov Matlab kod.

Kao prirodan nastavak prethodnih glava, u četvrtoj glavi slijedi proučavanje nelinearnih tehnika za redukciju dimenzionalnosti signala. Nelinearne tehnike su pogodne za rješavanje problema višedimenzionalnih podataka, kada podaci leže na strukturama koje nisu linearne, konkretno kada leže na nelinearnim mnogostrukostima (eng. manifold), što je često slučaj. Pošto svaka od tehnika ima svoje prednosti i nedostatke, u zavisnosti od problema sa kojim smo se susreli biraćemo tehniku koja će nam što efikasnije riješiti problem. Upravo iz tog razloga, u poslednjih nekoliko godina javlja se veliki broj uopštenja već razvijenih tehnika za redukciju dimenzionalnosti. U ovom radu odlučili smo se za dvije tehnike i to Isomap i Lokalno linearno potapanje (LLE). Kao i kod linearnih tehnika i ovdje smo u oba poglavlja, dali opis Isomap, odnosno LLE metoda, kao i opis algoritma obe tehnike sa Matlab kodom. S obzirom da je literatura iz ove oblasti obimna, može se naći mnogo interesantnih primjera i tehnika koje su vrijedne pomena, ali zbog obima ovog rada nije ih moguće navesti. Rad smo završili primjerima primjene nelinearnih tehnika na vještačke površi.

★ ★ ★

Na samom kraju bih se zahvalio svojim roditeljima, Željani i Goranu, kao i sestri Vladani, na bezgraničnoj ljubavi, podršci i pomoći tokom svih ovih godina. Takođe, zahvaljujem se svojim prijateljima, kao i kolegama koji su moje studentske dane učinili ljepšim. Zahvaljujem se svojim profesorima na stečenom znanju kako tokom osnovnih tako i tokom master studija, kao i dr Aleksandri Pižurici i dr Dejanu Vukobratoviću, koji su pristali da budu članovi komisije za odbranu ovog master rada. Na kraju, moram da istaknem da mi je posebna čast bila što sam za profesora imao i dr Sanju Konjik. Njoj dugujem najveću zahvalnost kako na znanju koje mi je pružila tako i na izboru teme, korisnim sugestijama, pomoći i podršci tokom pisanja ovog master rada.

Novi Sad, jul 2016.

Srđan Lazendić

Sadržaj

Predgovor	i
Uvod	1
1 Osnovni pojmovi linearne algebre i statistike	3
1.1 Osnovni pojmovi linearne algebre	3
1.1.1 Vektori i matrice	3
1.1.2 Vektorske i matrice norme	4
1.1.3 Simetrične i pozitivno definitne matrice	6
1.1.4 Dekompozicija singularnim vrijednostima (SVD)	7
1.1.5 Projekcija na potprostor	8
1.2 Osnovni pojmovi vjerovatnoće i statistike	9
1.2.1 Slučajne promjenljive i slučajni vektori	9
1.2.2 Numeričke karakteristike slučajnih promjenljivih	10
1.2.3 Osnovni pojmovi statistike i multivarijacione analize	12
2 Osnovni pojmovi diferencijalne geometrije i teorije grafova	16
2.1 Osnovni pojmovi diferencijalne geometrije	16
2.1.1 Tangentni prostor i tangentni vektori	17
2.1.2 Rimanove mnogostrukosti	18
2.2 Osnovni pojmovi teorije grafova	19
2.2.1 Sistem okolina	19
2.2.2 Graf kao model za opis strukture podataka	20
2.2.3 Laplasijan grafa	22
3 Linearne tehnike za redukciju dimenzionalnosti signala	24
3.1 Problem dimenzionalnosti	24
3.2 Analiza glavnih komponenti - PCA	25
3.2.1 Geometrijski pristup PCA	25
3.2.2 Statistički pristup PCA	29
3.2.3 Matlab kod za PCA algoritam i primjeri	31
3.2.4 Primjena PCA	34
3.3 Multidimenzionalno skaliranje - MDS	36
3.3.1 Euklidska i Gramova matrica	37
3.3.2 Klasično multidimenzionalno skaliranje - CMDS	41
3.3.3 Matlab kod za CMDS	43

4	Nelinearne tehnike za redukciju dimenzionalnosti	46
4.1	Zajednički pojmovi za nelinearne tehnike	46
4.2	Isomap	48
4.2.1	Isomap metod	48
4.2.2	Isomap algoritam	51
4.2.3	Matlab kod za Isomap algoritam	52
4.3	Lokalno linearno potapanje - LLE	54
4.3.1	Baricentrične koordinate	54
4.3.2	LLE metod	55
4.3.3	LLE algoritam	59
4.3.4	Matlab kod za LLE algoritam	61
	Zaključak	64
	Literatura	65
	Biografija	67

Uvod

Sa razvojem računara i tehnologije omogućeno je da se različiti objekti mogu predstaviti elektronskim putem u obliku višedimenzionalnih podataka. Takvi su npr. zvučni i video signali, slike, tekstualni dokumenti, otisci prstiju, hiperspektralno snimanje itd. Kako bismo dobili određene informacije potrebno je analizirati veliku količinu podataka. Npr. kada se vrši analiza otiska prsta, pretraga putem interneta pomoću ključnih riječi ili izdvajanje određenih uzoraka na nekoj slici potrebno je pretraživati velike baze podataka. Stoga, potrebno je razviti sisteme koji će omogućiti da se vrši obrada tih podataka. S obzirom na veliku količinu podataka, direktna pretraga bila bi nestabilna ili čak neizvodljiva. Veliki broj sistema radi samo sa podacima koji imaju prilično malu dimenziju, pa ukoliko je dimenzija podataka veća nego što to sistem dozvoljava, ne može se dobiti odgovor, tako da je redukcija dimenzije neophodna. Razmotrićemo sada neke od primjera u kojima se javlja potreba za redukcijom dimenzionalnosti.

[1] Posmatrajmo primjer analize digitalne slike otisaka prstiju. Analiza otiska prsta naziva se daktiloskopija. Prsti su ispresijecani mnogobrojnim linijama, takozvanim papilarnim linijama. Otisak tih papilarnih linija je jedinstven i na osnovu njih utvrđujemo kome pripada otisak prsta. Pretpostavimo da je rezolucija slike 64×64 , tada je dimenzija otiska prsta 4096. Skup podataka koji sadrži slike otisaka prstiju može da ima i više od milion uzoraka. Slika 1 se može naći na veb-stranici https://www.fer.unizg.hr/_download/repository/projekt.pdf



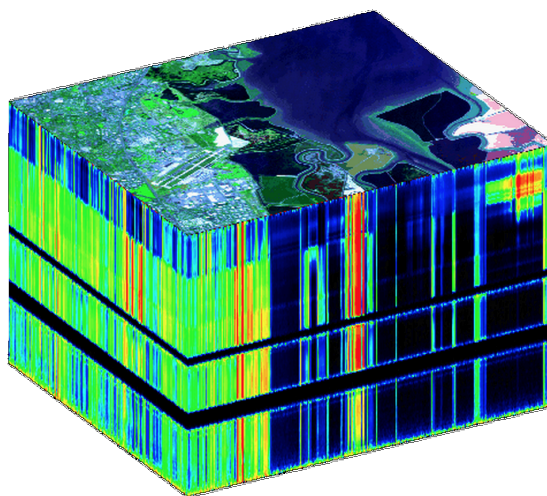
Slika 1: Pet kategorija otisaka prstiju.

[2] Razmotrimo sljedeći primjer u prepoznavanju lica. Baze podataka sadrže veliki broj slika ljudskih lica iste rezolucije, recimo 256×256 . Ako ne bismo imali tehnike za redukciju dimenzije sistemi za prepoznavanje lica morali bi da procesiraju realne vektore dimenzije 65536 koje dobijamo određenim transformacijama sa slika lica. Dimenzija je prevelika da bismo mogli da vršimo pretraživanje pomoću standardnih sistema za prepoznavanje. Stoga, ovo je još jedan primjer gdje je neophodna redukcija dimenzije. Osim toga, slike ljudi mogu da budu napravljene pod različitim uslovima osvjetljenja, u različitim pozama ili sa različitim izrazima lica. U bazama podataka sve slike imaju istu dimenziju (takozvana rezolucija). Na slici ispod imamo primjer slika koje su izdvojene iz neke baze podataka koja sadrži ljudska lica. Ovaj primjer, kao i mnogi drugi mogu se naći na veb-stranici <http://www.cs.nyu.edu/~roweis/data.html>. Na slici imamo 25 likova od 5 različitih ljudi u različitim pozama i različitim izrazima lica. Rezolucija svake slike je 64×64 .



Slika 2: Olliveti likovi.

[3] Hiperspektralni senzori se koriste da se sakupe geološki ili geografski podaci, tako što snimaju objekte u širokom dijelu elektromagnetnog spektra. Pomoću hiperspektralnog snimanja moguće je registrovati ultraljubičasto i infracrveno zračenje koje nije vidljivo ljudskom oku. Hiperspektralni senzori bilježe informacije u obliku skupa snimaka. Svaki snimak predstavlja dio elektromagnetnog spektra, koji se naziva spektralni kanal. Ovi snimci se zatim kombinuju, formirajući trodimenzionalnu hiperspektralnu kocku ili HSI (hyperspectral imagery cube), koja služi za procesiranje i analizu. Procesiranje ovih snimaka omogućava pretraživanje i korišćenje svih informacija koje oni sadrže. Više o hiperspektralnog snimanju, kao i Slika 3, može se naći na veb-stranici <http://large.stanford.edu/courses/2015/ph240/islam1/>. Na Slici 3 svaka vertikalna linija predstavlja spektralni vektor, a svaka horizontalna linija je spektralni kanal. Za analiziranje hiperspektralnih podataka potrebni su brzi računari i veliki kapaciteti za skladištenje analiziranih podataka, tako da bi obrada ovakvog skupa podataka bila skoro nemoguća bez primjene tehnika za redukciju dimenzionalnosti.



Slika 3: Hiperspektralna kocka.

Glava 1

Osnovni pojmovi linearne algebre i statistike

U ovoj glavi navešćemo osnovne pojmove na kojima su zasnovane tehnike za redukciju dimenzionalnosti. Prvo ćemo navesti pojmove linearne algebre koji se tiču vektora i matrica, matricnih reprezentacija, karakterističnih korijena i vektora. Zatim ćemo navesti osnovne pojmove vjerovatnoće i statistike (vjerovatnoća, slučajne promjenljive i njihova statistička karakterizacija, slučajni vektori, numeričke karakteristike slučajnih promjenljivih itd.), da bismo na samom kraju naveli osnovne pojmove multivarijacione analize koja se bavi višedimenzionalnim podacima. Većina dokaza navedenih tvrđenja u ovoj glavi je izostavljena, a više se može naći u [7–9, 12–14].

1.1 Osnovni pojmovi linearne algebre

1.1.1 Vektori i matrice

Neka je $\mathbb{R}^n$ skup n -dimenzionalnih realnih vektora. Za svako $x \in \mathbb{R}^n$ sa x_i ćemo označavati i -tu koordinatu odnosno i -tu komponentu vektora. Vektore u $\mathbb{R}^n$ ćemo posmatrati kao kolone, osim ako drugačije ne naglasimo. Sa x' ćemo označavati *transponovani* vektor vektora x , tj. n -dimenzionalni vektor vrste. Za proizvoljna dva vektora $x, y \in \mathbb{R}^n$ definišemo njihov *skalarni proizvod* $\langle x, y \rangle$ sa $\langle x, y \rangle = x'y = y'x = \sum_{i=1}^n x_i y_i$. Vektori x, y su *ortogonalni* ako je $\langle x, y \rangle = 0$, a *ortonormirani* ako su jedinični i ortogonalni. Ako imamo vektor $x \in \mathbb{R}^n$ onda oznaka $x > 0$ znači da je svaka komponenta vektora $x_i > 0$ za $i \in \{1, \dots, n\}$. Ako imamo dva vektora onda $x > y$ znači da je $x - y > 0$. Analogno definišemo $x \geq 0$, $x < 0$ i $x \leq 0$.

Matricu A sa elementima a_{ij} ćemo označavati sa $A = [a_{ij}]$. *Transponovana* matrica A' matrice A je definisana sa $A' = [a_{ji}]$. Dakle, operacija transponovanja mijenja mjesta vrstama i kolonama. Za dvije matrice A, B kompatibilnih dimenzija važi da je $(AB)' = B'A'$.

Neka je A kvadratna matrica. Kažemo da je A *simetrična* ako je $A' = A$. Matrica A je *dijagonalna* ako je $a_{ij} = 0$ za $i \neq j$. A je *tridijagonalna* ako je $a_{ij} = 0$ za sve i, j za koje je $|i - j| > 1$. A je *donja trougaona* ako je $a_{ij} = 0$ za $i < j$ i slično matrica A je *gornja trougaona* ako je $a_{ij} = 0$ za $i > j$. Kažemo da je matrica A *ortogonalna* ako su sve vrste i sve kolone matrice A jedinični ortogonalni vektori, tj. $A'A = AA' = I$. Jediničnu matricu označavaćemo sa I .

Podskup $S \subset \mathbb{R}^n$ je *potprostor* od $\mathbb{R}^n$ ako $ax + by \in S$ za sve $x, y \in S$ i sve $a, b \in \mathbb{R}$. Neka je dat neki konačan skup vektora $\{x_1, \dots, x_k\} \in \mathbb{R}^n$. Tada definišemo *linearni omotač* - *lineal* skupa F , u oznaci $\text{span}(F)$ kao skup svih *linearnih kombinacija* vektora iz F , tj. skup svih vektora y oblika $y = \sum_{i=1}^k a_i x_i$ gdje su a_i skalari iz $\mathbb{R}$. Kažemo da su vektori $x_1, \dots, x_k \in \mathbb{R}^n$ *linearno nezavisni* ako za proizvoljne skalare $a_1, \dots, a_k$ važi implikacija $\sum_{i=1}^k a_i x_i = 0 \Rightarrow a_i = 0$ za svako $i = 1, \dots, k$. Neka je S potprostor od $\mathbb{R}^n$, tada je *baza* za S skup svih linearno nezavisnih vektora čiji je lineal baš

S . Svaka baza datog potprostora S ima isti broj vektora i taj broj zovemo *dimenzija*. Svaki konačno dimenzionalan vektorski prostor sa skalarnim proizvodom ima ortogonalnu (ortonormiranu) bazu.

1.1.2 Vektorske i matrične norme

1.1.1 Definicija Norma na $\mathbb{R}^n$ je preslikavanje koje svakom vektoru $x \in \mathbb{R}^n$ dodjeuje skalar $\|x\|$ sa sljedećim osobinama:

1. $\|x\| \geq 0$, za sve $x \in \mathbb{R}^n$;
2. $\|x\| = 0$ akko je $x = 0$;
3. $\|\lambda x\| = |\lambda| \cdot \|x\|$, za sve $\lambda \in \mathbb{R}$ i sve $x \in \mathbb{R}^n$;
4. $\|x + y\| \leq \|x\| + \|y\|$, za sve $x, y \in \mathbb{R}^n$.

Norma može biti zadana na različite načine. U ovom radu, osim ako nije drugačije naglašeno, radićemo sa *euklidskom normom* definisanom sa

$$\|x\| = \sqrt{x'x} = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}.$$

Prostor sa euklidskom normom zovemo *Euklidski prostor*. Euklidska norma zadovoljava *Švarcovu*¹ nejednakost, tj. važi

$$|x'y| \leq \|x\| \cdot \|y\|,$$

za sve vektore x, y . Jednakost važi akko je $x = \lambda y$ za neki skalar λ . Spomenućemo još dvije važne norme, a to su *norma beskonačno* (supremum norma ili maksimum norma ili ℓ_∞), koja je data sa

$$\|x\|_\infty = \max_i |x_i|,$$

i ℓ_1 norma koja je definisana sa

$$\|x\|_1 = \sum_{i=1}^n |x_i|.$$

Norma na skupu $n \times n$ matrica je preslikavanje koje svakoj matrici A dodjeljuje realan broj $\|A\|$ sa istim osobinama koje ima i vektorska norma kada matricu A gledamo kao element prostora $\mathbb{R}^{n^2}$. Posebno su interesantne *indukovane (operatorske) norme*. Ako je data vektorska norma $\|\cdot\|$ onda je indukovana norma koju označavamo sa na isti način $\|\cdot\|$ definisana sa

$$\|A\| := \max_{\|x\|=1} \|Ax\|.$$

Na osnovu Švarcove nejednakosti imamo da je

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=\|y\|=1} |y'Ax|.$$

Obrnemo li uloge x i y dobijamo iz prethodne jednakosti koristeći da je $y'Ax = x'A'y$ imamo da je $\|A\| = \|A'\|$.

Kao što smo imali vektorsku normu 1 i normu ∞ imamo i indukovane matrične norme, koje označavamo na isti način. Tačnije, imamo

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| \quad \text{i} \quad \|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

¹Hermann A. Schwarz (1843-1921) - njemački matematičar

Očigledno je da važi $\|A\|_1 = \|A'\|_\infty$. Sve smo diskutovali za kvadratne matrice, ali isto važi i za matrice formata $m \times n$. Navedimo još jednu bitnu matricnu normu koju ćemo često koristiti. To je *Frobenijusova*² *norma* i za realnu matricu formata $m \times n$ je definisana sa

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\text{tr}(A'A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2},$$

gdje su σ_i karakteristične vrijednosti matrice A , a sa $\text{tr}(A'A)$ označili smo *trag* matrice $A'A$.

1.1.2 Definicija Neka je data matrica $A = [a_{ij}]_{n \times n}$. **Trag** matrice A u oznaci $\text{tr } A$ je suma elemenata na dijagonali, tj. $\text{tr } A = \sum_{i=1}^n a_{ii}$.

1.1.3 Propozicija Neka je $\|\cdot\|$ indukovana norma na skupu svih $n \times n$ matrica. Tada važi:

1. $\|A\| = \max_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}$;
2. $\|Ax\| \leq \|A\| \cdot \|x\|$ za sve $x \in \mathbb{R}^n$;
3. $\|AB\| \leq \|A\| \cdot \|B\|$ za matrice A, B , $n \times n$ matrice.

1.1.4 Definicija Kvadratna matrica A je **singularna** ako je njena determinanta jednaka nuli. U suprotnom kažemo da je matrica A **regularna**.

1.1.5 Teorema

1. Neka je data kvadratna matrica formata $n \times n$. Tada su sljedeća tvrđenja ekvivalentna:

- (a) Matrica A je regularna.
- (b) Matrica A' je regularna.
- (c) Za svaki nenula vektor $x \in \mathbb{R}^n$ važi da je $Ax \neq 0$.
- (d) Za svaki vektor $y \in \mathbb{R}^n$ postoji jedinstven vektor $x \in \mathbb{R}^n$ takav da je $Ax = y$.
- (e) Postoji kvadratna matrica B formata $n \times n$ takva da je $AB = I = BA$.
- (f) Kolone matrice A su linearno nezavisni vektori.
- (g) Vrstice matrice A su linearno nezavisni vektori.

2. Ako je A regularna matrica, onda matrica B za koju važi da je $AB = BA = I$ je jedinstvena i zovemo je **inverzna matrica** matrice A , a označavamo sa $B = A^{-1}$. Ako je matrica A realna matrica, onda je i A^{-1} realna matrica.

3. Za bilo koje dvije kvadratne invertibilne matrice A i B istog reda važi $(AB)^{-1} = B^{-1}A^{-1}$.

1.1.6 Definicija **Karakteristični polinom** $n \times n$ matrice A je definisan sa $\phi(\lambda) = \det(\lambda I - A)$, gdje je I jedinična matrica iste dimenzije kao i matrica A . n korijena polinoma ϕ zovemo **karakteristične vrijednosti** matrice A . Vektor x (koji može da ima i kompleksne koordinate) za koji važi da je $Ax = \lambda x$, gdje je λ karakteristična vrijednost matrice A , zovemo **karakteristični vektor** koji odgovara korijenu λ . Skup svih karakterističnih vrijednosti matrice A nazivamo **spektar** matrice A .

1.1.7 Propozicija Trag matrice A jednak je sumi karakterističnih vrijednosti matrice A pri čemu višestruke korijene brojimo onoliko puta kolika im je višestrukost.

²Ferdinand G. Frobenius (1849-1917) - njemački matematičar

1.1.3 Simetrične i pozitivno definitne matrice

Simetrične matrice imaju neka svojstva koja će biti od velikog značaja u daljem radu. One su u velikoj vezi sa prethodno navedenim karakterističnim vrijednostima i vektorima.

1.1.8 Propozicija *Neka je A simetrična matrica reda $n \times n$. Tada:*

1. *Karakteristične vrijednosti matrice A su realne;*
2. *Matrica A ima n realnih, nenula, međusobno ortogonalnih karakterističnih vektora $x_1, \dots, x_n$ koji odgovaraju karakterističnim vrijednostima $\lambda_1, \dots, \lambda_n$.*
3. *Pretpostavimo da su vektori u 2. normalizovani tako da je $\|x_i\| = 1$ za svako i . Tada je*

$$A = \sum_{i=1}^n \lambda_i x_i x_i'.$$

*Ovakav razvoj simetrične matrice pomoću njenih karakterističnih vrijednosti i vektora naziva se **spektralna dekompozicija**.*

1.1.9 Primjer Neka je data matrica

$$A = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix}$$

tada je $\det(\lambda I - A) = \lambda^2 - 5\lambda + 6.16 - 0.16 = \lambda^2 - 5\lambda + 6 = (\lambda - 2)(\lambda - 3)$, pa su karakteristični vektori matrice A baš $\lambda_1 = 2$ i $\lambda_2 = 3$. Odgovarajući karakteristični vektori su $e'_1 = [1/\sqrt{5}, 2/\sqrt{5}]$ i $e'_2 = [2/\sqrt{5}, -1/\sqrt{5}]$, redom. Tada

$$A = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix} = 3 \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} + 2 \begin{bmatrix} \frac{2}{\sqrt{5}} \\ -1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} 0.6 & 1.2 \\ 1.2 & 2.4 \end{bmatrix} + \begin{bmatrix} 1.6 & -0.8 \\ -0.8 & 0.4 \end{bmatrix}.$$

1.1.10 Definicija *Kvadratna realna, simetrična matrica A dimenzije $n \times n$ je **pozitivno definitna** ako je A realna i $x'Ax > 0$ za sve $x \in \mathbb{R}^n, x \neq 0$. Kažemo da je A **pozitivno semidefinitna** ako je realna i $x'Ax \geq 0$ za sve $x \in \mathbb{R}^n$.*

1.1.11 Propozicija

1. *Za svaku realnu matricu A , matrica $A'A$ je simetrična i pozitivno semidefinitna. $A'A$ je pozitivno definitna ako je A regularna.*
2. *Simetrična realna kvadratna matrica je pozitivno semidefinitna (resp. pozitivno definitna) ako su sve karakteristične vrijednosti nenegativne (resp. pozitivne).*
3. *Inverzna matrica simetrične pozitivno definitne matrice je simetrična i pozitivno definitna.*

1.1.12 Propozicija *Neka je A kvadratna simetrična pozitivno semidefinitna matrica.*

1. *Tada postoji matrica $A^{1/2}$ sa osobinom da je $A^{1/2}A^{1/2} = A$, koju zovemo **kvadratni korijen matrice A** .*
2. *Matrica $A^{1/2}$ je regularna ako je A regularna. Njen inverz označavamo sa $A^{-1/2}$.*
3. *Važi da je $A^{-1/2}A^{-1/2} = A^{-1}$.*
4. *Važi da je $AA^{1/2} = A^{1/2}A$.*

Dokaz. 1. Neka su $\lambda_1, \dots, \lambda_n$ karakteristične vrijednosti matrice A i neka su $x_1, \dots, x_n$ odgo-
varajući nenula, međusobno ortogonalni karakteristični vektori koji su normalizovani tako da je
 $\|x_i\| = 1$ za sve i . Neka je

$$A^{1/2} = \sum_{i=1}^n \lambda_i^{1/2} x_i x_i',$$

gdje je $\lambda_i^{1/2}$ nenegativan kvadratni korijen od λ_i . Tada imamo

$$A^{1/2} A^{1/2} = \sum_{k=1}^n \sum_{i=1}^n \lambda_k^{1/2} \lambda_i^{1/2} x_k x_k' x_i x_i' = \sum_{i=1}^n \lambda_i x_i x_i' = A.$$

Druga jednakost slijedi iz ortogonalnosti različitih karakterističnih vektora. Poslednja jednakost
slijedi iz Propozicije 1.1.8.

2. Ovo slijedi iz činjenice da su karakteristične vrijednosti matrice A kvadrati karakterističnih
vrijednosti matrice $A^{1/2}$.

3. Lako vidimo da je $(A^{-1/2} A^{-1/2}) A = A^{-1/2} (A^{-1/2} A^{1/2}) A^{1/2} = A^{-1/2} E A^{1/2} = E$.

4. Slično prethodnom imamo $AA^{1/2} = A^{1/2} A^{1/2} A^{1/2} = A^{1/2} A$. Time smo dokazali tvrđenje. $\square$

1.1.13 Napomena Simetrični kvadratni korijen matrice A nije jedinstven. Npr. znamo da
 $A^{1/2} A^{1/2} = A$, ali i za matricu $-A^{1/2}$ važi da je $(-A^{1/2})(-A^{1/2}) = A$. Ali, ako je A pozitivno
definitna matrica onda se može pokazati da je $A^{1/2}$ jedini simetričan pozitivno definitan kvadratni
korijen matrice A .

1.1.4 Dekompozicija singularnim vrijednostima (SVD)

Ideja koju smo koristili kod spektralne dekompozicije može da se proširi do dekompozicije pravo-
ugaonih matrica. Koristićemo takozvanu SVD dekompoziciju odnosno dekompoziciju singularnim
vrijednostima.

1.1.14 Teorema Neka je A realna matrica formata $m \times n$. Tada postoje ortogonalne matrice V ,
formata $m \times m$ i U formata $n \times n$ takve da važi

$$A = V \Lambda U'$$

gdje je $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{\min\{m,n\}})$ pri čemu je $\lambda_i \geq 0$ za $i = 1, 2, \dots, \min\{m, n\}$. Brojeve
 λ_i zovemo **singularnim vrijednostima** matrice A . Kolone matrice $V = [v_1, \dots, v_m]$ zovemo
lijevim, a kolone matrice $U = [u_1, \dots, u_n]$ **desnim singularnim vektorima** matrice A .

Dekompozicija singularnim vrijednostima može da se izrazi kao razvoj matrice koji zavisi od
ranga r matrice A . Posebno, postoji r pozitivnih konstanti $\lambda_1, \dots, \lambda_r$, r ortogonalnih $m \times 1$ jedi-
ničnih vektora $v_1, \dots, v_r$ i r ortogonalnih $n \times 1$ jediničnih vektora $u_1, \dots, u_r$, takvih da važi

$$A = \sum_{i=1}^r \lambda_i v_i u_i' = V_r \Lambda_r U_r',$$

gdje su $V_r = [v_1, \dots, v_r]$, $U_r = [u_1, \dots, u_r]$ i Λ_r je $r \times r$ dijagonalna matrica čiji su elementima λ_i
na dijagonali. Ovdje su λ_i^2 (u_i) karakteristični korijeni (vektori) matrice AA' pa imamo da je

$$AA' v_i = \lambda_i^2 v_i,$$

gdje su $\lambda_1^2, \dots, \lambda_r^2 > 0$ i $\lambda_{r+1}^2 = \lambda_{r+2}^2 = \dots = \lambda_m^2 = 0$ za $m > n$. Tada $u_i = \lambda_i^{-1} A' v_i$. S druge
strane, u_i su karakteristični vektori matrice $A'A$ sa istim nenula karakterističnim vrijednostima λ_i^2 .

Dakle, sada možemo predstaviti SVD dekompoziciju sa

$$A_{[m \times n]} = V_{[m \times m]} \Lambda_{[m \times n]} U'_{[n \times n]}$$

gdje V ima m ortogonalnih karakterističnih vektora matrice AA' kao svoje kolone, a matrica U ima n ortogonalnih karakterističnih vektora matrice $A'A$, a Λ smo već opisali u prethodnoj teoremi.

1.1.15 Primjer Neka je data matrica

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}.$$

Tada je

$$AA' = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}.$$

Odatle slijedi da je $\det(\gamma E - AA') = \gamma^2 - 22\gamma + 120 = (\gamma - 12)(\gamma - 10)$, pa su karakteristični korijeni matrice AA' upravo $\gamma_1 = \lambda_1^2 = 12$ i $\gamma_2 = \lambda_2^2 = 10$. Odgovarajući karakteristični vektori su $v'_1 = [1/\sqrt{2}, 1/\sqrt{2}]$ i $v'_2 = [1/\sqrt{2}, -1/\sqrt{2}]$, redom. Slično je

$$A'A = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}.$$

Ovdje imamo da je $\det(\gamma E - A'A) = \gamma^3 - 22\gamma^2 + 120\gamma = \gamma(\gamma - 12)(\gamma - 10)$. Stoga su odgovarajući karakteristični korijeni $\gamma_1 = \lambda_1^2 = 12$, $\gamma_2 = \lambda_2^2 = 10$ i $\gamma_3 = \lambda_3^2 = 0$. Vidimo da se nenula karakteristične vrijednosti poklapaju sa karakterističnim vrijednostima matrice AA^T . Lako dobijamo da su odgovarajući karakteristični vektori jednaki $u'_1 = [1/\sqrt{6}, 2/\sqrt{6}, 1/\sqrt{6}]$, $u'_2 = [2/\sqrt{5}, -1/\sqrt{5}, 0]$ i $u'_3 = [1/\sqrt{30}, 2/\sqrt{30}, -5/\sqrt{30}]$. Uzmemo li sada da je $\lambda_1 = \sqrt{12}$ i $\lambda_2 = \sqrt{10}$ dobijamo da je SVD dekompozicija matrice data sa

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \sqrt{12} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{\sqrt{2}}{1} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} + \sqrt{10} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{\sqrt{2}}{-1} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \end{bmatrix}.$$

1.1.5 Projekcija na potprostor

Neka je $\mathcal{B} = \{b_1, \dots, b_k\}$ ortonormirana baza k -dimenzionalnog potprostora $S \subset \mathbb{R}^m$. Skup $\mathcal{B}$ možemo predstaviti u obliku matrice $B = [b_1, \dots, b_k]$ čiji su elementi ortogonalni vektori kolone. Ortogonalna projekcija $P: \mathbb{R}^m \rightarrow S$ može se zapisati u obliku

$$y = P(x) = BB'x, \quad x \in \mathbb{R}^m.$$

Za vektor $y \in S$ je $y = P(y) = BB'y$. S obzirom da je dimenzija potprostora S jednaka k , onda postoji izometrični izomorfizam između S i prostora $\mathbb{R}^k$, i matrica B predstavlja ortogonalno potapanje iz $\mathbb{R}^k$ u $\mathbb{R}^m$ dato sa

$$y = Bz, \quad z \in \mathbb{R}^k.$$

Osim toga imamo da važi

$$d_2(Bz_1, Bz_2) = d_2(z_1, z_2), \quad z_1, z_2 \in \mathbb{R}^k,$$

gdje je $d_2(x, y)$ euklidska metrika (vidjeti Potpoglavlje 3.3.1). Neka je sada dat skup $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^m$ i neka je $S = \text{span}\{x_1, \dots, x_n\}$. Tada je ortogonalna projekcija P data sa

$$y = P(x) = XX^-x$$

gdje je X^- generalizovani inverz matrice X , i $x \in \mathbb{R}^m$. Ako je skup vektora $\mathcal{X}$ linearno zavisn, onda X^- nije jedinstveno određeno, u suprotnom $X^- = (X'X)^{-1}X'$, tj. X^- je lijevi inverz sa X .

Znamo da se svaka matrica X može zapisati u obliku

$$X = QR,$$

tj. postoji njena QR dekompozicija, gdje je matrica $Q = [u_1, \dots, u_k]_{m \times k}$ ortogonalna matrica, a R je $k \times n$ gornja trougaona matrica. Tada je skup $\{u_1, \dots, u_k\}$ ortonormirana baza za $S = \text{span}\{x_1, \dots, x_n\}$. Stoga je QQ' ortogonalna projekcija iz $\mathbb{R}^m$ na S .

Neka je sada $S \subset \mathbb{R}^m$ k -dimenzionalni potprostor i $a \in \mathbb{R}^m$. Tada je *linearna mnogostrukost* (*hiperravan*) definisana kao

$$H = a + S = \{x + a : x \in S\}.$$

Pored toga, svaka hiperravan H ima jedinstveni paralelni potprostor S koji dobijamo na sljedeći način: Neka je $a \in H$, tada

$$S = H - a = \{x - a : x \in H\}$$

i dimenzija hiperravni je jednaka dimenziji odgovarajućeg paralelnog potprostora.

Za svaki realan broj $c \neq 0$, operator *dilatacije* skupa $\mathcal{X} \subset \mathbb{R}^m$ je definisan sa

$$c\mathcal{X} = \{cx : x \in \mathcal{X}\}.$$

Rotacija skupa $\mathcal{X}$ je skup $\mathcal{X}_O = \{Ox_1, \dots, Ox_n\}$, gdje je O realna ortogonalna $m \times m$ matrica. Lako se vidi da rotacija, translacija i dilatacija ne mijenjaju dimenziju potprostora.

1.2 Osnovni pojmovi vjerovatnoće i statistike

1.2.1 Slučajne promjenljive i slučajni vektori

1.2.1 Definicija *Familija $\mathcal{F}$ podskupova od $\Omega \neq \emptyset$ je σ -algebra ako zadovoljava aksiome:*

1. $\Omega \in \mathcal{F}$;
2. Ako $A \in \mathcal{F}$ onda i komplement $\bar{A} \in \mathcal{F}$;
3. Ako su $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, tada i $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

*Elementarni događaj ω je osnovni pojam koji se ne definiše. Neka je skup Ω skup elementarnih događaja i neka je na njemu data σ -algebra $\mathcal{F}$. Skupove iz σ -algebre zovemo *slučajnim događajima*, skup Ω *siguran događaj*, a $\emptyset$ *nemoguć događaj*. Uređen par $(\Omega, \mathcal{F})$ zovemo *mjerljiv prostor događaja*.*

1.2.2 Definicija *Neka je $(\Omega, \mathcal{F})$ mjerljiv prostor događaja. Skupovna funkcija $P : \mathcal{F} \rightarrow [0, 1]$ zove se **vjerovatnoća** ako ima osobine:*

1. $P(\Omega) = 1$;
2. Ako su $A_1, A_2, \dots$ disjunktni događaji iz $\mathcal{F}$ tada važi $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

*Uređena trojka $(\Omega, \mathcal{F}, P)$ naziva se **prostor vjerovatnoće**.*

1.2.3 Definicija *Preslikavanje $X : \Omega \rightarrow \mathbb{R}$ iz prostora vjerovatnoće $(\Omega, \mathcal{F}, P)$ u mjerljiv prostor $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ naziva se **slučajna promjenljiva** ako inverzna slika svakog Borelovog skupa pripada σ -algebri $\mathcal{F}$, odnosno za proizvoljno $B \in \mathcal{B}(\mathbb{R})$ važi $X^{-1}(B) \in \mathcal{F}$.*

1.2.4 Definicija Funkciju $F_X : \mathbb{R} \rightarrow [0, 1]$ definisanu sa $F_X(x) = P\{\omega \in \Omega : X(\omega) \leq x\}$ nazivamo **funkcija raspodjele slučajne promjenljive X** .

Za naš rad biće prilično važne slučajne promjenljive. Razlikujemo slučajne promjenljive diskretnog i slučajne promjenljive apsolutno neprekidnog tipa. S tim u vezi imamo sljedeću definiciju.

1.2.5 Definicija Slučajna promjenljiva $I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$, za događaj $A \in \mathcal{F}$ naziva se **indikator događaja A** . Ako postoji disjunktno razbijanje sigurnog događaja $\Omega = A_1 \cup A_2 \cup \dots$ i ako postoji najviše prebrojiv skup $\{x_1, x_2, \dots\} \subset \mathbb{R}$ tada slučajnu promjenljivu oblika

$$X(\omega) = \sum_{i=1}^{\infty} x_i I_{A_i}(\omega) \quad (1.1)$$

zovemo **diskretna slučajna promjenljiva**.

1.2.6 Definicija Slučajna promjenljiva X je **apsolutno neprekidnog tipa** ako postoji nenegativna integrabilna funkcija $\varphi_X(x) : \mathbb{R} \rightarrow \mathbb{R}_+$ takva da za svaki skup $B \in \mathcal{B}(\mathbb{R})$ važi

$$P\{X \in B\} = \int_B \varphi_X(x) dx. \quad (1.2)$$

Funkcija $\varphi_X(x)$ zove se **gustina raspodjele slučajne promjenljive X** .

1.2.7 Definicija Preslikavanje $X : \Omega \rightarrow \mathbb{R}^n$ iz prostora vjerovatnoće $(\Omega, \mathcal{F}, P)$ u mjerljiv prostor $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ naziva se **slučajan vektor** (n -dimenzionalna slučajna promjenljiva), ako inverzna slika svakog Borelovog skupa pripada σ -algebri $\mathcal{F}$, tj. za svako $B \in \mathcal{B}(\mathbb{R}^n)$ važi $X^{-1}(B) \in \mathcal{F}$.

Za nas će u daljem radu biti važna sljedeća teorema.

1.2.8 Teorema Preslikavanje $X : \Omega \rightarrow \mathbb{R}^n$ čija su komponentna preslikavanja $X = (X_1, \dots, X_n)$ je slučajan vektor ako i samo ako je svako $X_i, i = 1, \dots, n$ jednodimenzionalna slučajna promjenljiva.

1.2.9 Definicija Slučajne promjenljive $X_1, \dots, X_n$ su **nezavisne** ako za svako $k \leq n$, za svaki izbor indeksa $(i_1, \dots, i_k)$ i za proizvoljne Borelove skupove $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ važi

$$P\left(\bigcap_{j=1}^k \{X_{i_j} \in B_j\}\right) = \prod_{j=1}^k P\{X_{i_j} \in B_j\}.$$

1.2.10 Teorema Slučajne promjenljive $X_1, \dots, X_n$ na prostoru vjerovatnoće $(\Omega, \mathcal{F}, P)$ su nezavisne ako i samo ako je

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

1.2.2 Numeričke karakteristike slučajnih promjenljivih

U ovom radu često će biti važni pojedini parametri koji će dobro okarakterisati osobine raspodjele vjerovatnoća slučajne promjenljive. Nama su važni pokazatelji centralne tendencije, specijalno *očekivanje slučajne promjenljive* i pokazatelji rasipanja, specijalno disperzija i *standardna devijacija slučajne promjenljive*.

1.2.11 Definicija *Matematičko očekivanje* diskretne slučajne promjenljive X date sa (1.1) definišemo sa

$$E(X) = \sum_{k=1}^{\infty} x_k p(x_k),$$

gdje smo koristili oznaku $p(x_k) = P(A_k) = P\{X = x_k\}$, $k = 1, \dots, n$. Očekivanje postoji ako i samo ako je $\sum_{k=1}^{\infty} |x_k| p(x_k) < \infty$.

Za apsolutno neprekidnu slučajnu promjenljivu X sa gustom $\varphi_X(x)$ datu sa (1.2) očekivanje je dato sa

$$E(X) = \int_{-\infty}^{\infty} x \varphi_X(x) dx,$$

i ono postoji ako integral apsolutno konvergira, odnosno ako je $\int_{-\infty}^{\infty} |x| \varphi_X(x) dx < \infty$. U multivarijacionoj analizi češće umjesto $E(X)$ pišemo $\mu(X)$.

1.2.12 Teorema Osobine matematičkog očekivanja su:

1. $E(c) = c$, gdje je $c \in \mathbb{R}$ proizvoljna konstanta;
2. $E(cX) = cE(X)$, gdje je $c \in \mathbb{R}$ proizvoljna konstanta;
3. $E(X + Y) = E(X) + E(Y)$.

1.2.13 Definicija Neka je X slučajna promjenljiva. **Disperzija** (varijansa) slučajne promjenljive X , u oznaci $D(X)$, $\sigma^2(X)$ ili $\text{Var}(X)$ definisana je sa

$$D(X) = E((X - E(X))^2) = D(X) = E(X^2) - E^2(X).$$

1.2.14 Definicija Očekivanje slučajnog vektora $X = (X_1, \dots, X_n)$, ukoliko postoji, je definisano sa

$$E(X) = E(X_1, \dots, X_n) = (E(X_1), \dots, E(X_n)).$$

Disperzija slučajnog vektora $X = (X_1, \dots, X_n)$, ukoliko postoji, je definisana sa

$$D(X) = D(X_1, \dots, X_n) = (D(X_1), \dots, D(X_n)).$$

1.2.15 Definicija **Kovarijansa** slučajne promjenljive (X, Y) je broj $\text{cov}(X, Y)$ definisan sa

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

Očigledno, $\text{cov}(X, X) = D(X)$. **Koeficijent korelacije** slučajne promjenljive (X, Y) je broj

$$\rho_{XY} = \frac{E(XY) - E(X)E(Y)}{\sqrt{D(X)D(Y)}} = \frac{\text{cov}(X, Y)}{\sqrt{D(X)D(Y)}}.$$

Kažemo da su slučajne promjenljive X i Y **nekorelirane** ako je $\rho_{XY} = 0$.

Uvešćemo još sljedeće oznake. Ako imamo slučajni vektor $X = (X_1, \dots, X_n)$ onda za proizvoljan par slučajnih promjenljivih X_i, X_j označavaćemo njihovu kovarijansu sa $\sigma_{ij} = \text{cov}(X_i, X_j)$. Pri tome važi kao i do sad $\text{cov}(X_i, X_i) = D(X_i) = \sigma_{ii} = \sigma_i^2$ i $\sigma_{ij} = \text{cov}(X_i, X_j) = \text{cov}(X_j, X_i) = \sigma_{ji}$. U daljem radu izuzetno će biti važna *slučajna matrica*. Kolekciju slučajnih vektora možemo zapisati u vidu matrice. Takvu matricu čiji su elementi (jednodimenzionalne) slučajne promjenljive nazivamo **slučajna matrica**. Neka je $X = [X_{ij}]$ slučajna matrica formata $n \times p$. Tada očekivanje slučajne matrice definišemo sa

$$E(X) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \dots & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & \dots & E(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \dots & E(X_{np}) \end{bmatrix}. \quad (1.3)$$

Disperziju slučajne matrice definišemo analogno. Ako su X i Y dvije slučajne matrice istih dimenzija, onda sve osobine koje smo definisali za očekivanje jednodimenzionalnih slučajnih promjenljivih važe i ovdje. Npr. $E(X + Y) = E(X) + E(Y)$ i ako su A, B dvije realne matrice onda je $E(AXB) = AE(X)B$ itd.

1.2.16 Definicija Neka je dat slučajan vektor $X = (X_1, \dots, X_n)$. **Kovarijaciona matrica** vektora X je matrica data sa $\Sigma = [\text{cov}(X_i, X_j)]_{n \times n}$. Očigledno, Σ je simetrična matrica.

1.2.17 Teorema Matrica Σ je kovarijaciona matrica nekog slučajnog vektora ako i samo ako je simetrična i pozitivno semidefinitna.

Definišimo sada matricu uzajamnih proizvoda odstupanja elemenata slučajnog vektora od odgovarajućeg očekivanja, tj. ako imamo slučajan vektor $X = (X_1, \dots, X_n)$ sa odgovarajućim očekivanjem $E(X) = (E(X_1), \dots, E(X_n))$ onda je odgovarajuća **matrica odstupanja** data sa $(X - E(X))(X - E(X))'$. Očekivana vrijednost matrice odstupanja je zapravo kovarijaciona matrica, tj. $E((X - E(X))(X - E(X))') = \Sigma$.

1.2.18 Teorema Neka su $X_1, \dots, X_n$ slučajne promjenljive. Tada je

$$D\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j) = \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j).$$

1.2.19 Definicija Neka je dat slučajan vektor $X = (X_1, \dots, X_n)$. **Korelaciona matrica** vektora X je matrica data sa $R = \left[\frac{\text{cov}(X_i, X_j)}{\sqrt{D(X_i)D(X_j)}} \right]_{n \times n}$.

Korelacionu matricu možemo dobiti iz kovarijacione matrice i to kao $R = (D^{1/2})^{-1} \Sigma (D^{1/2})^{-1}$, gdje je matrica D dijagonalna matrica koja na glavnoj dijagonali sadrži elemente sa glavne dijagonale kovarijacione matrice Σ . Matricu D zovemo **matrica standardnih devijacija**. Slično, kovarijacionu matricu možemo izraziti preko korelacione kao $\Sigma = D^{1/2} R D^{1/2}$.

1.2.3 Osnovni pojmovi statistike i multivarijacione analize

Statistika se bavi prikupljanjem, prikazivanjem, interpretacijom, prezentacijom i korišćenjem podataka kako bismo izveli određene zaključke. Multivarijaciona analiza predstavlja skup statističkih metoda pomoću kojih se analiziraju višedimenzionalna mjerenja dobijena za svaku jedinicu nekog skupa objekata koje posmatramo i ispituje.

Osnovni skup, tj. skup na kome proučavamo određene pojave, zovemo **populacija**. Osnovne osobine po kojima se razlikuju elementi neke populacije zovemo **obilježja**. Osnovni problem je da odredimo raspodjelu obilježja na nekoj populaciji. U praksi je veoma teško ispitivati cijelu populaciju, pa se uzima neki podskup populacije koji zovemo **uzorak**.

U skladu sa prethodnom pričom populaciju možemo posmatrati kao skup svih mogućih ishoda, tj. u prethodnoj notaciji skup Ω , a svako obilježje X kao slučajnu promjenljivu, tj. $X : \Omega \rightarrow \mathbb{R}$. Nakon što izaberemo uzorak obima n na svakom elementu populacije posmatramo obilježje X . To sakupljanje podataka prilikom nekog mjerenja zovemo eksperiment. Niz dobijenih vrijednosti $(x_1, \dots, x_n)$ zovemo **realizovan uzorak**. Prirodno je posmatrati sljedeće slučajne promjenljive: X_1 je vrijednost obilježja na prvom elementu populacije koji smo izabrali u uzorak, $\dots$, X_n je vrijednost obilježja na posljednjem elementu populacije koji smo izabrali u uzorak. Svaka od promjenljivih ima istu raspodjelu kao i osnovno obilježje X i sve slučajne promjenljive $X_1, \dots, X_n$ su nezavisne. **Prost slučajan uzorak** obima n je n -dimenzionalna slučajna promjenljiva $(X_1, X_2, \dots, X_n)$ u kojoj sve promenljive X_i , $i \in \mathbb{N}$, imaju istu raspodjelu kao i obilježje X i međusobno su nezavisne.

Na osnovu uzorka $(X_1, \dots, X_n)$ potrebno je ocijeniti očekivanje, odnosno **srednju vrijednost** (sredinu), i disperziju. Koristimo sljedeće veličine:

1. *aritmetička sredina uzorka (uzoračka srednja vrijednost)* $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$;
2. *uzoračka disperzija* $\bar{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2$;
3. *korigovana uzoračka disperzija* $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$;
4. *standardna devijacija uzorka* $\bar{S}_n = \sqrt{\bar{S}_n^2}$;
5. Ako su dati uzorci $(X_1, \dots, X_n)$ obilježja X i $(Y_1, \dots, Y_n)$ obilježja Y tada je *uzoračka kovarijansa* $C_{XY} = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n$.

Pretpostavimo da smo prilikom eksperimenta sakupili podatke za p osobina (promjenljivih) za nekih n objekata (uzoraka). Dobijene podatke predstavljamo u obliku **matrice podataka** koja ima n vrsta i p kolona. Vrste matrice podataka zovemo *objektima* i oni predstavljaju entitete na kojima smo vršili mjerenje (npr. osobe), a kolone zovemo *promjenljivama* i one predstavljaju skup vrijednosti mjerenih na datom objektu. U praksi je n često mnogo veće od p .

Matricu podataka možemo posmatrati i kao vektor i to na dva načina u zavisnosti da li X posmatramo po vrstama ili po kolonama. Ako posmatramo po vrstama onda svaka vrsta predstavlja jednu višedimenzionalnu opservaciju (zapažanje) slučajnih promjenljivih $X_1, \dots, X_p$. U tom slučaju matrica X sadrži n opservacija višedimenzionalne slučajne promjenljive u oznaci $X_1, \dots, X_n$. Pod vektorom X_i podrazumijevamo $p \times 1$ slučajni vektor opservacija promjenljivih $X_1, \dots, X_p$ za i -tu jedinicu posmatranja.

Ako matricu podataka posmatramo po kolonama, tada svaka kolona predstavlja n opservacija jedne od p promjenljivih, odnosno svaka kolona matrice X je $n \times 1$ vektor opservacija jedne od p promjenljivih koji označavamo sa $X_1, \dots, X_p$ i $X = [X_1, \dots, X_p]$, gdje je

$$X_j = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix}, \quad j = 1, 2, \dots, p.$$

Neka imamo matricu podataka formata $n \times p$, tj. imamo n objekata i p promjenljivih. Uzoračka sredina u oznaci $\bar{X}$ definiše se kao $(p \times 1)$ slučajan vektor čiji su elementi uzoračke sredine promjenljivih, tj.

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_n \end{bmatrix} \quad \text{gdje je } \bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}, \quad j = 1, 2, \dots, p.$$

Definišimo **matricu centriranih podataka** u oznaci X^* sa

$$X^* = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 & \dots & X_{1p} - \bar{X}_p \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 & \dots & X_{2p} - \bar{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \bar{X}_1 & X_{n2} - \bar{X}_2 & \dots & X_{np} - \bar{X}_p \end{bmatrix} \quad (1.4)$$

jer pomoću nje možemo definisati uzoračku kovarijacionu matricu kao simetričnu slučajnu matricu $p \times p$ datu sa

$$S = \frac{1}{n-1} (X^*)' X^* = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})',$$

gdje je $(X^*)'X^*$ simetrična slučajna matrica

$$(X^*)'X^* = \begin{bmatrix} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) & \dots & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{i1} - \bar{X}_1) & \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 & \dots & \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{ip} - \bar{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i1} - \bar{X}_1) & \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i2} - \bar{X}_2) & \dots & \sum_{i=1}^n (X_{ip} - \bar{X}_p)^2 \end{bmatrix}.$$

Elementi uzoračke kovarijacione matrice su dati sa

$$S_{jk} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k), \quad j, k = 1, 2, \dots, p.$$

Sada uzoračku korelacionu matricu možemo odrediti pomoću elemenata uzoračke kovarijacione matrice. Njeni elementi su uzorački koeficijenti korelacije

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}}, \quad j, k = 1, 2, \dots, p.$$

Sama korelaciona matrica je data sa

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{12} & 1 & r_{23} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{1p} & r_{2p} & r_{3p} & \dots & 1 \end{bmatrix}.$$

1.2.20 Primjer Neka su realizovane vrijednosti slučajnog uzorka od $n = 3$ elementa iz dvodimenzionalnog osnovnog skupa $p = 2$ date matricom podataka

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} -2 & 3 \\ 4 & 2 \\ 1 & 1 \end{bmatrix}.$$

Odredimo realizovanu vrijednost sredine $\bar{X}$. Tada je

$$\bar{x} = \begin{bmatrix} \frac{1}{n} \sum_{j=1}^n x_{j1} \\ \frac{1}{n} \sum_{j=1}^n x_{j2} \end{bmatrix} = \begin{bmatrix} \frac{-2+4+1}{3} \\ \frac{3+2+1}{3} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Oдавде čitamo da je uzoračka srednja vrijednost (sredina) prve promjenljive $\bar{x}_1 = 1$ i da je sredina druge promjenljive $\bar{x}_2 = 2$. Matrica centriranih podataka je

$$X^* = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 \\ X_{31} - \bar{X}_1 & X_{32} - \bar{X}_2 \end{bmatrix} = \begin{bmatrix} -2-1 & 3-2 \\ 4-1 & 2-2 \\ 1-1 & 1-2 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix}.$$

Zatim lako dobijamo da je matrica uzajamnih proizvoda centriranih podataka

$$(X^*)'X^* = \begin{bmatrix} -3 & 1 \\ 3 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -3 & 3 & 0 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 18 & -3 \\ -3 & 2 \end{bmatrix}.$$

Odatle imamo da je realizovana vrijednost uzoračke kovarijacione matrice data sa

$$S = \frac{1}{n-1}(X^*)'X^* = \frac{1}{3-1} \begin{bmatrix} 18 & -3 \\ -3 & 2 \end{bmatrix} = \begin{bmatrix} 9 & -3/2 \\ -3/2 & 1 \end{bmatrix},$$

odakle vidimo da je disperzija (varijansa) prve promjenljive $s_1^2 = s_{11} = 9$, druge $s_2^2 = s_{22} = 1$, a njihova kovarijansa je $s_{12} = s_{21} = -3/2$. Pošto smo rekli da je veza između kovarijacione i korelacione matrice $R = (D^{1/2})^{-1}S(D^{1/2})^{-1}$ imamo da je za dijagonalnu matricu

$$D = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$$

realizovana vrijednost uzoračke korelacione matrice

$$R = \begin{bmatrix} 1/3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 9 & -3/2 \\ -3/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}.$$

Prethodni primjer je od velike važnosti. S obzirom da ćemo u ovom radu posmatrati višedimenzionalne centrirane podatke, vidimo da to nije neko ograničenje, jer se proizvoljan skup podataka može centrirati. Osim toga vidjeli smo kako se računa vrijednost uzoračke kovarijacione i uzoračke korelacione matrice.

Glava 2

Osnovni pojmovi diferencijalne geometrije i teorije grafova

Ovu glavu smo namjerno odvojili od Glave 1, s obzirom da ćemo ovdje navesti pojmove koji će se više koristiti kod nelinearnih tehnika za redukciju dimenzionalnosti, a koje ćemo razmatrati u Glavi 4, dok se linearne tehnike za redukciju dimenzionalnosti, kao što ćemo vidjeti, baziraju na multivarijacionoj analizi i statistici. U ovoj glavi navešćemo osnovne pojmove diferencijalne geometrije kao i teorije grafova. Za literaturu iz diferencijalne geometrije izabrali smo [4, 25], dok smo za teoriju grafova koristili [3, 25].

2.1 Osnovni pojmovi diferencijalne geometrije

Prilikom redukcije dimenzionalnosti, pretpostavljamo da se podaci nalaze na nižedimenzionalnoj mnogostrukosti čija dimenzija i oblik nisu poznati. Mnogostrukost je moguće predstaviti pomoću njenih koordinata u euklidskom prostoru, tako da tehnike za redukciju dimenzionalnosti traže koordinate za elemente skupa podataka. Mi se ovdje nećemo baviti globalnim svojstvima mnogostrukosti, dovoljno će nam biti lokalne osobine mnogostrukosti, jer će nam koordinatne karte, o kojima ćemo kasnije više reći, biti dovoljne da lokalno opišemo podatke. S obzirom da je većina dokaza iz ove oblasti prilično složena, ovdje ćemo navesti potrebna tvrđenja bez dokaza, sa referencama na literaturu gdje se izostavljeni dokazi mogu naći.

Diferencijabilne mnogostrukosti predstavljaju uopštenja glatkih krivih i površi.

2.1.1 Definicija *Neka su $X \subset \mathbb{R}^n$ i $Y \subset \mathbb{R}^m$ dva neprazna skupa. Preslikavanje $f : X \rightarrow Y$ je **difeomorfizam** ako je f bijekcija i f i f^{-1} su glatka preslikavanja¹. Ako postoji difeomorfizam između X i Y , kažemo da su X i Y **difeomorfni**.*

Intuitivno, mnogostrukosti su prostori koji su lokalno euklidske strukture, tj. svaka tačka sa mnogostrukosti ima okolinu koja je lokalno difeomorfna otvorenom skupu u euklidskom prostoru $\mathbb{R}^n$. Klasična definicija mnogostrukosti može se naći u [10], dok ćemo ovdje navesti definiciju koja će biti dovoljna za rad sa nelinearnim tehnikama.

2.1.2 Definicija *Neka je $M \subset \mathbb{R}^m$ neprazan skup. Ako za svaku tačku $x \in M$ postoji otvoren skup $W \subset M$ takav da je W difeomorfno otvorenom skupu $U \subset \mathbb{R}^k$, tada kažemo da je M **k -dimenzionalna diferencijabilna mnogostrukost** (glatka mnogostrukost), difeomorfizam $g : U \rightarrow W$ je **parametrizacija** za W , inverzno preslikavanje $h := g^{-1} : W \rightarrow U$ zovemo **koordinanto preslikavanje**, dok W zovemo **koordinatna okolina** tačke x . Uređen par (W, h) zovemo **koordinatni sistem** ili **karta** na mnogostrukosti M .*

¹Preslikavanje f je *glatko* (eng. smooth) ako svi parcijalni izvodi postoje i neprekidne su funkcije, tj. $f \in C^\infty$.

Na karti (W, h) tačka $x \in W$ ima sljedeću parametrizaciju:

$$h(x) = [h^1(x), \dots, h^m(x)],$$

pri čemu h^i zovemo i -ta koordinatna funkcija. Dakle, mnogostrukost se sastoji od karata gdje je svaka karta euklidski prostor.

2.1.3 Definicija *Diferencijabilna struktura tzv. atlas $\mathcal{D}$ na k -dimenzionalnoj mnogostrukosti M je kolekcija svih karata $\{(W_i, h_i)\}$ na M , tako da važe uslovi:*

1. *Unija svih W_i prekriva M , tj. $M = \cup_i W_i$;*
2. *Za svaki par (i, j) preslikavanje $h_j \circ h_i^{-1}$ je glatko preslikavanje na $h_i(W_i \cap W_j)$.*

2.1.1 Tangentni prostor i tangentni vektori

Neka je $M \subset \mathbb{R}^m$ mnogostrukost. U tački $p \in M$ možemo konstruisati hiperravan koja će najbolje aproksimirati M u okolini tačke p . Intuitivno, tangentni prostor T_p će biti potprostor paralelan sa hiperravni H .

Podsjetimo se, ako je $f = [f_1, \dots, f_l]' : U \rightarrow V$ diferencijabilno preslikavanje, gdje su $U \subset \mathbb{R}^k$ i $V \subset \mathbb{R}^l$ otvoreni skupovi, onda izvod preslikavanja f u tački $x \in U$ je linearno preslikavanje $df_x : \mathbb{R}^k \rightarrow \mathbb{R}^l$ definisano sa

$$df_x = \lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t}, \quad h \in \mathbb{R}^k,$$

ili u matricnom obliku

$$df_x = \left[\frac{\partial f_i(x)}{\partial x_j} \right]_{i,j=1}^{l,k} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_l(x)}{\partial x_1} & \dots & \frac{\partial f_l(x)}{\partial x_k} \end{bmatrix}.$$

Sada dolazimo do definicije tangentnog prostora.

2.1.4 Definicija *Neka je $M \subset \mathbb{R}^m$ k -dimenzionalna mnogostrukost, $U \subset \mathbb{R}^k$ otvoren skup i $g : U \rightarrow M$ parametrizacija okoline $g(U) \subset M$. Neka je $p \in M$, $u \in U$ i $g(u) = p$. Tada je sliku linearne transformacije d_{g_u} , u oznaci $T_p M := d_{g_u} \mathbb{R}^k$, zovemo **tangentni prostor** za M u tački p . Vektore iz $T_p M$ zovemo **tangentni vektori**.*

Pisaćemo

$$\frac{\partial g}{\partial u^i} = \left[\frac{\partial g_1}{\partial u^i}, \dots, \frac{\partial g_m}{\partial u^i} \right]', \quad 1 \leq i \leq k,$$

pa koristeći prethodnu definiciju imamo da je skup

$$\left\{ \frac{\partial g}{\partial u^1}, \dots, \frac{\partial g}{\partial u^k} \right\} \tag{2.1}$$

baza tangentnog prostora $T_p M$. Ako definišemo

$$\frac{\partial g}{\partial u} := \left[\frac{\partial g_i}{\partial u^j} \right]_{i,j=1}^{m,k} = \begin{bmatrix} \frac{\partial g_1}{\partial u^1} & \dots & \frac{\partial g_1}{\partial u^k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial u^1} & \dots & \frac{\partial g_m}{\partial u^k} \end{bmatrix}, \tag{2.2}$$

tada je tangentni vektor u tački p oblika

$$X_p = \sum_{i=1}^k \alpha_i \frac{\partial g}{\partial u^i} = \frac{\partial g}{\partial u} \alpha$$

gdje je $\alpha = [\alpha_1, \dots, \alpha_k]' \in \mathbb{R}^k$.

2.1.2 Rimanove mnogostrukosti

Kako bi mnogostrukost bila metrički prostor, potrebno je definisati metriku. Kao i do sad, nećemo davati opštu definiciju Rimanove² mnogostrukosti, nego samo onu koja će nama biti dovoljna za rad sa nelinearnim tehnikama.

Neka je M k -dimenzionalna mnogostrukost i T_p tangentni prostor u tački $p \in M$. Neka je g parametrizacija za M i neka je (W, h) karta na M , gdje je $h = g^{-1}$. Za sve $p \in M$ neka je $u = h(p) \in \mathbb{R}^k$, tada g definiše bazu (2.1) za $T_p M$ koju možemo predstaviti u obliku matrice (2.2). Tada definišemo metriku $G_g(p)$ sa

$$G_g(p) = \left(\frac{\partial g}{\partial u} \right)' \frac{\partial g}{\partial u}, \quad p \in M. \quad (2.3)$$

Matrica $G_g(p)$ je pozitivno definitna i naziva se **Rimanova metrika** na M . Ako su $X, Y \in T_p(M)$ dva tangentna vektora data sa $X = \frac{\partial g}{\partial u} x$ i $Y = \frac{\partial g}{\partial u} y$, tada definišemo njihov skalarni proizvod sa

$$\langle X, Y \rangle_p = x' G_g(p) y. \quad (2.4)$$

2.1.5 Lema Za proizvoljne tangentne vektore $X, Y, Z \in T_p$ i realne brojeve α, β , važi sljedeće:

1. *Linearnost*: $\langle \alpha X + \beta Y, Z \rangle_p = \alpha \langle X, Z \rangle_p + \beta \langle Y, Z \rangle_p$;
2. *Simetričnost*: $\langle X, Y \rangle_p = \langle Y, X \rangle_p$;
3. *Pozitivna definitnost*: $\langle X, X \rangle_p \geq 0$.

Skalarni proizvod definiše normu vektora X sa

$$\|X\|_p^2 = x' G_g(p) x.$$

2.1.6 Definicija Neka je M k -dimenzionalna mnogostrukost. Ako je na M definisana metrika $G_g(p)$ i ako je za svaki par tangentnih vektora skalarni proizvod definisan sa (2.4), kažemo da je M **Rimanova mnogostrukost** u oznaci (M, G) .

U ovom radu korišćićemo uvijek diferencijabilne Rimanove mnogostrukosti, tako da to nećemo posebno naglašavati.

2.1.7 Definicija Neka je M povezana Rimanova mnogostrukost i $\gamma \subset M$ glatka kriva na M koja je data parametarski sa $\gamma = [\gamma^1(t), \dots, \gamma^m(t)]'$, $t \in [a, b]$. Dužina luka krive γ je definisana sa

$$\|\gamma\| = \int_a^b \|\dot{\gamma}(t)\| dt = \int_a^b \sqrt{\sum_{i=1}^m (\dot{\gamma}^i)^2(t)} dt,$$

gdje je $\dot{\gamma}$ izvod po t krive γ .

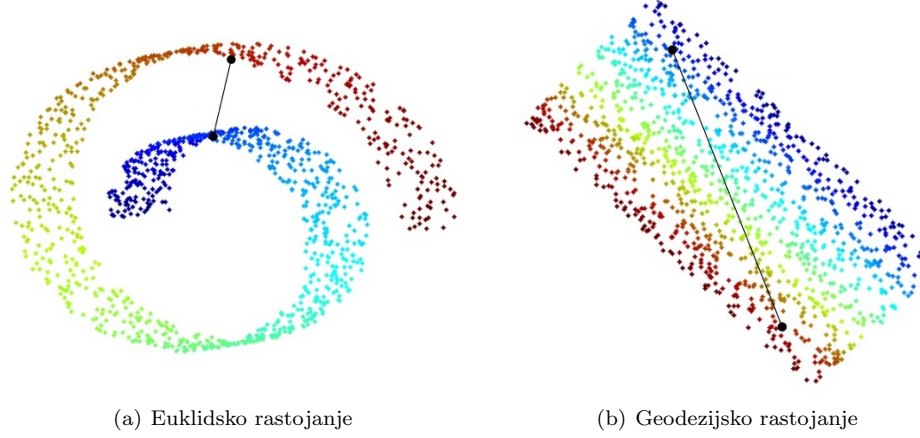
Prethodna definicija je dobra, tj. dužina luka krive ne zavisi od izbora parametrizacije krive γ . Reći ćemo da je kriva parametrizovana *dužinom luka* ili prirodno parametrizovana, ako je $\|\dot{\gamma}(t)\| = 1$ za sve $t \in [a, b]$. Bitno je spomenuti, da svaku regularnu krivu možemo parametrizovati dužinom luka.

2.1.8 Definicija Neka su date tačke $p, q \in M$. **Geodezijsko rastojanje** (geodezijska metrika) između tačaka p i q definisano je sa

$$d_g(p, q) = \inf \{\|\gamma\|\}$$

gdje se infimum dostiže po svim glatkim krivama koje povezuju tačke p i q .

²Bernhard Riemann (1826-1866) - njemački matematičar



Slika 2.1: Euklidsko i geodezijsko rastojanje na primjeru rolne.

2.2 Osnovni pojmovi teorije grafova

2.2.1 Sistem okolina

Standardan način da predstavimo eksperimentalne podatke (informacije) je pomoću *oblaka podataka* koji je definisan kao skup vektora $\mathcal{X} = \{x_k\}_{k=1}^n \subset \mathbb{R}^D$ ili u matičnom obliku kao $X = [x_1, \dots, x_n]_{D \times n}$. Redukcija dimenzionalnosti je bazirana na pretpostavci da uočeni podaci $\mathcal{X} = \{x_1, \dots, x_n\} \in \mathbb{R}^D$ leže na nižedimenzionalnoj d -dimenzionalnoj mnogostrukosti $M \subset \mathbb{R}^D$, $d \ll D$. U većini slučajeva, skupovi podataka su konačni, tj. radimo sa diskretnim skupovima tačaka, pri čemu ne znamo eksplicitno mnogostrukost na kojoj se nalaze podaci. Stoga, potrebno je nekako da opišemo sličnosti između skupova podataka, što se matematički vrši uvođenjem *sistema okolina*.

2.2.1 Definicija Neka je $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ dati skup podataka i neka je dato $\varepsilon > 0$. Tada je ε -okolina tačke $x_i \in \mathcal{X}$ podskup od $\mathcal{X}$ definisan sa

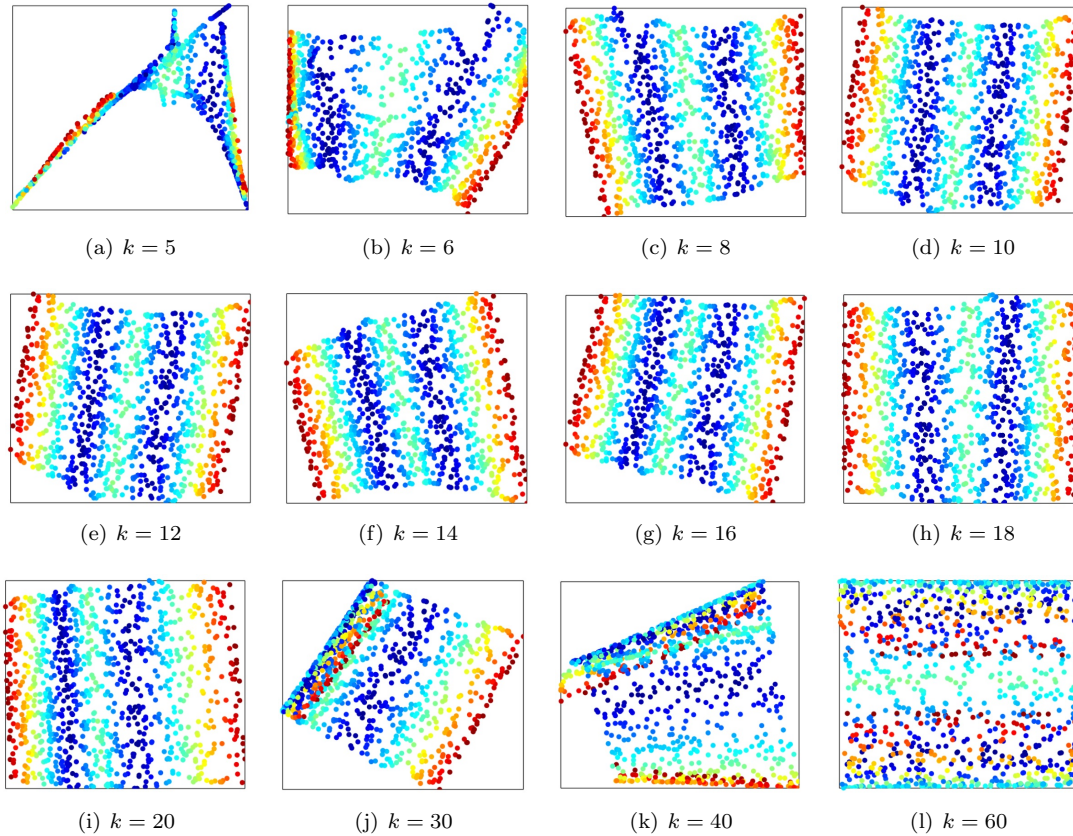
$$O_i = O_i(\varepsilon) = \mathcal{X} \cap (B_\varepsilon(x_i) \setminus \{x_i\})$$

gdje je $B_\varepsilon(x_i) \subset \mathbb{R}^D$ otvorena lopta sa centrom u x_i i poluprečnikom ε .

Iz prethodne definicije lako vidimo da ako $x_i \in O_j$ onda i $x_j \in O_i$. Ako je moguće okolinu svake tačke projektovati na s -dimenzionalnu mnogostrukost, tada je unutrašnja dimenzija skupa podataka s , pa je moguće smanjiti dimenziju podataka do dimenzije s , bez da izgubimo veliku količinu podataka. Ako bismo izabrali malu okolinu, tj. ako izaberemo malo ε onda ćemo izgubiti određene osobine podataka, dok s druge strane ako izaberemo veliko ε dobijamo veliku unutrašnju dimenziju, pa projekcija na mnogostrukost može da bude nestabilan postupak (primjer za ovo može se naći u [25]). Stoga, kako je u praksi teško definisati ε -okolinu skupa podataka, često se koristi k -okolina kako bi se opisale sličnosti, odnosno različitosti između objekata skupa podataka. k -okolina tačke $x_i \in \mathcal{X}$ sadrži k najbližih tačaka tačke x_i , ne uključujući tačku x_i . Označimo sa $N_i = N_k(i)$ k -okolinu tačke x_i . Tada,

$$\max_{x \in N(i)} d_2(x_i, x) \leq \max_{x \notin N(i) \cup \{x_i\}} d_2(x_i, x).$$

Primjetimo da ovdje ne važi da ako $x_i \in N_j$ onda $x_j \in N_i$. Kao i kod ε -okoline i za k -okolinu važi da nije zgodno izabrati ni veliko, ni malo k . To smo predstavili na primjeru na Slici 2.2, koji se može naći u [17]. U [25] mogu se naći algoritmi sa Matlab kodovima za konstrukciju sistema okolina.



Slika 2.2: Primjer primjene LLE metoda (vidjeti Poglavlje 4.2) za različite k -okoline.

2.2.2 Graf kao model za opis strukture podataka

Geometrija diskretne strukture podataka, opisana je sistemom okolina, koji se može predstaviti koristeći grafove. Graf je apstraktni matematički objekat, koji predstavljamo pomoću tačaka, koje predstavljaju podatke, povezanih linijama. Formalno, imamo sljedeću definiciju. Slike koje smo koristili u ovom poglavlju mogu se naći na veb-stranici: <http://poincare.matf.bg.ac.rs/~jelenagr/ASP/grafovi.htm>.

Neorijentisan graf

2.2.2 Definicija Neka je V konačan skup. **(Neorijentisan) graf** G je uređen par $G = [V, E]$, gdje je E skup dvoelementnih podskupova (neuređenih parova) od V . Elemente skupa V zovemo **čvorovi grafa** G , a elementi skupa E su **grane grafa** G .

Neka su $a, b \in V$ dva čvora. Granu između čvorova a i b označavaćemo sa (a, b) i u tom slučaju kažemo da su a i b povezani. Ako je (a, b) isto što i (b, a) onda kažemo da je graf **neorijentisan**. **Petlju**, tj. granu koja počinje i završava u istom čvoru označavaćemo sa (a, a) . **Stepen čvora** a u grafu je broj grana grafa koji imaju kraj u tom čvoru i označavamo ga sa d_a . Petlju brojimo dva puta. Ako ne naglasimo drugačije smatraćemo da nemamo petlji, tj. da imamo *prosti graf*.

2.2.3 Definicija Dva grafa G i H su *izomorfna*, u oznaci $G \sim H$, ako i samo ako postoji bijektivna funkcija čvorova i grana, tako da se očuvava susjednost svih čvorova, tj. dva čvora u G su povezana ako i samo ako su odgovarajući čvorovi u H povezani.

Neka je $G = [V, E]$ graf i $n = |V|$ broj čvorova grafa. Moguće je konstruisati graf $H = [V_H, E_H]$ koji je ekvivalentan sa G i $V_H = \{1, \dots, n\}$. Tada izomorfizam T sa G na H zovemo obilježavanje grafa G jedinstvenim oznakama $1, 2, \dots, n$. Radi jednostavnosti ćemo identifikovati graf $G = [V, E]$ sa ekvivalentnim grafom oznaka H , pa ćemo umjestio $(v_i, v_j) \in E$ pisati $(i, j) \in E$.

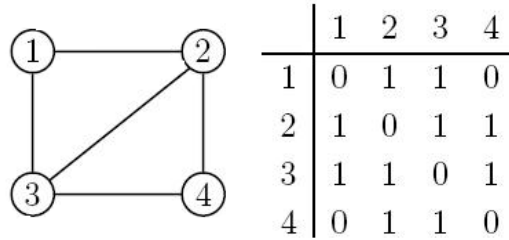
Put između dva čvora u grafu je niz grana koji spaja ova dva čvora, u kome se svaka grana iz grafa pojavljuje najviše jedanput. **Povezan graf** je takav neorijentisani graf kod koga su bilo koja dva čvora povezana putem. Ako postoje dva čvora koja se ne mogu povezati, graf je **nepovezan**.

Kako bismo predstavili čvorove u grafu koristi se **matrica povezanosti** (eng. adjacency matrix). Ako je broj čvorova grafa $G = [V, E]$ jednak n , matrica povezanosti A grafa G je $n \times n$ simetrična matrica definisana sa

$$A(i, j) = \begin{cases} 1, & (i, j) \in E \text{ ili } (j, i) \in E; \\ 0, & \text{inače.} \end{cases} \quad (2.5)$$

Skup čvorova V grafa $G = [V, E]$ je jedinstveno određen pomoću pridružene matrice povezanosti. S obzirom da je za matematička računanja zgodnije raditi sa matricom povezanosti, ubuduće ćemo pisati $G = [V, A]$ umjesto $G = [V, E]$. Na Slici 2.3 dali smo primjer neusmjerenog grafa sa odgovarajućom matricom povezanosti.

Sada ćemo definisati uopštenje matrice povezanosti, a to je **težinska matrica**. U težinskoj matrici, element (i, j) takozvana **težina** (težinski koeficijent) je uglavnom pozitivan realan broj. Težinska matrica $W = [w_{i,j}]_{n \times n}$ grafa $G = [V, E]$ je simetrična matrica gdje je $w_{i,j} \neq 0$ ako $(i, j) \in E$, a inače je $w_{i,j} = 0$. Graf kojem smo dodijelili težinsku matricu zovemo **težinski graf**.



Slika 2.3: Matrica povezanosti neusmjerenog grafa.

Kada imamo skup podataka, koji generiše graf G , uvijek ćemo identifikovati graf $G = [\mathcal{X}, A]$ sa izomorfnim grafom obilježavanja $[V, A]$, gdje je izomorfizam dat sa $x_i \mapsto i$. Okolina čvora $i \in V$ je skup čvorova

$$N(i) = \{j \in V : A(i, j) = 1\},$$

a stepen čvora i je kardinalnost od $N(i)$, tj. $|N(i)|$.

Orijentisan graf (digraf)

2.2.4 Definicija Uređen par $G = [V, E]$ je **usmjeren graf (digraf)** ako je V skup čvorova i E je skup uređenih parova čvorova iz V .

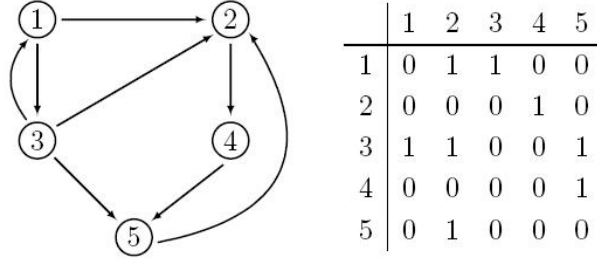
Dok smo kod neorijentisanih grafova imali da je $(a, b) = (b, a)$ kod orijentisanih to nije slučaj, tako razlikujemo (a, b) i (b, a) . Intuitivno, kod neorijentisanih nemamo „strelice” između čvorova, dok kod orijentisanih imamo, odakle je onda jasno, da ako imamo „strelicu” od a ka b , onda ne moramo imati i strelicu od b ka a .

Pojam izomorfizma kao i grafa obilježavanja se uvodi analogno kao i kod neorijentisanog grafa, tako da ćemo odmah preći na matricu povezanosti i težinsku matricu, jer one imaju najveću važnost za nas.

2.2.5 Definicija Neka je $G = [V, E]$ digraf sa $V = \{1, \dots, n\}$. Matrica povezanosti grafa G je $n \times n$ matrica takva da je:

$$A(i, j) = \begin{cases} 1, & (i, j) \in E; \\ 0, & \text{inače} . \end{cases} \quad (2.6)$$

Kao i kod neorijentisanog grafa i ovdje važi da matrica povezanosti orijentisanog grafa jedinstveno određuje graf, s tim što ovdje ona može biti asimetrična matrica (vidjeti primjer na Slici 2.4). Digraf $G = [V, A]$ je simetričan, ako je matrica A simetrična. Očigledno, simetričan digraf je ekvivalentan sa neorijentisanim grafom, stoga je neorijentisan graf specijalan slučaj orijentisanog grafa.



Slika 2.4: Matrica povezanosti usmjerenog grafa.

2.2.6 Definicija Neka je $G = [V, E]$ orijentisan graf sa $V = \{1, \dots, n\}$. Težinska matrica $W := [W_{i,j}]$ na G je definisana sa

$$W_{i,j} = \begin{cases} w_{i,j}, & A(i, j) = 1; \\ 0, & \text{inače} \end{cases} \quad (2.7)$$

pri čemu je $w_{i,j} > 0$ težina na čvoru (i, j) .

2.2.3 Laplasijan grafa

2.2.7 Definicija Neka je $G = [V, A]$ graf, gdje je V skup svih čvorova, kardinalnosti $|V| = n$, i A matrica povezanosti grafa G . **Laplas³-ova matrica** L_G grafa G je data sa

$$L_G(i, j) = \begin{cases} -1, & A(i, j) = 1; \\ d_i, & i = j; \\ 0, & \text{inače} , \end{cases} \quad (2.8)$$

gdje je d_i stepen i -tog čvora.

Označimo sa D dijagonalnu matricu takvu da je $D_{i,i} = d_i$. Tada se Laplasova matrica u (2.8) može zapisati kao

$$L_G = L = D - A.$$

Češće se koristi normalizovana Laplasova matrica definisana sa:

$$\mathcal{L}(i, j) = \begin{cases} \frac{-1}{\sqrt{d_i d_j}}, & A(i, j) = 1, \\ 1, & i = j \text{ i } d_i \neq 0; \\ 0, & \text{inače} , \end{cases} \quad (2.9)$$

³Pierre-Simon Laplace (1745-1827) - francuski matematičar, fizičar i astronom

Imamo da važi:

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2},$$

pri čemu se definiše da je $D^{-1}(i, i) = 0$ ako je $d_i = 0$. Po definiciji stepena čvora, imamo da je $d_i = 0$ ako je i -ti čvor izolovani čvor, tj. nema grana koje ga povezuju sa ostalim čvorovima. Sljedeću teoremu navodimo bez dokaza, koji se može naći u [25].

2.2.8 Teorema *Za Laplasovu matricu $\mathcal{L}$ grafa $G = [V, E]$ važi da postoji matrica S takva da je $\mathcal{L} = S' S$.*

Sada ćemo da uvedemo pojam Laplasove matrice za težinske grafove.

2.2.9 Definicija *Neka je $G = [W, E]$ težinski graf (moguće je da ima i petlje). **Težinska Laplasova matrica** L na G je definisana sa*

$$L(i, j) = \begin{cases} -w_{i,j}, & (i, j) \in E, i \neq j, \\ d_i - w_{i,i}, & i = j, \\ 0, & \text{inače,} \end{cases} \quad (2.10)$$

gdje je $d_i = \sum_{j \in N(i) \cup \{i\}} w_{i,j} > 0$.

Poseban značaj teorije grafova kao i Laplasijana grafa vidjeće se u Poglavlju 4.3, prilikom formulacije LLE metoda.

Glava 3

Linearne tehnike za redukciju dimenzionalnosti signala

U ovoj glavi bavićemo se tehnikama za redukciju, odnosno smanjenje dimenzionalnosti signala, tačnije linearnim tehnikama za redukciju dimenzionalnosti signala. Na samom početku razjasnićemo šta predstavlja pojam *dimenzionalnosti*, a zatim kako i na koji način se vrši smanjenje te dimenzionalnosti. Razmatraćemo različite tehnike, koje grubo možemo podijeliti u dvije grupe i to na *linearne* i *nelinearne* tehnike za redukciju dimenzionalnosti. Potreba za ovim tehnikama je velika, tako da se može naći obimna literatura iz ove oblasti, a u ovom radu najviše smo se bazirali na [6, 25]. Ova glava je posvećena linearnim tehnikama, a u Glavi 4 posvetićemo se nelinearnim tehnikama.

3.1 Problem dimenzionalnosti

Problem dimenzionalnosti (eng. curse of dimensionality) formulisao je 1961. godine Bellman¹ prilikom rješavanja problema iz dinamičke optimizacije. Bellman je primjetio sljedeće: Mreža sa korakom $1/10$ u jediničnoj kocki u $\mathbb{R}^D$ sastoji se od 10^D tačaka, što raste eksponencijalno kada se dimenzija prostora D povećava. U poslednjih nekoliko godina veoma mnogo se radi na razvoju tehnika i alata koji će moći da riješe problem sa nevjerovatnim povećanjem količine podataka iz raznih oblasti nauke i tehnike. Bitno je da imamo mogućnost efikasnijeg skladištenja velike količine podataka, ali istovremeno javlja se problem sa eksponencijalnim porastom količine podataka. Stoga, potrebno je da iskoristimo različite matematičke alate u kombinaciji sa alatima numeričke analize. Značajan doprinos se pojavio pod nazivom *redukcija dimenzionalnosti signala* sa ciljem da se analiziraju višedimenzionalni skupovi podataka. Koriste se analitički i geometrijski pristup ovim problemima koji se mogu smatrati kao dodatak klasičnim statističkim metodama koje se koriste u mašinskom učenju i analizi podataka.

U praksi, velika dimenzija nekog skupa podataka je samo prividno tako velika. Pretpostavimo da se podaci $\mathcal{X} = \{x_\alpha\}_A$ nalaze na s -dimenzionalnoj mnogostrukosti M koja je smještena u $\mathbb{R}^D$, tj. $M \subset \mathbb{R}^D$. Tada kažemo da je D *spoljašnja dimenzija*, a s *unutrašnja dimenzija* skupa podataka $\mathcal{X}$. Sa stanovišta statistike, $\mathcal{X}$ možemo posmatrati kao skup uzoraka slučajnog vektora $X \in \mathbb{R}^D$. Ako je vektor X određen sa s nezavisnih promjenljivih, tj. postoji slučajan vektor $Y \in \mathbb{R}^s$ i invertibilna analitička funkcija $f : \mathbb{R}^s \rightarrow \mathbb{R}^D$ tako da je $F(Y) = X$, tada kažemo da vektor X ima *unutrašnju dimenziju* s . Mala unutrašnja dimenzija višedimenzionalnog skupa podataka je ključna za primjenu redukcije dimenzionalnosti. Zahvaljujući maloj unutrašnjoj dimenziji, možemo smanjiti spoljašnju dimenziju bez da izgubimo važne informacije. Intuitivno, redukcija dimenzionalnosti ima za cilj da pronađe parametrizaciju mnogostrukosti na kojoj se nalaze posmatrani podaci.

¹Richard E. Bellman (1920-1984) - američki matematičar

U opštem slučaju problem *redukcije dimenzionalnosti* je sljedeći: Dat je višedimenzionalni skup podataka $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$, cilj je da pronađemo nižedimenzionalnu reprezentaciju $\mathcal{Y} = \{y_1, \dots, y_n\} \subset \mathbb{R}^d$ gdje je $d \ll D$, ali tako da se analiza podataka skupa $\mathcal{X}$ može izvršiti analizom podataka skupa $\mathcal{Y}$ sa određenom greškom. Probleme redukcije dimenzionalnosti dijelimo u dvije grupe:

- **Teški problemi redukcije dimenzionalnosti.** Kod problema ovog tipa imamo da se spoljašnja dimenzija skupa podataka kreće od nekoliko stotina do nekoliko stotina hiljada komponenti, pa je potrebno izvršiti jako veliko smanjenje dimenzije. U ovoj kategoriji se nalaze problemi prepoznavanja oblika i klasifikacije uzoraka koji uključuju rad sa slikama (prepoznavanja likova, prepoznavanje znakova (slova, oblika) itd.) i zvukom.
- **Laki problemi redukcije dimenzionalnosti.** U ovoj kategoriji nalaze se problemi, kod kojih spoljašnja dimenzija skupa podataka nije pretjerano velika, tako da ni smanjenje dimenzije nije pretjerano veliko. Mnogi problemi statističke analize u oblastima psihologije, sociologije i drugih društvenih nauka spadaju u ovu grupu problema. Multivarijaciona analiza je ključna za rješavanje problema ovog tipa.

Tehnike koje rješavaju probleme redukcije dimenzionalnosti možemo da klasifikujemo na različite načine. Jedna od podjela je na konveksne i nekonveksne tehnike, ali kao što smo već prethodno naveli ovdje ćemo posmatrati podjelu na linearne i nelinearne tehnike. Više o samoj klasifikaciji i kratak pregled tehnika za redukciju dimenzionalnosti može se naći u [24].

Linearne tehnike za redukciju dimenzionalnosti razvile su se u sklopu statistike i mašinskog učenja i postale su nezaobilazan alat za analizu višedimenzionalnih skupova podataka. Neke od najpoznatijih linearnih tehnika za redukciju dimenzionalnosti signala su: analiza glavnih komponenti (eng. Principal Component Analysis - PCA), multidimenzionalno skaliranje (eng. Multidimensional Scalling - MDS), Fisher-ova linearna diskriminantna analiza (eng. Linear Discriminant Analysis - LDA), kanonička korelaciona analiza (eng. Canonical Correlation Analysis - CCA), latentno semantičko indeksiranje (eng. Latent Semantic Indexing - LSI), slučajna projekcija (eng. Random Projection - RP), analiza nezavisnih komponenti (eng. Independent Component Analysis - ICA) kao i faktorska analiza (Factor Analysis - FA). U poslednjih sto godina, počevši sa Pearsonom i definicijom PCA 1901. godine, pojavio se veoma veliki broj linearnih tehnika za redukciju dimenzije. Više o linearnim tehnikama može se naći u [5]. U ovom radu odlučili smo se za dvije linearne tehnike i to PCA i MDS.

3.2 Analiza glavnih komponenti - PCA

Najpoznatija i najvažnija linearna tehnika za redukciju dimenzionalnosti jeste *Analiza glavnih komponenti - PCA*. Metod je uveo Pearson² 1901. godine prilikom pronalaženja linija i ravni koje najbolje fituju dati skup tačaka nekog prostora.

Postoje dva različita pristupa ovoj tehnici. Jedan je geometrijski, a drugi je statistički, i u ovom radu ćemo razmotriti oba. Neka je početni skup podataka $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ i neka je $\mathcal{Y}$ skup dimenzije d dobijen od skupa $\mathcal{X}$ prilikom redukcije dimenzije. U odnosu na euklidsko rastojanje, PCA traži linearnu projekciju $T : \mathbb{R}^D \rightarrow \mathbb{R}^d$ tako da $\mathcal{Y} = T(\mathcal{X})$ maksimizira energiju skupa podataka. Šta to u stvari znači vidjećemo u sljedećem poglavlju.

3.2.1 Geometrijski pristup PCA

Poznato je da translacija ne mijenja geometriju samog skupa podataka. Neka je $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ dati skup podataka i $a \in \mathbb{R}^D$. Sa $\mathcal{X}_a$ označićemo translaciju za vektor a skupa $\mathcal{X}$, tj.

$$\mathcal{X}_a = \{x - a \in \mathbb{R}^D : x \in \mathcal{X}\}.$$

²Karl Pearson (1875-1936) - engleski matematičar i statističar

S obzirom da je translacija za vektor a skupa $\mathcal{X}$ geometrijski ekvivalentna sa $\mathcal{X}$ onda ćemo kod PCA uvijek posmatrati translaciju za vektor

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (3.1)$$

takozvani *centar*, tj. posmatramo *centrirane podatke* $\hat{\mathcal{X}} = \mathcal{X}_{\bar{x}}$ (vidjeti (1.4)), odnosno izvršimo translaciju u koordinatni početak. Definišimo energiju skupa $\mathcal{X}$ sa

$$\mathcal{E}(\mathcal{X}) = \sum_{i=1}^n \|x_i\|^2 = \|X\|_F^2,$$

gdje je $X = [x_1, \dots, x_n]$ matrica podataka, a $\|X\|_F$ je Frobenijusova norma matrice X . Tada centrirani podaci $\hat{\mathcal{X}}$ imaju minimalnu energiju u odnosu na sve translacije skupa $\mathcal{X}$:

$$\mathcal{E}(\hat{\mathcal{X}}) = \min_{a \in \mathbb{R}^D} \mathcal{E}(\mathcal{X}_a).$$

PCA je metod koji nalazi potprostor gdje su dati podaci najviše koncentrisani. Kako bismo opisali *koncentraciju* skupa $\mathcal{X}$ posmatrajmo sljedeći problem maksimizacije: Duž koje prave projekcija skupa $\mathcal{X}$ ima najveću energiju? Prvo, označimo tu pravu sa S_1 . S obzirom da smo rekli da posmatramo centrirane podatke $\mathcal{X}$, prvo smo translirali skup podataka u koordinatni početak, odakle onda dobijamo da prava S_1 prolazi kroz koordinatni početak, tj. S_1 je jednodimenzionalni potprostor od $\mathbb{R}^D$. Označimo jedinični vektor pravca prave S_1 sa v_1 , ortogonalni operator projekcije sa $T_{v_1} : \mathbb{R}^D \rightarrow S_1$, a skup svih jediničnih vektora pravaca u $\mathbb{R}^D$ je jedinična sfera $\mathbb{S}^{D-1} \subset \mathbb{R}^D$. Tada je vektor v_1 rješenje sljedećeg problema maksimizacije:

$$v_1 = \arg \max_{a \in \mathbb{S}^{D-1}} \mathcal{E}(T_a(\mathcal{X})). \quad (3.2)$$

Označimo sa

$$Y_1 = [y_{1,1}, \dots, y_{1,n}]' = T_{v_1}(\mathcal{X}) = [T_{v_1}(x_1), \dots, T_{v_1}(x_n)]' \in \mathbb{R}^n.$$

Kažemo da je $v_1 \in \mathbb{S}^{D-1}$ *prvi glavni (principalni) vektor* (eng. first principal direction), a Y_1 je *prva glavna (principalna) komponenta* (eng. first principal component) skupa $\mathcal{X}$. Matematičkom indukcijom sada definišemo sukcesivno *glavne vektore* i *glavne komponente* skupa $\mathcal{X}$ na sljedeći način: Pretpostavimo da je prvih $s-1$ glavnih vektora $\mathcal{V}_{s-1} = \{v_1, \dots, v_{s-1}\} \subset \mathbb{S}^{D-1}$ i odgovarajućih $s-1$ glavnih komponenti $\{Y_1, \dots, Y_{s-1}\} \subset \mathbb{R}^n$, za $Y_j = [y_{j,1}, \dots, y_{j,n}]'$, $j = 1, \dots, s-1$, skupa $\mathcal{X}$ dobro definisano, gdje je $\mathcal{V}_{s-1}$ po konstrukciji ortonormiran skup vektora u $\mathbb{R}^D$. Definišimo:

$$\mathcal{X}_{s-1} = \left\{ x_i - \sum_{j=1}^{s-1} v_j y_{j,i}, \quad 1 \leq i \leq n \right\}$$

i

$$S_{s-1} = \text{span}(v_1, \dots, v_{s-1}), \quad s = 2, 3, \dots$$

Tada s -ti glavni vektor definišemo sa

$$v_s = \arg \max_{a \in \mathbb{S}_{s-1}^\perp \cap \mathbb{S}^{D-1}} \mathcal{E}(T_a(\mathcal{X}_{s-1})). \quad (3.3)$$

Dakle, na ovaj način smo *koncentraciju* opisali pomoću glavnih vektora i komponenti.

Prethodni problem možemo posmatrati i kao sljedeći problem minimizacije. To je pogodno jer onda možemo iskoristiti određene rezultate iz numeričke optimizacije. Umjesto (3.2) i (3.3) posmatrajmo sljedeće probleme minimizacije:

$$v_1 = \arg \min_{a \in \mathbb{S}^{D-1}} \mathcal{E}((I - T_a)(\mathcal{X})), \quad (3.4)$$

i

$$v_s = \arg \min_{a \in \mathbb{S}_{s-1}^\perp \cap \mathbb{S}^{D-1}} \mathcal{E}((I - T_a)(\mathcal{X}_{s-1})), \quad (3.5)$$

gdje smo sa I označili jediničnu matricu. Sada ćemo iskoristiti indukciju, tj. pretpostavimo da važi hipoteza za d , što dalje implicira:

- $\mathcal{V} = \{v_1, \dots, v_d\}$ je ortonormirana baza d -dimenzionalnog potprostora $S_d \subset \mathbb{R}^D$ i
- Vektor

$$y_i = \sum_{j=1}^d v_j y_{j,i} \quad (3.6)$$

je ortogonalna projekcija vektora x_i na potprostor S_d takav da skup vektora $\mathcal{Y} = \{y_1, \dots, y_n\}$ očuvava maksimalnu energiju skupa $\mathcal{X}$ u odnosu na sve ostale d -dimenzionalne podatke.

Dakle, PCA je metod koji traži d -dimenzionalnu projekciju skupa $\mathcal{X}$ koji maksimizira energiju skupa podataka. Reći ćemo da je $\mathcal{Y}$ redukcija dimenzionalnosti za $\mathcal{X}$. Prethodno dobijene rezultate $\mathcal{V}$ i $\mathcal{Y}$ možemo dobiti i koristeći sljedeću teoremu iz numeričke optimizacije poznatiju kao Teorema Eckart-Young-Mirsky.

3.2.1 Teorema (Eckart-Young-Mirsky 1936.) *Neka je X realna matrica formata $D \times n$, ranga r i neka je njena SVD dekompozicija data sa*

$$X = V \Sigma U' = \sum_{i=1}^r \sigma_i v_i u_i'. \quad (3.7)$$

Neka je

$$X_* = \sum_{i=1}^d \sigma_i v_i u_i', \quad d \leq r,$$

gdje su $U = [u_1, \dots, u_r]_{n \times r}$ i $V = [v_1, \dots, v_r]_{D \times r}$ matrice čije su kolone ortonormirani vektori i $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ pri čemu je $\sigma_1 \geq \sigma_2 \geq \dots \sigma_r > 0$. Tada je realna $D \times n$ matrica X_ , ranga d , najbolja aproksimacija matrice X ranga d u Frobenijusovoj normi, tj.*

$$\|X - X_*\|_F = \min_B \|X - B\|_F$$

gdje je matrica B realna $D \times n$ matrica ranga d , a ocjena greške je data sa

$$\|X - X_*\|_F = \sqrt{\sum_{l=d+1}^r \sigma_l^2}.$$

Dokaz. Neka je B realna $D \times n$ matrica. Iskoristićemo činjenicu da je Frobenijusova norma invarijantna u odnosu na unitarne transformacije, tj. da važi $\|D\|_F = \|VDU'\|_F$. Zaista, kako su matrice U i V unitarne (ortogonalne u realnom slučaju, tj. $UU' = U'U = I$) i kako D i VDU' imaju iste karakteristične vrijednosti, imamo da je

$$\begin{aligned} \|VDU'\|_F^2 &= \text{tr}((VDU')'VDU) = \text{tr}(UD'V'VDU) = \text{tr}(UD'DU') \\ &= \text{tr}(U'UD'D) = \text{tr}(D'D) = \|D\|_F^2. \end{aligned}$$

Definišimo matricu $D = V'BU$. Tada, kako je $V^{-1} = V'$ slijedi da je $B = VDU'$, odakle direktnim računom dobijamo da je

$$\begin{aligned}\|X - B\|_F^2 &= \|V\Sigma U' - VDU'\|_F^2 = \|\Sigma - D\|_F^2 = \sum_{i,j} |\sigma_{i,j} - D_{i,j}|^2 \\ &= \sum_{i=1}^r |\sigma_i - D_{ii}|^2 + \sum_{i>r} |D_{ii}|^2 + \sum_{i \neq j} |D_{i,j}|^2 \\ &\geq \sum_i |\sigma_i - D_{ii}|^2 \geq \sum_{i>d} |\sigma_i - D_{ii}|^2.\end{aligned}$$

Stoga, matrica D ranga d koja minimizira $\|X - B\|_F^2$ je dijagonalna matrica sa nenula dijagonalnim elementima $D_{ii} = \sigma_i$ za $i = 1, \dots, d$. Odgovarajuća matrica B je tada matrica X_* , što je i trebalo dokazati. Ocjena greške onda slijedi direktno. $\square$

Iskoristimo sada prethodnu teoremu kako bismo dokazali sljedeće tvrđenje, koje omogućava da redukciju dimenzionalnosti $\mathcal{Y}$ skupa $\mathcal{X}$ izračunamo direktno iz SVD dekompozicije matrice $\mathcal{X}$.

3.2.2 Teorema *Neka je dat centrirani skup podataka $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ i neka je SVD dekompozicija skupa podataka data sa $X = V\Sigma U' = \sum_{i=1}^r \sigma_i v_i u_i'$, gdje su V, Σ i U kao u Teoremi 3.2.1. Neka je $\mathcal{V}_d = \{v_1, \dots, v_d\} \subset \mathbb{R}^D$, $\Sigma_d = \text{diag}(\sigma_1, \dots, \sigma_d)$ i $U_d = [u_1, \dots, u_d]$. Tada je d glavnih vektora skupa X baš $\mathcal{V}_d$, d odgovarajućih glavnih komponenti su vektori vrste matrice*

$$Y = [y_1, \dots, y_n] = \Sigma_d(U_d)' \quad (3.8)$$

*i redukcija dimenzionalnosti je skup $\mathcal{Y} = \{y_1, \dots, y_n\}$. Matricu Y zovemo **težinska matrica**, a u težinskoj matrici i -ta kolona $y_i = [y_{1i}, \dots, y_{di}]'$ se naziva **težinski vektor**.*

Dokaz. Neka je $r = \text{rang } X$. Bez umanjenja opštosti, pretpostavimo da je $d \leq r$. Za $1 \leq s \leq d$, neka je $V_s = [v_1, \dots, v_s]$, $\Sigma_s = \text{diag}(\sigma_1, \dots, \sigma_s)$, $U_s = [u_1, \dots, u_s]$ i $Y_s = \Sigma_s(U_s)'$. Očigledno da se onda Y_s sastoji od s prvih vektora vrste matrice Y . Tada je matrica $B_s := V_s Y_s$ realna matrica formata $D \times n$, ranga s . Tada na osnovu Teoreme 3.2.1 imamo da je

$$\|X - B_s\|_F = \min_B \|X - B\|_F, \quad 1 \leq s \leq d,$$

što implicira da su $\{v_1, \dots, v_d\} \subset \mathbb{R}^D$ sukcesivni glavni vektori skupa $\mathcal{X}$ i vektori vrste matrice Y su d odgovarajućih glavnih komponenti od $\mathcal{X}$. Time smo pokazali tvrđenje. $\square$

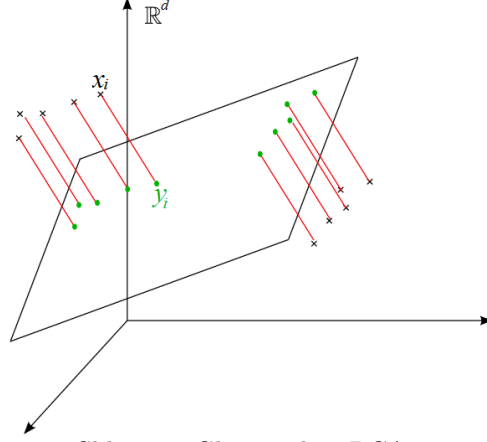
Na Slici 3.1. predstavili smo geometrijski ideju koju smo prethodno iznijeli. Za dati skup tačaka $\{x_i\}_{i=1}^n$ potrebno je da nađemo hiperravan (linearnu mnogostrukost) takvu da su projekcije $\{T(x_i)\}_{i=1}^n = \{y_i\}_{i=1}^n$ određene u skladu sa maksimalnom varijansom ili kao što smo vidjeli, suma rastojanja tačaka i njihovih projekcija je minimalna. Ta tražena hiperravan je određena karakterističnim vektorima kovarijacione matrice S koja može da se dobije kao SVD dekompozicija matrice $X = [x_1, \dots, x_n] \in \mathbb{R}^{D \times n}$.

Zaista, Teorema 3.2.2 omogućava da skup $\mathcal{Y}$ dobijemo iz SVD dekompozicije matrice podataka X koristeći formulu (3.8). U praksi, matrica Y se nesto nalazi računanjem spektralne dekompozicije kovarijacione matrice skupa X . Vidjeli smo da se kovarijaciona matrica može izračunati kao

$$S = \frac{1}{n-1} X X'. \quad (3.9)$$

Neka je spektralna dekompozicija (Propozicija 1.1.8) matrice S data sa

$$S = V \Lambda V'.$$



Slika 3.1: Glavna ideja PCA.

Tada je matrica V ista kao i u (3.7), jer ako je SVD za $X = V\Sigma U'$ onda uvrstimo li to u (3.9) dobijamo

$$S = \frac{1}{n-1} (V\Sigma U')' V\Sigma U' = V \frac{\Sigma \Sigma'}{n-1} V' = V\Lambda V'.$$

Stoga, možemo izdvojiti podmatricu V_d matrice V . S obzirom da su kolone matrice V_d ortonormirana baza prostora koji najbolje aproksimira potprostor S_d , tada je ortonormirana projekcija vektora $x_i \in \mathcal{X}$ na potprostor S_d vektor $(V_d)'x_i$, što implicira da je

$$Y = (V_d)'X. \quad (3.10)$$

U mnogim primjenama broj tačaka n je mnogo veći nego dimenzija D . Stoga, kovarijaciona matrica S ima prilično malu dimenziju, npr. $D \times D$, pa je algoritam za računanje spektralne dekompozicije prilično brz.

3.2.2 Statistički pristup PCA

Sa statističke tačke gledišta, osnovni zadatak metode glavnih komponenti jeste određivanje nekoliko linearnih kombinacija originalnih promjenljivih koje treba da imaju maksimalnu varijansu, da budu međusobno nekorelisane ali da pritom izgube što je moguće manje informacija koje su sadržane u skupu originalnih promjenljivih. Prilikom ovog postupka početne promjenljive transformišu se u nove promjenljive koje nazivamo *glavne komponente*. Prva glavna komponenta konstruisana je tako da obuhvati najveći dio varijanse originalnog skupa podataka, a naredne komponente treba da obuhvate onaj dio varijanse originalnog skupa podataka koji nije obuhvaćen prethodno izdvojenim glavnim komponentama. Cilj je da početni skup podataka opišemo sa što manjim brojem glavnih komponenti.

Neka je $\mathcal{X} \subset \mathbb{R}^D$ dati skup podataka i $X = [x_1, \dots, x_n]_{D \times n}$ njegova matrična reprezentacija. U statistici, prvi glavni vektor slučajne matrice X (sa očekivanjem jednakim 0) definišemo kao vektor u smjeru najveće varijanse. Drugim riječima, imamo sljedeću definiciju.

3.2.3 Definicija Neka je X slučajna matrica formata $D \times n$. Tada je prvi glavni vektor jedinični $D \times 1$ vektor v_1 definisan sa

$$v_1 = \arg \max_{\|w\|=1} \text{Var}(w'X). \quad (3.11)$$

Ako sada raspišemo desnu stranu dobijamo:

$$\begin{aligned} v_1 &= \arg \max_{\|w\|=1} \text{Var}(w'X) \\ &= \arg \max_{\|w\|=1} w'Sw \end{aligned}$$

pri čemu koristimo činjenicu da imamo podatke sa očekivanjem jednakim 0 i S je uzoračka kovarijaciona matrica. Dakle cilj je da nađemo vektor w tako da v_1 ima najveću vrijednost, pri čemu imamo uslov da je $\|w\| = 1$ ili ekvivalentno $w'w = 1$. To ćemo uraditi koristeći Lagranžove množioce, tj. uvodimo novu promjenljivu λ i rješavamo problem:

$$\begin{aligned} \mathcal{L} &= w'Sw - \lambda(w'w - 1) \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= w'w - 1 \\ \frac{\partial \mathcal{L}}{\partial w} &= 2Sw - 2\lambda w. \end{aligned}$$

Nakon što izjednačimo izvode sa 0 dobijamo

$$\begin{aligned} w'w &= 1 \\ Sw &= \lambda w \end{aligned}$$

Dakle, dobijamo da je traženi glavni vektor karakteristični vektor koji odgovara najvećoj karakterističnoj vrijednosti λ kovarijacione matrice S . Odgovarajuća glavna komponenta je definisana kao slučajna promjenljiva data sa

$$P_1 = v_1'X.$$

Analogno, neka imamo da su glavni vektori $v_1, \dots, v_{k-1}$ i odgovarajuće glavne komponente $P_1, \dots, P_{k-1}$ vektora X dobro definisani. Tada k -ti glavni vektor je jedinični vektor određen sa

$$v_k = \arg \max_{\|w\|=1} \text{Var} \left(w' \left(X - \sum_{i=1}^{k-1} v_i P_i \right) \right),$$

a k -ta glavna komponenta je slučajna promjenljiva

$$P_k = v_k'X.$$

Dakle, k karakterističnih vektora kovarijacione matrice S je k glavnih glavnih vektora, za koje onda računamo glavne komponente. Sa stanovišta statistike, PCA metod traži d glavnih komponenti datog skupa podataka X . Pretpostavimo da je $\text{rang}(X) = r$. Neka je SVD dekompozicija matrice podataka X data sa $X = V\Sigma U' = \sum_{i=1}^r \sigma_i v_i u_i'$. Matematičkom indukcijom može se pokazati da d vodećih glavnih vektora skupa X su upravo $v_1, \dots, v_d$ i da su odgovarajuće glavne komponente

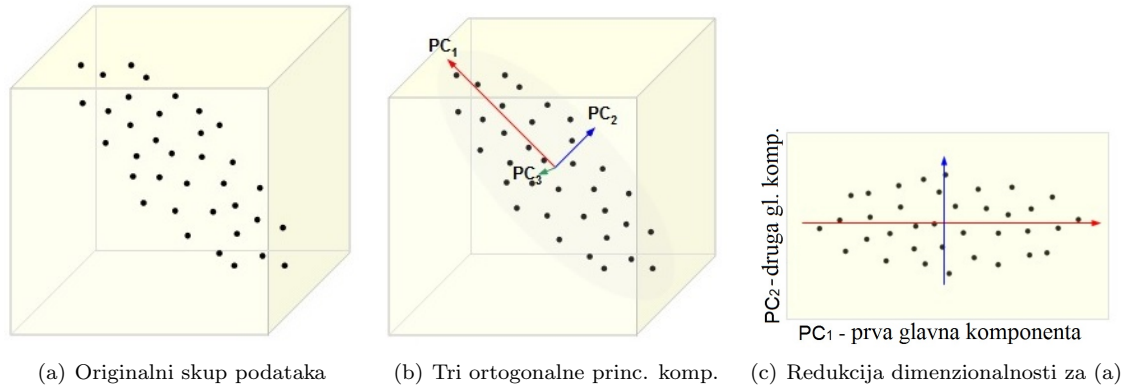
$$P_i = (v_i)'X, \quad 1 \leq i \leq d.$$

Ako predstavimo matricu svih principalnih komponenti Y u matričnom obliku dobijamo

$$Y = (V_d)'X,$$

što se poklapa sa redukcijom dimenzionalnosti $\mathcal{Y}$ u (3.10).

Da sumiramo, glavni vektori su određeni smjerom maksimalne varijanse početnog skupa podataka. Osim toga, dobijeni vektori čine ortonormiranu bazu, a ovo je važno jer onda glavne komponente su nekorelisane. Na Slici 3.2. dali smo ilustraciju glavnih komponenti. Konkretno, na Slici 3.2 (a) dat je trodimenzionalni skup podataka. Na slici 3.2 (b) vidimo tri ortogonalne glavne komponente za dati skup podataka uređene po „veličini varijanse”, dok smo na slici (c) odbacili treću komponentu i posmatramo projekciju pomoću prve dvije glavne komponente. Na taj način (odbacivanjem trećeg vektora) izvršili smo redukciju dimenzije skupa podataka sa 3 na 2. Prethodnu sliku kao i primjenu PCA u biološkim procesima zainteresovani čitalac može pronaći na veb-stranici <http://cnx.org/contents/Av9d0v4w@10/Dimensionality-Reduction-Metho>.



Slika 3.2: Ilustracija PCA.

3.2.3 Matlab kod za PCA algoritam i primjeri

Neka je data realna matrica X formata $D \times n$. Hoćemo da napišemo PCA algoritam koji će izračunavati d glavnih vektora i odgovarajuće glavne komponente, predstavljene u obliku matrice Y formata $d \times n$. Postoji nekoliko različitih verzija PCA algoritama. Mi ćemo ovdje navesti dva algoritma.

Prvi PCA algoritam se može naći na veb-stranici: http://www.cs.stevens.edu/~mordohai/classes/cs559_s09/PCA_in_MATLAB.pdf. Pomoću ovog algoritma izračunavamo nižedimenzionalne podatke koristeći kovarijacionu matricu i njene karakteristične vektore, na način na koji smo to opisali u prethodnom poglavlju.

```

1 function [Y,PC,V] = PCA(X)
2
3 % PCA: Odredjivanje PCA koristeći kovarijacionu matricu.
4
5 % X — Dxn matrica ulaznih podataka
6 % Y — dxn matrica nizedimenzionalnih podatata
7 % PC — matrica čije su kolone principalne komponente
8 % V — Dxl matrica koja sadrži varijanse
9
10 [D,n] = size(X);
11
12 % Centrirajmo podatke:
13 mn = mean(data,2);
14 data = data - repmat(mn,1,n);
15
16 % Izracunajmo kovarijacionu matricu:
17 covariance = 1 / (N-1) * data * data';
18
19 % Izracunajmo karakteristicne korijene i vektore kovarijacione matrice:
20 [PC, V] = eig(covariance);
21
22 % Izdvoj dijagonalne elemente matrice u vektor:
23 V = diag(V);
24
25 % Poredajmo varijanse u opadajući poredak:
26 [junk, rindices] = sort(-1*V);
27 V = V(rindices);
28 PC = PC(:,rindices);
29
30 % Nadjimo nizedimenzionalni skup podataka:
31 Y = PC'*X;
```

Drugi algoritam koji ćemo ovdje navesti razvio je 1988. godine S. Roweiss. Algoritam prvo traži potprostor koji formiraju glavni vektori, zatim se vrši projekcija početnog skupa podataka na prethodno dobijeni potprostor i na kraju se vrši SVD dekompozicija preslikanih (utopljenih) podataka, pri čemu dobijamo tražene glavne komponente. Ovaj algoritam je poznat i pod nazivom EM (expectation-maximization) PCA algoritam i to iz razloga što se svaka iteracija sastoji iz dva koraka:

- E-korak: U ovom koraku data je realna matrica formata $D \times n$ i linearna transformacija T kao matrica formata $D \times n$ ranga d i podaci Y za koje važi $TX = Y$;
- M-korak: Ovdje je data matrica X i dati su podaci Y , a najbolja linearna transformacija se određuje tako što minimiziramo $\|TY - X\|_F$, tj. tražimo preslikavanje tako da imamo što bolje poklapanje početnih i transformisanih podataka.

Na ovaj način dolazimo do traženog potprostora. U ovom radu nećemo se upuštati u dokaze dobre definisanosti i konvergencije datog PCA algoritma, što se se može naći u [15], a još jedan primjer PCA algoritma može se naći u [25].

```

1 function [Y, v, Vd, mu] = PCA(X, d, iter)
2
3 % PCA za redukciju dimenzionalnosti skupa podataka X
4
5 % Sintaksa: [Y, v, Vd, mu] = PCA(X, d, iter)
6
7 % ULAZNE VRIJEDNOSTI:
8 % X: (dim,n) – matrica podataka, tj. kolone matrice su podaci;
9 % d: zeljena dimenzija nakon redukcije dimenzionalnosti;
10 % iter – broj iteracija;
11
12 % IZLAZNE VRIJEDNOSTI:
13 % Y: (d, n) – matrica redukcije dimenzionalnosti skupa X;
14 % v: (d, 1) – vektor, d karakteristicnih vrijednosti kovar. matrice S ;
15 % Vd:(dim, d) – matrica koja sadrzi d karakt. vektora koji odg. preth. vrijednostima;
16 % mu:(d, 1) – vektor koji predstavlja ocekivanje skupa podataka;
17
18 % Inicijalizacija:
19 [dim,n] = size(X);
20 if nargin<4
21 T = randn(dim,d);
22 if nargin<3
23 iter =20;
24 end
25 end
26 assert(d<=dim);
27 if isempty(T)
28 T = rand(dim,d);
29 else
30 assert(d ==size(T,2));
31 assert(dim ==size(T,1));
32 end
33
34 % Centriramo podatke:
35 mu = mean(X,2);
36 X = X - repmat(mu,1,n);
37
38 % Iteracije:
39 for i=1:iter
40 % E korak – preslikavamo X u R^d.
41 % (Rjesavamo TY=X po Y).
42 Y = T\X;
43 % M korak – pronalazimo najbolje slaganje sa podacima.
44 % (Rjesavamo TY=X po T).

```

```

45 T = X/Y;
46 end
47
48 % Ortonormirana baza potprostora:
49 T = orth(T);
50
51 % Ortonormirana projekcija (elemenata) skupa X:
52 tx = T'*X;
53
54 % Kovarijaciona matrica za tx':
55 ctx = cov(tx',1);
56
57 % Spektralna dekompozicija kovarijacione matrice ctx:
58 [vects,eigv] = eig(ctx);
59 S = fliplr(vects);
60 eigv = diag(eigv);
61 v = eigv(d:-1:1);
62 Vd= T*S;
63 Y = S'*tx;

```

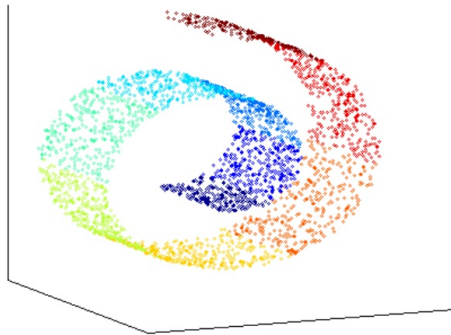
Jedan od najpoznatijih primjera koji se koristi kako bi se ilustrovale metode za redukciju dimenzionalnosti jeste nelinearna 2-dimenzionalna mnogostrukost u $3D$, poznata pod nazivom *Švajcarska rolna* (eng. Swiss roll) ili samo rolna, koja je parametarski data sa

$$\begin{cases} x = \left(\frac{3\pi}{2}(1+2t)\right) \cos\left(\frac{3\pi}{2}(1+2t)\right), \\ y = s, \\ z = \left(\frac{3\pi}{2}(1+2t)\right) \sin\left(\frac{3\pi}{2}(1+2t)\right). \end{cases} \quad 0 \leq s \leq L, \quad |t| \leq 1, \quad (3.12)$$

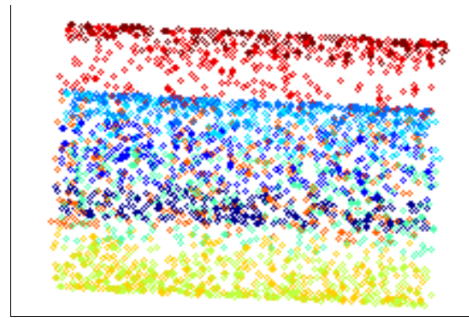
Kako bismo iskoristili PCA za redukciju rolne u 2D možemo iskoristiti sljedeće:

1. $N = 1000$;
2. $tt = (3 * \pi / 2) * (1 + 2 * rand(1, N))$;
3. $height = 21 * rand(1, N)$;
4. $X = [tt * \cos(tt); height; tt * \sin(tt)]$;
5. $d = 2$;

i pozvati prethodno napravljeni algoritam sa $[Y, v, Vd, mu] = PCA(X, d, 30)$, gdje smo proizvoljno izabrali da je broj iteracija 30. Napomenimo još da se na veb-stranici <https://lvdmaaten.github.io/> može besplatno preuzeti Matlab toolbox za redukciju dimenzionalnosti koji sadrži 34 tehnike za redukciju dimenzionalnosti, uključujući i PCA. Više o samom toolbox-u kao i uputstvo za upotrebu može se naći na navedenoj veb-stranici.



(a) Skup podataka rolne



(b) Primjena PCA na (a)

Slika 3.3: Primjer primjene PCA na švajcarsku rolnu.

Vidjeli smo da je PCA linearna tehnika, stoga kada podaci leže na linearnoj mnogostrukosti, tj. hiperravni, ova tehnika je pogodna. Ali ako se podaci nalaze na nelinearnim mnogostrukostima, često ne dobijemo pravi odgovor. Na Slici 3.3 vidimo da PCA nije dao najbolji rezultat, s obzirom da je došlo do preklapanja velikog broja podataka, a to je iz razloga što su naši podaci u blizini nelinearne mnogostrukosti, u ovom slučaju rolne.

3.2.4 Primjena PCA

Uz pomoć PCA algoritma vrši se automatsko prepoznavanje ljudskog lica uz pomoć računara. U posljednje vrijeme koristi se prepoznavanje lica na društvenim mrežama, prilikom otključavanja android mobilnih telefona, kao i u sistemima za nadzor, ali kao što smo već u Uvodu naveli, prilikom prepoznavanja ljudskog lica javljaju se određeni problemi kao što su promjene u osvjetljenju, uglu posmatranja i sl.

Pristup koji se koristi je sljedeći: Posmatra se lice kao jedna cjelina i prethodno napravljeni šabloni se nalaze u bazi podataka, a potom se vrši poređenje sa datim šablonima. Postoji nekoliko algoritama koji funkcionišu na istom principu, a najpoznatiji jeste *metod svojstvenih lica* (eng. eigenface). Eigenface algoritam se zasniva na PCA metodi.

Pretpostavimo da imamo veliku bazu podataka koja sadrži šablone ljudskih lica, ali tako da svaka osoba ima više različitih šablona. U mašinskom učenju bazu podataka često zovemo *skup za učenje* (eng. training set). Kada se opisuju lica često se veliki broj slika može odbaciti. To je iz razloga što je svaki element slike lica usko povezan sa ostalim elementima. Prvo, računamo glavne komponente slika lica iz skupa za učenje. U prethodnoj terminologiji, ove komponente su karakteristični vektori kovarijacione matrice (svojstvena lica - eigenfaces) skupa slika lica koje posmatramo. Npr. ako je dimenzija slike 256×256 onda je posmatramo kao matricu formata 256×256 , ali svaku sliku možemo posmatrati i kao vektor dimenzije $65536 = 256 \cdot 256$ i na taj način smo sliku pretvorili u niz tačaka u višedimenzionalnom prostoru. Mi ćemo nadalje svaku sliku posmatrati kao vektor u višedimenzionalnom prostoru, tj. kao slučajnu promjenljivu. Nakon što izračunamo *srednje lice* (vidjeti (3.1)) s obzirom da su svojstvena lica glavne komponente skupa slika lica, svako lice iz skupa možemo predstaviti kao težinsku sumu svojstvenih lica sa određenim gubitkom energije. Prostor koji razapinju svojstvena lica (vektori) zovemo *prostor lica* (eng. face space). Svako lice možemo da predstavimo kao sumu srednjeg lica, težinskih svojstvenih lica i ostatka. Ostatak mjeri rastojanje lica od ulazne slike lica do prostora lica, sa ciljem da se provjeri da li je ulazna slika lica. Ako slika jeste lice, onda se vrši klasifikacija da li je to poznato lice (dakle šablon se nalazi u bazi) ili nepoznato lice. Uzorak težine se poredi sa poznatim težinama kako bi se pronašlo odgovarajuće lice.

Neka je dat skup za učenje X , gdje je svaki vektor vrste matrice X slučajna promjenljiva koja predstavlja sliku lica. Predstavljemo sada jedan algoritam za prepoznavanje ljudskog lica. Algoritam ćemo izložiti u tri koraka.

1. **Konstrukcija svojstvenih lica.** Neka slučajan vektor $X = [X_1, \dots, X_n]$ predstavlja skup za učenje gdje svaka promjenljiva X_j predstavlja sliku lica. Primjenom PCA algoritma na X dobijamo srednje lice $\bar{X}$, svojstvena lica $F_1, \dots, F_d$ i težinsku matricu $Y = [y_1, \dots, y_n]$ formata $d \times n$. Svojstvena lica formiraju ortonormiranu bazu prostora lica S_d . U težinskoj matrici Y , svaki težinski vektor y_i približno predstavlja i -to lice X_i u prostoru za učenje tako da je $X_i \approx \sum_{j=1}^d y_{ji} F_j$.
2. **Računanje ključnog vektora lica.** Svako novo lice koje dobijamo se projektuje u prostor lica pa se dobija novi vektor z , koji zovemo *ključno lice*. Pretpostavimo da je Z ulazna slika (predstavljena vektorom kolone). Tada ključno lice ulazne slike Z u odnosu na svojstvena lica računamo kao

$$z = F'(Z - \bar{X}).$$

Rastojanje ulazne slike Z od prostora lica S_d računamo kao

$$d_1 = \|Z - (z'F + \bar{X})\|.$$

3. **Prepoznavanje lica.** Neka je ε_1 vrijednost pomoću koje ćemo određivati da li je ulazna slika lice i neka je ε_2 vrijednost koja govori da li je to poznato lice. Imaćemo sljedeće:

3.1. $d_1 > \varepsilon_1$ - tada Z nije lice;

3.2. $d_1 < \varepsilon_1$ - tada je Z lice, pa onda računamo sumu rastojanja između ključnog vektora z i težinskih vektora kao

$$d_2 = \sum_{i=1}^n (z - y_i)^2.$$

Ako je $d_2 < \varepsilon_2$, onda je Z poznato lice, tj. lice iz skupa za učenje. Ako je $d_2 > \varepsilon_2$ onda je Z novo lice.

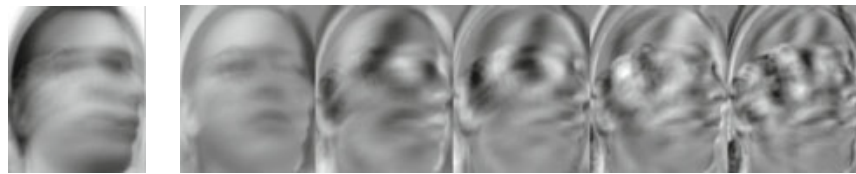
3.2.4 Primjer Razmotrimo na kraju jedan primjer primjene metoda svojstvenih lica. Na Slici 3.4 (a) je dat skup za učenje, dok smo na (b) predstavili isti skup koji je rekonstruisan pomoću srednjeg lica i svojstvenih lica koji su predstavljeni na Slici 3.5, koristeći eigenface algoritam.



(a) Slike lica iz skupa za učenje

(b) Lica dobijena pomoću eigenface algoritma za (a)

Slika 3.4: Ilustracija primjene eigenface algoritma.



(a) Srednje lice

(b) Svojstvena lica

Slika 3.5: Srednje lice i svojstvena lica za lica sa Slike 2.4. (a).

Mnoštvo drugih primjera kao i prethodno korišćene slike mogu se naći u [25]. Osim toga, baze podataka sa ljudskim licima mogu se naći na veb-stranicama <http://vis-www.cs.umass.edu/lfw/> i <http://www.face-rec.org/databases/>.

3.3 Multidimenzionalno skaliranje - MDS

Izraz *multidimenzionalno skaliranje* - MDS se odnosi na familiju tehnika za redukciju dimenzionalnosti koje se koriste da se višedimenzionalni podaci predstave u nižedimenzionalnom prostoru, ali tako da se razmatraju rastojanja dobijenih tačaka, koja predstavljaju geometrijsku sliku početnog skupa podataka. MDS tehnike prvi put se javljaju 1952. godine u psihometriji, tj. oblasti psihologije koja se bavi mjerenjem mentalnih procesa i sposobnosti, kao što su znanje i umijeća, ali tako da se ti procesi predstave matematičkim sredstvima, kako bi se dobili objektivni i precizni rezultati. Danas se multidimenzionalno skaliranje koristi kao neizbježna tehnika u različitim oblastima nauke i tehnike kao što su biologija, sociologija, fizika, psihometrija, političke nauke, a veliki dio primjene MDS tehnika može se naći u [19].

MDS daje sliku strukture skupa objekata iz početnog skupa podataka tako što tu strukturu predstavlja kao rastojanja između parova tih objekata. Termin *sličnost* (*različitost*) ćemo koristiti kao zamjenu za *mjeru sličnosti* (*mjeru različitosti*) između objekata datog skupa podataka. Konkretno, pretpostavimo da imamo n objekata u nekom skupu podataka i da smo izmjerili sličnosti između svih parova. Tada ćemo izmjerene sličnosti između parova objekata predstaviti u obliku $n \times n$ *guste* (eng. dense) matrice D , tj. matrice koja ima mali broj nula, i matricu D ćemo zvati *matrica sličnosti*. Pošto se često ne mogu izmjeriti sličnosti između svih objekata, onda umjesto *guste* matrice koristimo *rijetku* (eng. sparse) matricu sličnosti koja ima veliki broj nula proizvoljno raspoređenih unutar matrice, pri čemu 0 ne-dijagonalni element implicira da mjera sličnosti između dva objekta nije izračunata. Dakle, naš cilj je da se odredi optimalan koordinatni sistem kako bismo imali dobru vizualizaciju tačaka. Svaki objekat je predstavljen kao tačka u višedimenzionalnom prostoru. Tačke u ovom prostoru su raspoređene tako da rastojanja između tačaka predstavljaju na najbolji mogući način sličnost između parova objekata. One tačke koje imaju veliku sličnost su u ovom prostoru predstavljene tako da se nalaze blizu jedna drugoj, dok su tačke koje imaju veliku različitost na velikom rastojanju u ovom prostoru. Najčešće taj prostor je dvodimenzionalan ili trodimenzionalan euklidski prostor, ali u opštem slučaju ne mora biti. Kako bismo razumjeli prethodno navedenu ideju razmotrimo sljedeći primjer.

3.3.1 Primjer Pretpostavimo da je dato 10 evropskih gradova i neka su njihova međusobna rastojanja data u sljedećoj tabeli.

	1 Barselona	2 Beograd	3 Berlin	4 Brisel	5 Bukurešt	6 Budimpešta	7 Kopenhagen	8 Dablin	9 Hamburg	10 Istanbul
Barselona	0	1528,13	1497,61	1062,89	1968,42	1498,79	1757,54	1469,29	1471,78	2230,42
Beograd	1528,13	0	999,25	1372,59	447,34	316,41	1327,24	2145,39	1229,93	809,48
Berlin	1497,61	999,25	0	651,62	1293,4	689,06	354,03	1315,16	254,51	1735,01
Brisel	1062,89	1372,59	651,62	0	1769,69	1131,52	766,67	773,2	489,76	2178,85
Bukurešt	1968,42	447,34	1293,4	1769,69	0	639,77	1571,54	2534,72	1544,17	445,62
Budimpešta	1498,79	316,41	689,06	1131,52	639,77	0	1011,31	1894,95	927,92	1064,76
Kopenhagen	1757,54	1327,24	354,03	766,67	1571,54	1011,31	0	1238,38	287,97	2017,17
Dablin	1469,29	2145,39	1315,16	773,2	2534,72	1894,95	1238,38	0	1073,36	2950,11
Hamburg	1471,78	1229,93	254,51	489,76	1544,17	927,92	287,97	1073,36	0	1983,75
Istanbul	2230,42	809,48	1735,01	2178,85	445,62	1064,76	2017,17	2950,11	1983,75	0

Tabela 3.1: Udaljenost 10 evropskih gradova.

Gradovi su objekti, a rastojanja su sličnosti između gradova. Primjenom MDS na skup podataka datih u prethodnoj tabeli, dobijamo "kartu" (Slika 3.6) na kojoj se nalaze lokacije datih gradova na osnovu podataka u tabeli. 10 tačaka predstavlja 10 gradova i gradovi koji imaju više sličnosti su bliže na "karti" i slično gradovi koji imaju više različitosti su više udaljeni. Primjetimo da nismo

dobili baš očekivano, tj. nismo dobili pozicije gradova kao na karti Evrope, ali s obzirom da ose mogu biti proizvoljno orijentisane (jer proizvoljnom rotacijom osa rastojanja između gradova će ostati ista), dobili smo približno traženo rješenje.



Slika 3.6: MDS za rastojanja između 10 gradova.

Prethodnu tabelu kao i scatterplot na Slici 3.6 dobili smo pomoću statističkog paketa *Statistica* 10.

Pošto smo već naveli da je MDS familija tehnika za redukciju dimenzionalnosti, može se izvršiti klasifikacija na osnovu toga da li su sličnosti između podataka predstavljene kvantitativnim (metrički MDS - MMDS) ili kvalitativnim (ne-metrički MDS - NMDS) vrijednostima. Pored ove klasifikacije, MDS tehnike možemo da klasifikujemo i u odnosu na broj matrica sličnosti koje je potrebno da iskoristimo kako bismo dobro opisali podatke, tj. nekad nije dovoljno da iskoristimo samo jednu matricu sličnosti. Npr. u prethodnom primjeru sličnost između gradova smo mjerili pomoću njihove udaljenosti pa smo dobili jednu matricu sličnosti, ali smo mogli da tražimo sličnost između gradova prema broju stanovnika pri čemu bismo dobili još jednu matricu sličnosti itd. Stoga, imamo i klasifikaciju na: *Klasično multidimenzionalno skaliranje* - CMDS, kod kojeg imamo samo jednu matricu sličnosti, *replicirano multidimenzionalno skaliranje* RMDS, kod kojeg imamo više matrica sličnosti na osnovu kojih se onda formira jedna matrica sličnosti na koju zatim primjenjujemo RMDS, i *težinsko multidimenzionalno skaliranje* WMDS koje kao i RMDS za ulaznu vrijednost ima više matrica sličnosti, ali se ne formira jedna matrica sličnosti kao kod RMDS, nego se formiraju skupovi tačaka pri čemu je svaki skup izveden na osnovu svoje matrice sličnosti, za razliku od RMDS gdje su dobijeni skupovi tačaka izvedeni iz jedne matrice sličnosti. U ovom poglavlju koristili smo [25, 26].

3.3.1 Euklidska i Gramova matrica

U ovom radu ćemo razmatrati samo klasično multidimenzionalno skaliranje - CMDS. Kod CMDS pretpostavlja se da su sličnosti između podataka opisane pomoću *euklidske metrike*, tj. ako imamo skup podataka $\mathcal{X} = \{x_1, \dots, x_n\}$ u D -dimenzionalnom prostoru $\mathbb{R}^D$ tada matricu euklidskog rastojanja $D = [d_{ij}]_{i,j=1}^n$ dobijamo tako što rastojanja između tačaka x_i i x_j računamo kao

$$d_{ij} = d_2(x_i, x_j) = \sqrt{\sum_{k=1}^D |x_{ki} - x_{kj}|^2}.$$

U CMDS matricu euklidskog rastojanja D zovemo i matricom sličnosti. Matrica euklidskog rastojanja definiše metriku na skupu $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ koju zovemo *euklidska metrika*. CMDS spada u klasu MMDS, tako da ako matrica sličnosti D nije matrica euklidskog rastojanja, nećemo moći da primijenimo CMDS, ali onda mogu da se iskoriste druge metrike, kao što su npr. metrika Minkovskog koja je za $p \geq 1$ data sa

$$d_{ij} = d_p(x_i, x_j) = \sqrt[p]{\sum_{k=1}^D |x_{ki} - x_{kj}|^p}.$$

S obzirom da smo se u ovom radu odlučili za geometrijski pristup tehnikama za redukciju dimenzionalnosti, tako ćemo i ovdje sa stanovišta geometrije imati da euklidsko rastojanje opisuje različitosti parova tačaka, dok će skalarni proizvod opisivati njihovu sličnost. Ponovo ćemo skup podataka $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ predstavljati u obliku realne $D \times n$ matrice $X = [x_1, \dots, x_n]$, gdje je svaka kolona matrice X tačka skupa $\mathcal{X}$. Osim toga, ponovo ćemo identifikovati $\mathcal{X}$ i X .

U ovom poglavlju cilj je da vidimo u kakvoj su vezi Gramova matrica i matrica euklidskog rastojanja. Pored matrice euklidskog rastojanja $D := [d_2(x_i, x_j)]_{i,j=1}^n$ biće potrebna i matrica kvadrata euklidskog rastojanja koja je data sa

$$S = [d_2^2(x_i, x_j)]_{i,j=1}^n.$$

Primjetimo da su matrice D i S simetrične matrice koje su invarijantne u odnosu na translaciju i rotaciju. Ove matrice su očigledno simetrične zbog definicije rastojanja d_2 . Dalje, za vektor $a \in \mathbb{R}^D$ vidjeli smo da je translacija skupa $\mathcal{X}$ data sa $\mathcal{X}_a = \{x - a : x \in \mathcal{X}\}$, a za ortogonalnu dijagonalnu kvadratnu matricu O dimenzije D , O rotacija skupa $\mathcal{X}$ data je sa $\mathcal{X}_O = \{Ox_1, \dots, Ox_n\}$. Direktnom provjerom vidi se da je matrica euklidskog rastojanja ista za skupove $\mathcal{X}$, $\mathcal{X}_a$ i $\mathcal{X}_O$, pa imamo invarijantnost matrice D . Invarijantnost matrice kvadrata euklidskog rastojanja P onda slijedi direktno.

U opštem slučaju euklidsku metriku definišemo na sljedeći način.

3.3.2 Definicija *Simetrična matrica D formata $n \times n$ naziva se matrica euklidskog rastojanja ili euklidska metrika ako postoji prirodan broj $m > 0$ i skup vektora $\mathcal{Z} = \{z_1, \dots, z_n\} \subset \mathbb{R}^m$ tako da je*

$$D := [d_{ij}] = [d_2(z_i, z_j)]_{i,j=1}^n.$$

Skup vektora $\mathcal{Z}$ zovemo konfiguracija za D .

3.3.3 Definicija *Gramova matrica skupa vektora $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ je definisana sa*

$$G = [G_{ij}] = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \dots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \dots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \dots & \langle x_n, x_n \rangle \end{bmatrix} = [\langle x_i, x_j \rangle]_{i,j=1}^n. \quad (3.13)$$

3.3.4 Teorema *Vektori $\{x_1, x_2, \dots, x_n\}$ su linearno zavisni ako i samo ako je determinanta njihove Gramove matrice jednaka 0, tj. $\det G = 0$. Za linearno nezavisne vektore važi da je $\det G > 0$. U prostoru $\mathbb{R}^n$ sa skalarnim proizvodom $\det G$ je kvadrat zapremine generalisanog paralelopipeda određenog vektorima $\{x_1, x_2, \dots, x_n\}$.*

3.3.5 Napomena Primjetimo da se Gramova matrica može predstaviti i kao $G = X'X$ gdje je $X = [x_1, \dots, x_n]$. Lako se vidi da je svaka Gramova matrica simetrična, pozitivno definitna matrica. Zaista, simetričnost slijedi iz simetričnosti skalarnog proizvoda, a ako uzmemo proizvoljan vektor $y \in \mathbb{R}^n$ onda je $y'Gy = y'X'Xy = (Xy)'Xy = \|Xy\|^2 \geq 0$, odakle slijedi i pozitivna semidefinitnost. S druge strane, svaka simetrična, pozitivno definitna matrica je Gramova matrica nekog skupa

podataka. Zaista, ako $n \times n$ simetrična pozitivno definitna matrica G ima rang m onda se ona može zapisati u obliku $G = X'X$ gdje je $X = \{x_1, \dots, x_n\}$ matrica formata $m \times n$.³ Napomenimo još da za Gramovu matricu ne važi invarijantnost u odnosu na translaciju.

Vidjećemo sada u kakvoj su vezi euklidska metrika i Gramova matrica. Neka je data euklidska metrika D i neka je G Gramova matrica skupa podataka $\mathcal{X}$. Neka su $x, y \in \mathbb{R}^D$. Tada koristeći kosinusnu teoremu imamo da važi

$$d_2(x, y) = \sqrt{\langle x, x \rangle - 2\langle x, y \rangle + \langle y, y \rangle}.$$

Odatle slijedi da za vektore iz skupa $\mathcal{X}$ važi

$$d_2(x_i, x_j) = D_{ij} = \sqrt{G_{ii} - 2G_{ij} + G_{jj}} \quad \text{za sve } i, j \in \{1, \dots, n\}.$$

Pretpostavimo da podaci $\mathcal{X}$ leže na d -dimenzionalnoj hiperravni $H \subset \mathbb{R}^D$. Tada se centar $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ nalazi u H i $Q = H - \bar{x}$ je d -dimenzionalan potprostor od $\mathbb{R}^D$ paralelan sa H . Kao i kod PCA označimo sa $\hat{\mathcal{X}}$ skup podataka transliranih za $\bar{x}$. Tada su podaci $\hat{\mathcal{X}}$ centrirani, a odgovarajuća matrica $\hat{X}$ je matrica centriranih podataka.

3.3.6 Definicija Neka je za dati skup podataka $\mathcal{X}$ sa $\hat{\mathcal{X}}$ označen skup centriranih podataka. Tada je **centrirana Gramova matrica** G^c za $\mathcal{X}$ Gramova matrica skupa $\hat{\mathcal{X}}$, tj.

$$G^c = [\langle \hat{x}_i, \hat{x}_j \rangle]_{i,j=1}^n = \hat{X}'\hat{X}.$$

Sada lako dobijamo da su centrirana Gramova matrica G^c i Gramova matrica G u istoj vezi sa euklidskom metrikom D , tj. i za G^c važi

$$D_{ij} = \sqrt{G_{ii}^c - 2G_{ij}^c + G_{jj}^c}.$$

3.3.7 Definicija Neka je $\mathbf{1} = [1, \dots, 1]' \in \mathbb{R}^n$, $E = \mathbf{1}\mathbf{1}'$ i neka je I jedinična $n \times n$ matrica. Tada matricu

$$H = I - \frac{1}{n}E$$

zovemo **centrirajuća matrica**.

3.3.8 Lema Centrirajuća matrica H ima sljedeće osobine:

1. $H^2 = H$;
2. $\mathbf{1}'H = H\mathbf{1} = 0$;
3. $\mathcal{X}$ je skup centriranih podataka ako i samo ako je $XH = X$;
4. Simetrična, pozitivno definitna matrica C je centrirana Gramova matrica ako i samo ako je $HCH = C$.

Dokaz.

1. Iz činjenice da je $E^2 = nE$ slijedi

$$H^2 = \left(I - \frac{1}{n}E\right)^2 = I - \frac{2}{n}E + \frac{1}{n^2}E^2 = I - \frac{2}{n}E + \frac{1}{n^2}nE = I - \frac{1}{n}E = H.$$

2. Kako je $\mathbf{1}'\mathbf{1} = n$ imamo

$$\mathbf{1}'H = \mathbf{1}'\left(I - \frac{1}{n}E\right) = \mathbf{1}' - \frac{1}{n}\mathbf{1}'\mathbf{1}\mathbf{1}' = \mathbf{1}' - \mathbf{1}' = 0,$$

i slično, koristeći da je $E\mathbf{1} = n\mathbf{1}$ dobijamo da je $H\mathbf{1} = 0$.

³Svaka pozitivno definitna matrica može da se zapiše u ovom obliku, tj. postoji njena Cholesky dekompozicija.

3. Prvo primjetimo da je $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} X \mathbf{1}$ i da se centrirana matrica podataka može zapisati kao $\hat{X} = X - \bar{x} \mathbf{1}'$. Odavde onda imamo da je

$$\hat{X} = X - \bar{x} \mathbf{1}' = X - \frac{1}{n} X \mathbf{1} \mathbf{1}' = X - \frac{1}{n} X E = X \left(I - \frac{1}{n} E \right) = XH.$$

Prema tome, podaci su centrirani ako i samo ako je $\hat{X} = XH = X$, što je trebalo dokazati.

4. Neka je C simetrična, pozitivno definitna matrica. Tada imamo:

$$\begin{aligned} C \text{ je centrirana Gramova matrica} &\iff C = \hat{X}' \hat{X} \\ &\stackrel{3.}{\iff} C = (XH)'(XH) \\ &\iff C = H' X' X H \\ &\stackrel{H'=H}{\iff} C = HCH. \end{aligned}$$

Time smo pokazali tvrđenje.

Direktna posljedica prethodne leme je sljedeće tvrđenje.

3.3.9 Lema *Neka je X matrica podataka i neka je G Gramova matrica. Tada je matrica centriranih podataka skupa X baš $\hat{X} = XH$ i centrirana Gramova matrica za X je data sa $G^c = HGH$.*

Sada u opštem slučaju centriranu matricu definišemo na sljedeći način.

3.3.10 Definicija *Neka je A simetrična matrica (ne nužno pozitivno definitna). Tada je **centrirana matrica** definisana sa $A^c = HAH$.*

Odavde sada lako zaključujemo da je simetrična matrica S centrirajuća ako i samo je $S = HSH$. Pojam simetrične centrirane matrice omogućava da izrazimo centriranu Gramovu matricu pomoću matrice kvadrata euklidskog rastojanja.

3.3.11 Teorema *Za matricu kvadrata euklidskog rastojanja S i centriranu Gramovu matricu G^c skupa podataka $\mathcal{X}$ važi:*

$$G^c = -\frac{1}{2} S^c.$$

Dokaz. Iz Leme 3.3.8 imamo da je $\sum_{i=1}^n G_{ij}^c = 0$. Dalje, koristeći da je $D_{ij} = \sqrt{G_{ii}^c - 2G_{ij}^c + G_{jj}^c}$, dobijamo

$$\sum_{i=1}^n D_{ij}^2 = nG_{jj}^c + \sum_{i=1}^n G_{ii}^c,$$

i

$$\sum_{j=1}^n D_{ij}^2 = nG_{ii}^c + \sum_{j=1}^n G_{jj}^c.$$

Stoga, (i, j) -element matrice S^c je dat sa

$$\begin{aligned} (S^c)_{ij} &= D_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^n D_{ij}^2 + \sum_{j=1}^n D_{ij}^2 - \frac{1}{n} \sum_{i,j=1}^n D_{ij}^2 \right) \\ &= D_{ij}^2 - G_{ii}^c - G_{jj}^c \\ &= -2G_{ij}^c. \end{aligned}$$

Time smo dokazali tvrđenje. □

3.3.12 Teorema *Neka je A simetrična matrica. Tada važi:*

1. *Matrica A je Gramova matrica skupa podataka ako i samo ako je simetrična pozitivno definitna matrica.*
2. *Matrica A je centrirana Gramova matrica skupa podataka ako i samo ako je centrirana simetrična pozitivno definitna matrica.*
3. *A je matrica kvadrata euklidskog rastojanja ako i samo ako je matrica $-\frac{1}{2}A^c$ centrirana simetrična pozitivno definitna matrica.*

Dokaz. U dokazu ćemo koristiti prethodno dokazana tvrđenja. 1. i 2. slijede direktno iz prethodnog.

3. Označimo $G = -\frac{1}{2}A^c$. Ako je G centrirana simetrična pozitivno definitna matrica tada imamo da postoji centrirana matrica $V = [v_1, \dots, v_n]_{m \times n}$ takva da je $G = V'V$ (Cholesky dekompozicija). Na osnovu Teoreme 3.3.11 je matrica A data sa $A^c = -2G^c$ matrica kvadrata euklidskog rastojanja. S druge strane, ako je A matrica kvadrata euklidskog rastojanja skupa podataka $\mathcal{X}$ tada je $G^c = -\frac{1}{2}A^c$ centrirana Gramova matrica skupa podataka $\mathcal{X}$, odakle imamo da je A centrirana simetrična pozitivno definitna matrica, čime smo pokazali tvrđenje. $\square$

3.3.2 Klasično multidimenzionalno skaliranje - CMDS

Neka je $D = [d_{ij}]_{i,j=1}^n$ data matrica rastojanja skupa od n objekata i neka je $d > 0$ prirodan broj. Metod CMDS je postupak koji traži konfiguraciju $\mathcal{Y} = \{y_1, \dots, y_n\} \subset \mathbb{R}^d$ tako da matrica rastojanja pridružena skupu $\mathcal{Y}$ je što "bliža" matrici D , tj. da za sve i, j važi

$$d_Y(y_i, y_j) \approx d_{ij}.$$

Kod CMDS za matricu D koristi se euklidska metrika.

3.3.13 Lema *Neka je matrica $D = [d_{ij}]_{n \times n}$ euklidska metrika i $P = [d_{ij}^2]$ odgovarajuća matrica kvadrata euklidskog rastojanja. Neka je $G^c = -\frac{1}{2}P^c$. Ako je rang matrice G^c jednak r , tada postoji centriran r -dimenzionalni skup vektora $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^r$ tako da za $1 \leq i, j \leq n$ važi*

$$d_2(x_i, x_j) = d_{ij}.$$

Dokaz. Iz Teoreme 3.3.12, G^c je centrirana Gramova matrica. S obzirom da je matrica G^c ranga r imamo da postoji $r \times n$ centrirana matrica podataka X takva da je $G^c = X'X$. Tada za centrirani skup podataka $\mathcal{X}$ važi traženo. $\square$

Broj r nazivamo *unutrašnja dimenzija* matrice D i kažemo da je $\mathcal{X}$ *egzaktna konfiguracija* za D . Ako hoćemo da iskoristimo konfiguraciju kako bismo vizualizovali podatke, a imamo da je r veliki broj, kako bismo to uspjeli potrebno je da nađemo nižedimenzionalnu konfiguraciju npr. d -dimenzionalnu konfiguraciju $\mathcal{Y}$ ($d \ll r$). Pridružene matrice X i Y posmatrajmo sada kao slučajne vektore. Intuitivno, hoćemo da Y sadrži d glavnih komponenti za X . Stoga, definišemo funkciju greške za CMDS sa

$$\eta(\mathcal{Y}) = \sum_{i,j=1}^n (d_{ij}^2 - d_2^2(y_i, y_j))$$

tako da je

$$T(\mathcal{X}) = \mathcal{Y},$$

gdje je $T : \mathbb{R}^r \rightarrow S_d$, $S_d \subset \mathbb{R}^r$ d -dimenzionalni potprostor, a $\mathcal{X}$ egzaktna konfiguracija od D . Tada je konfiguracija $\mathcal{Y}$ rješenje problema minimizacije

$$Y = \arg \min_Y \eta(\mathcal{Y}) \quad \text{gdje je} \quad T(\mathcal{X}) = \mathcal{Y},$$

i je Y realna matrica formata $d \times n$. Cilj je da riješimo ovaj problem minimizacije. Za početak dokaćemo nekoliko tvrđenja koja će olakšati rješavanje ovog problema.

3.3.14 Lema *Neka je $\mathcal{Z} \subset \mathbb{R}^r$ dati skup podataka sa odgovarajućom matricom kvadrata euklidskih rastojanja $S_Z = [s_{ij}]$ gdje smo označili $s_{ij} = d_2^2(z_i, z_j)$, i neka je onda G_Z^c odgovarajuća centrirana Gramova matrica. Tada je*

$$\text{tr}(G_Z^c) = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n s_{ij}.$$

Dokaz. Radi jednostavnosti, uvedimo oznaku da je $\tilde{s} = \sum_{i=1}^n \sum_{j=1}^n s_{ij}$. Neka su matrice E i H formata $n \times n$ kao u Definiciji 3.3.7. Označimo sa $\hat{\mathcal{Z}} = \{\hat{z}_i\}_{i=1}^n$ centrirani skup podataka za skup $\mathcal{Z}$. Koristeći Teoremu 3.3.12 imamo

$$G_Z^c = -\frac{1}{2}S_Z^c = -\frac{1}{2}HS_ZH = -\frac{1}{2}\left(S_Z - \frac{1}{n}ES_Z\frac{1}{n}S_ZE + \frac{1}{n^2}ES_ZE\right),$$

odakle koristeći da je $s_{ii} = 0$ dobijamo

$$\langle \hat{z}_i, \hat{z}_i \rangle = -\frac{1}{2}\left(s_{ii} - \frac{1}{n} \sum_{k=1}^n s_{ik} - \frac{1}{n} \sum_{k=1}^n s_{ki} + \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n s_{kl}\right) = \frac{1}{n} \sum_{k=1}^n s_{ik} - \frac{1}{2n^2} \tilde{s}.$$

Stoga, dobijamo

$$\text{tr}(G_Z^c) = \sum_{i=1}^n \langle \hat{z}_i, \hat{z}_i \rangle = \sum_{i=1}^n \left(\frac{1}{n} \sum_k s_{ik} - \frac{1}{2n^2} \tilde{s} \right) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^n s_{ik} - \frac{1}{2n} \tilde{s} = \frac{1}{2n} \tilde{s},$$

čime smo pokazali tvrđenje. $\square$

3.3.15 Posljedica *Neka je $D_Z = [d_2(z_i, z_j)]_{i,j=1}^n$ i $\hat{\mathcal{Z}} = [\hat{z}_1, \dots, \hat{z}_n]$. Tada je*

$$\|\hat{\mathcal{Z}}\|_F = \frac{1}{\sqrt{2n}} \|D_Z\|_F.$$

Dokaz. Kako je $\|D_Z\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n s_{ij}$ i kako je $\|\hat{\mathcal{Z}}\|_F^2 = \text{tr}(G_Z^c)$ imamo traženo na osnovu Leme 3.3.14. $\square$

Sada smo pripremili sve što je potrebno da bismo dokazali sljedeće tvrđenje koje daje opis metoda CMDS.

3.3.16 Teorema *Neka je dat centrirani skup podataka $\mathcal{X} \subset \mathbb{R}^r$ koji predstavlja konfiguraciju za D iz Leme 3.3.13 i neka je SVD dekompozicija skupa podataka data sa*

$$X = V \Sigma_r U',$$

gdje su $U = [u_1, \dots, u_r]_{n \times r}$ i $V = [v_1, \dots, v_r]_{r \times r}$ matrice čije su kolone ortonormirani vektori i $\Sigma_r = \text{diag}(\sigma_1, \dots, \sigma_r)$ realna $r \times r$ matrica. Za dato $d \leq r$ neka je $U_d = [u_1, \dots, u_d]$ i $Y = XU_d'$. Tada je Y rješenje problema minimizacije

$$Y = \arg \min_Y \eta(Y) \quad \text{gdje je} \quad T(\mathcal{X}) = \mathcal{Y}, \quad (3.14)$$

i je Y realna matrica formata $d \times n$. Ocjena greške data je sa

$$\eta(Y) = \sum_{i=d+1}^r \sigma_i^2. \quad (3.15)$$

Dokaz. Neka je S_d d -dimenzionalan potprostor od $\mathbb{R}^r$ i neka je B $r \times d$ ortonormirana matrica, tj. matrica čije su kolone ortonormirani vektori koji formiraju bazu potprostora S_d . Lako se vidi da je

$$\sum_{i,j=1}^n d_2^2(i, j) - d_2^2(B'x_i, B'x_j) \leq \sum_{i,j=1}^n \left| \|x_i - x_j\|^2 - \|Tx_i - Tx_j\|^2 \right|$$

i da je

$$\|(I - BB')(x_i - x_j)\|^2 = d_2^2(i, j) - d_2^2(B'x_i - B'x_j).$$

Stoga, da bismo riješili problem minimizacije (3.14) potrebno je naći $r \times d$ ortonormiranu matricu B_* tako da skup vektora $B_* X$ minimizira $\eta(\mathcal{Z})$, tj.

$$B_* = \arg \min_B \sum_{i,j=1}^n \|(I - BB')(x_i - x_j)\|^2. \quad (3.16)$$

Označimo li sa $\hat{Z} = (I - BB')X$ na osnovu Posljedice 3.3.15 imamo da važi

$$\|D_{\hat{Z}}\|_{\hat{Z}}^2 = 2n\|\hat{Z}\|_F^2$$

i

$$\hat{Z}_F = \|X - B_*(B_*)'X\|_F.$$

Iskoristimo sada Teoremu 3.2.1 (Eckart-Young-Mirsky), odakle dobijamo da je matrica

$$B_* = U_d$$

rješenje problema (3.16). Odavde, automatski dobijamo da je skup

$$\mathcal{Y} = U_d'X$$

rješenje problema (3.14). Ostalo je još da pokažemo ocjenu greške. To dobijamo iz

$$\eta(U_d\mathcal{X}) = \|X - U_dU_d'X\|_F^2 = \sum_{i=d+1}^r \sigma_i^2.$$

Time smo pokazali tvrdjenje. □

3.3.3 Matlab kod za CMDS

Pretpostavimo da imamo n objekata i da je data simetrična matrica euklidskog rastojanja kao matrica sličnosti D . Cilj algoritma je da nađemo d -dimenzionalnu konfiguraciju matrice D . Algoritam ćemo predstaviti u tri koraka.

1. **Formiranje centrirane Gramove matrice.** Prvo, konstruišemo matricu rastojanja D . Centrirana matrica kvadrata rastojanja S^c može se lako zapisati u obliku $S^c = H(D \odot D)E$, gdje je $D \odot D$ tačkasti (Adamarov)⁴ proizvod matrica. Zatim koristeći formulu $G^c = -\frac{1}{2}S^c = -\frac{1}{2}H(D \odot D)E$ dobijamo centriranu Gramovu matricu.
2. **Izvršimo spektralnu dekompoziciju matrice G .** Pretpostavimo da je rang matrice G jednak r . Neka je spektralna dekompozicija matrice G data sa $G = U\Lambda U'$, gdje je $U = [u_1, \dots, u_r]$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ sa $\lambda_1 \geq \dots \geq \lambda_r$.
3. **Pronađemo nižedimenzionalni konfiguracijski skup (konfiguraciju).** Neka je $U_d = [u_1, \dots, u_d]$ i $\Sigma_d = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_d})$, tada je tražena konfiguracija $Y = \Sigma_d U_d'$.

⁴Jacques Hadamard (1865-1963) - francuski matematičar

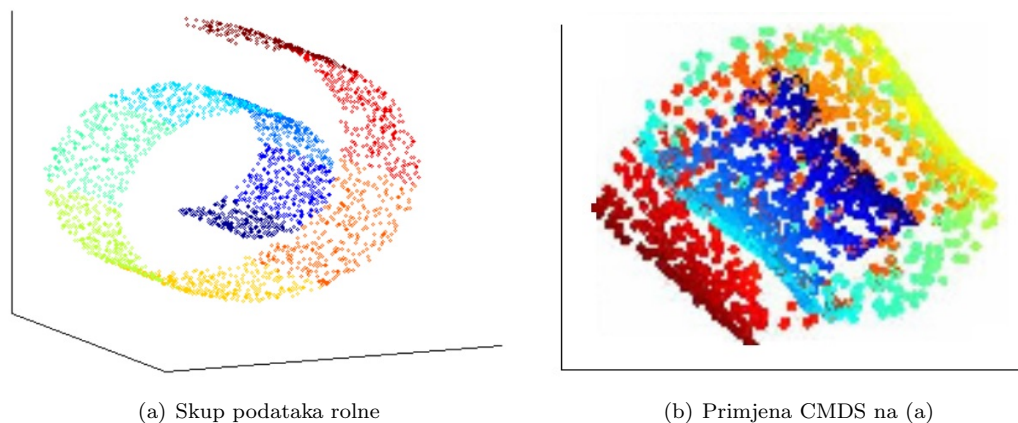
```

1 function [Y, ev] = CMDS(X, ndims, options)
2
3 % Klasicno multidimenzionalno skaliranje
4
5 % Sintaksa: Y = CMDS(D, ndims, options);
6
7 % ULAZNE VRIJEDNOSTI:
8 % 'X': N x N simetricna matrica.
9 % ndims: zeljena dimenzija izlaznog skupa podataka.
10 % OPTIONS: .verb: prikazimo postupak tekstualno.
11
12 % IZLAZNE VRIJEDNOSTI:
13 % Y: d x N dimenzija matrice podataka.
14 % ev: odgovarajuće karakteristicne vrijednosti.
15
16 % Inicijalizacija:
17 if nargin<3
18     options.null=0;
19 end
20 if isfield(options, 'verb')
21     verb = option.verb;
22 else
23     verb = 1;
24 end
25
26 % Provjera ulaznih podataka:
27 [N,M]=size(X);
28 if N~=M
29     error('Ulazna matrica nije kvadratna.');
```

```

30 end
31 if any(any(X~=X'))
32     mes1=sprintf('Ulazna matrica nije simetricna.\n');
33     mes2='Koristimo simetricnu matricu 1/2*(X+X') inace.';
34     warning([mes1,mes2]);
35     X = 1/2*(X+X');
36 end
37 if any(diag(X))
38     mes1=sprintf('Dijagonalni elementi nisu nula.\n');
39     mes2='Staviti dijagonalne elemente jednake 0.';
40     warning([mes1,mes2]);
41     X=X-diag(diag(X));
42 end
43
44 % 1. dio: Konstruisemo centriranu Gramovu matricu:
45 if verb
46     disp('— konstruisemo Gramovu matricu');
47 end
48 DG = X.^2;
49 GC = -.5*(DG - sum(DG)'*ones(1,N)/N - ones(N,1)*sum(DG)/N + sum(DG(:))/(N^2));
50
51 % 2. dio: Dekompozicija:
52 if verb
53     disp('— Nizedimenzionalni izlazni podaci');
54 end
55 opt DISP = 0;
56 opt ISREAL = 1;
57 opt ISSYM = 1;
58 [Y, ev] = eigs(GC, ndims, 'LA', opt);
59 for i=1:ndims
60     Y(:,i) = Y(:,i)*sqrt(val(i,i));
61 end
62 Y = Y';

```



Slika 3.7: Primjer primjene CMDS na švajcarsku rolnu.

Na Slici 3.7 vidimo da ni CMDS nije pogodna tehnika kada radimo sa podacima koji leže na ili u blizini nelinearnih mnogostrukosti, jer se kao i kod PCA veliki broj podataka "poklopio". To je iz razloga što ni PCA, ni CMDS nemaju mogućnost da "razmotaju" rolnu. U sljedećoj glavi vidjećemo tehnike koje će pomoći u prevazilaženju tog problem, tj. radićemo sa nelinearnim tehnikama za redukciju i pokazaćemo takođe njihovu primjenu na primjeru švajcarske rolne.

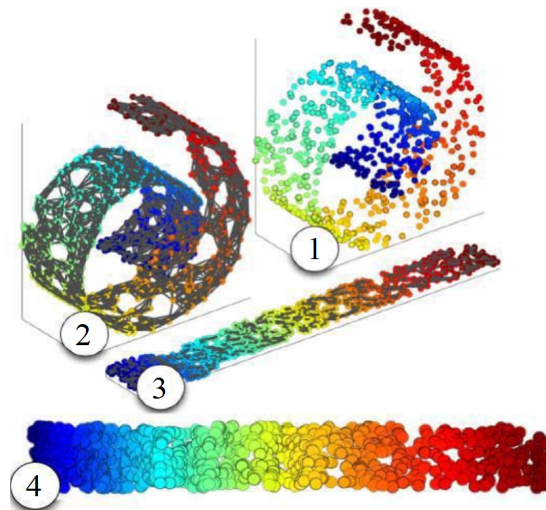
Glava 4

Nelinearne tehnike za redukciju dimenzionalnosti

Kada podaci leže na d -dimenzionalnoj nelinearnoj mnogostrukosti, onda linearne tehnike nisu efikasne prilikom redukcije dimenzionalnosti, stoga ćemo u ovoj glavi razmotriti nelinearne tehnike za redukciju dimezionalnosti signala. Najpoznatije nelinearne tehnike za redukciju dimenzionalnosti su: Isomap (eng. Isometric Feature Mapping - Isomap), lokalno linearno potapanje (eng. Local Linear Embedding - LLE), Laplasovo karakteristično preslikavanje (eng. Laplacian Eigenmaps), Hesijanovo karakteristično preslikavanje (eng. Hessian Eigenmaps), a detaljniji pregled nelinearnih tehnika i više o njima se može naći u [11].

4.1 Zajednički pojmovi za nelinearne tehnike

Kao što smo vidjeli na primjeru rolne, najveći problem kod linearnih tehnika, bio je što nemaju mogućnost da „razmotaju” rolnu, tako da nisu bile pogodne za primjenu kada su se podaci nalazili na nelinearnim mnogostrukostima. Na Slici 4.1 vidimo primjer primjene jedne od nelinearnih tehnika, tačnije MVU (eng. Maximum Variance Unfolding) tehnike, na primjeru rolne. Više o ovoj tehnici može se naći u [18, 25].



Slika 4.1: Primjena MVU algoritma na primjeru rolne.

Kod linearnih tehnika smo različitosti između tačaka skupa podataka opisivali pomoću euklidskog rastojanja, dok kod nelinearnih tehnika različitosti datog skupa objekata opisujemo pomoću ne-euklidskih metrika. Osim toga, različitosti između tačaka opisivaćemo koristeći sistem ε - ili k -okolina (vidjeti Potpoglavlje 2.2.1).

Za dvije tačke koje ne pripadaju istoj okolini smatraćemo da su različite, a metriku na skupu podataka definisaćemo tako da mjeri različitosti koje su opisane sistemom okoline. Metrika koju ćemo definisati lokalno će da opisuje euklidsko rastojanje, tj. u svakoj okolini definisana metrika će (približno) da predstavlja euklidsku metriku. Različite nelinearne DR tehnike koriste različite metrike kako bi opisali različitosti na skupu podataka. Tako Isomap tehnika, koju ćemo razmotriti u sljedećem poglavlju, koristi *geodezijsku* metriku (vidjeti Definiciju 2.1.8) na projektovanoj mnogostrukosti kako bi mjerila različitosti, odnosno sličnosti početnog skupa podataka.

Pretpostavimo da početni podaci $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ leže na d -dimenzionalnoj Riemann-ovoj mnogostrukosti $M \subset \mathbb{R}^D$, pri čemu je $d \ll D$. Označimo geodezijsko rastojanje na M sa d_M . Tada postoji koordinatno preslikavanje $f: M \rightarrow \mathbb{R}^d$ dato sa $f(x_i) = y_i$ koje očuvava geodezijsko rastojanje, tj. za sve $i, j \in \{1, \dots, n\}$ važi

$$d_2(y_i, y_j) = d_M(x_i, x_j).$$

Pošto je sličnost između skupa podataka $\mathcal{Y} = \{y_1, \dots, y_n\}$ ista kao i sličnost skupa podataka $\mathcal{X}$ možemo zaključiti da je onda $\mathcal{Y}$ redukcija dimenzionalnosti za $\mathcal{X}$. Pošto ne znamo mnogostrukost na kojoj leže podaci, onda ne možemo eksplicitno definisati izometriju za redukciju dimenzionalnosti, tj. ne možemo izmjeriti geodezijska rastojanja između tačaka mnogostrukosti. Umjesto toga koristi se aproksimativna geodezijska metrika d_M na skupu podataka $\mathcal{X}$, tj. koristi se takozvana *metrika grafa*, koju definišemo na sljedeći način.

4.1.1 Definicija *Neka je na skupu podataka $\mathcal{X}$ dobro definisan sistem okolina, koji formira graf $G = [\mathcal{X}, E]$ tako da $(x_i, x_j) \in E$ ako i samo ako x_i i x_j pripadaju istoj okolini. Tada je **metrika grafa** d_G definisana sa:*

- Ako $(x, y) \in E$ tada je $d_G(x, y) = d_2(x, y)$.
- Ako $(x, y) \notin E$ za granu

$$\gamma = (x_0, x_1, \dots, x_{s+1})$$

koja povezuje tačke $x = x_0$ i $y = x_{s+1}$ definišimo rastojanje d_γ pomoću grana sa

$$d_\gamma = d_2(x_0, x_1) + \dots + d_2(x_s, x_{s+1}).$$

Ako je Γ skup svih grana koji povezuju tačke x i y , onda je

$$d_G(x, y) = \min_{\gamma \in \Gamma} d_\gamma(x, y).$$

Lako je uočiti da je sa d_G zaista definisana metrika na G . Kada su tačke na mnogostrukosti M dovoljno guste, onda pod određenim uslovima metrika grafa d_G aproksimira geodezijsku metriku d_M . Kako bismo došli do tog rezultata potrebno je da uvedemo dodatne pojmove.

4.1.2 Definicija *Neka je data povezana d -dimenzionalna Riemann-ova mnogostrukost M . Za mnogostrukost M kažemo da je **konveksna** ako postoji injektivno preslikavanje iz M u $\mathbb{R}^d$, tako da je slika mnogostrukosti M konveksna oblast u $\mathbb{R}^d$.*

Veliki broj eksperimenata je pokazao da višedimenzionalni podaci leže na konveksnim mnogostrukostima. S obzirom da posmatramo konačne skupove podataka koji leže na nepoznatim mnogostrukostima ne možemo direktno izračunati geodezijska rastojanja između tačaka skupa. Stoga, računamo metriku grafa, koristeći sistem okolina (ε -okoline ili k -okoline).

Za krivu $\gamma(s)$, gdje je s prirodan parametar, *poluprečnik krivine* u $\gamma(s)$, u oznaci $r(\gamma, s)$, se definiše kao

$$\frac{1}{r(\gamma, s)} = \|\ddot{\gamma}(s)\|.$$

Minimalni poluprečnik krivine mnogostrukosti M definišemo sa:

$$r_0 = r_0(M) = \min_{\gamma \in \Gamma(G)} (\min_s r(\gamma, s)),$$

gdje je $\Gamma(G)$ skup svih geodezijskih krivih na M i s pripada domenu krive γ . *Minimalno gransko razdvajanje* (eng. minimum branch separation) mnogostrukosti M definišemo sa

$$s_0 = s_0(M) = \arg \sup_{x, y \in M} \{s : \|x - y\| < s \Rightarrow d_M(x, y) < \pi r_0(M)\}.$$

Kada je M kompaktna mnogostrukost, $r_0(M)$ i $s_0(M)$ su pozitivni brojevi.

Dokaz ključne teoreme, koju ovdje samo navodimo, a koja garantuje da se geodezijska metrika, pod određenim uslovima, može aproksimirati metrikom grafa, tj. da je $d_G \approx d_M$, može se naći u [1].

4.1.3 Teorema *Neka je $G = [\mathcal{X}, E]$ povezan graf i neka je $M \subset \mathbb{R}^D$ kompaktna, povezana mnogostrukost, na kojoj se nalaze podaci skupa $\mathcal{X}$. Za date $\varepsilon > 0$, $\delta > 0$, i brojeve $0 < \lambda_1, \lambda_2 < 1$ takve da je $4\delta < \lambda_2\varepsilon$ pretpostavimo da važi sljedeće:*

1. *Ako za $x, y \in \mathcal{X}$ važi $\|x - y\| \leq \varepsilon$ onda grana $(x, y) \in E$, tj. graf G sadrži sve grane (x, y) za koje važi $\|x - y\| \leq \varepsilon$.*
2. *Za svaku tačku $p \in M$ postoji $x \in \mathcal{X}$ tako da je $d_M(x, p) \leq \delta$.*
3. $\max_{(x, y) \in E} \|x - y\| < \min\left(s_0(M); \frac{2\pi r_0(M)}{\sqrt{24\lambda_1}}\right).$

Tada za sve $x, y \in \mathcal{X}$ važi sljedeća nejednakost:

$$(1 - \lambda_1)d_M(x, y) \leq d_G(x, y) \leq (1 + \lambda_2)d_M(x, y). \quad (4.1)$$

4.2 Isomap

Za preslikavanje kažemo da je *izometrija* (izometrično preslikavanje) ako očuvava rastojanje između tačaka skupa podataka. Prilikom Isomap (eng. Isometric mapping) metoda euklidsko rastojanje između dobijenih nižedimenzionalnih podataka jednako je geodezijskom rastojanju između podataka početnog skupa podataka. U narednim potpoglavljima daćemo opis Isomap metoda.

4.2.1 Isomap metod

Ako podaci skupa podataka nisu dovoljno gusti na mnogostrukosti M onda aproksimacija geodezijskih rastojanja metrikom grafa neće biti dobra, pa je potrebno primjeniti tehniku transliranja za pogodno izabranu konstantu vrijednost.

Pretpostavimo da podaci $\mathcal{X}$ leže na mnogostrukosti $M \subset \mathbb{R}^D$ i neka je $f : M \rightarrow \mathbb{R}^D$ izometrija takva da je $f(x_i) = y_i$. Za dati graf $G = [\mathcal{X}, E]$ označimo matricu metrike grafa $D_G = [d_G(i, j)]$, čiji su elementi $d_G(x_i, x_j)$ metrike grafa između tačaka x_i i x_j . Ako označimo sa $d_M(x_i, x_j)$ geodezijsku metriku između x_i i x_j , onda ako se tačke x_i, x_j nalaze u istoj okolini očigledno je da $d_G(x_i, x_j) \leq d_M(x_i, x_j)$. Dakle, naš cilj je da i matrica metrike grafa D_G dobro aproksimira matricu geodezijskog rastojanja D_M . Kao rezultat, matrica D_G dobro aproksimira i euklidsku metriku na skupu $\mathcal{Y} = \{y_1, \dots, y_n\}$, tj.

$$D_G \approx D := [d_2(y_i, y_j)].$$

Neka je $S = [d_G^2(x_i, x_j)]$ i $S^c = HCH$ njena centrirana matrica (Definicija 3.3.10). Sada iz Teoreme 3.3.11 imamo da je centrirana Gramova matrica skupa podataka $\mathcal{Y}$ data sa $G^c = -\frac{1}{2}S^c$. Ako primijenimo PCA metod na matricu G^c , dobijamo DR skup $\mathcal{Y}$. Matricu G^c zovemo *Isomap jezgro* (eng. Isomap kernel).

Kada početni skup podataka $\mathcal{X}$ nije dovoljno gust na mnogostrukosti M onda metrika grafa između pojedinih tačaka može biti prilično manja nego geodezijska rastojanja između istih. Stoga, matrica D_G neće biti dobra aproksimacija matrice D_M . Kao rezultat, možemo dobiti da matrica G^c nije pozitivno semi-definitna, odakle dobijamo da ona ne može biti Gramova matrica ni za jedan skup podataka. U tom slučaju hoćemo da primijenimo tehniku koja se koristi kada je potrebno da simetričnu matricu pretvorimo u simetričnu pozitivno definitnu matricu. Tu tehniku zovemo *translacija za konstantnu vrijednost* (eng. constant-shift technique) koju ćemo izložiti u nastavku. Ideja je da se doda nenegativan broj τ metrici grafa $d_G(x_i, x_j)$ i da posmatramo

$$d_\tau(x_i, x_j) = \begin{cases} d_G(x_i, x_j) + \tau, & i \neq j \\ 0, & i = j. \end{cases} \quad (4.2)$$

Nakon modifikacije, dobijamo da je

$$S_\tau = [d_\tau^2(x_i, x_j)]$$

odakle je

$$G_\tau^c = -\frac{1}{2}S_\tau^c.$$

Cilj je da pokažemo da je moguće izabrati $\tau \geq 0$ tako da je matrica G_τ^c simetrična, pozitivno definitna matrica i da je matricu G_τ^c modifikacija Isomap jezgra G^c . U nastavku, pokazaćemo da postoji takvo τ i daćemo formulu za određivanje te vrijednosti. Zapišimo $S_\tau = S + 2\tau D + \tau^2(E - I)$ gdje je kao i do sad matrica I jedinična matrica, a E je matrica čiji su svi elementi jednaki 1. Tada

$$\begin{aligned} G_\tau^c &= -\frac{1}{2}(HSH + 2\tau HDH + \tau^2 H(E - I)H) \\ &= \frac{1}{2}(\tau^2 H - 2\tau D^c - S^c) \end{aligned} \quad (4.3)$$

što je kvadratna forma po $\tau \geq 0$. Za dovoljno veliko τ matrica G_τ^c je pozitivno semidefinitna i riješili smo problem.

Kako bismo pronašli donje ograničenje za τ posmatrajmo karakteristične vrijednosti matrice G_τ^c . S obzirom na centriranost, uvijek postoji jedan karakteristični vektor $\mathbf{1} = [1, \dots, 1]'$ matrice G_τ^c koji odgovara karakterističnoj vrijednosti $\lambda_1 = 0$ (vidjeti Lemu 3.3.8). Znamo da onda svi ostali karakteristični vektori moraju biti ortogonalni¹ vektoru $\mathbf{1}$. Neka je λ_{min} najmanja karakteristična vrijednost matrice $-S^c$ i neka je η_{min} najmanja karakteristična vrijednost matrice $-D^c$. S obzirom na (4.3) neka je τ^+ pozitivan korijen kvadratne jednačine

$$\tau^2 + 2\tau\eta_{min} + \lambda_{min} = 0. \quad (4.4)$$

Rješavajući jednačinu (4.4) dobijamo

$$\tau^+ = \sqrt{\eta_{min}^2 - \lambda_{min}} - \eta_{min}.$$

Za svako $\tau \geq \tau^+$ matrica G_τ^c je pozitivno semidefinitna. Zaista, za svaki centriran vektor z imamo da je

$$z'G_\tau^c z \geq \|z\|^2(\tau^2 + 2\tau\eta_{min} + \lambda_{min}) \geq 0.$$

Odavde zaključujemo da minimalna vrijednost τ za koju je matrica G_τ^c pozitivno semidefinitna, nije veća od τ^+ . Tražena minimalna vrijednost se može odrediti analitički koristeći sljedeću teoremu.

¹Karakteristični vektori realne simetrične matrice su ortogonalni.

4.2.1 Teorema *Minimalna vrijednost τ^* za koju je matrica G_τ^c simetrična pozitivno definitna matrica je najveća karakteristična vrijednost matrice*

$$B = \begin{bmatrix} 0 & -2S^c \\ -I & 2D^c \end{bmatrix}.$$

Dokaz. Za dati vektor z definišimo polinom $h(\tau) := z'G_\tau^c z$. Tada postoji $\tau(z)$ tako da je $h(\tau(z)) = 0$ i $h(\tau) > 0$ za sve $\tau > \tau(z)$. Ako je $h(\tau) > 0$ za sve τ onda ćemo uzeti da je $\tau(z) = -\infty$. Takođe očigledno je da je $\tau(z) = \tau(kz)$ za sve $k \in \mathbb{R} \setminus \{0\}$. S obzirom da G^c nije pozitivno semidefinitna, onda postoji bar jedno z_0 za koje je $z_0'G^c z_0 < 0$ i $\tau(z_0) > 0$. S druge strane, iz prethodnog izvođenja je $\tau(z) \leq \tau^*$ za sve z . Neka je

$$\tau^* = \sup_x \tau(x). \quad (4.5)$$

Tada je $0 < \tau^* \leq \tau^+$. Pošto je polinom neprekidna funkcija, onda je i $\tau(z)$ neprekidna funkcija po z pa postoji z^* tako da je $\tau(z^*) = \tau^*$. Imamo sljedeće:

$$z'G_\tau^c z \geq 0 \text{ za sve } z, \tau \geq \tau^*, \quad (4.6)$$

$$(z^*)'G_{\tau^*}^c z^* = 0. \quad (4.7)$$

Ostalo je još pokazati da je τ^* najveća karakteristična vrijednost matrice B .

1. τ^* je karakteristična vrijednost matrice B . Neka je $w = [y, Hz^*]'$ gdje je $y = -\tau^* S^c z^*$. Tada imamo da je

$$-S^c z^* = \tau^* y, \quad (4.8)$$

odakle koristeći (4.3) i (4.7) dobijamo

$$(\tau^{*2}H - 2\tau^*D^c - S^c)z^* = 0. \quad (4.9)$$

Iz (4.8) i (4.9) dobijamo

$$y + \tau^* Hz^* - 2D^c Hz^* = 0.$$

Iz (4.7) i (4.9) slijedi

$$\begin{bmatrix} 0 & -2S^c \\ -I & 2D^c \end{bmatrix} \begin{bmatrix} y \\ Hz^* \end{bmatrix} = \tau^* \begin{bmatrix} y \\ Hz^* \end{bmatrix},$$

čime smo dokazali da je τ^* karakteristična vrijednost matrice B .

2. τ^* je najveća karakteristična vrijednost matrice B . Neka je σ proizvoljna karakteristična vrijednost matrice B kojoj odgovara karakteristični vektor $[v, u]'$:

$$-S^c u = \sigma v, \quad (4.10)$$

$$-v + 2D^c u = \sigma u. \quad (4.11)$$

Iz (4.10) imamo da je vektor v centriran jer pripada $-S^c$. Slično, kako su u (4.11) $2D^c u$ i v centrirani i u je centriran vektor. Stoga, $u = Hu$ pa imamo da je

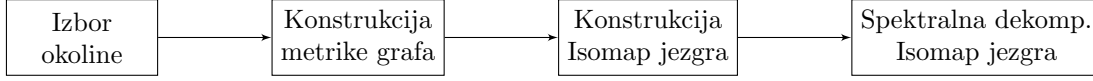
$$(-S^c - 2\sigma D^c + \sigma^2 H)u = 0,$$

odakle imamo da je $u'G_{\sigma}^c u = 0$, tj. $\tau(u) = \sigma$. Tada zbog (4.5) imamo da je $\sigma \leq \tau^*$.

Time smo pokazali tvrđenje. $\square$

4.2.2 Isomap algoritam

Isomap algoritam koristi ideju prikazanu na Slici 4.2. Algoritam se sastoji od četiri koraka. Generalno, cilj je da izračunamo matricu metrike grafa na skupu $\mathcal{X}$ koja aproksimira geodezijska rastojanja na mnogostrukosti, a zatim primijenimo neki linearan metod, npr. PCA, kako bismo našli nižedimenzionalni skup podataka koji očuvava metriku grafa.



Slika 4.2: Koraci Isomap algoritma.

Isomap algoritam sastoji se od četiri koraka.

1. **Izbor okoline.** Okolinu tačke x_i skupa $\mathcal{X} = \{x_1, \dots, x_n\}$ možemo konstruisati birajući ε -ili k -okolinu, pa stoga imamo ε -Isomap ili k -Isomap, u zavisnosti koju okolinu izaberemo. Okolinu tačke x_i , kao i do sad, označavamo sa O_i . U praksi se češće koristi sistem k -okolina.
2. **Računanje matrice metrike grafa.** Izabrani sistem okolina formira graf $G = [\mathcal{X}, E]$ na skupu podataka $\mathcal{X}$. Zatim, za svaki par tačaka x_i, x_j iz $\mathcal{X}$ računamo metriku grafa $d_G(x_i, x_j)$. Ako izaberemo k -okoline onda $d_G(x_i, x_j)$ ne mora biti jednako sa $d_G(x_j, x_i)$ jer je poznato da $x_i \in O_j \not\Rightarrow x_j \in O_i$ u slučaju k -okolina. Kako bismo osigurali da je $d_G(x_i, x_j) = d_G(x_j, x_i)$ možemo definisati da je metrika grafa data sa $\min(d_G(x_i, x_j), d_G(x_j, x_i))$ ili možemo posmatrati neusmjeren graf umjesto digrafa, što je u praksi češće. Ako ne postoji grana koja spaja tačke x_i i x_j smatraćemo da je $d_G(x_i, x_j) = \infty$.
3. **Konstrukcija Isomap jezgra.** Pretpostavimo da su podaci skupa $\mathcal{X}$ dovoljno gusti. Tada je matrica metrike grafa D_G dobra aproksimacija euklidskog rastojanja. Izračunamo matrice $S_G = [d_G^2(x_i, x_j)]$ i $G^c = -\frac{1}{2}HS_GH$, gdje je G^c Isomap jezgro. Ako podaci ipak nisu dovoljno gusti, vidjeli smo da je onda potrebno da translirano podatke i da obezbjedimo da je Isomap jezgro G^c pozitivno definitna matrica. U tom slučaju se 3. korak ovog algoritma sastoji iz dva dijela:

- 3.1. **Računanje konstante transliranja.** Prvo, centriramo matrice D_G i S_G koristeći formule

$$\begin{aligned} F^c &= -\frac{1}{2}HD_GH \\ G^c &= -\frac{1}{2}HS_GH, \end{aligned}$$

a zatim ćemo izračunati najveću karakterističnu vrijednost t matrice

$$B = \begin{bmatrix} 0 & 2G^c \\ -I & -4F^c \end{bmatrix}.$$

- 3.2. **Računanje matrice G_t^c .** Izračunamo

$$G_t^c = G^c + 2tF^c + \frac{1}{2}t^2H$$

što će biti zamjena za Isomap jezgro.

4. **Spektralna dekompozicija Isomap jezgra.** Neka je spektralna dekompozicija matrice G^c data sa

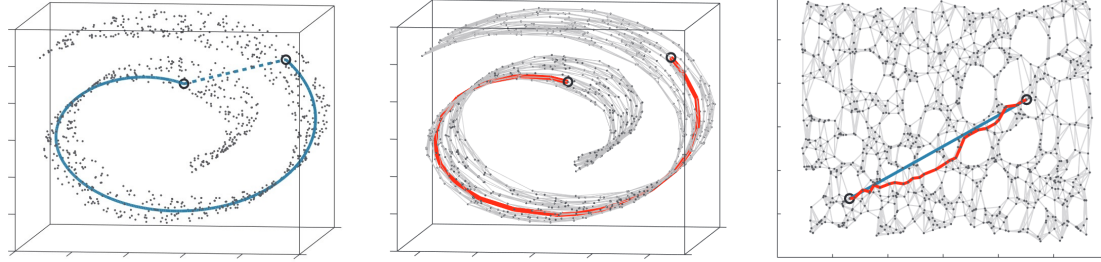
$$G^c = V\Lambda V'$$

gdje je $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d, \dots, \lambda_n)$ i $V = [v_1, \dots, v_n]$. Neka je $Y = [y_1, \dots, y_n]$ određeno sa

$$Y' = [\sqrt{\lambda_1}v_1, \dots, \sqrt{\lambda_d}v_d].$$

Tada je Y redukcija dimenzionalnosti za X . Koristeći Teoremu 3.3.16 dobijamo da je ocjena greške:

$$\frac{1}{2n} \sum_{i,j=1}^n (d_G^2(x_i, x_j) - d_2^2(x_i, x_j)) \leq \sum_{l=d+1}^n \lambda_l.$$



(a) Euklidsko i geodezijsko rastojanje na primjeru rolne

(b) Graf G za $n = 1000$ tačaka koristeći k -okoline za $k = 7$ konstruisan u 1. koraku Isomap algoritma

(c) U 3. i 4. koraku Isomap algoritma dobijamo sliku podataka u $2D$ gdje je plavom bojom označeno euklidsko rastojanje, a crvenom bojom aproksimacija metrikom grafa

Slika 4.3: Ilustracija Isomap algoritma.

4.2.3 Matlab kod za Isomap algoritam

Kako bismo izračunali metriku grafa, tj. rastojanja između svaka dva čvora u dobijenom grafu koristi se nekoliko različitih algoritama. Jedan od njih je Floyd-ov algoritam koji traži najkraće rastojanje između svaka dva čvora u grafu, ali ako imamo n tačaka potrebno je $O(n^3)$ računanja, odnosno računaska složenost je $O(n^3)$. Postoji i drugi algoritam, tzv. Dijkstra algoritam koji zahtijeva $O(n^2 \log n)$ računanja, tj. ima manju računsku složenost od Floyd-ovog algoritma. Dijkstra algoritam se koristi u oblast linearnog programiranja i detaljan opis može se naći u [23, 25].

Na veb-stranici <http://isomap.stanford.edu/> može se naći rad pod nazivom „A Global Geometric Framework for Nonlinear Dimensionality Reduction” kao i detaljan dokaz konvergencije Isomap algoritama (vidjeti [22]). Osim toga, može se naći i Matlab toolbox sa uputstvom za upotrebu kao i skupovi podataka na kojima se mogu primjeniti Isomap algoritmi. Slika 4.3 se takođe nalazi na navedenoj veb-stranici. Pored standardnog Isomap algoritma, nalazi se i Dijkstra algoritam. U ovom radu ćemo navesti modifikaciju algoritma iz ovog rada, a može se naći u [25].

```

1 function Y = Isomap(X_data, ndims, options)
2 % Isomap DR
3 % Sintaksa: Y = Isomap(X_data, ndims, options);
4
5 % ULAZNE VRIJEDNOSTI:
6 % 'X': D x N matrica (D = dimension, N = #points)
7 % 'ndims': zeljena dimenzija izlaznog skupa podataka.
8 % OPTIONS: :
9 % 'options.epsilon': koristimo 'epsilon'-okoline;
10 % 'options.k': koristimo 'k'-okoline;
11 % 'options.verb': prikazimo postupak tekstualno
12 %                 (1 = DA (default), 0 = NE);

```

```

13 % POZVATI: L2_distance
14 %           dijkstra
15 %           dijk
16
17 % Reference:
18 % Tenenbaum, de Silva, and Langford,
19 % Isomap code — (c)1998–2000, Josh Tenenbaum
20
21 % Inicijalizacija:
22 options.symmetric = 1;
23 if isfield(options, 'verb')
24     verb = options.verb;
25 else
26     verb = 1;
27 end
28
29 % 1. korak: Konstrukcija grafa podataka.
30 if verb
31     disp('—Konstruisemo graf');
32 end
33 D2 = L2_distance(X_data(:,1:1000), X_data(:,1:1000), 1)
34
35 % 2. korak: Racunanje matrice metrike grafa.
36 if verb
37     disp('— Racunamo matricu metrike grafa');
38 end
39 N = size(D2,1);
40 D1 = sqrt(D2);
41
42 % Iskoristimo .dll za brze izracunavanje:
43 if exist('dijkstra.dll','file')
44     DG = dijkstra(D1, 1:N);
45 else
46     DG = dijk(D1, 1:N);
47 end
48
49 % 3. korak: Generisanje Isomap jezgra:
50 if verb
51     disp('— Generisemo Isomap jezgro');
52 end
53 DG = DG.^2;
54 GC = -.5*(DG - sum(DG)'*ones(1,N)/N - ones(N,1)*sum(DG)/N + sum(DG(:))/(N^2));
55
56 % 4. korak: Spektralna dekompozicija Isomap jezgra:
57 if verb
58     disp('— Izlazni nizedimenzionalni podaci');
59 end
60 opt.display = 0;
61 opt.isreal = 1;
62 opt.issym = 1;
63 [Y, val] = eigs(GC, ndims, 'LR', opt);
64 for i=1:ndims
65     Y(:,i) = Y(:,i)*sqrt(val(i,i));
66 end

```

Primjetimo da smo u prethodnom kodu iskoristili gotove Matlab kodove `L2_distance`, `dijkstra` i `dijk` koji se nalaze u sklopu fajla "IsomapR1.tar", koji se takođe nalazi na gore navedenoj veb-stranici. Osim toga, skup podataka rolne nalazi se u fajlu "swiss_roll_data". Jedan primjer je moguće pokrenuti na sljedeći način:

```

>> load swiss_roll_data;
>> options.verb = 1;
>> Y = Isomap(X_data, 2, options).

```

4.3 Lokalno linearno potapanje - LLE

Druga nelinearna tehnika koju ćemo razmatrati u ovom master radu jeste *lokalno linearno potapanje - LLE* (eng. Locally Linear Embedding). Tehnike su razvili Roweis S.T. i Saul L.K. u radovima [16, 17]. Ideja LLE metoda je u sljedećem: Ako se podaci nalaze na glatkoj nelinearnoj mnogostrukosti, tada sve tačke koje pripadaju istoj okolini, nakon redukcije dimenzionalnosti ponovo ostaju blizu i u nižedimenzionalnom prostoru. Pošto je poznato da su mnogostrukosti objekti koji lokalno imaju strukturu euklidskog prostora, onda ćemo nižedimenzionalne podatke konstruisati tako da se očuvaju lokalna svojstva originalnog skupa podataka.

Pretpostavimo da podaci iz skupa $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ leže na glatkoj mnogostrukosti i da je struktura okoline skupa podataka definisana tako da svaka okolina leži na ili u blizini linearne karte mnogostrukosti. Kod LLE metoda, lokalna geometrija svake karte je opisana pomoću linearnih koeficijenata koji približno rekonstruišu svaku tačku skupa podataka $\mathcal{X}$ koristeći njenu okolinu. LLE algoritam računa nižedimenzionalni skup podataka ali tako da se očuva lokalna konfiguracija skupa podataka.

Primjetimo da smo i kod Isomap metoda koristili lokalnu geometriju skupa podataka, ali smo Isomap jezgro računali koristeći se globalnom strukturom kako bismo očuvali rastojanja između tačaka. Pored toga, koristili smo geodezijska rastojanja, koja smo morali izračunati između svake dvije tačke skupa podataka. Takođe smo kao rezultat dobijali gustu matricu jezgra. S druge strane, LLE metod konstruiše matricu jezgra koristeći direktno lokalnu geometriju skupa podataka i kao rezultat dobija se rijetka matrica jezgra.

4.3.1 Baricentrične koordinate

LLE metod koristi *baricentrične koordinate* svake okoline kako bi konstruisao nižedimenzionalnu reprezentaciju.

4.3.1 Definicija Skup tačaka $\{p_1, \dots, p_{m+1}\} \subset \mathbb{R}^m$ je u opštem položaju ako i samo ako je skup $\{p_2 - p_1, p_3 - p_1, \dots, p_{m+1} - p_1\}$ linearno nezavisan.

4.3.2 Definicija Ako je skup tačaka $\{p_1, \dots, p_{m+1}\} \subset \mathbb{R}^m$ u opštem položaju, skup

$$\sigma = (p_1, \dots, p_{m+1}) = \left\{ \sum_{i=1}^{m+1} w_i p_i : \sum_{i=1}^{m+1} w_i = 1, w_i \geq 0 : \forall i \leq m+1 \right\}$$

zovemo $(m+1)$ -*dimenzionalni simpleks* u $\mathbb{R}^m$. Tačke p_i su **tjemena**, a $m+1$ je dimenzija simpleksa σ .

Pretpostavimo da skup $\mathcal{P} = \{p_1, \dots, p_{m+1}\} \subset \mathbb{R}^m$ generiše simpleks u $\mathbb{R}^m$ sa tjemena $p_1, \dots, p_{m+1}$. Tada svaka tačka p unutar politopa² može na jedinstven način da se izrazi pomoću koeficijenata $w_1, \dots, w_{m+1}$ kao

$$p = w_1 p_1 + \dots + w_{m+1} p_{m+1}$$

uz uslov da je $\sum_{i=1}^{m+1} w_i = 1$. Koeficijente $w_1, \dots, w_{m+1}$ zovemo **baricentrične koordinate** tačke p (u odnosu na simpleks $\mathcal{P}$). $\frac{\sum_{i=1}^{m+1} p_i}{m+1} \in \mathcal{P}$ je **baricentar** (težište) simpleksa $\mathcal{P}$.

Baricentrične koordinate su invarijantne u odnosu na izometrije euklidskog prostora i izometrična potapanja. Ako posmatramo svaku baricentričnu koordinatu w_i kao funkciju od p , tj. $w_i = w_i(p)$ tada funkcije $w_i(p)$ ima sljedeće osobine:

1. *Linearnost*: $w_i(\lambda p + (1 - \lambda)q) = \lambda w_i(p) + (1 - \lambda)w_i(q)$;

²Politop je ograničena konveksna oblast n -dimenzionalnog prostora ograničeno sa konačno mnogo ravni. Politop za $n = 2$ je poligon, a za $n = 3$ poliedar.

2. *Pozitivnost:* Ako je tačka p strogo unutar simpleksa onda je $w_i(p) > 0$;
3. *Lagranžovo³ svojstvo interpolacije:* Za svako tjeme p_j , $w_i(p_j) = \delta_{ij}$, gdje je δ_{ij} Kronecker-ov δ -simbol definisan sa $\delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases}$

Kako baricentrične koordinate w_i , $i \in \{1, \dots, m+1\}$, „privlače” tačku p u blizinu tačke p_i , imamo da one opisuju lokalne sličnosti podataka. S obzirom na navedene osobine, u baricentričnim koordinatama svako linearno preslikavanje F u tački p ima sljedeću reprezentaciju:

$$F(p) = \sum_{i=1}^{m+1} w_i F(p_i).$$

Stoga, ako imamo linearnu transformaciju ϕ maksimalnog ranga, koja preslikava skup $\{p_1, \dots, p_{m+1}\}$ na skup $\{q_1, \dots, q_{m+1}\}$ sa $\phi(p_i) = q_i$, $i \in \{1, \dots, m+1\}$, tada je

$$q := \phi(p) = \sum_{i=1}^{m+1} w_i q_i,$$

pa se baricentrične koordinate očuvavaju prilikom linearnih transformacija.

4.3.2 LLE metod

Kao što smo kod Isomap algoritma konstruisali Isomap jezgro, slično ćemo i kod LLE algoritma prvo razmotriti kako da konstruišemo LLE jezgro. Neka su dati podaci $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$ koji se nalaze na d -dimenzionalnoj mnogostrukosti $M \subset \mathbb{R}^D$ i neka je matricna reprezentacija skupa podataka $\mathcal{X}$ data kao $X = [x_1, \dots, x_n]_{D \times n}$. Pretpostavimo da za svaku tačku iz skupa podataka $\mathcal{X}$ imamo dobro definisan sistem okolina (ε - ili k -okoline). Tada okoline skupa tačaka definišu graf $G = [\mathcal{X}, A]$, gdje je A matrica povezanosti grafa G . Indeksni skup, susjednih tačaka tačke x_i , označavaćemo sa $N(i) = N_i$ i predstavljace i -tu vrstu matrice A tako da je

$$N(i) = \{j : A_{i,j} \neq 0\}$$

i kao i do sad, okolina tačke x_i je skup

$$O(i) = \{x_j \in \mathcal{X} : j \in N(i)\}.$$

Dalje, konstruišemo težinsku $n \times n$ matricu W na grafu kako bismo mogli definisati sličnosti između skupova podataka. S obzirom da tačku $x_j \notin O_i$ smatramo različitom od tačke x_i , definišaćemo da je $w_{i,j} = 0$ ako je $A_{i,j} = 0$. Kako bismo definisali $w_{i,j}$ za svako $j \in N_i$, biraćemo da su $w_{i,j}$ baricentrične koordinate tačke x_i u odnosu na svoju okolinu $\{x_j\}_{j \in O(i)}$, gdje su tačke x_i i x_j „blizu” jedna drugoj na mnogostrukosti M . Geometrijska struktura okoline O_i može da se aproksimira svojom projekcijom na tangentni prostor T_{x_i} . Neka je f ortogonalna projekcija sa M na T_{x_i} i označimo da je $\tilde{x} = f(x)$. Skup $\{\tilde{x}_j - \tilde{x}_i : j \in N(i)\}$ razapinje potprostor tangentnog prostora T_{x_i} , odakle zaključujemo da je skup težinskih koeficijenata dat sa

$$\left\{ w_{i,j} \in \mathbb{R} : j \in N(i) \text{ i } \sum_{j \in N(i)} w_{i,j} = 1 \right\},$$

pri čemu je $\sum_{j \in N(i)} w_{i,j} \tilde{x}_j = \tilde{x}_i$. Primjetimo da težinski koeficijenti mogu biti i negativni, ako tačka $\tilde{x}_i$ nije u politopu koji formira skup $\{\tilde{x}_j\}_{j \in N(i)}$. Neka je stepen čvora x_i označen sa $d_i = |N(i)|$

³Joseph-Louis Lagrange (1736-1813) - italijanski matematičar i astronom

i neka je i -ta vrsta matrice W označena sa $w_i = [w_{i,1}, \dots, w_{i,n}]$, što je rijetki vektor koji ima najviše d_i nenula elemenata. Radi jednostavnosti, definišimo d_i -dimenzionalni podvektor vektora w_i sa

$$w_i^{sub} = [w_{i,1}, \dots, w_{i,j_{d_i}}], \quad j_s \in N(i). \quad (4.12)$$

Skup podvektora $\{w_i^{sub}\}_{i=1}^n$ zajedno sa indeksnim skupom $\{N(i)\}_{i=1}^n$ jedinstveno određuje težinsku matricu W . Kako bismo olakšali izračunavanje težinskih koeficijenata, nećemo koristiti projekciju f , nego ćemo tražiti podvektor w_i^{sub} koji minimizira grešku:

$$(w_i^{sub}) = \arg \min_{a \in \mathbb{R}^{d_i}} \left\| x_i - \sum_{j \in N(i)} a_j x_j \right\| \quad \text{tako da je} \quad \sum_{j=1}^{d_i} a_j = 1, \quad (4.13)$$

gdje je $a = [a_1, \dots, a_{d_i}]$. Dakle, težinski koeficijenti $w_{i,j}$ koristeći $x_j \in N(i)$ dovoljni su da rekonstruišu x_i . Otuda, minimiziramo funkciju cilja u (4.13) uz uslove:

1. *Uslov rijetkosti:* Svaka tačka x_i je rekonstruisana koristeći samo tačke iz svoje okoline $O(i)$ pa je $w_{i,j} = 0$ ako $x_j \notin O(i)$.
2. *Uslov invarijantnosti:* Suma svake vrste težinske matrice W je jednaka 1, tj. $\sum_{j=1}^n w_{i,j} = 1$.

Uslov invarijantnosti je važan kako bi se očuvala lokalna geometrija skupa podataka. Daćemo sada matematičko objašnjenje uloge ovog uslova. Intuitivno, težinski koeficijenti bi trebali da budu invarijantni u odnosu na koordinate prostora $\mathbb{R}^D$.

4.3.3 Teorema *Težinski koeficijenti za x_i koje dobijamo rješavajući (4.13) su invarijantni u odnosu na translaciju, dilataciju i rotaciju skupa podataka $\mathcal{X}$.*

Dokaz. Označimo rotaciju sa T i za $\lambda \neq 0$ operator dilatacije sa D_λ , koji je definisan kao $D_\lambda x = \lambda x$. S obzirom na linearnost operatora T i D_λ imamo da je

$$\sum_{j \in N(i)} w_{i,j} (Tx_j) = T \left(\sum_{j \in N(i)} w_{i,j} x_j \right)$$

i

$$\sum_{j \in N(i)} w_{i,j} (D_\lambda x_j) = \lambda \left(\sum_{j \in N(i)} w_{i,j} x_j \right) = D_\lambda \left(\sum_{j \in N(i)} w_{i,j} x_j \right).$$

Označimo sa $\{w_{i,j}^*\}_{j \in N(i)}$ skup težinskih koeficijenata koji su rješenje problema (4.13). Tada po definiciji rotacije, znamo da je T ortogonalna matrica pa je

$$\begin{aligned} (\tilde{w}_{i,j}) &= \arg \min_{\sum w_{i,j}=1} \left\| Tx_i - \sum_{j \in N(i)} w_{i,j} Tx_j \right\| \\ &= \arg \min_{\sum w_{i,j}=1} \left\| T \left(x_i - \sum_{j \in N(i)} w_{i,j} x_j \right) \right\| \\ &= \arg \min_{\sum w_{i,j}=1} \left\| x_i - \sum_{j \in N(i)} w_{i,j} x_j \right\| = (w_{i,j}^*) \end{aligned}$$

i slično za dilataciju imamo da je

$$\begin{aligned}
 (\tilde{w}_{i,j}) &= \arg \min_{\sum w_{i,j}=1} \left\| \lambda x_i - \sum_{j \in N(i)} w_{i,j} \lambda x_j \right\| \\
 &= |\lambda| \arg \min_{\sum w_{i,j}=1} \left\| x_i - \sum_{j \in N(i)} w_{i,j} x_j \right\| \\
 &= \arg \min_{\sum w_{i,j}=1} \left\| x_i - \sum_{j \in N(i)} w_{i,j} x_j \right\| = (w_{i,j}^*).
 \end{aligned}$$

Time smo pokazali tvrđenje za rotaciju i dilataciju, pa ostaje još da pokažemo i za translaciju. Pretpostavimo da je svaki vektor u $\mathbb{R}^D$ transliran za vektor $a \in \mathbb{R}^D$. Dobijamo da je

$$\begin{aligned}
 (\tilde{w}_{i,j}) &= \arg \min_{\sum w_{i,j}=1} \left\| (x_i - a) - \sum_{j \in N(i)} w_{i,j} (x_j - a) \right\| \\
 &= \arg \min_{\sum w_{i,j}=1} \left\| x_i - a - \sum_{j \in N(i)} w_{i,j} x_j + \left(\sum_{j \in N(i)} w_{i,j} \right) a \right\| \\
 &= \arg \min_{\sum w_{i,j}=1} \left\| x_i - a - \sum_{j \in N(i)} w_{i,j} x_j + a \right\| \\
 &= \arg \min_{\sum w_{i,j}=1} \left\| x_i - \sum_{j \in N(i)} w_{i,j} x_j \right\| = (w_{i,j}^*).
 \end{aligned}$$

Time smo pokazali tvrđenje, da su težinski koeficijenti za LLE metod invarijantni u odnosu na translaciju, rotaciju i dilataciju. $\square$

Sumu

$$\sum_{j \in N(i)} w_{i,j} x_j \tag{4.14}$$

zovemo *lokalno linearno potapanje* (LLE) za x_i . Težinski koeficijenti u prethodnoj sumi su jedinstveni. Zaista, definišimo sljedeće:

$$\mathcal{E}_i(W) := \left\| x_i - \sum_{j \in N(i)} w_{i,j} x_j \right\|^2 = \left\| \sum_{j \in N(i)} w_{i,j} (x_i - x_j) \right\|^2$$

i

$$\mathcal{E}(W) := \sum_{i=1}^n \mathcal{E}_i(W) = \sum_{i=1}^n \left\| \sum_{j \in N(i)} w_{i,j} (x_i - x_j) \right\|^2.$$

Neka je kao i do sad, $\mathbf{1} = [1, \dots, 1]'$ n -dimenzionalni vektor. Tada je LLE težinska matrica W^* rješenje problema minimizacije:

$$W^* = \arg \min_{W \mathbf{1} = \mathbf{1}} \mathcal{E}(W),$$

ili ekvivalentno rješavamo problem

$$\arg \min_{w_{i,j}} \mathcal{E}_i(W) \text{ uz uslov } \sum_{j \in N(i)} w_{i,j} = 1 \tag{4.15}$$

za svako $1 \leq i \leq n$. Dakle, možemo da rješavamo problem posebno za svako i . Rješavanje problema (4.15) vodi nas u varijacioni račun, tačnije dobijamo Ojler⁴-Lagranžovu jednačinu

$$\sum_{j \in N(i)} c_{s,j} w_{i,j} - \lambda = 0, \quad s \in N(i) \quad (4.16)$$

gdje je λ Lagranžov množilac i $c_{s,j} = \langle x_j - x_i, x_s - x_i \rangle, j, s \in N(i)$.

Označimo smo sa $k := k(i) = |N(i)|$. Tada je po definiciji okoline $k > d$ gdje je d željena dimenzija nakon redukcije dimenzionalnosti. Neka je $C = (c_{s,j})_{k \times k}$ i definišimo $(w_i^{sub}) = [w_{ij_1}, \dots, w_{ij_k}]$ kao u (4.12). Tada jednačinu (4.16) možemo zapisati u matricnom obliku kao

$$C(w_i^{sub})' = \lambda \mathbf{1}, \quad \mathbf{1} = [1, \dots, 1] \in \mathbb{R}^k. \quad (4.17)$$

Diskutovaćemo jedinstvenost rješenja u (4.17) u zavisnosti od ranga matrice C .

1. **C je maksimalnog ranga.** U ovom slučaju, w_i^{sub} možemo dobiti stavljajući da je $\lambda = 1$ i normiranjem težinskog vektora, tj. dijeljenjem sa $\sum_{j \in N(i)} w_{i,j}$. Ekvivalentno je da izaberemo $\lambda = 1/(\sum_{j \in N(i)} w_{i,j})$. U ovom slučaju skup $\{x_j\}_{j \in N(i)}$ se nalazi na $(k-1)$ -dimenzionalnoj hiperravni $S \subset \mathbb{R}^D$ i x_i se ne nalazi na S . Označimo projekciju vektora x_i na S sa $\tilde{x}_i$. Tada su težinski koeficijenti $\{w_{i,j}\}_{j \in N(i)}$ baricentrične koordinate tačke $\tilde{x}_i$ u odnosu na skup $\{x_j\}_{j \in N(i)}$.
2. **rang(C) = k - 1.** Ovdje imamo da je 0 karakteristična vrijednost matrice C . Tada biramo da je $(w_i^{sub})'$ karakteristični vektor matrice C koji odgovara 0-karakterističnoj vrijednosti, normalizovan sa $\sum_{j \in N(i)} w_{i,j} = 1$. Pod ograničenjem $\sum_{j \in N(i)} w_{i,j} = 1$ rješenje w_i^{sub} je jedinstveno. U ovom slučaju, skup $\{x_j\}_{j \in N(i)}$ se nalazi na $(k-1)$ -dimenzionalnoj hiperravni S , ali se i tačka x_i sada nalazi na S . Stoga, težinski koeficijenti $\{w_{i,j}\}_{j \in N(i)}$ su baricentrične koordinate tačke x_i u odnosu na skup $\{x_j\}_{j \in N(i)}$.
3. **rang(C) < k - 1.** U ovom slučaju imamo da je 0 višetruka karakteristična vrijednost matrice C . Ponovo možemo da nađemo vektor koji odgovara 0-karakterističnoj vrijednosti koristeći uslov normiranosti $\sum_{j \in N(i)} w_{i,j} = 1$, ali u ovom slučaju nemamo jedinstveno rješenje.

Konačno, cilj nam je da dobijemo nižedimenzionalne podatke $\mathcal{Y} = \{y_1, \dots, y_n\} \subset \mathbb{R}^d$. Njih određujemo tako da se lokalno linearno potapanje očuva što više. Odnosno, hoćemo da $\mathcal{Y}$ minimizira grešku dobijenu prilikom lokalnog linearnog potapanja (vidjeti (4.14)):

$$\mathcal{Y} = \arg \min_{\mathcal{Y} \in \mathbb{R}^d} \sum_{i=1}^n \left\| y_i - \sum_{j \in N(i)} w_{i,j} y_j \right\|^2 \quad (4.18)$$

Prvo, po konstrukciji sistema okolina znamo da se u k -okolini svake tačke nalazi više tačaka nego što je dimenzije d , tj. $k > d$. Stoga, graf $G = [\mathcal{X}, A]$ nema izolovanih tačaka. Osim toga, pretpostavićemo da je graf povezan, pa je 0 karakteristična vrijednost Laplasijana $L = I - W$, a odgovarajući karakteristični vektor je $\frac{1}{\sqrt{n}} \mathbf{1}$. Ako je $\mathcal{Y}$ rješenje problema (4.18), zapišimo podatke u obliku realne matrice $Y = [y_1, \dots, y_n]_{d \times n}$. S obzirom da je rješenje problema invarijantno u odnosu na translaciju, možemo smatrati da su podaci centrirani, odnosno da je $\sum_{i=1}^n y_i = 0$, što je ekvivalentno sa uslovom

$$Y \mathbf{1} = 0. \quad (4.19)$$

Dalje, s obzirom na to da na koordinatne karte možemo gledati kao na način da svaku tačku iz neke okoline, sa mnogostrukosti na kojoj leže podaci, izrazimo kao koordinate u euklidskom prostoru,

⁴Leonhard Euler (1707-1783) - švajcarski matematičar, astronom i fizičar

tražićemo da za $x_i \in \mathcal{X}$ koordinate budu baš y_i za $1 \leq i \leq n$. Stoga, možemo da zahtjevamo da vektori vrste matrice Y formiraju ortonormirani sistem. Odatle, dobijamo drugi uslov

$$YY' = I_d,$$

gdje je I_d jedinična $d \times d$ matrica. Sada ćemo riješiti problem (4.18) uz navedena ograničenja. Neka je

$$\begin{aligned} \Psi &:= \sum_{i=1}^n \left\| y_i - \sum_{j \in N(i)} w_{i,j} y_j \right\|^2 = \sum_{i=1}^n \left\langle y_i - \sum_{j=1}^n w_{i,j} y_j, y_i - \sum_{j=1}^n w_{i,j} y_j \right\rangle \\ &= \sum_{i=1}^n \left\langle \sum_{j=1}^n w_{i,j} (y_i - y_j), \sum_{j=1}^n w_{i,j} (y_i - y_j) \right\rangle \\ &= \text{tr} (Y(I - W)'(I - W)Y') \end{aligned}$$

odakle, ako označimo Laplasijan sa $L = I - W$, dobijamo

$$\Psi = \text{tr} (YL'LY'), \quad (4.20)$$

pa je LLE jezgro dato sa

$$K = L'L = (I - W)'(I - W).$$

Neka je spektralna dekompozicija matrice K data sa

$$K = U\Lambda U'$$

gdje je $\Lambda = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_{n-1})$ sa karakterističnom vrijednostima u opadajućem poretku tako da $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$. Imamo da je

$$\Psi = \text{tr} (YL'LY) = \text{tr} (YU\Lambda U'Y') = \text{tr} (YU\Lambda(YU)').$$

Neka je $U_d = [u_1, \dots, u_d]$ podmatrica matrice U , koja se sastoji od karakterističnih vektora koji odgovaraju karakterističnim vrijednostima počevši od druge pa do $(d+1)$ najmanje karakteristične vrijednosti matrice K . S obzirom da je $K\mathbf{1} = 0$ imamo da skup vektora $\{u_1, \dots, u_d\}$ zadovoljavaju tražena ograničenja, pa je tražena redukcija dimenzionalnosti baš $Y = (U_d)'$, jer se tada dostiže minimum za Ψ .

4.3.3 LLE algoritam

LLE algoritam ćemo izložiti u četiri koraka.

1. **Definicija okoline.** Za okolinu svake tačke možemo birati ε - ili k -okoline. Broj tačaka iz okoline biramo da bude veći od željene dimenzije d . Konstruisani graf označimo sa $G = [\mathcal{X}, A]$, gdje je A matrica povezanosti. U praksi se uglavnom koristi sistem k -okolina. U tom slučaju matrica A može biti asimetrična.
2. **Konstrukcija težinske matrice.** Potrebno je da na grafu $G = [\mathcal{X}, A]$ generišemo težinsku matricu W . Ako je $A_{i,j} = 0$ stavimo $w_{i,j} = 0$, a ako je $A_{i,j} = 1$ izračunavamo $w_{i,j}$ rješavajući problem minimizacije (4.13). Za rješavanje problema (4.13) koristi se metod najmanjih kvadrata. Tačnije, neka je G_i Gramova matrica za skup $O(i)$, tj.

$$G_i = [\langle x_j - x_i, x_k - x_i \rangle]_{j,k \in N(i)}.$$

Onda u zavisnosti od ranga matrice G_i , ako je d_i stepen čvora i u grafu G , rješavamo sistem linearnih jednačina:

$$G_i w'_i = \begin{cases} 0, & \text{rang}(G_i) < d_i \\ 1, & \text{rang}(G_i) = d_i \end{cases}$$

i na kraju normalizujemo vektor dijeljenjem sa $\sum_{j \in N(i)} w_{i,j} = 1$ (vidjeti Poglavlje 2.2.3).

3. **Konstrukcija LLE jezgra.** Uzmemo li rijetku matricu $L = I - W$, tada je jezgro dato sa

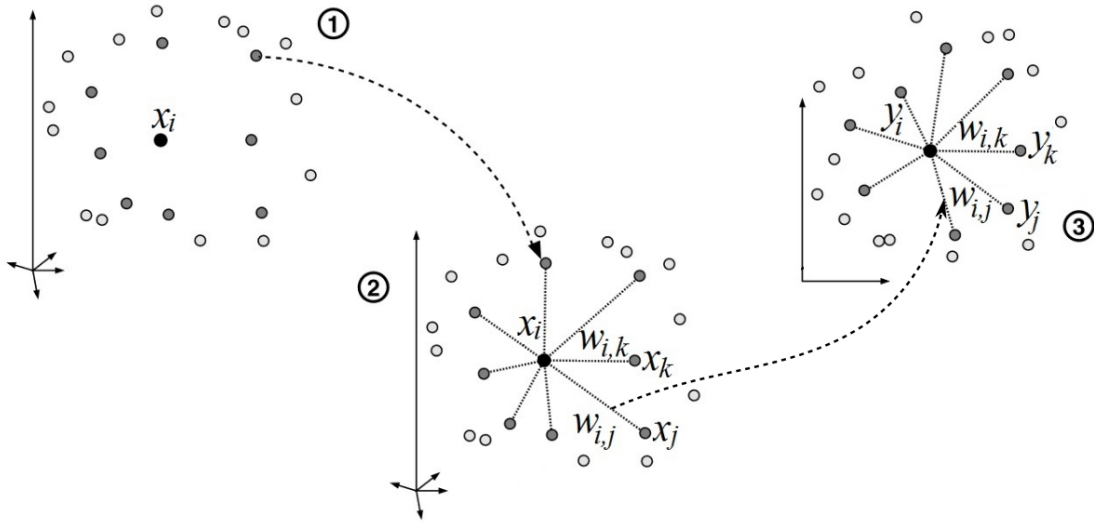
$$K = L'L = (I - W)'(I - W),$$

gdje je W težinska matrica iz Koraka 2. U [25] se može naći detaljno opisana konstrukcija jezgra za nelinearne tehnike za redukciju dimenzionalnosti.

4. **Spektralna dekompozicija LLE jezgra.** Neka je spektralna dekompozicija LLE jezgra data sa

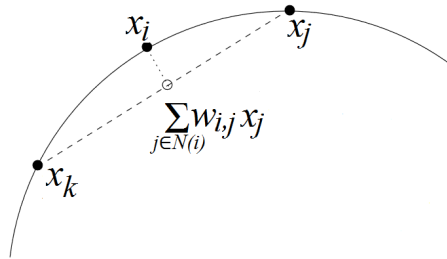
$$K = U\Lambda U',$$

gdje je $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{n-1})$ sa $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$. Pretpostavimo da je graf $G = [\mathcal{X}, A]$ povezan. Tada je $\lambda_1 > 0$. Odredimo $d+1$ karakterističnih vektora, koji odgovaraju najmanjim $d+1$ karakterističnim vrijednostima i odbacimo karakteristični vektor $[1, \dots, 1]'$ koji odgovara 0 karakterističnoj vrijednosti. Tada neka su ti vektori $u_1, \dots, u_d$ vektori kolone podmatrice U_d . Tada je tražena redukcija dimenzionalnosti $Y = (U_d)'$.



Slika 4.4: Ilustracija LLE algoritma.

Na Slici 4.4 predstavili smo ilustraciju LLE algoritma. Na Slici 4.4 u koraku ① nalazimo k -najbližih tačaka tačke x_i u $\mathbb{R}^D$. Zatim u koraku ② računamo težinske koeficijente $w_{i,j}$ koji najbolje rekonstruišu tačku x_i koristeći njene susjede, tj. tačke iz prethodno definisane okoline odnosno $x_i \approx \sum_{j \in N(i)} w_{i,j} x_j$. Na kraju u ③ nalazimo potopljene nižedimenzionalne podatke $y_i \in \mathbb{R}^d$ koje rekonstruišemo koristeći LLE jezgro i težinske koeficijente, dobijamo $y_i \approx \sum_{j \in N(i)} w_{i,j} y_j$.



Slika 4.5: Tačka x_i i njeni susjedi x_j i x_k , kao i lokalna linearna rekonstrukcija $\sum_{j \in N(i)} w_{i,j} x_j$, pri čemu važi ograničenje $\sum w_{i,j} = 1$.

4.3.4 Matlab kod za LLE algoritam

Na veb-stranici <https://www.cs.nyu.edu/~roweis/lle/>, mogu se naći reprezentativni radovi autora Sam T. Roweis-a i Lawrence K. Saul-a u vezi sa LLE metodom. Osim toga, može se naći i opis LLE algoritma, kao i Matlab kod za LLE algoritam, koji ćemo i predstaviti u ovom radu. Na stranici se nalazi i galerija slika na kojima može da se primjeni LLE algoritam, kao i Matlab kod za rolnu, na koju se primjeni prethodno navedeni LLE algoritam.

```

1 function [Y] = lle(X, K, d)
2
3 % LLE ALGORITAM (koristeci sistem 'k' okolina)
4 %
5 % [Y] = lle(X, K, dmax)
6 %
7 % X = podaci u obliku D x n matrice (D = dimenzija, n = broj tacaka)
8 % K = broj clanova svake okoline (broj susjeda)
9 % dmax = maksimalna zeljena dimenzija nakon redukcije
10 % Y = nizedimenzionalni podaci u obliku dmax x n matrice
11
12 % Reference:
13 % S. T. Roweis and L. K. Saul
14 % Nonlinear dimensionality reduction by locally linear embedding
15 % Science v.290 no.5500, pp.2323—2326, Dec. 22, 2000.
16
17 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
18
19 [D,n] = size(X);
20 fprintf(1, 'Primjenjujemo LLE metod na %d points u %d dimenzija\n', n, D);
21
22 % Korak 1: IZRACUNAJ RASTOJANJA PAROVA TACAKA I ODREDI SUSJEDE
23 fprintf(1, '—>Nalazenje %d najblizih susjeda.\n', K);
24
25 X2 = sum(X.^2,1);
26 rastojanje = repmat(X2,n,1)+repmat(X2',1,n)-2*X'*X;
27
28 [sorted,index] = sort(distance);
29 neighborhood = index(2:(1+K),:);
30
31 % Korak 2: IZRACUNAVANJE TEZINSKIH KOEFICIJENATA
32 fprintf(1, '—>Izracunavanje tezinskih koeficijenata.\n');
33
34 if (K>D)
35 fprintf(1, ' [napomena: K>D; koristi se regularizacija]\n');
36 tol=1e-3;
37 else
38 tol=0;
39 end
40
41 W = zeros(K,n);
42 for ii=1:n
43 z = X(:,neighborhood(:,ii))-repmat(X(:,ii),1,K); % translacija u koordinatni pocitak i-te tacke
44 C = z'*z; % local covariance
45 C = C + eye(K,K)*tol*trace(C); % regularizacija (K>D)
46 W(:,ii) = C\ones(K,1); % rijesi Cw=1
47 W(:,ii) = W(:,ii)/sum(W(:,ii)); % normalizuj sum(w)=1
48 end;
49
50 % Korak 3: Konstrukcija LLE jezgra
51 M=(I-W)'(I-W)
52 fprintf(1, '—>Konstrukcija jezgra.\n');
53
54 % M=eye(n,n);
55 % koristimo rijetku matricu sa 4Kn nenula elemenata

```

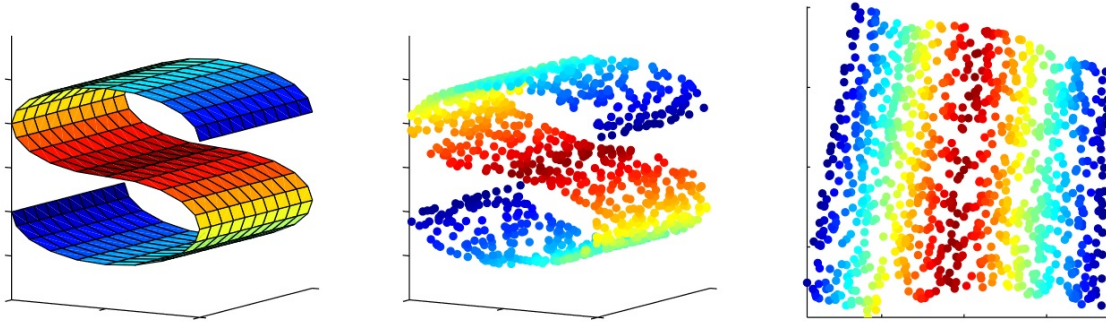
```

56 M = sparse(1:n,1:n,ones(1,n,n,4*K*n));
57 for ii=1:n
58     w = W(:,ii);
59     jj = neighborhood(:,ii);
60     M(ii,jj) = M(ii,jj) - w';
61     M(jj,ii) = M(jj,ii) - w;
62     M(jj,jj) = M(jj,jj) + w*w';
63 end;
64
65 % IZRACUNAVANJE NIZEDIMENZIONALNIH PODATAKA
66 options.disp = 0; options.isreal = 1; options.issym = 1;
67 [Y,eigenvals] = eigs(M,d+1,0,options);
68 Y = Y(:,2:d+1)'*sqrt(n); % odbacimo k.v. [1,1,1,1...] koji odgovara k.k. 0
69
70 fprintf(1, 'Kraj.\n');

```

Na samom kraju ovog rada navodimo još dva primjera *vještačkih površi* (eng. artificial surface) na kojima se često pokreću algoritmi za redukciju dimenzionalnosti, kako bi se vizualizovali dobijeni rezultati. U ovom radu smo za primjer primjene algoritama za redukciju dimenzionalnosti izabrali Švajcarsku rolnu, ali u mnogo radova mogu se naći i *S-kriva* (eng. *S-curve*) koja je data parametarskom jednačinom

$$\begin{cases} x = -\cos(1.5\pi t), \\ y = s, \\ z = \begin{cases} -\sin(1.5\pi t), & 0 \leq t \leq 1, \\ 2 + \sin(1.5\pi t), & 1 < t \leq 2, \end{cases} \end{cases} \quad 0 \leq s \leq L, \quad 0 \leq t \leq 2, \quad (4.21)$$



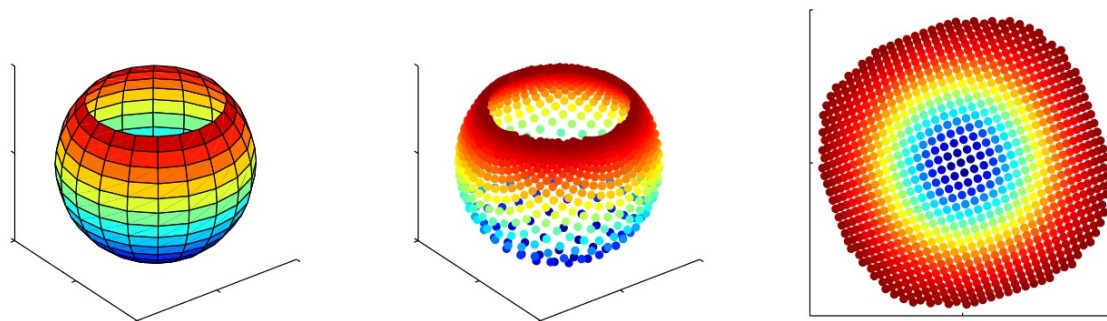
Slika 4.6: Primjena LLE algoritma na S-krivu.

i koja je data na Slici 4.6, kao i *Odsječena sfera* (eng. Punched sphere), tj. sfera kod koje je odsječen poklopac, tj. gornji dio sfere. Ona je data parametarski sa

$$\begin{cases} x = \cos(s) \sin(t), \\ y = \sin(s) \sin(t), \\ z = \cos(t) \end{cases} \quad 0 \leq s \leq 2\pi, \quad \theta \leq t \leq \pi. \quad (4.22)$$

i prikazana je na Slici 4.7.

Na Slikama 4.6 i 4.7 smo pored navedenih površi naveli i podatke koji leže na navedenim nelinearnim mnogostrukostima, kao i rezultat primjene LLE algoritma na date skupove podataka.



Slika 4.7: Primjena LLE algoritma na Odsječenu sferu.

LLE tehnika ima veliku primjenu, a neki od primjera mogu se naći u [25]. LLE algoritam koristi rijetku matricu, tj. matricu u kojoj se nalazi mnogo nula, tako da je algoritam prilično brz. Pored toga, LLE algoritam daje prilično dobru linearnu sliku podataka na mnogostrukosti. Najveći nedostatak ovog algoritma je što je ponekad računski (kompjuterski) nestabilan.

Zaključak

Na kraju samog rada, važno je da sumiramo prethodno izloženu teoriju.

Na osnovu do sad rečenog i izloženih primjera očigledno je da tehnike za redukciju dimenzionalnosti imaju veliku primjenu prilikom rješavanja različitih problema koji se javljaju u analizi podataka, mašinskom učenju, vještačkoj inteligenciji i sl. Svjedoci smo brzog razvoja tehnike, naročito Android telefona i laptop računara, kao i razvoja nauke, što sa sobom nosi veliku količinu informacija i podataka. Kao što smo vidjeli u primjerima navedenim u ovom radu, tehnike za redukciju dimenzionalnosti imaju važnu ulogu u rješavanju takve vrste problema.

U [21] može se naći pregled najvažnijih radova koji su objavljeni od 2003. do 2012. godine iz oblasti redukcije dimenzionalnosti. Najviše se primjenjuju linearne tehnike (PCA, ICA, FA, MDS), iako se u posljednje vrijeme razvio veliki broj tehnika koje mogu da analiziraju mnogostrukosti na kojima leže skupovi podataka, a neke od njih smo vidjeli i u ovom radu.

Osim toga, kada je u pitanju analiza glavnih komponenti PCA, mnogo se radi na uopštenju klasičnog metoda. Pojavili su se kernel PCA, rijetki PCA, robusni PCA koji takođe u osnovi imaju ideju koju smo koristili i u ovom radu, traže se projekcije, odnosno glavne komponente.

Kada su u pitanju nelinearne metode, one su najvažnije oruđe koje se pojavilo u poslednjih 15 godina. U ovom radu mi smo vidjeli dvije nelinearne tehnike, Isomap i LLE, ali pored njih, Laplasovo i Hesijanovo karakteristično preslikavanje, takođe daju dosta dobre rezultate. Ove tehnike mogu dosta detaljno da ispituju strukturu mnogostrukosti, tako da se smanjila upotreba nelinearnih metoda kao što su glavne krive, glavne površi i glavne mnogostrukosti koje takođe predstavljaju uopštenja PCA metoda.

S obzirom da svaka tehnika ima svoje prednosti i mane, stalno se radi na poboljšanju već postojećih tehnika, ali i na razvoju novih. Neki od najnovijih rezultata tiču se primjene algebarske topologije i diferencijalne geometrije, ali i geometrijske (Clifford-ove) algebre, kako bi se izvršila modifikacija geometrije posmatranog skupa podataka i na taj način primjenili određeni filtri za izdvajanje odgovarajućih komponenti.

S obzirom na prethodno izloženo, očigledno je da je potreba za tehnikama za redukciju dimenzionalnosti velika. Stoga, nesumnjivo je da će se sve više raditi na razvoju ovih tehnika, jer s obzirom na broj tehnika koje su se pojavile u poslednjih 55 godina (od Bellmana i formulacije problema dimenzionalnosti), jasno je da su one neiscrpan izvor za naučnike širom svijeta već dugi niz godina, i da novi rezultati tek slijede.

Literatura

- [1] Bernstein, M., De Silva, V., Langford, J., Tenenbaum, J. B. *Graph approximations to geodesics on embedded manifolds*, Technical report, Department of Psychology, Stanford University, 2000.
- [2] Bertsekas, D. P. *Nonlinear Programming*, 2nd Edition Athena Scientific, Belmont, 1999.
- [3] Bondy, J., Murty U. *Graph Theory.*, Springer, 2008.
- [4] Boothby, W.M. *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, 1975.
- [5] Cunningham, J.P., Ghahramani, Z. *Linear dimensionality reduction: survey, insights, and generalizations*, Journal of Machine Learning Research, 16(Dec):2858–2900, 2015.
- [6] Guillemard, M. *On Dimensional Reduction Techniques in Signal Processing and Applications in EMG Analysis*, Diplomarbeit, Department Mathematik, Universität Hamburg, 2008.
- [7] Hadžić, O. *Numeričke i statističke metode u obradi eksperimentalnih podataka I deo*, drugo izdanje, Institut za matematiku, Novi Sad, 1992.
- [8] Johnson, A. R., Wichern, D. W. *Applied Multivariate Statistical Analysis*, 6th Edition, Pearson, New Jersey, 2007.
- [9] Kovačić, Z. J. *Multivarijaciona analiza*, Ekonomski fakultet, Beograd, 1994.
- [10] Kunzinger, M. *Differential Geometry 1*, Lecture notes, University of Vienna, 2008.
- [11] Lee, J., Verleysen, M. *Nonlinear Dimensionality Reduction*, Springer, Information Science and Statistics series, 2007.
- [12] Lozanov Crvenković, Z. *Statistika*, Prirodno-matematički fakultet, Novi Sad, 2012.
- [13] Mladenović, P. *Verovatnoća i statistika*, VESTA - Matematički fakultet, Beograd, 1995.
- [14] Rajter Ćirić, D. *Verovatnoća*, treće izdanje, Prirodno-matematički fakultet, Novi Sad, 2013.
- [15] Roweis, S., *EM algorithms for PCA and SPCA*, In: anips, vol. 10, pp. 626 – 632, 1998.
- [16] Roweis, S.T., Saul, L.K. *Nonlinear dimensionality reduction by locally linear embedding*, Science v.290 no.5500, pp.2323–2326, Dec. 22, 2000.
- [17] Saul, L.K., Roweis, S.T. *Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds*, Journal of Machine Learning Research, v4, pp. 119-155, 2003.
- [18] Saul, L.K., Roweis, S.T., Weinberger, K.Q., Sha, F., Packer, B. *Unfolding a manifold by semidefinite programming*, Presentation, Computer and Information Science, University of Pennsylvania, 2005.

- [19] Schiffman, S. S., Reynolds, M. L., Young, F. W. *Introduction to Multidimensional Scaling*, Academic Press, New York, 1981.
- [20] Shalizi, C. R., *Advanced Data Analysis from an Elementary Point of View*, <http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf>, 2016.
- [21] Sorzano, C. O. S., Vargas, J., Pascual-Montano, A. *A survey of dimensionality reduction techniques*, arXiv:1403.2877
- [22] Tenenbaum, J. B., De Silva, V., Langford, J. *A global geometric framework for nonlinear dimensionality reduction*, Science 290 (5500): 2319-2323, Dec. 22, 2000.
- [23] Vanderbei, R. J. *Linear Programming Foundations and Extensions*, Springer US, New York, 2014.
- [24] Van der Maaten, L.J.P., Postma, E.O., Van den Herik, H.J. *Dimensionality Reduction: A Comparative Review.*, Tilburg University Technical Report, TiCC-TR 2009-005, 2009.
- [25] Wang, J. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, Higher Education Press (China) and Springer, Beijing, 2011.
- [26] Young, F. W. *Multidimensional scaling*, In: Kotz, S. and Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 5. 649-659, New York: Wiley, 1985.

Biografija



Srđan Lazendić je rođen 19.6.1992. godine u Vršcu, Srbija. Odrastao je u Doboju, Republika Srpska - BiH, gdje je 2007. godine završio osnovnu školu „Vuk Stefanović Karadžić”. Zatim je upisao opšti smjer Gimnazije „Jovan Dučić” u Doboju. Istu je završio 2011. godine nakon čega je upisao osnovne akademske studije matematike na Departmanu za matematiku i informatiku, Prirodno-matematičkog fakulteta u Novom Sadu, smjer Diplomirani profesor matematike. Studije je završio zaključno sa junskim rokom 2015. godine sa prosječnom ocjenom 9.61. Master studije upisao je iste godine na istom fakultetu, smjer Master matematika (MA). Sve ispite položio je u junskom roku 2016. godine i time stekao uslov za odbranu ovog master rada. Od 1. septembra 2016. godine student je doktorskih studija iz oblasti matematičkog inženjerstva na Univerzitetu u Gentu, Belgija.

Novi Sad, jul 2016.

Srđan Lazendić

UNIVERZITET U NOVOM SADU
PRIRODNO–MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: Monografska dokumentacija

TD

Tip zapisa: Tekstualni štampani materijal

TZ

Vrsta rada: Master rad

VR

Autor: Srđan Lazendić

AU

Mentor: Dr Sanja Konjik, vanr. prof.

MN

Naslov rada: Matematički alati za redukciju dimenzionalnosti signala

NR

Jezik publikacije: srpski (latinica)

JP

Jezik izvoda: srpski/engleski

JI

Zemlja publikovanja: Republika Srbija

ZP

Uže geografsko područje: Vojvodina

UGP

Godina: 2016.

GO

Izdavač: Autorski reprint

IZ

Mesto i adresa: Departman za matematiku i informatiku, Prirodno–matematički fakultet, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 4, Novi Sad

MA

Fizički opis rada: (6, 67, 26, 1, 21, 0, 0)

FO

Naučna oblast: Matematika
NO

Naučna disciplina: Matematička analiza
ND

Ključne reči: Redukcija dimenzionalnosti signala, višedimenzionalni podaci, PCA, MDS, Isomap, LLE
PO

UDK:

Čuva se: Biblioteka Departmana za matematiku i informatiku Prirodno-matematičkog fakulteta Univerziteta u Novom Sadu
ČU

Važna napomena:
VN

Izvod:
IZ

U ovom master radu bavimo se tehnikama za redukciju dimenzionalnosti signala, tačnije, posmatramo tehnike za redukciju dimenzionalnosti proizvoljnih višedimenzionalnih skupova podataka. Rad se sastoji od četiri glave, pri čemu smo u prve dvije glave iznijeli osnovne pojmove na kojima se zasnivaju tehnike za redukciju dimenzionalnosti. Nakon uvodnih pojmova, u trećoj i četvrtoj glavi ispitivali smo linearne (PCA i CMDS) i nelinearne (Isomap i LLE) tehnike. Dali smo kratak pregled i opis navedenih tehnika, nakon čega smo ilustrovali primjenu navedenih tehnika na primjerima vještačkih površi. Vještački konstruisani skupovi tačaka se često koriste kako bi se lakše uočile prednosti i nedostaci različitih metoda. Ukazali smo na razlike između tehnika, ali pored toga, za svaku tehniku dali smo opis algoritma posmatranog metoda, kao i Matlab kod za svaki od njih.

Datum prihvatanja teme od strane NN veća: 13.4.2016.
DP

Datum odbrane:
DO

Članovi komisije:
KO

Predsjednik:	Dr Aleksandra Pižurica, redovni profesor, TELIN, Univerzitet u Gentu
Član:	Dr Dejan Vukobratović, vanredni profesor, Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Član:	Dr Sanja Konjik, vanredni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, mentor

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
KEY WORDS DOCUMENTATION

Accession number:

ANO

Identification number:

INO

Document type: Monographic type

DT

Type of record: Printed text

TR

Contents Code: Master thesis

CC

Author: Srđan Lazendić

AU

Mentor: Sanja Konjik, PhD

MN

Title: Mathematical tools for signal dimensionality reduction

TI

Language of text: Serbian

LT

Language of abstract: Serbian/English

LA

Country of publication: Republic of Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2016.

PY

Publisher: Author's reprint

PU

Publ. place: Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, Novi Sad

PP

Physical description:(6, 67, 26, 1, 21, 0, 0)

PD

Scientific field: Mathematics

SF

Scientific discipline: Mathematical analysis

SD

Key words: Dimensionality reduction, High-dimensional data, PCA, MDS, Isomap, LLE

SKW

UC:

Holding data: Library of the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad

HD

Note:

N

Abstract:

AB

In this master thesis, techniques for signal dimensionality reduction are examined. More general, techniques for dimensionality reduction of arbitrary multidimensional data sets, are examined. The thesis consists of four chapters. In the first two chapters the basic concepts of dimensionality reduction techniques are presented. After the introductory part, in the third and fourth chapter we examine linear (PCA and CMDS) and nonlinear (Isomap and LLE) techniques. We give a short overview and description of mentioned techniques, after which we illustrate the application of these techniques on examples of artificial surfaces. Artificially created data sets are often used to illustrate the advantages and drawbacks of the various methods. We point out the differences between techniques, but in addition, for each technique we give a description of the algorithm of the observed methods, and Matlab code for each of them.

Accepted by the Scientific Board on: 13. April 2016.

ASB

Defended:

DE

Thesis defend board:

DB

President: Dr. Aleksandra Pižurica, Full professor, TELIN, Ghent University

Member: Dr. Dejan Vukobratović, Associate professor, Faculty of Technical Sciences,
University of Novi Sad,

Member: Dr. Sanja Konjik, Associate professor, Faculty of Sciences, University of Novi Sad