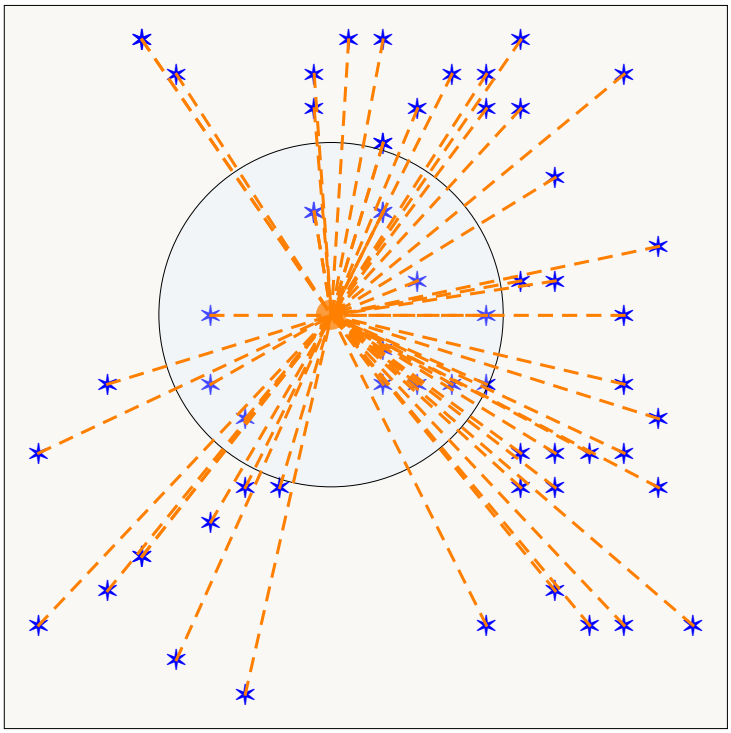
Sveučilište Josipa Jurja Strossmayera u Osijeku
Odjel za matematiku

Rudolf Scitovski, Kristian Sabo

Klaster analiza i prepoznavanje geometrijskih objekata



Osijek, 2020.

Prof. dr. sc. Rudolf Scitovski
Prof. dr. sc. Kristian Sabo
Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku
Trg Ljudevita Gaja 6
HR-31 000 Osijek

Izdavač:
Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku

Recenzenti:
Prof. dr. sc. Robert Manger
Matematički odsjek Prirodoslovno-matematičkog fakulteta, Sveučilište u Zagrebu

Prof. dr. sc. Ivan Slapničar
Fakultet elektrotehnike, strojarstva i brodogradnje, Sveučilište u Splitu

Izv. prof. dr. sc. Snježana Majstorović
Odjel za matematiku, Sveučilište Josipa Jurja Strossmayera u Osijeku

Lektorica:
Ivanka Ferčec
Fakultet elektrotehnike, računarstva i informacijskih tehnologija, Sveučilište Josipa Jurja Strossmayera u Osijeku

CIP zapis dostupan je u računalnom katalogu Gradske i sveučilišne knjižnice Osijek pod brojem 150112078.

ISBN 978-953-8154-11-9

Ovaj udžbenik objavljuje se uz suglasnost Senata Sveučilišta Josipa Jurja Strossmayera u Osijeku pod brojem 43/19.

Ovaj udžbenik objavljuje se uz financijsku pomoć Ministarstva znanosti i obrazovanja Republike Hrvatske.

Tisak: STUDIO HS Internet d.o.o. Osijek

PREDGOVOR

Sadržaj ovog udžbenika izvodio se posljednjih nekoliko godina u okviru kolegija Grupiranje podataka i primjene u 8. semestru sveučilišnog nastavničkog studija Matematike i informatike i u 2. semestru sveučilišnog diplomskog studija Matematika, smjer Matematika i računarstvo i izbornog predmeta Grupiranje podataka: pristupi, metode i primjene u 2. semestru sveučilišnog diplomskog studija Matematika, smjer Financijska matematika i statistika na Odjelu za matematiku Sveučilišta u Osijeku (30 sati predavanja i 30 sati seminara, 6 ECTS bodova).

Tekst je pisan tako da podrazumijeva poznavanje osnova matematičke analize, linearne algebre, numeričke matematike i programiranja, a namijenjen je studentima diplomskih studijskih programa u `STEM` područjima.

Brojni ilustrativni primjeri doprinose razumijevanju izložene materije. Osim toga, u okviru svakog poglavlja nalaze se brojni zadaci: od sasvim jednostavnih, do onih koji mogu poslužiti kao teme seminarskih i sličnih radova. Izrada većeg broja zadataka povezana je s mogućnošću korištenja programskih sustava *Mathematica* ili `Matlab`.

Sadržaj naveden u ovom udžbeniku, kao i priloženi *Mathematica*-programi, mogu poslužiti i u nekim praktičnim istraživanjima. Za sve metode navedene u udžbeniku izrađeni su odgovarajući moduli, a *Mathematica*-kodovi slobodno su dostupni na adresi: `https://www.mathos.unios.hr/images/homepages/scitowsk/Clustering.rar`.

Opsežna recentna literatura navedena na kraju udžbenika, koja obuhvaća brojne knjige i odgovarajuće članke iz renomiranih međunarodnih znanstvenih časopisa, daje neophodan pregled najvažnijih spoznaja, a također može poslužiti i za nastavak samostalnog rada u ovom znanstvenom području.

U ovom udžbeniku razmatra se problem grupiranja konačnog skupa to-

čaka $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Kako $\mathbb{R}^n$ možemo promatrati i kao realni vektorski prostor, njegove elemente označavat ćemo s $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ te ih ponekad zvati točkama, a ponekad vektorima u ovisnosti o kontekstu. Pri tome to nećemo uvijek posebno naglašavati.

Zbog specifičnih zahtjeva matematičkog teksta i sličnih zahtjeva u programskim sustavima *Mathematica*, `Matlab`, i `FORTRAN`, cijeli dio od decimalnog dijela decimalnog broja odvajan je decimalnom točkom (.), a ne decimalnim zarezom (,).

Svi teoremi, leme, definicije, slike, tablice, primjedbe, primjeri i zadaci u tekstu imaju svoju jedinstvenu oznaku i na taj način pozivaju se u cijelom tekstu. Zbog toga su pisani velikim početnim slovom[1].

Zahvaljujemo recenzentima *Robertu Mangeru, Ivanu Slapničaru* i *Snježani Majstorović* te lektorici *Ivanki Ferčec*, koji su svojim primjedbama i prijedlozima značajno pomogli da ovaj tekst bude bolji. Također zahvaljujemo kolegi *Ivanu Soldi* koji je u tehničkom smislu pomogao podizanju kvalitete ovog udžbenika.

Osijek, 10. siječnja 2020.

Rudolf Scitovski
Kristian Sabo

[1] Za pregledavanje ovog udžbenika Adobe Readerom možete koristiti sljedeće pogodnosti:

- klikom na naslov nekog poglavlja u Sadržaju dolazite na to poglavlje. Povratak (na isto mjesto odakle ste krenuli) je držanjem tipke Alt pa nakon toga pritisnuti tipku <;
- klikom na oznaku nekog teorema, leme, definicije, slike, tablice, primjedbe, primjera ili zadatka u tekstu odlazite na taj objekt. Povratak je na prethodno opisani način;
- klikom na oznaku neke reference u tekstu odlazite na tu referencu u Literaturi na kraju knjige. Povratak je na prethodno opisani način;
- klikom na oznaku stranice u Indeksu na kraju knjige odlazite na taj pojam u knjizi. Povratak je na prethodno opisani način.

Sadržaj

Poglavlje 1

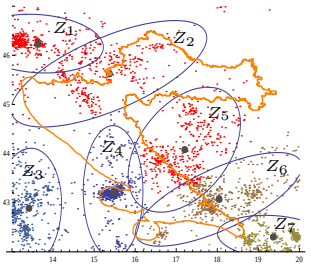
Uvod

Namjena ovog udžbenika je upoznati studente s najnovijim spoznajama u području grupiranja podataka i mogućnostima primjena. Različite mogućnosti grupiranja podataka koriste se u brojnim znanstvenim disciplinama. Navedimo neke od njih.

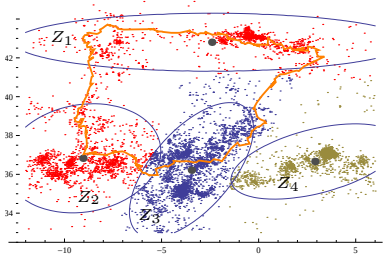
Primjer 1.1. *Na Slici 1.1a točkicama su prikazane geografske lokacije u širem području Republike Hrvatske na kojima se od 1900. godine dogodio potres magnitude* ≥ 3. *Ovi podaci slobodno su dostupni na* U.S. Geological Survey `http://earthquake.usgs.gov/` *u obliku:* Godina/Mjesec/Dan/Sat/Min./Sec./Latituda (φ)/Longituda (λ)/Dubina/Magnituda/. *Iz ovih podataka definiramo skup*

$$\mathcal{A} = \{a^i = (\lambda_i, \varphi_i) \in \mathbb{R}^2 \colon L_\lambda \leq \lambda_i \leq U_\lambda, \quad L_\varphi \leq \varphi_i \leq U_\varphi\},$$

(a) Hrvatska



(b) Pirenejski poluotok



Slika 1.1: Određivanje zona seizmičkih aktivnosti

koji sadrži lokacije potresa određene longitudom λ_i *i latitudom* φ_i *u pravokutniku* $[L_\lambda, U_\lambda] \times [L_\varphi, U_\varphi]$ *koji definira šire područje Republike Hrvatske.*

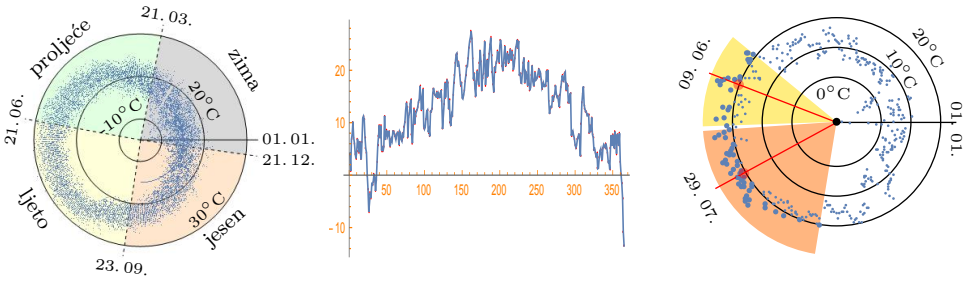
Nadalje, svakoj točki a^i *pridružuje se težina* $w_i > 0$ *definirana kao magnituda* M_i *potresa u točki* a^i. *U našem je slučaju* $\mathcal{A} \subset [12.5, 21] \times [41.5, 47.5]$, *a broj podataka je* $|\mathcal{A}| = 8\,744$. *Budući da je ovo relativno maleno geografsko područje, neće biti potrebna transformacija u Gauss-Krügerov koordinatni sustav. U svrhu određivanje zona seizmičkih aktivnosti koristit će se* $5\,324$ *podataka magnitude* ≥ 3, *između kojih se nalazi* $4\,051$ *podataka magnitude* < 4, $1\,124$ *podataka magnitude između* 4 *i* 5, *i* 149 *podataka magnitude* ≥ 5.

Primjenom Mahalanobis inkrementalnog algoritma (točka 6, str. 99) i Mahalanobis k-means algoritma (točka 6.4.1, str. 111) uz korištenje Mahalanobis indeksa (točka 6.5, str. 114) za izbor particije s najprikladnijim brojem elipsoidnih klastera, određene su dominantne zone seizmičkih aktivnosti u širem području Republike Hrvatske (elipsoidna područja na Slici 1.1a): Slovenija i sjeverozapadna Italija (Z_1), *sjeverna Hrvatska* (Z_2), *istočno područje Italije* (Z_3), *Jadransko more i Dalmacija* (Z_4), *Dinara i Bosna i Hercegovina* (Z_5), *područje Ston–Metković, južno jadransko područje i južna Bosna i Hercegovina* (Z_6) *i Dubrovnik i Crna Gora* (Z_7).

Slični podaci prikazani su za Pirenejski poluotok od 1978. godine na Slici 1.1b. Težine podataka definirane su također kao magnitude potresa. Geografske pozicije potresa grupirane su u zone seizmičkih aktivnosti, koje su u formi elipsi također prikazane na Slici 1.1.

Detaljan opis problema, način izračunavanja i diskusija rezultata dani su u [62, 98].

(a) Osijek: 1985.-2014. (b) Osijek, 2014. (c) Osijek, 2014.



Slika 1.2: Prosječne dnevne temperature klastera toplih dana u Osijeku od 1985. do 2014. i detaljnija analiza za 2014. godinu

Primjer 1.2. *U cilju istraživanja klimatskih promjena u Osijeku promatrani su podaci o prosječnim dnevnim temperaturama od 1985. godine (vidi Burnov dijagram na Slici 1.2a). Specijalno, na Slici 1.2b prikazane su prosječne dnevne temperature izražene u* $^\circ C$ *od 21. ožujka 2014. do 21. ožujka 2015. godine (Sunčeva godina), a na Slici 1.2c primjenom Burnovog dijagrama*

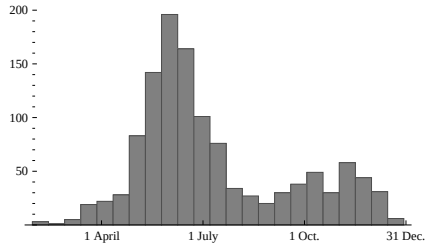
(vidi točku 2.4.2, str. 31) prikazani su klasteri toplih dana iste godine.

Uočava se da je 2014. godina imala jedno dugo, toplo i stabilno razdoblje koje je trajalo od 21. svibnja do 21. rujna. Prvih 43 dana do 2. srpnja prosječna dnevna temperatura bila je $20.2°\,C$*, a nakon toga čak 80 dana prosječna dnevna temperatura bila je* $20.6°\,C$*. Najviša prosječna dnevna temperatura 2014. godine zabilježena je već 11. lipnja i iznosila je* $27.4°\,C$*.*

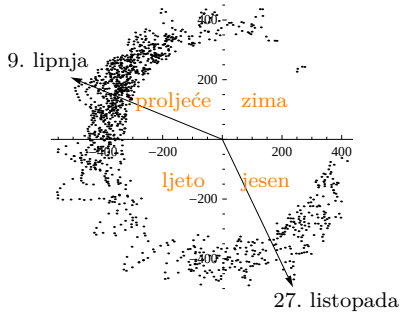
U ovom istraživanju radi se o periodičnim podacima pa ih je prikladno promatrati kao podatke na kružnici (vidi točku 2.4, str. 28). Detaljan opis problema, način izračunavanja i diskusija rezultata dani su u [86].

Primjer 1.3. *Na Slici 1.3a prikazan je histogram visokog vodostaja Drave kod Donjeg Miholjca od 1900. godine, a na Slici 1.3b odgovarajući Burnov dijagram.*

(a) Histogram



(b) Burnov dijagram



Slika 1.3: Vodostaj Drave kod D. Miholjca od 1900. godine

Točke na Burnovom dijagramu prikazuju godišnje trenutke visokog vodostaja i njihove vrijednosti kao mjeru udaljenosti točke do ishodišta. Točke na Burnovom dijagramu grupirane su u dva klastera primjenom specijalne kvazimetričke funkcije za periodične podatke (vidi točku 2.4, str. 28). Primijetite da se najviši vodostaj Drave kod D. Miholjca može očekivati početkom lipnja i krajem listopada (centri klastera!).

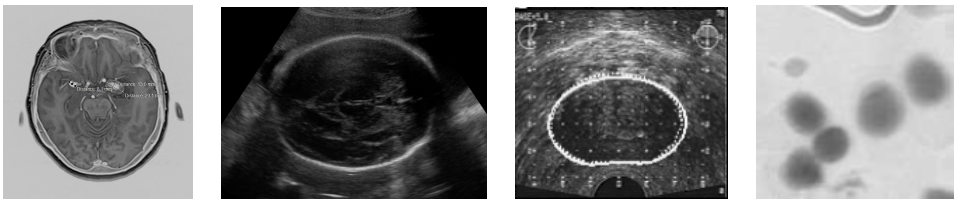
U ovom istraživanju korištena je jedna modifikacija Metode najbližih susjeda*. Detaljan opis problema, način izračunavanja i diskusija rezultata dani su u [88].*

Primjer 1.4. *Na Slici 1.4 prikazano je nekoliko tipičnih medicinskih slika: crno-bijela slika ljudskog mozga*[1] *(Slika 1.4a), slika glave ljudskog embrija nakon 28 tjedana (Slika 1.4b (vidi [37, 76]), slika prostate (Slika 1.4c) i*

[1]Sliku pripremio dr. Salha Tamer, KBC Osijek, Klinički zavod za dijagnosticku i intervencijsku radiologiju.

tkivo raka dojke (Slika 1.4d). Primjenom klaster analize cilj je prepoznati anomalije i tendencije njihovih povećanja ili smanjivanja. U ovom slučaju radi se o prepoznavanju jedne ili više elipsi na slici.

(a) Ljudska glava (b) Glava fetusa (c) Prostata (d) Rak dojke

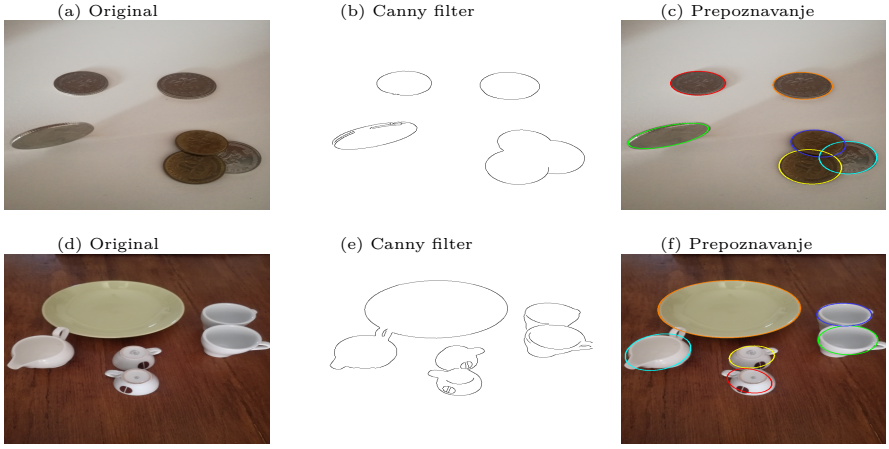


Slika 1.4: Prepoznavanje objekata na medicinskim slikama

Nešto jednostavniji primjeri prepoznavanja više elipsi na slici iz realnog svijeta prikazani su na Slici 1.5.

(a) Original (b) Canny filter (c) Prepoznavanje

(d) Original (e) Canny filter (f) Prepoznavanje



Slika 1.5: Prepoznavanje elipsi na slikama iz realnog svijeta

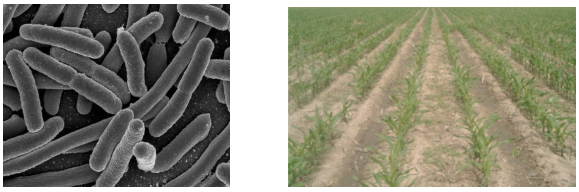
Za svaku od navedenih slika najprije je potrebno provesti postupak predobrade slike primjenom nekog od poznatih filtera, kao što je primjerice Canny filter (vidi [18]). Na taj način nastoji se izdvojiti važnije obrise na slici (vidi primjerice Sliku 1.5b ili Sliku 1.5e).

Nakon toga, na ovako pripremljenu sliku može se primijeniti metoda[2] *navedena u točki 8.9, str. 165 ili u točki 8.10, str. 170. Za slike prikazane na Slici 1.5 rezultati su prikazani na Slici 1.5c, odnosno na Slici 1.5f.*

[2]Za navenu metodu student Odjela za matematiku, Sveučilišta u Osijeku Patrick Nikić izradio je demo verziju programa koja je javno dostupna na `http://cs.mathos.unios.hr/~pnikic/ells/`.

Primjer 1.5. *Tipični oblik generalizirane kružnice, čije je središte segment, imaju Enterobacter cloacae i Enterobacter aerogenes, koje možemo naći u ljudskom gastrointestinalnom traktu, i Escherichia coli (vidi Sliku 1.6a). Detekcija ovakvih bića može se provesti primjenom grupiranja podataka u klastere čiji su centri generalizirane kružnice. Detaljan opis problema, način izračunavanja i diskusija rezultata dani su u [96].*

(a) Escherichia Coli (b) Polje kukuruza



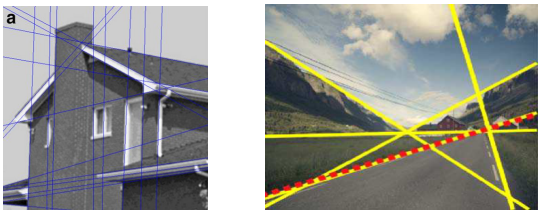
Slika 1.6: Slika Escherichia coli potječe iz National Institutes of Health, agencije United States Department of Health and Human Services, a slika polja kukuruza je iz rada [113]

Primjer 1.6. *Procesi sadnje, gnojidbe i zaštite te konačno žetva plodova, najvažniji su procesi u poljoprivrednoj proizvodnji koji mogu biti automatizirani. Prilikom bilo kojeg od navedenih procesa poljoprivredne proizvodnje čovjek mora upravljati strojem (npr. traktorom) s visokom preciznošću te ponavljati istu radnju više sati, što može biti jako zamorno. Prihvatljivo točno i u realnom vremenu detekcijom redova zasijanih biljaka (primjerice redove kukuruza na Slici 1.6b) moguće je automatizirati strojni rad pomoću kojeg se obavljaju za čovjeka zamorni, a nekada i suviše zahtjevni, dijelovi proizvodnje.*

Postupkom koji je sličan postupku opisanom u Primjeru 1.4 najprije je potrebno provesti postupak predobrade primjenom Canny filtera, a nakon toga primjenom neke od metoda prepoznavanja više pravaca na slici koje su navedene u točki 8.5, str. 145 u što kraćem vremenu što preciznije odrediti redove zasijanih biljaka.

Primjer 1.7. *Prethodno spomenuti problem detekcije više pravaca na slici također se često javlja u građevinarstvu prilikom određivanja glavnih pravaca neke zgrade (Slika 1.7a), prilikom određivanja smjera kretanja ceste (Slika 1.7b) itd.*

(a) Pravci zgrade (b) Pravci ceste



Slika 1.7: Pravci

Primjer 1.8. *Problem optimalnog definiranja izbornih jedinica u nekoj zemlji općenito je težak optimizacijski problem, pogotovo u državama s velikim brojem teritorijalnih jedinica. Iz tog je razloga ovaj problem i dalje popularan u znanstvenoj i stručnoj literaturi gdje se mogu pronaći algoritmi za generiranje raspodjele izbornih jedinica utemeljeni na raznim pristupima [17, 26, 41, 71, 72]. Postoji mnoštvo primjera u kojima su političke stranke pristrano generirale izborne jedinice u skladu sa svojim političkim interesima. Najpoznatiji takav primjer dogodio se 1812. godine kada je država Massachusetts pristrano podijeljena u neprirodne izborne jedinice. Prema tadašnjem guverneru Massachusettsa Elbridgu Gerryju, ovakvo prekrajanje izbornih jedinica poznato je u literaturi kao* Gerrymandering *(vidi primerice [17, 26]). Prema trenutno važećim zakonima u Republici Hrvatskoj, izborne jednice trebale bi se sastojati od približno jednakog broja birača (do na 5%), koje povezuje zajednički interes kroz ekonomsku, prometnu, povijesnu i drugu povezanost. Ovakav problem moguće je rješavati kao problem grupiranja podataka. Tako se primjerice u radu [79] primjenjuje metoda koja kombinira elemente klaster analize i metode kvadratne optimizacije. U tom radu promatra se područje Republike Hrvatske organizirano u* $m = 556$ *teritorijalnih jedinica (gradova ili općina), koje su određene svojim geografskim položajem u tzv. Gauss–Krügerovom koordinatnom sustavu točkama* $a_i = (x_i, y_i)$, $i = 1, \ldots, m$. *Nadalje, pretpostavlja se da teritorijalna jedinica* a_i *ima* q_i *birača i da je ukupni broj birača* Q *određen sa* $\sum_{i=1}^{m} q_i = Q$. *Teritorij Republike Hrvatske želimo podijeliti u* k, $(1 < k < m)$, *izbornih jednica (klastera)* $\pi_1, \ldots, \pi_k$, *tako da*

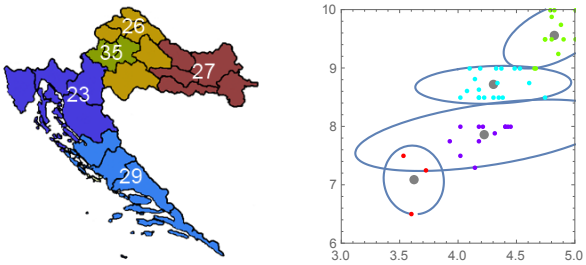
(i) *opću izbornu jednicu sačinjavaju teritorijalne jednice (gradovi i općine) koje su u nekom razdaljinskom smislu međusobno bliske,*

(ii) *je zadovoljen uvjet da se broj birača u izbornim jedinicama ne smije razlikovati više od* $\pm 5\%$.

Odgovarajuća kriterijska funkcija kao i rezultati dobiveni temeljem podataka iz 2007. godine dani su u [79].

Spomenimo još da je ovaj problem moguće promatrati i tako da izborne jedinice ne budu jednake veličine te da imaju različiti broj zastupnika, kao što je to napravljeno u radu [56] (jedno od mogućih rješenja na primjeru Republike Hrvatske prikazano je na Slici 1.8a).

(a) Izborne jedinice (b) Klasteri projekata



Slika 1.8: Neke primjene grupiranja podataka u društvenim znanostima

Primjer 1.9. *U radu [114] razmatra se problem grupiranja i rangiranja istraživačkih projekata prijavljenih na raspisani natječaj. Projekti se grupiraju u klastere na temelju ocjena dobivenih u recenzentskom postupku uz primjenu adaptivne Mahalanobis kvazimetričke funkcije. Posebno se analizira klaster najbolje ocijenjenih projekata. Navedeno je nekoliko mogućnosti korištenja podataka dobivenih recenzentskim postupkom, a predložena metoda ilustrirana je na primjeru internih istraživačkih projekata na Sveučilištu u Osijeku. Na Slici 1.8b prikazana je jedna mogućnost razvrstavanja projekata.*

(a) Originalna fotografija (b) Reprezentacija s 2 nijanse (c) Reprezentacija s 8 nijansi

Slika 1.9: Segmentacija crno-bijele slike „Elaine”

Primjer 1.10. *Grupiranje podataka može se primijeniti u svrhu segmentacije slike [85]. Na Slici 1.9a prikazana je „crno-bijela”* 512×512 *slika*

„Elaine" i njena segmentacija u 2 *(Slika 1.9b) i* 8 *(Slika 1.9c) klastera (nijansi). U ovom slučaju podaci* $\mathcal{A} = \{a^i \in \mathbb{R} \colon i = 1, \ldots, 262\,144\}$ *imaju samo jedan atribut (gray level). Redni broj (indeks) podatka* a^i *definira njegovu poziciju na slici.*

Metoda grupiranja podataka također se često koristi u robotici [23, 24], u računalnoj simulaciji [2, 37], prilikom upravljanja energetskim resursima [82, 119, 120], u bioinformatici i analizi velikih skupova podataka (engl.: big data analysis) [83, 106], kod segmentacije astronomskih i geoloških oblika, za kontrolu kvalitete vina [73], za detektiranje predmeta i oblika [37], u svrhu detekcije instalacija pod zemljom i u zidu, za klasifikaciju i obradu teksta [28], itd.

Programska podrška

Postoje različiti izvori programske podrške za traženje i grafičko prikazivanje grupiranja podataka. Navedimo neke koji su korišteni u ovom udžbeniku:

- Programski sustav *Mathematica*: naredba `FindClusters` s kvazimetričkim funkcijama:
  $DistanceFunction \to SquaredEuclideanDistance$ (default),
  $DistanceFunction \to ManhattanDistance$, ...

- *Mathematica*-moduli uz ovaj udžbenik:
  `https://www.mathos.unios.hr/images/homepages/scitowsk/Clustering.rar`

Poglavlje 2

Reprezentant

Često je u primijenjenim istraživanjima potrebno dani skup podataka reprezentirati jednim podatkom koji na neki način obuhvaća većinu svojstava promatranog skupa. Najčešće korištena veličina u tom smislu je dobro poznata aritmetička sredina podataka. Primjerice, prosječna ocjena svih položenih kolegija nekog studenta može se izraziti aritmetičkom sredinom, ali stopu ekonomskog rasta tijekom nekoliko godina ne bi bilo dobro predstaviti na takav način (vidi [87]). U ovom poglavlju razmotrit ćemo dva najčešće korištena reprezentanta nekog skupa podataka: *aritmetičku sredinu* i *medijan skupa*.

2.1 Reprezentant podataka s jednim obilježjem

Skup podataka s jednim obilježjem obično interpretiramo kao konačan podskup $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}$ skupa realnih brojeva[1]. U cilju određivanja realnog broja $c^\star \in \mathbb{R}$ koji će što bolje reprezentirati skup podataka $\mathcal{A}$ (reprezentant skupa $\mathcal{A}$) najprije je potrebno definirati neku mjeru međusobne udaljenosti elemenata skupa $\mathcal{A}$. Naravno, to može biti neka od poznatih metričkih funkcija, ali raznovrsne primjene pokazuju da je kao mjeru udaljenosti korisnije uvesti funkciju koja će imati samo svojstvo pozitivne definitnosti.

[1] U cijelom tekstu elementi $a^i \in \mathbb{R}^n$ skupa $\mathcal{A} \subset \mathbb{R}^n$ označavaju se gornjim indeksom jer je donji indeks rezerviran za komponente elementa a^i.

Definicija 2.1. *[105] Funkciju* $d\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$, $\mathbb{R}_+=\{x\in\mathbb{R}\colon x\geq 0\}$, *koja ima svojstvo* pozitivne definitnosti*:*

$$\begin{aligned} &(i) \quad d(x,y)\geq 0 \quad \forall x,y\in\mathbb{R},\\ &(ii) \quad d(x,y)=0 \quad \Leftrightarrow \quad x=y, \end{aligned}$$

zovemo kvazimetrička funkcija*.*

Dvije najčešće korištene kvazimetričke funkcije na $\mathbb{R}$ su *kvazimetrička funkcija najmanjih kvadrata* (engl.: least squares distance like function) i ℓ_1-metrička funkcija koja se u literaturi (vidi primjerice [30, 33, 49, 81]) često naziva *Manhattan metrička funkcija*:

$$\begin{aligned} &d_{LS}(x,y)=(x-y)^2 && \text{[least squares (LS) kvazimetrička funkcija]}\\ &d_1(x,y)=|x-y| && \text{[}\ell_1\text{-metrička funkcija (Manhattan metrika)]} \end{aligned}$$

Zadatak 2.1. *Provjerite vrijedi li na skupu realnih brojeva:*

$$d_1(x,y)=d_2(x,y)=d_\infty(x,y)=d_p(x,y),\quad p\geq 1,\quad \forall x,y\in\mathbb{R},$$

gdje je d_p *metrika na* $\mathbb{R}$ *(vidi primjerice [53, 106]).*

Zadatak 2.2. *Pokažite da funkcija* d_{LS} *iz prethodnog primjera nije metrika na* $\mathbb{R}$*, a da je funkcija* d_1 *metrika na* $\mathbb{R}$*.*

Definicija 2.2. *Neka je* $d\colon\mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$ *kvazimetrička funkcija.* Najbolji reprezentant *skupa podataka* $\mathcal{A}=\{a^1,\ldots,a^m\}\subset\mathbb{R}$ *u odnosu na kvazimetričku funkciju* d *je točka* $c^\star\in\mathbb{R}$ *u kojoj se postiže globalni minimum funkcije* $F\colon\mathbb{R}\to\mathbb{R}_+$*,*

$$F(x)=\sum_{i=1}^{m} d(x,a^i), \tag{2.1}$$

što formalno zapisujemo na sljedeći način:

$$c^\star\in\operatorname*{argmin}_{x\in\mathbb{R}}\sum_{i=1}^{m} d(x,a^i). \tag{2.2}$$

Primijetite da za točku globalnog minimuma $c^\star\in\mathbb{R}$ i za svaki $x\in\mathbb{R}$ vrijedi

$$F(x)=\sum_{i=1}^{m} d(x,a^i)\geq\sum_{i=1}^{m} d(c^\star,a^i)=F(c^\star), \tag{2.3}$$

pri čemu jednakost vrijedi onda i samo onda ako je $x=c^\star$. Zapis (2.2) sugerira da može postojati više točaka u kojima se postiže globalni minimum funkcije F.

2.1.1 Najbolji LS-reprezentant

Funkcija (2.1) u slučaju LS-kvazimetričke funkcije

$$F_{LS}(x) := \sum_{i=1}^{m}(x - a^i)^2 \tag{2.4}$$

postiže jedinstveni globalni minimum za

$$c^\star_{LS} = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^{m} d_{LS}(x, a^i) = \tfrac{1}{m}\sum_{i=1}^{m} a^i, \tag{2.5}$$

jer je F_{LS} konveksna funkcija te vrijedi $F'_{LS}(c^\star_{LS}) = 0$ i $F''_{LS}(x) = 2m > 0$, $\forall x \in \mathbb{R}$. Dakle, najbolji LS-reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}$ je običnа aritmetička sredina[2], koja ima svojstvo da je suma kvadrata odstupanja do svih podataka minimalna:

$$\sum_{i=1}^{m}(x - a^i)^2 \geq \sum_{i=1}^{m}(c^\star_{LS} - a^i)^2, \tag{2.6}$$

pri čemu jednakost vrijedi za $x = c^\star_{LS}$.

Kao mjera raspršenosti skupa podataka $\mathcal{A}$ oko aritmetičke sredine $c^\star_{LS}$ u statističkoj literaturi [12] koristi se varijanca podataka (prosječno kvadratno odstupanje):

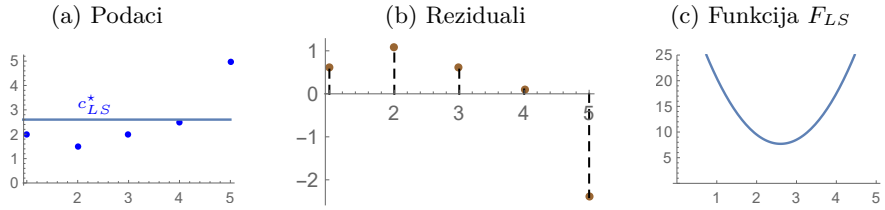
$$s_m^2 = \tfrac{1}{m-1}\sum_{i=1}^{m}(c^\star_{LS} - a^i)^2. \tag{2.7}$$

Broj s_m zovemo standardna devijacija.

Primjer 2.1. *Zadan je skup podataka* $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$. *Njegova aritmetička sredina je* $c^\star_{LS} = 2.6$.

Na Slici 2.1a vidljivi su podaci i aritmetička sredina $c^\star_{LS}$, na Slici 2.1b tzv. „reziduali" podataka (brojevi $c^\star_{LS} - a^i$), a na Slici 2.1c graf funkcija F_{LS}. Primijetite da je njezin graf parabola i da je $F_{LS}(c^\star_{LS}) = 7.7$. Kolika je varijanca, a kolika standardna devijacija ovog skupa?

[2] Problem traženja najboljeg LS-reprezentanta skupa podataka pojavljuje se u literaturi kao poznati princip najmanjih kvadrata, koji je 1795. godine postavio njemački matematičar Carl Friedrich Gauss (1777.–1855.) prilikom izučavanja kretanja nebeskih tijela, što je objavio 1809. godine u radu *Teoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, Perthes and Besser, Hamburg. Treba također napomenuti da je 1805. godine francuski matematičar Adrien-Marie Legendre (1752.-1833.) prvi objavio algebarski postupak metode najmanjih kvadrata.



Slika 2.1: Aritmetička sredina skupa $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$

Što bi se dogodilo ako bi se među podacima pojavio barem jedan „outlier" (jako stršeći podatak)? Kako bi se u tom slučaju promijenio najbolji LS-reprezentant (aritmetička sredina) skupa $\mathcal{A}$? Koji biste rezultat biste dobili ako bi umjesto podatka 5 stajao broj 10?

Zadatak 2.3. *Neka je $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}$ skup podataka, a $c^\star_{LS}$ njegova aritmetička sredina. Pokažite da tada vrijedi:*

$$\sum_{i=1}^{m} (c^\star_{LS} - a^i) = 0.$$

Provjerite ovo svojstvo na podacima iz Primjera 2.1.

Zadatak 2.4. *Neka su $\mathcal{A} = \{a^1, \ldots, a^p\}$, $\mathcal{B} = \{b^1, \ldots, b^q\} \subset \mathbb{R}$ disjunktni skupovi, a $a^\star_{LS}$ i $b^\star_{LS}$ njihove aritmetičke sredine. Pokažite da je tada aritmetička sredina skupa $\mathcal{C} = \mathcal{A} \bigcup \mathcal{B}$:*

$$c^\star_{LS} = \tfrac{p}{p+q} a^\star_{LS} + \tfrac{q}{p+q} b^\star_{LS}.$$

Provjerite formulu na nekoliko primjera. Kako bi glasila generalizacija ove formule za n skupova podataka $\mathcal{A}_1, \ldots, \mathcal{A}_n$ s po $p_1, \ldots, p_n$ elemenata?

2.1.2 Najbolji ℓ_1-reprezentant

Najprije ćemo razmotriti jednostavniji slučaj kada skup $\mathcal{A}$ sadrži međusobno različite brojeve. Slučaj kada među podacima može biti i jednakih razmatramo u točki 2.1.3, str. 15.

Kao što pokazuje sljedeća lema, funkcija (2.1) u slučaju ℓ_1-metričke funkcije

$$F_1(x) := \sum_{i=1}^{m} |x - a^i|, \tag{2.8}$$

postiže svoj globalni minimum na *medijanu* skupa $\mathcal{A}$ (vidi primjerice [12, 77, 106]).

Lema 2.1. *Neka je* $\mathcal{A} = \{a^i \in \mathbb{R} \colon i = 1, \ldots, m\}$ *skup međusobno različitih podataka. Funkcija* F_1 *zadana s* (2.8) *postiže svoj globalni minimum na medijanu skupa* $\mathcal{A}$.

Dokaz. Zbog jednostavnosti, a bez smanjenja općenitosti, možemo pretpostaviti da je $a^1 < a^2 < \cdots < a^m$. Primijetite da je funkcija F_1 konveksna po dijelovima linearna funkcija (vidi Sliku 2.2c) pa zato globalni minimum može postići u jednoj točki iz $\mathcal{A}$ ili u svakoj točki intervala između dviju točaka iz $\mathcal{A}$.

Pretpostavimo da je $x \in (a_k, a_{k+1})$. Tada vrijedi

$$F_1(x) = \sum_{i=1}^{k}(x - a^i) - \sum_{i=k+1}^{m}(x - a^i) = (2k - m)x - \sum_{i=1}^{k} a^i + \sum_{i=k+1}^{m} a^i,$$

$$F_1'(x) = 2k - m.$$

Prema tome, funkcija F_1 opada na intervalu (a_k, a_{k+1}) ako je $k < \frac{m}{2}$, odnosno raste ako je $k > \frac{m}{2}$.

Zato moramo razmotriti dva slučaja:

- Ako je m neparan, tj. $m = 2p + 1$, funkcija F_1 postiže svoj globalni minimum na srednjem podatku a^p,
- Ako je m paran, tj. $m = 2p$, funkcija F_1 postiže minimum u svakoj točki intervala $[a^p, a^{p+1}]$.

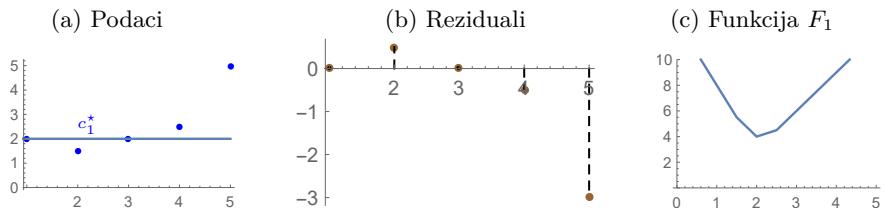
Dakle, najbolji ℓ_1-reprezentant skupa $\mathcal{A} \subset \mathbb{R}$ je medijan skupa $\mathcal{A}$. □

Primijetite da medijan skupa $\mathcal{A}$ može biti skup (segment realnih brojeva) ili samo jedan realni broj. Ako je medijan skupa $\mathcal{A}$ skup, označavat ćemo ga s $\operatorname{Med}\mathcal{A}$, a njegove elemente s $\operatorname{med}\mathcal{A}$. $\operatorname{med}\mathcal{A}$ je broj koji ima svojstvo da je suma apsolutnih odstupanja do svih podataka minimalna.[3]

$$\sum_{i=1}^{m} |x - a^i| \geq \sum_{i=1}^{m} |\operatorname{med}\mathcal{A} - a^i|, \tag{2.9}$$

[3]Problem traženja najboljeg ℓ_1-reprezentanta skupa podataka pojavljuje se u literaturi kao poznati princip najmanje sume apsolutnih odstupanja i pripisuje se hrvatskom znanstveniku Josipu Ruđeru Boškoviću (1711.–1787.), koji je ovaj princip iznio još 1757. godine u radu [16]. Dugo je vremena ovaj princip zbog složenosti računskih procesa zapostavljan u odnosu na Gaussov princip najmanjih kvadrata. Tek dolaskom modernih računala ovaj princip ponovo je zauzeo važno mjesto u znanstvenim istraživanjima, posebno zbog svojstva svoje robusnosti: ovaj princip, za razliku od Gaussovog principa najmanjih kvadrata, u značajnoj mjeri ignorira jako stršeće podatke („outliers") u skupu podataka. U švicarskom gradu Neuchâtelu još uvijek se redovito održavaju znanstvene konferencije posvećene ℓ_1 metodama i primjenama, a na naslovnici zbornika radova nalazi se slika hrvatske novčanice s likom Josipa Ruđera Boškovića [29].

pri čemu vrijedi jednakost za $x = \operatorname{med} \mathcal{A}$.



Slika 2.2: Medijan skupa $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$

Primjer 2.2. *Zadan je skup podataka* $\mathcal{A} = \{2, 1.5, 2, 2.5, 5\}$. *Njegov je medijan* $\operatorname{med} \mathcal{A} = 2$. *Koliko je prosječno apsolutno odstupanje?*

Na Slici 2.2a vidljivi su podaci i medijan $c_1^\star$, na Slici 2.2b prikazani su reziduali podataka (brojevi $c_1^\star - a^i$), a na Slici 2.2c prikazan je graf funkcije F_1. Primijetite da je F_1 konveksna po dijelovima linearna funkcija i da je $F_1(c_1^\star) = 4$.

Koliki bi bio medijan ovog skupa kada bi se među podacima pojavio jedan outlier? Koliki bi bio medijan ovog skupa kada bi umjesto podatka 5 stajao broj 10, a koliki da umjesto podatka 5 stoji broj 100? Usporedite ove rezultate s onima iz Primjera 2.1.

Medijan skupa $\mathcal{A}$ dobije se tako da se njegovi elementi najprije sortiraju. Tada, ako skup $\mathcal{A}$ ima neparan broj elemenata, medijan je srednji element, a ako skup $\mathcal{A}$ ima paran broj elemenata, medijan je bilo koji broj između dva srednja elementa. Primjerice[4],

$$\operatorname{Med}\{3, 1, 4, 5, 9\} = \{4\},$$
$$\operatorname{Med}\{-1, 1, -2, 2, -5, 5, -9, 9\} = [-1, 1],$$

ali $\operatorname{med}\{3, 1, 4, 5, 9\} = 4$ i $\operatorname{med}\{-1, 1, -2, 2, -5, 5, -9, 9\} \in [-1, 1]$.

Primjedba 2.1. *Primijetite da se medijan skupa podataka* $\mathcal{A}$ *uvijek može izabrati iz samog skupa* $\mathcal{A}$. *To znači da medijan kao reprezentant skupa ujedno može biti i element tog skupa, što nije slučaj kod aritmetičke sredine kao najboljeg LS-reprezentanta. Ova činjenica može biti korisna u nekim primjenama.*

Primijetite također da se po 50% *elemenata skupa* $\mathcal{A}$ *nalazi lijevo, odnosno desno od medijana skupa* $\mathcal{A}$.

Kao mjera raspršenosti skupa podataka $\mathcal{A}$ oko medijana u statističkoj literaturi [74, 75] koristi se Medijan apsolutnog odstupanja od medijana (engl.

[4]Medijan skupa $\mathcal{A}$ može se izračunati primjenom *Mathematica*–naredbe: `Median[]`. Pri tome, ako je medijan podataka interval, naredba `Median[]` daje polovište tog intervala.

Median of Absolute Deviations from Median (MAD)):

$$\mathtt{MAD}\,\mathcal{A} = 1.483 \operatorname*{med}_{i=1,\ldots,m} |a^i - \operatorname*{med}_{j=1,\ldots,m} a^j|. \tag{2.10}$$

Primjer 2.3. *Relativnu veličinu elemenata skupa*

$$\mathcal{A} = \{9.05,\ 2.83,\ 3.00,\ 3.16,\ 4.12,\ 3.00,\ 3.50\}$$

možemo bolje usporediti ako skup preslikamo na jedinični interval $[0, 1]$ *pomoću linearnog preslikavanja*

$$\varphi(x) = \tfrac{x-a}{b-a}, \qquad \text{gdje je } a = \min \mathcal{A},\ b = \max \mathcal{A}. \tag{2.11}$$

Dobivamo $\varphi(\mathcal{A}) = \{1., 0., 0.027, 0.053, 0.207, 0.027, 0.108\}$. *Vidi se da je* $a^1 \in \mathcal{A}$ *značajno najveći element u skupu* $\mathcal{A}$.

Prema [74], to egzaktnije možemo ustanoviti tako da prema (2.10) *najprije odredimo* $\mathtt{MAD} = 0.489$ *i novi skup*

$$\begin{aligned}\tilde{\mathcal{A}} =& \{\tilde{a}^i = |a^i - \operatorname*{med}_{j=1,\ldots,m} a^j| / \mathtt{MAD} \colon a \in \mathcal{A}\} \\ =& \{12.04,\ 0.67,\ 0.33,\ 0,\ 1.96,\ 0.33,\ 0.69\}.\end{aligned}$$

Tada element $a^i \in \mathcal{A}$ *za koji je* $\tilde{a}^i > 2.5$ *smatramo „jako stršećim podatkom" (outlier). Dakle, u ovom je slučaju samo prvi element* $a^1 = 9.05$ *outlier u skupu* $\mathcal{A}$.

U statističkoj literaturi [12] uz pojam medijana vežu se i pojmovi donjeg kvartila (element skupa $\mathcal{A}$ koji se nalazi na mjestu $1/4$ sortiranih podataka) i gornjeg kvartila (element skupa $\mathcal{A}$ koji se nalazi na mjestu $3/4$ sortiranih podataka). Koliki bi bio donji i gornji kvartil skupa podataka iz prethodnog primjera?

2.1.3 Najbolji reprezentant težinskih podataka

Ponekad je u praktičnim primjenama podacima potrebno dodijeliti težine (pondere). Na taj način svakom podatku pridružujemo faktor utjecaja ili učestalost njegova pojavljivanja. Primjerice u Primjeru 3.8 [86, str. 30], u kojemu se analizira problem pojave visokog vodostaja rijeke Drave na mjernom mjestu Donji Miholjac, težine podataka su visine vodostaja.

Definicija 2.3. *Neka je* $d \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *kvazimetrička funkcija. Najbolji reprezentant skupa podataka* $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}$ *s odgovarajućim težinama*

$w_1, \ldots, w_m > 0$ u odnosu na kvazimetričku funkciju d je točka $c^\star \in \mathbb{R}$ u kojoj se postiže globalni minimum funkcije $F\colon \mathbb{R} \to \mathbb{R}_+$,

$$F(x) = \sum_{i=1}^{m} w_i d(x, a^i), \tag{2.12}$$

što formalno zapisujemo na sljedeći način:

$$c^\star \in \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^{m} w_i d(x, a^i). \tag{2.13}$$

Primijetite da za točku globalnog minimuma $c^\star \in \mathbb{R}$ vrijedi:

$$F(x) = \sum_{i=1}^{m} w_i d(x, a^i) \geq \sum_{i=1}^{m} w_i d(c^\star, a^i) = F(c^\star), \tag{2.14}$$

pri čemu jednakost vrijedi onda i samo onda ako je $x = c^\star$.

Slično kao u slučaju podataka bez težina, može se pokazati da funkcija

$$F_{LS}(x) = \sum_{i=1}^{m} w_i (x - a^i)^2.$$

postiže jedinstveni globalni minimum u točki

$$c^\star_{LS} = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{i=1}^{m} w_i d_{LS}(x, a^i) = \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i, \quad W = \sum_{i=1}^{m} w_i,$$

koju zovemo *težinska aritmetička sredina* [77].

U slučaju ℓ_1-metričke funkcije, funkcija (2.12) glasi:

$$F_1(x) = \sum_{i=1}^{m} w_i |x - a^i|, \tag{2.15}$$

i postiže svoj globalni minimum na težinskom medijanu $\operatorname*{Med}_i(w_i, a^i)$ skupa $\mathcal{A}$, kao što pokazuje sljedeća lema.

Lema 2.2. *[77] Neka je $I = \{1, \ldots, m\}$ skup indeksa, a $a^1 < \cdots < a^m$ skup podataka s težinama $w_1, \ldots, w_m > 0$. Označimo*

$$J := \{\nu \in I\colon \sum_{i=1}^{\nu} w_i \leq \sum_{i=\nu+1}^{m} w_i\},$$

a za $J \neq \emptyset$, označimo $\nu_0 = \max J$. Tada vrijedi:

(i) Ako je $J = \emptyset$*, (tj.* $w_1 > \sum_{i=2}^{m} w_i$*), minimum funkcije* F_1 *postiže se u točki* $\alpha^\star = a^1$*.*

(ii) Ako je $J \neq \emptyset$ *i* $\sum_{i=1}^{\nu_0} w_i < \sum_{i=\nu_0+1}^{m} w_i$*, minimum funkcije* F_1 *postiže se u točki* $\alpha^\star = a^{\nu_0+1}$*.*

(iii) Ako je $J \neq \emptyset$ *i* $\sum_{i=1}^{\nu_0} w_i = \sum_{i=\nu_0+1}^{m} w_i$*, minimum funkcije* F_1 *postiže se u svakoj točki* $\alpha^\star$ *iz segmenta* $[a^{\nu_0}, a^{\nu_0+1}]$*.*

Dokaz. Primijetite da je na svakom od intervala

$$(-\infty, a^1),\ [a^1, a^2), \ldots, [a^{m-1}, a^m),\ [a^m, \infty)$$

funkcija F_1 linearna funkcija s koeficijentima smjera redom $d_\nu, \nu = 0, \ldots, m$, gdje je

$$\begin{aligned} d_0 &= -\sum_{i=1}^{m} w_i, \\ d_\nu &= \sum_{i=1}^{\nu} w_i - \textstyle\sum_{i=\nu+1}^{m} w_i = d_{\nu-1} + w_\nu + w_{\nu+1}, \quad \nu = 1, \ldots, m-1, \\ d_m &= \sum_{i=1}^{m} w_i. \end{aligned}$$

Ako je $J = \emptyset$, onda za svaki $\nu = 1, \ldots, m$, $\sum_{i=1}^{\nu} w_i - \sum_{i=\nu+1}^{m} w_i > 0$ i $d_0 < 0 < d_\nu$. To znači da je funkcija F_1 strogo padajuća na $(-\infty, a^1)$ i strogo rastuća na (a^1, ∞) i da svoj globalni minimum postiže u točki $\alpha^\star = a^1$.

Ako je $J \neq \emptyset$, primijetite da je $\nu_0 = \max\{\nu \in I \colon d_\nu \leq 0\}$. Budući da je $d_{\nu+1} - d_\nu = w_{\nu+1} + w_{\nu+2} > 0$ i $d_0 < 0$ i $d_m > 0$, niz (d_ν) je rastući i vrijedi:

$$d_0 < d_1 \ldots < d_{\nu_0} \leq 0 < d_{\nu_0+1} < \ldots < d_m. \tag{2.16}$$

Ako je $d_{\nu_0} < 0$, tj. $\sum_{i=1}^{\nu_0} w_i < \sum_{i=\nu_0+1}^{m} w_i$, iz (2.16) zaključujemo da je funkcija F_1 strogo padajuća na intervalu $(-\infty, a^{\nu_0+1})$ i strogo rastuća na intervalu (a^{ν_0+1}, ∞), pa zato svoj globalni minimum postiže u točki $\alpha^\star = a^{\nu_0+1}$.

Ako je $d_{\nu_0} = 0$, tj. $\sum_{i=1}^{\nu_0} w_i = \sum_{i=\nu_0+1}^{m} w_i$, iz (2.16) zaključujemo da je funkcija F_1 strogo padajuća na $(-\infty, a^{\nu_0})$, konstantna na $[a^{\nu_0}, a^{\nu_0+1}]$ i strogo rastuća na (a^{ν_0+1}, ∞). Zato se minimum funkcije F_1 u ovom slučaju postiže u svakoj točki $\alpha^\star$ iz segmenta $[a^{\nu_0}, a^{\nu_0+1}]$. □

Dakle, najbolji ℓ_1-reprezentant skupa podataka s težinama je težinski medijan $\underset{i}{\text{Med}}(w_i, a^i)$. Primijetite da težinski medijan također može biti skup (segment realnih brojeva) ili samo jedan realni broj. Težinski medijan $\underset{i}{\text{med}}(w_i, a^i)$ je broj koji ima svojstvo da je suma težinskih apsolutnih odstupanja do svih podataka minimalna:

$$\sum_{i=1}^{m} w_i|x - a^i| \geq \sum_{i=1}^{m} w_i|\underset{i}{\text{med}}(w_i, a^i) - a^i|, \tag{2.17}$$

pri čemu vrijedi jednakost za $x = \underset{i}{\text{med}}(w_i, a^i)$.

Sljedeći korolar pokazuje da je Lema 2.1 specijalni slučaj Leme 2.2.

Korolar 2.1. *Neka je* $a^1 \leq a^2 \leq \ldots \leq a^m$, $m > 1$, *skup podataka s težinama* $w_1 = \cdots = w_m = 1$. *Tada:*

(i) Ako je m *neparan broj* ($m = 2k + 1$), *minimum funkcije* F_1 *postiže se u točki* $\alpha^\star = a^{k+1}$,

(ii) Ako je m *paran broj* ($m = 2k$), *minimum funkcije* F_1 *postiže se u svakoj točki* $\alpha^\star$ *iz segmenta* $[a^k, a^{k+1}]$.

Dokaz. Najprije primijetite da je u ovom slučaju skup J iz Leme 2.2 uvijek neprazan.

Ako je $m = 2k + 1$, prema Lemi 2.2 (ii) je

$$\begin{aligned}\nu_0 &= \max\{\nu \in I : 2\nu - m \leq 0\} = \max\{\nu \in I : \nu \leq k + \tfrac{1}{2}\} = k,\\ d_{\nu_0} &= d_k = 2k - m = 2k - 2k - 1 < 0,\end{aligned}$$

što znači da se minimum funkcije F_1 postiže u točki $\alpha^\star = a^{k+1}$.

Ako je $m = 2k$, prema Lemi 2.2 (iii) je

$$\begin{aligned}\nu_0 &= \max\{\nu \in I : 2\nu - m \leq 0\} = \max\{\nu \in I : \nu = k\} = k,\\ d_{\nu_0} &= d_k = 2k - m = 2k - 2k = 0,\end{aligned}$$

što znači da se minimum funkcije F_1 postiže u svakoj točki $\alpha^\star$ iz segmenta $[a^k, a^{k+1}]$. □

Određivanje težinskog medijana u općem slučaju vrlo je složeni numerički postupak [39, 77]. U stručnoj literaturi postoje brojni algoritmi za njegovo određivanje [39].

Primjer 2.4. *Težinski medijan skupa $\mathcal{A} \subset \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$, u slučaju kada su težine w_i cijeli brojevi, određuje se slično kao medijan skupa podataka bez težina. Medijan ovakvog skupa dobije se tako da najprije sortiramo elemente skupa $\mathcal{A}$ s odgovarajućom frekvencijom pojavljivanja i nakon toga odredimo srednji element. Primjerice, medijan skupa $\mathcal{A} = \{3, 1, 4, 5, 9\}$ s težinama $3, 1, 3, 2, 2$ srednji je element u nizu*

$$1, 3, 3, 3, 4, 4, 4, 5, 5, 9, 9.$$

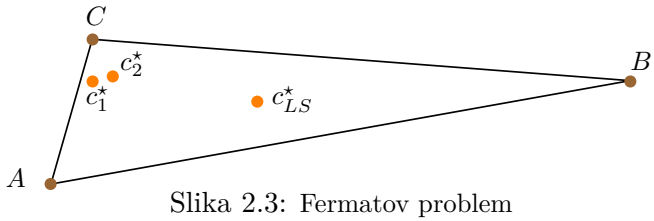
U ovom slučaju, $\operatorname{med} \mathcal{A} = 4$. *Koliki su donji i gornji kvartil skupa $\mathcal{A}$?*

2.2 Reprezentant podataka s dva obilježja

Skup podataka s dva obilježja obično interpretiramo kao konačan podskup $\mathcal{A} = \{a^1 = (x_i, y_i)^T : i = 1, \ldots, a^m\} \subset \mathbb{R}^2$, a geometrijski ga možemo promatrati kao skup točaka u ravnini. U sljedećoj točki dat ćemo kratki povijesni pregled traženja najboljeg reprezentanta skupa podataka s dva obilježja i moguće primjene.

2.2.1 Fermat–Torricelli–Weberov problem

Neka su $A, B, C \in \mathbb{R}^2$ tri nekolinearne točke u ravnini (vidi Sliku 2.3).



Slika 2.3: Frematov problem

Problem određivanja točke $c_2^\star \in \mathbb{R}^2$, za koju je suma euklidskih udaljenosti do vrhova trokuta $\triangle ABC$ minimalna, u stručnoj literaturi naziva se *Fermatov problem.*

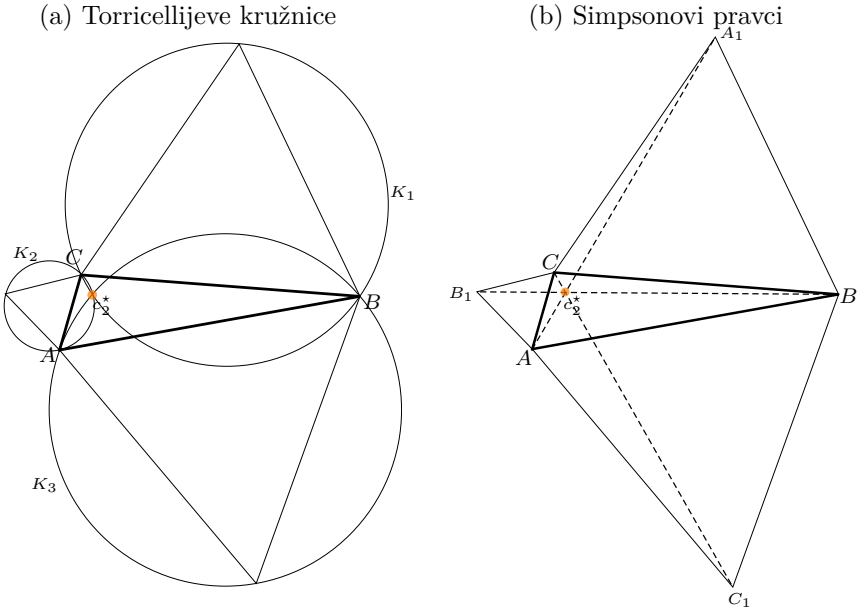
Točka $c_2^\star$ naziva se *geometrijski medijan* točaka A, B, C i može se dobiti [60] na presjeku tzv. Torricellijevih kružnica (vidi Sliku 2.4a) ili na presjeku tzv. Simpsonovih pravaca (vidi Sliku 2.4b). Problem se također može promatrati za različite kvazimetričke funkcije, i to u fizikalnom smislu (Torricellijev problem) ili u ekonometrijskom smislu (Weberov problem) [30].

Specijalno, točka $c_{LS}^\star \in \mathbb{R}^2$ (vidi Sliku 2.3), za koju je suma LS-udaljenosti (suma kvadrata euklidskih udaljenosti) do vrhova trokuta minimalna zove se

centroid ili Steinerova točka (povezano s pojmom centra masa u fizici). Geometrijski gledano, to je težište trokuta koje se dobije na presjeku težišnica trokuta (spojnice vrhova trokuta s polovištima nasuprotnih stranica).

Točka $c_1^\star \in \mathbb{R}^2$ (vidi Sliku 2.3), za koju je suma ℓ_1 udaljenosti do vrhova trokuta A, B, C minimalna, zove se medijan skupa točaka $\{A, B, C\}$.

(a) Torricellijeve kružnice (b) Simpsonovi pravci



Slika 2.4: Frematov problem

Općenito, može se promatrati konačni skup točaka iz $\mathbb{R}^n$ i proizvoljna kvazimetrička funkcija d. Problem određivanja najboljeg d-reprezentanta ima brojne primjene u najrazličitijim područjima: telekomunikacije (problem optimalnog antenskog pokrivanja, problem diskretne mreže), javni sektor (problem optimalnog pokrivanja), ekonomija (optimalna lokacija potrošačkih centara), problem lokacije hubova, robotika, problem optimalne asignacije, problem satne prognoze potrošnje energenata, itd. [30, 82].

2.2.2 Centroid skupa točaka u ravnini

Neka je $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, m\} \subset \mathbb{R}^2$ skup točaka u ravnini. Centroid $c_{LS}^\star$ skupa $\mathcal{A}$ rješenje je optimizacijskog problema

$$\operatorname*{argmin}_{c \in \mathbb{R}^2} \sum_{i=1}^{m} d_{LS}(c, a^i), \tag{2.18}$$

gdje je[5] $d_{LS}(a,b) = d_2^2(a,b) = \|a-b\|_2^2$. Točka $c_{LS}^\star$ točka je u kojoj se postiže globalni minimum funkcije

$$F_{LS}(x,y) = \sum_{i=1}^{m} \|c - a^i\|_2^2 = \sum_{i=1}^{m} [(x-x_i)^2 + (y-y_i)^2], \quad c = (x,y)^T.$$

Funkcija F_{LS} predstavlja sumu kvadrata euklidskih ℓ_2 udaljenosti točaka $a^i \in \mathcal{A}$ do neke točke $c = (x,y)^T$ u ravnini $\mathbb{R}^2$. Prema (2.6), str. 11 vrijedi:

$$F_{LS}(x,y) = \sum_{i=1}^{m} [(x-x_i)^2 + (y-y_i)^2] \geq \sum_{i=1}^{m} (\bar{x}-x_i)^2 + \sum_{i=1}^{m} (\bar{y}-y_i)^2, \quad (2.19)$$

gdje je

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i,$$

pri čemu jednakost u (2.19) vrijedi onda i samo onda ako je $x = \bar{x}$ i $y = \bar{y}$. Zato je rješenje globalno optimizacijskog problema (2.18) centroid skupa točaka $\mathcal{A}$ koji se može eksplicitno zapisati s $c_{LS}^\star = (\bar{x}, \bar{y})^T$.

Dakle, centroid skupa točaka $\mathcal{A}$ u ravnini je točka čija je apscisa aritmetička sredina svih apscisa točaka iz $\mathcal{A}$, a ordinata aritmetička sredina svih ordinata točaka iz $\mathcal{A}$.

Primjer 2.5. *Zadane su tri točke:* $a^1 = (0,0)^T$, $a^2 = (1,3.5)^T$ *i* $a^3 = (14,2.5)^T$. *Centroid skupa* $\{a^1, a^2, a^3\}$ *je točka* $c_{LS}^\star = (5,2)$. *Provjerite!*

2.2.3 Medijan skupa točaka u ravnini

Medijan skupa točaka $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, m\} \subset \mathbb{R}^2$ rješenje je optimizacijskog problema

$$\operatorname*{argmin}_{c \in \mathbb{R}^2} \sum_{i=1}^{m} d_1(c, a^i). \quad (2.20)$$

To je svaka točka u kojoj se postiže globalni minimum funkcije

$$F_1(x,y) = \sum_{i=1}^{m} \|c - a^i\|_1 = \sum_{i=1}^{m} (|x-x_i| + |y-y_i|), \quad c = (x,y)^T.$$

[5] Budući da se u cijelom tekstu najviše koristi euklidska norma $\|\cdot\|_2$, nadalje, ako ne postoji mogućnost nesporazuma, u tom slučaju jednostavno ćemo pisati $\|\cdot\|$.

Funkcija F_1 predstavlja sumu ℓ_1 udaljenosti točaka $a^i \in \mathcal{A}$ do neke točke $c = (x, y)^T$ u ravnini $\mathbb{R}^2$. Prema (2.8), vrijedi:

$$F_1(x,y) = \sum_{i=1}^{m}(|x-x_i|+|y-y_i|) \geq \sum_{i=1}^{m}|\operatorname*{med}_k x_k - x_i| + \sum_{i=1}^{m}|\operatorname*{med}_k y_k - y_i|, \quad (2.21)$$

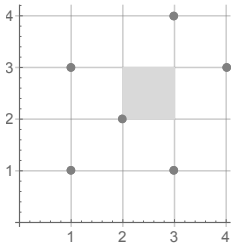
pri čemu jednakost u (2.21) vrijedi onda i samo onda ako je $x = \operatorname*{med}_k x_k$ i $y = \operatorname*{med}_k y_k$. Zato je rješenje globalno optimizacijskog problema (2.20) medijan skupa točaka $\mathcal{A}$ koji se može eksplicitno zapisati s

$$(\operatorname*{med}_k x_k, \operatorname*{med}_k y_k)^T. \quad (2.22)$$

Dakle, medijan skupa točaka $\mathcal{A}$ u ravnini je točka čija je apscisa medijan svih apscisa točaka iz $\mathcal{A}$, a ordinata medijan svih ordinata točaka iz $\mathcal{A}$. Primijetite da medijan skupa $\mathcal{A}$ može biti jedna točka, segment ili čitavi pravokutnik.

Primjer 2.6. *Zadane su tri točke:* $A_1 = (0,0)^T$, $A_2 = (1, 3.5)^T$ *i* $A_3 = (14, 2.5)^T$. *Medijan skupa* $\{A_1, A_2, A_3\}$ *je točka* $c_1^\star = (1, 2.5)^T$. *Provjerite!*

Primjer 2.7. *Medijan skupa* $\mathcal{A} = \{(1,1)^T, (1,3)^T, (2,2)^T, (3,1)^T, (3,4)^T, (4,3)^T\}$ *u ravnini je bilo koja točka iz kvadrata* $[2,3]\times[2,3]$ *(vidi Sliku 2.5) jer je medijan apscisa podataka* $\operatorname{med}\{1,1,2,3,3,4\} \in [2,3]$, *a medijan ordinata podataka* $\operatorname{med}\{1,3,2,1,4,3\} \in [2,3]$.



Slika 2.5: Medijan skupa $\mathcal{A} = \{(1,1)^T, (1,3)^T, (2,2)^T, (3,1)^T, (3,4)^T, (4,3)^T\}$

Zadatak 2.5. *Promijenite poziciju položaja samo jedne točke tako da medijan skupa točaka* $\mathcal{A}$ *iz prethodnog primjera bude točka, segment ili pravokutnik.*

2.2.4 Geometrijski medijan skupa točaka u ravnini

Geometrijski medijan $c^\star$ skupa $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, m\} \subset \mathbb{R}^2$ rješenje je optimizacijskog problema

$$c^\star = \underset{c \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{m} d_2(c, a^i). \tag{2.23}$$

Točka $c^\star$ točka je u kojoj se postiže globalni minimum funkcije

$$F_2(x, y) = \sum_{i=1}^{m} \|c - a^i\|_2 = \sum_{i=1}^{m} \sqrt{(x - x_i)^2 + (y - y_i)^2}, \quad c = (x, y)^T. \tag{2.24}$$

Funkcija F_2 predstavlja sumu euklidskih ℓ_2 udaljenosti točaka $a^i \in \mathcal{A}$ do neke točke $c = (x, y)^T$ u ravnini $\mathbb{R}^2$ i ne može se separirati po varijablama x i y kao u prethodnim slučajevima. Zato se rješenje globalno optimizacijskog problema (2.23) ne može eksplicitno zapisati.

Primjer 2.8. *Zadane su tri točke* $a^1 = (0, 0)^T$, $a^2 = (1, 3.5)^T$ *i* $a^3 = (14, 2.5)^T$. *U cilju pronalaženja geometrijskog medijana treba riješiti sljedeći optimizacijski problem:*

$$\underset{(x,y)^T \in \mathbb{R}^2}{\operatorname{argmin}} \; F_2(x, y),$$
$$F_2(x, y) = \sqrt{x^2 + y^2} + \sqrt{(x - 1)^2 + (y - 3.5)^2} + \sqrt{(x - 14)^2 + (y - 2.5)^2}.$$

Ovaj optimizacijski problem možemo riješiti primjenom programskog sustava *Mathematica*. Najprije definiramo funkciju

```
In[1]:= F2[x_, y_]:= Sqrt[x^2 + y^2] + Sqrt[(x-1)^2 + (y-3.5)^2]
                     + Sqrt[(x-14)^2 + (y-2.5)^2]
```

Problem možemo pokušati riješiti kao problem globalne optimizacije pozivanjem `Mathematica`-modula

```
In[2]:= NMinimize[F2[x, y], {x, y}]
```

Prema [116], modul `NMinimize[]` neki puta može pronaći samo lokalni minimum. U tom slučaju problem možemo pokušati riješiti kao problem lokalne optimizacije, i to pozivanjem `Mathematica`-modula

```
In[2]:= FindMinimum[F2[x, y], {x, 1}, {y, 2}]
```

uz primjenu dobre početne aproksimacije bliske rješenju. U ovom primjeru dobivamo $c_2^\star = (1.51827, 2.5876)^T$.

Primjedba 2.2. *Najpoznatiji algoritam za traženje geometrijskog medijana rješavanjem optimizacijskog problema* (2.23) *je tzv.* Weiszfeldov algoritam *(vidi primjerice [44, 97]). To je iterativni postupak koji je nastao kao specijalni slučaj metode jednostavnih iteracija za rješavanje sustava nelinearnih jednadžbi.*

Odredimo najprije obje parcijalne derivacije funkcije (2.24) *i izjednačimo ih s nulom:*

$$\frac{\partial F_2}{\partial x} = \sum_{i=1}^{m} \frac{x - x_i}{\|c - a^i\|_2} = x \sum_{i=1}^{m} \frac{1}{\|c - a^i\|_2} - \sum_{i=1}^{m} \frac{x_i}{\|c - a^i\|_2} = 0,$$
$$\frac{\partial F_2}{\partial y} = \sum_{i=1}^{m} \frac{y - y_i}{\|c - a^i\|_2} = y \sum_{i=1}^{m} \frac{1}{\|c - a^i\|_2} - \sum_{i=1}^{m} \frac{y_i}{\|c - a^i\|_2} = 0,$$

što možemo zapisati kao:

$$x = \Phi(x, y), \quad y = \Psi(x, y), \tag{2.25}$$

gdje je:

$$\Phi(x, y) = \frac{\sum\limits_{i=1}^{m} \frac{x_i}{\|c-a^i\|_2}}{\sum\limits_{i=1}^{m} \frac{1}{\|c-a^i\|_2}}, \quad \Psi(x, y) = \frac{\sum\limits_{i=1}^{m} \frac{y_i}{\|c-a^i\|_2}}{\sum\limits_{i=1}^{m} \frac{1}{\|c-a^i\|_2}}. \tag{2.26}$$

Ako izaberemo neku početnu aproksimaciju $(x_0, y_0) \in \text{conv}(\mathcal{A})$ *iz konveksne ljuske skupa* $\mathcal{A}$*, onda sustav* (2.25) *možemo rješavati metodom sukcesivnih iteracija*

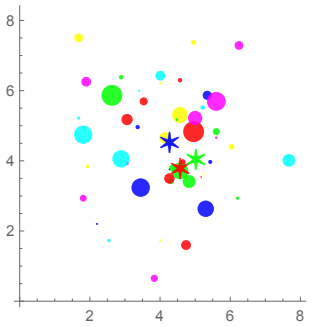
$$x_{k+1} = \Phi(x_k, y_k), \quad y_{k+1} = \Psi(x_k, y_k), \quad k = 0, 1, \ldots \tag{2.27}$$

Analogno Definiciji 2.3, str. 15 u slučaju podataka iz $\mathbb{R}^2$ također se mogu definirati težinska aritmetička sredina, težinski medijan i težinski geometrijski medijan skupa podataka $\mathcal{A}$ (vidi [77, 81]).

Primjer 2.9. *Skup podataka* $\mathcal{A} \subset \mathbb{R}^2$ *u ravnini definiran je na sljedeći način:*

```
In[1]:= SeedRandom[13]
        sig = 1.5; m1 = 50; cen = {4,5};
        podT = Table[cen + RandomReal[NormalDistribution[0, sig], {2}],
                     {i, m1}];
        podW = RandomReal[{0, 1}, m1];
```

Svakom podatku prikazanom na Slici 2.6 pridružena je težina proporcionalna veličini kružića. Na Slici 2.6 zelenom zvjezdicom označen je centroid, crvenom medijan, a plavom geometrijski medijan podataka.



Slika 2.6: Težinski reprezentanti skupa $\mathcal{A}$

Zadatak 2.6. *Zadan je skup* $\mathcal{A} = \{(x_i, y_i)^T \in \mathbb{R}^2 : i = 1, \ldots, 10\}$, *gdje je*

i	1	2	3	4	5	6	7	8	9	10
x_i	9	6	8	1	1	4	4	3	9	10
y_i	5	5	5	2	5	8	1	8	8	4

Nacrtajte skup $\mathcal{A}$ *u koordinatnoj ravnini i odredite centroid, medijan i geometrijski medijan ovog skupa.*

Uputa: Poslužite se niže navedenim *Mathematica*-programom.

```
In[1]:= SeedRandom[2]
        A = RandomInteger[{1, 10}, {10, 2}]
        ListPlot[A, ImageSize -> Small]
        Print["Centroid = ", Mean[A]]
        Print["Medijan = ", Median[A]]
        Psi[x_, y_] := Sum[Norm[{x, y} - A[[i]]], {i, Length[A]}]
        Print["Geometrijski medijan:"]
        NMinimize[Psi[x, y], {x, y}]
```

Rješenje: $c_{LS}^\star = (5.5, 5.1)^T$, $c_1^\star = (5, 5)^T$, $c_2^\star = (6, 5)^T$.

2.3 Reprezentant podataka s više obilježja

U praktičnim primjenama podaci mogu imati veći broj obilježja kao što je spomenuto na početku Uvoda, str. 1, gdje je i navedeno nekoliko takvih primjera. Budući da broj obilježja podataka predstavlja njihovu dimenziju $n \geq 1$, bit će potrebno znati odrediti reprezentant i za podatke proizvoljno visoke dimenzije.

Pretpostavimo da je zadan skup točaka $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i)^T \in \mathbb{R}^n : i = 1, \ldots, m\}$. Treba odrediti točku koja će što bolje reprezentirati taj skup.

Definicija 2.4. *Neka je* $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ *kvazimetrička funkcija.* Najbolji reprezentant *(centar* $c^\star \in \mathbb{R}^n$*) skupa* $\mathcal{A} \subset \mathbb{R}^n$ *u odnosu na kvazimetričku funkciju* d *je*

$$c^\star \in \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} d(c, a^i). \tag{2.28}$$

Točka $c^\star \in \mathbb{R}^n$ *točka je globalnog minimuma funkcije* $F\colon \mathbb{R}^n \to \mathbb{R}_+$ *zadane s*

$$F(c) = \sum_{i=1}^{m} d(c, a^i). \tag{2.29}$$

Specijalno,

(a) u slučaju LS-kvazimetričke funkcije, najbolji reprezentant skupa $\mathcal{A}$ je centroid (težište) skupa

$$c^\star_{LS} = \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} d_{LS}(c, a^i) = \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} \|c - a^i\|_2^2 = \tfrac{1}{m} \sum_{i=1}^{m} a^i,$$

a odgovarajuća minimizirajuća funkcija glasi:

$$F_{LS}(c) = \sum_{i=1}^{m} \|c - a^i\|_2^2;$$

(b) u slučaju ℓ_1-metričke funkcije, najbolji reprezentant skupa $\mathcal{A}$ je medijan skupa

$$c_1 = \operatorname*{med}_i a^i = \left(\operatorname*{med}_i a^i_1, \ldots, \operatorname*{med}_i a^i_n\right)^T \in \operatorname{Med}\mathcal{A} = \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} \|c - a^i\|_1,$$

a odgovarajuća minimizirajuća funkcija glasi:

$$F_1(c) = \sum_{i=1}^{m} \|c - a^i\|_1.$$

2.3.1 Reprezentant težinskih podataka

Definicija 2.5. *Neka je* $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ *kvazimetrička funkcija.* *Najbolji reprezentant* *skupa* $\mathcal{A} \subset \mathbb{R}^n$ *s težinama* $w_1, \ldots, w_m > 0$ *u odnosu na kvazimetričku funkciju* d *je*

$$c^\star \in \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i d(c, a^i). \tag{2.30}$$

Točka $c^\star \in \mathbb{R}^n$ *točka je globalnog minimuma funkcije* $F\colon \mathbb{R}^n \to \mathbb{R}_+$ *zadane s*

$$F(c) = \sum_{i=1}^{m} w_i d(c, a^i). \tag{2.31}$$

(a) Specijalno, ako je d LS-kvazimetrička funkcija, najbolji reprezentant skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ je težinski centroid (težište) skupa

$$c^\star_{LS} = \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i d_{LS}(c, a^i) = \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i \|c - a^i\|_2^2 = \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i, \text{ tj.}$$

$$c^\star_{LS} = \Big(\tfrac{1}{W} \sum_{i=1}^{m} w_i a^i_1, \ldots, \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i_n\Big)^T \quad \text{[po koordinatama]}, \tag{2.32}$$

gdje je $W = \sum_{i=1}^{m} w_i$, a odgovarajuća minimizirajuća funkcija glasi:

$$F_{LS}(c) = \sum_{i=1}^{m} w_i \|c - a^i\|_2^2; \tag{2.33}$$

(b) Ako je d, ℓ_1-metrička funkcija, najbolji reprezentant skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ je težinski medijan skupa

$$\begin{aligned} c^\star_1 &= \operatorname*{med}_i(w_i, a^i) = \big(\operatorname*{med}_i(w_i, a^i_1), \ldots, \operatorname*{med}_i(w_i, a^i_n)\big)^T \in \operatorname{Med}\mathcal{A} \\ &= \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^{m} w_i \|c - a^i\|_1, \end{aligned} \tag{2.34}$$

a odgovarajuća minimizirajuća funkcija glasi:

$$F_1(c) = \sum_{i=1}^{m} w_i \|c - a^i\|_1. \tag{2.35}$$

Naime, vrijedi:

$$\begin{aligned} F_1(c) &= \sum_{i=1}^{m} w_i \|c - a^i\|_1 = \sum_{i=1}^{m} w_i \Big(\sum_{k=1}^{n} |c_k - a_k^i| \Big) \\ &= \sum_{k=1}^{n} \Big(\sum_{i=1}^{m} w_i |c_k - a_k^i| \Big) = \sum_{k=1}^{n} \sum_{i=1}^{m} w_i |c_k - a_k^i| \\ &\geq \sum_{k=1}^{n} \sum_{i=1}^{m} w_i |\operatorname*{med}_j(w_j, a_k^j) - a_k^i| = \sum_{i=1}^{m} \sum_{k=1}^{n} w_i |\operatorname*{med}_j(w_j, a_k^j) - a_k^i| \\ &= \sum_{i=1}^{m} w_i \|c_1^\star - a^i\|_1 = F(c_1^\star), \end{aligned}$$

gdje je $\operatorname*{med}_j(w_j, a_k^j)$ težinski medijan podataka $\{a_k^1, \ldots, a_k^m\}$ s težinama $w_1, \ldots, w_m > 0$.

Zadatak 2.7. *Slično kao što je pokazano za težinski medijan, pokažite da funkcija F_{LS} zadana s* (2.33) *postiže svoj globalni minimum na težinskom centroidu $c_{LS}^\star$ zadanom s* (2.32).

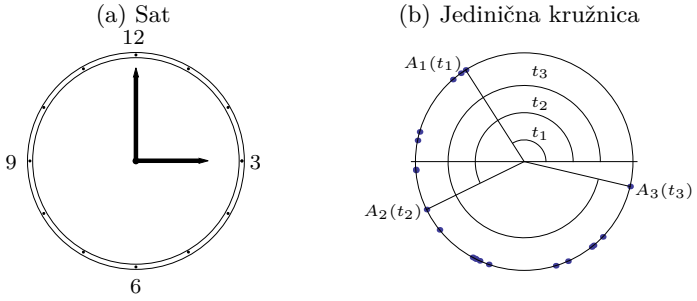
2.4 Reprezentant periodičnih podataka

Problem određivanja najboljeg reprezentanta skupa podataka u slučaju pojava koje pokazuju periodičnost u ponašanju također je često prisutan u literaturi [62]. Temperatura zraka na nekom mjernom mjestu tijekom godine, vodostaj rijeke na nekom mjernom mjestu, seizmičke aktivnosti tijekom više godina na nekom području, količina svjetla tijekom dana, itd. primjeri su takvih pojava. Matematički gledano, treba promatrati skup točaka $\mathcal{A}$ na kružnici. Naime, ako bismo takav skup podataka prikazali kao ranije na brojevnom pravcu, onda bi primjerice podaci s početka i kraja iste godine bili međusobno daleko, a zapravo pripadaju istom godišnjem dobu. I za takav skup treba definirati kvazimetričku funkciju i odrediti centar podataka.

Primjer 2.10. *Neka točke $t_i \in \mathcal{A}$ predstavljaju pozicije male kazaljke na satu s* 12 *oznaka (vidi Sliku 2.7a). Funkciju koja će pokazivati udaljenost na ovom skupu definirat ćemo kao* proteklo vrijeme *od trenutka t_1 do trenutka t_2:*

$$d(t_1, t_2) = \begin{cases} t_2 - t_1, & \text{ako } t_1 \leq t_2 \\ 12 + (t_2 - t_1), & \text{ako } t_1 > t_2 \end{cases}.$$

Udaljenost $d(t_1, t_2)$ predstavlja proteklo vrijeme na satu od trenutka "t_1" do trenutka "t_2". Primjerice: $d(2, 7) = 5$, ali $d(7, 2) = 12 + (-5) = 7$. Primijetite da ova funkcija nema svojstvo simetričnosti.



Slika 2.7: Skup podataka na kružnici

Primjer 2.11. *Neka točke $t_i \in \mathcal{A}$ predstavljaju pozicije male kazaljke na satu s* 12 *oznaka (vidi Sliku 2.7a). Funkciju koja će pokazivati udaljenost na ovom skupu definirat ćemo kao* duljinu vremenskog intervala *od trenutka t_1 do trenutka t_2:*

$$d(t_1, t_2) = \begin{cases} |t_2 - t_1|, & \text{ako } |t_2 - t_1| \leq 6 \\ 12 - |t_2 - t_1|, & \text{ako } |t_2 - t_1| > 6 \end{cases}.$$

Udaljenost $d(t_1, t_2)$ predstavlja duljinu vremenskog intervala na satu s 12 oznaka od trenutka "t_1" do trenutka "t_2". Primjerice, $d(2, 9) = 12 - 7 = 5$ i $d(2, 7) = 7 - 2 = 5$. Provjerite je li ovako definirana funkcija metrika na skupu $\mathcal{A}$.

2.4.1 Reprezentant podataka na jediničnoj kružnici

Općenito, neka je zadan skup podataka (T_i, w_i), $i = 1, \ldots, m$, gdje je T_i, trenutak unutar $M \geq 1$ sukcesivnih godina u kojemu se promatrana pojava dogodila, a $w_i > 0$ neka je intenzitet te pojave u trenutku T_i. Vremenski trenuci T_i mogu biti dani (primjerice, kod podataka o vodostaju neke rijeke na nekom mjestu), sati (temperature zraka na nekom mjestu) ili sekunde (trenuci potresa). Želimo identificirati moment u godini u kojemu je ta pojava najprisutnija.

Ako se trenuci $T_1, \ldots, T_m$ promatraju kao obični vremenski niz, onda bi primjerice podaci s početka i kraja iste godine bili međusobno daleko, a zapravo pripadaju istom godišnjem dobu. Zato najprije svakoj godini

pridružujemo interval duljine 2π, a nizu od M sukcesivnih godina interval $[0, 2\pi M]$. Na taj način niz $T_1, \ldots, T_m$ prelazi u niz $T_1', \ldots, T_m' \in [0, 2\pi M]$.

Budući da je u našem razmatranju važan samo trenutak u godini, a ne i godina u kojoj se pojavljuje podatak, onda ćemo umjesto niza (T_i') definirati novi niz $t_i \in [0, 2\pi]$, $i = 1, \ldots, m$, gdje je:

$$t_i = 2\pi T_i' (\text{mod } 2\pi), \quad i = 1, \ldots, m, \tag{2.36}$$

(ostatak pri dijeljenju broja $2\pi T_i'$ s 2π). Broj $t_i \in [0, 2\pi]$ predstavlja trenutak u godini koji je od 1. siječnja „udaljen" $\frac{t_i}{2\pi}$ dijela godine.

Pomoću niza (2.36) definiramo skup podataka

$$\mathcal{A} = \{a(t_i) = (\cos t_i, \sin t_i)^T \in \mathbb{R}^2 \colon t_i \in [0, 2\pi],\ i = 1, \ldots, m\} \subset K, \tag{2.37}$$

gdje je $K = \{(x, y)^T \in \mathbb{R}^2 \colon x^2 + y^2 = 1\}$ jedinična kružnica sa središtem u ishodištu pravokutnog koordinatnog sustava.

U sljedećoj lemi definirana je metrika na jediničnoj kružnici i dana su njena osnovna svojstva (vidi također [45, 54]).

Lema 2.3. *Neka je* $K = \{a(t) = (\cos t, \sin t)^T \in \mathbb{R}^2 \colon t \in [0, 2\pi]\}$ *jedinična kružnica sa središtem u ishodištu koordinatnog sustava. Funkcija* $d_K \colon K \times K \to \mathbb{R}_+$ *zadana s:*

$$d_K(a(t_1), a(t_2)) = \begin{cases} |t_1 - t_2|, & \text{ako } |t_1 - t_2| < \pi, \\ 2\pi - |t_1 - t_2|, & \text{ako } |t_1 - t_2| > \pi, \end{cases} \tag{2.38}$$

metrika je na K *i može se također zapisati na sljedeći način:*

$$d_K(a(t_1), a(t_2)) = \pi - \left| |t_1 - t_2| - \pi \right|, \quad t_1, t_2 \in [0, 2\pi]. \tag{2.39}$$

Dokaz. Direktnom provjerom lako je vidjeti da su (2.38), odnosno (2.39), ekvivalentni oblici funkcije d_K. Pokažimo da je d_K metrika na K.

Najprije pokažimo da je $d_K(a(t_1), a(t_2)) \geq 0$, za sve $t_1, t_2 \in [0, 2\pi]$. Neka je $t_1, t_2 \in [0, 2\pi]$. Vrijedi:

$$\begin{aligned} &0 \leq |t_1 - t_2| \leq 2\pi \ \Rightarrow\ -\pi \leq |t_1 - t_2| - \pi \leq \pi \ \Rightarrow\ \left| |t_1 - t_2| - \pi \right| \leq \pi \\ &\Rightarrow\ d_K(a(t_1), a(t_2)) = \pi - \left| |t_1 - t_2| - \pi \right| \geq 0. \end{aligned}$$

Također vrijedi $d_K(a(t_1), a(t_2)) = 0 \Leftrightarrow t_1 = t_2$. Naime, iz $t_1 = t_2$ slijedi $d_K(a(t_1), a(t_2)) = 0$. Obrnuto, ako je $d_K(a(t_1), a(t_2)) = 0$, vrijedi

$$\pi = \left| |t_1 - t_2| - \pi \right|. \tag{2.40}$$

Ako je $|t_1 - t_2| - \pi \leq 0$, onda iz (2.40) slijedi $\pi = \pi - |t_1 - t_2|$, odakle slijedi $t_1 = t_2$. Ako je $|t_1 - t_2| - \pi \geq 0$, onda iz (2.40) slijedi $\pi = |t_1 - t_2| - \pi$, iz čega slijedi $2\pi = |t_1 - t_2|$, a to je moguće onda i samo onda ako $t_1 = t_2$.

Konačno, vrijedi $d_K(a(t_1), a(t_2)) \leq d_K(a(t_1), a(t_3)) + d_K(a(t_3), a(t_2))$ za sve $t_1, t_2, t_3 \in K$. Pri tome, ako $a(t_3)$ leži na luku između $a(t_1)$ i $a(t_2)$, vrijedi jednakost. U suprotnom vrijedi stroga nejednakost. □

Koristeći metriku (2.38) tj. (2.39), možemo definirati najbolji reprezentant skupa $\mathcal{A}$ na jediničnoj kružnici.

Definicija 2.6. *Najbolji reprezentant skupa* $\mathcal{A} = \{a(t_i) \in K : t_i \in [0, 2\pi], i = 1, \ldots, m\}$ *s težinama* $w_1, \ldots, w_m > 0$ *u odnosu na metriku* d_K *definiranu s* (2.38)*, odnosno* (2.39)*, je točka* $c^\star(t^\star) = (\cos t^\star, \sin t^\star)^T \in K$*, gdje je:*

$$t^\star = \underset{\tau \in [0,2\pi]}{\operatorname{argmin}} \sum_{i=1}^{m} w_i d_K(a(\tau), a(t_i)), \quad a(\tau) = (\cos\tau, \sin\tau)^T \in K, \tag{2.41}$$

tj. $t^\star \in [0, 2\pi]$ *točka je u kojoj funkcija* $\Phi : [0, 2\pi] \to \mathbb{R}_+$ *zadana s:*

$$\Phi(\tau) = \sum_{i=1}^{m} w_i d_K(a(\tau), a(t_i)) \tag{2.42}$$

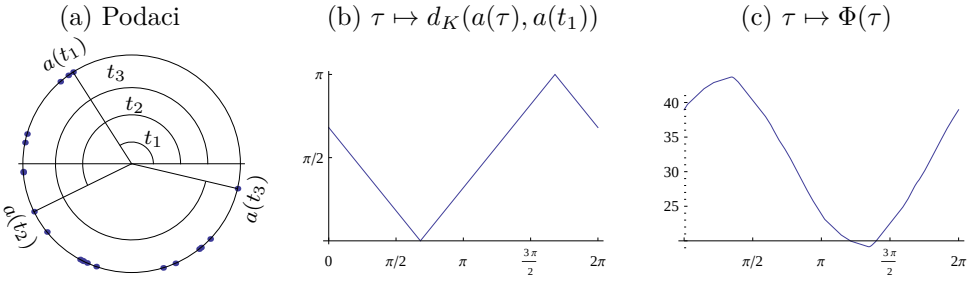
postiže svoj globalni minimum.

Primijetite da funkcija Φ nije konveksna ni diferencijabilna i općenito može imati više lokalnih minimuma. Za rješavanje GOP (2.41) možemo primijeniti optimizacijski algoritam `DIRECT` [38, 46, 85].

Primjer 2.12. *Neka je* $t_1, \ldots, t_m$ *konačni niz slučajnih brojeva iz* $\mathcal{N}(4, 1.2)$*. Definirajmo skup podataka* $\mathcal{A} = \{a(t_i) = (\cos t_i, \sin t_i)^T \in \mathbb{R}^2 : i = 1, \ldots, m\}$ *s težinama* $w_i > 0$, $i = 1, \ldots, m$*. Skup* $\mathcal{A}$ *označen je crnim točkicama na Slici 2.8a, a funkcija* $\tau \mapsto d_K(a(\tau), a(t_1))$ *i odgovarajuća funkcija* Φ *prikazane su na Slici 2.8b i Slici 2.8c.*

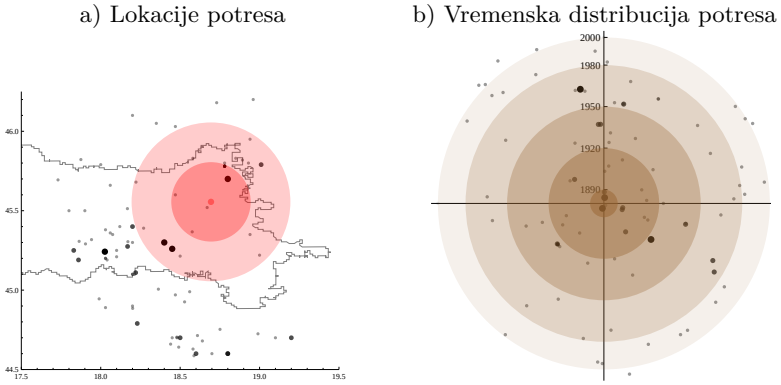
2.4.2 Burnov dijagram

Za grafičko prikazivanje periodičnih pojava zgodno je koristiti tzv. Burnov dijagram (vidi primjerice [98]). Neka točka T na Burnovom dijagramu prikazana je kao $T = r(\cos t, \sin t)^T$, gdje je t mjera kuta u radijanima kojeg zatvara radij vektor točke T s pozitivnim smjerom apscise, a r udaljenost točke T do ishodišta koordinatnog sustava.



Slika 2.8: Podaci i udaljenost na jediničnoj kružnici

Primjer 2.13. *Na Slici 2.9a prikazane su lokacije potresa u okolici Osijeka od 1880. godine. Točke na Burnovom dijagramu (Slika 2.9b) identificiraju pojedine potrese, tako da udaljenost od ishodišta odgovara godini kad se potres dogodio, pozicija na kružnici trenutak u godini, a veličina točke odgovara magnitudi. Na Slici 2.9 vidi se da se posljednji jači potres u neposrednoj okolici Osijeka dogodio krajem zime 1922. godine na geografskoj poziciji* $(18.8, 45.7)$ *(u blizini mjesta Lug, dvadesetak kilometara sjeveroistočno od Osijeka) i da je imao magnitudu* 5.1.



Slika 2.9: Potresi u okolici Osijeka od 1880. godine

Poglavlje 3

Grupiranje podataka

Definicija 3.1. *Neka je* $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ *skup koji sadrži* $m \geq 2$ *elemenata. Rastav skupa* $\mathcal{A}$ *na* $1 \leq k \leq m$ *disjunktnih nepraznih podskupova* $\pi_1, \ldots, \pi_k$*, takvih da je*

$$\bigcup_{j=1}^{k} \pi_j = \mathcal{A}, \qquad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \qquad |\pi_j| \geq 1, \quad j = 1, \ldots, k, \tag{3.1}$$

zovemo k*-particija skupa* $\mathcal{A}$ *i označavamo s* $\Pi = \{\pi_1, \ldots, \pi_k\}$*. Elemente particije zovemo klasteri, a skup svih particija skupa* $\mathcal{A}$ *sastavljenih od* k *klastera koje zadovoljavaju* (3.1) *označavamo s* $\mathcal{P}(\mathcal{A}; k)$*.*

Nadalje, kad god budemo govorili o particiji skupa $\mathcal{A}$, podrazumijevat ćemo da je ona sastavljena od ovakvih podskupova skupa $\mathcal{A}$. Sljedeći teorem daje broj svih particija skupa $\mathcal{A}$ iz Definicije 3.1.

Teorem 3.1. *Broj svih particija skupa* $\mathcal{A}$ *sastavljenih od* k *klastera jednak je Stirlingovom broju druge vrste*

$$|\mathcal{P}(\mathcal{A}, k)| = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^m. \tag{3.2}$$

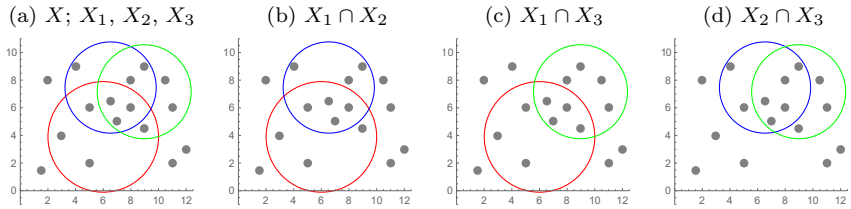
U dokazu Teorema 3.1 koristit ćemo poznatu formulu uključivanja isključivanja zapisanu u sljedećoj lemi.

Lema 3.1. *(Formula uključivanja isključivanja) Neka su* $X_1, \ldots, X_k$ *podskupovi konačnog skupa* X*. Broj elemenata od* X *koji ne pripadaju niti jednom od podskupova* $X_1, \ldots, X_k$ *jednak je*

$$\Big| \bigcap_{i=1}^{k} \overline{X}_i \Big| = |X| - \sum_{1 \leq i \leq k} |X_i| + \sum_{1 \leq i < j \leq k} |X_i \cap X_j| + \cdots + (-1)^k |X_1 \cap \cdots \cap X_k|.$$

Umjesto dokaza Leme 3.1 pokažimo sljedeću ilustraciju. Promatramo skup X sa 16 elemenata i tri njegova podskupa: X_1 (7 elemenata u crvenoj kružnici), X_2 (7 elemenata u plavoj kružnici) i X_3 (8 elemenata u zelenoj kružnici) prikazane na Slici 3.1a. Presjeci $X_1 \cap X_2$ i $X_1 \cap X_3$ imaju po 4 elementa, a presjek $X_2 \cap X_3$ ima 5 elemenata (vidi Slike 3.1b, c, d). Konačno, presjek $X_1 \cap X_2 \cap X_3$ ima 3 elementa (vidi Sliku 3.1a). Zato je

$$\begin{aligned} |\overline{X}_1 \cap \overline{X}_2 \cap \overline{X}_3| &= |X| - (|X_1| + |X_2| + |X_3|) + (|X_1 \cap X_2| + |X_1 \cap X_3| + |X_2 \cap X_3|) \\ &\quad - |X_1 \cap X_2 \cap X_3| \\ &= 16 - (7+7+8) + (4+4+5) - 3 = 4. \end{aligned}$$



Slika 3.1: Broj elemenata skupa X koji ne pripadaju niti jednom od podskupova X_1, X_2, X_3

Dokaz. (Teorema 3.1) Zbog jednostavnosti, a bez gubitka općenitosti, pretpostavimo da je $\mathcal{A} = \{1, \ldots, m\}$, a njegova k-particija $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$, gdje su $\pi_j \subset \mathcal{A}$ neprazni disjunktni podskupovi skupa $\mathcal{A}$. Označimo broj svih ovakvih particija skupa $\mathcal{A}$ s $|\mathcal{P}(\mathcal{A}; k)|$ i definirajmo funkcije $f\colon \mathcal{A} \to J$, $J = \{1, \ldots, k\}$, formulom

$$f(x) = j, \quad \text{ako je } x \in \pi_j.$$

Ovakvih funkcija ima $|\mathcal{P}(\mathcal{A}; k)|$, a permutiramo li ovih k skupova $\pi_1, \ldots, \pi_k$, dobivamo ukupni broj svih surjekcija iz $\mathcal{A}$ na J:

$$k!\,|\mathcal{P}(\mathcal{A}; k)|. \tag{3.3}$$

Broj svih surjekcija iz $\mathcal{A}$ na J možemo izračunati i na drugi način tako da od broja svih funkcija iz $\mathcal{A}$ u J oduzmemo broj funkcija iz $\mathcal{A}$ u J koje nisu surjekcije.

Neka je X skup svih funkcija iz $\mathcal{A}$ u J. Broj $|X|$ jednak je broju svih varijacija s ponavljanjem k-tog razreda od m elemenata: k^m.

Funkcija iz $\mathcal{A}$ u J nije surjekcija ako:

1. u slici nema jednog od elemenata iz J. Skup X_i svih takvih funkcija, koje u slici nemaju element $i \in J$ ima $(k-1)^m$ (broj svih varijacija s ponavljanjem od m elemenata $(k-1)$-og razreda). Skup svih takvih funkcija $\bigcup_{1\le i\le k} X_i$ ima $\binom{k}{1}(k-1)^m$ elemenata,

2. u slici nema jednog od parova međusobno različitih elemenata iz J. Skup $\bigcup_{1\le i<j\le k}(X_i\cap X_j)$ svih takvih funkcija ima $\binom{k}{2}(k-2)^m$ elemenata (broj svih varijacija s ponavljanjem od m elemenata $(k-2)$-og razreda za svaki od međusobno različitih parova elemenata skupa J),

   itd.

Funkcija iz X je surjekcija ako i samo ako ne leži niti u jednom od skupova $X_1,\dots,X_k$, odnosno onda i samo onda ako pripada skupu $\bigcap_{i=1}^{k} \bar{X}_i$. Koristeći Lemu 3.1 dobivamo broj svih surjekcija skupa $\mathcal{A}$ na J:

$$\begin{aligned}
\Big|\bigcap_{i=1}^{k}\overline{X}_i\Big| &= |X| - \sum_{1\le i\le k}|X_i| + \sum_{1\le i<j\le k}|X_i\cap X_j| + \cdots + (-1)^k|X_1\cap\cdots\cap X_k| \\
&= k^m - \binom{k}{1}(k-1)^m + \binom{k}{2}(k-2)^m + \cdots + (-1)^k\binom{k}{k}(k-k)^m \\
&= \sum_{j=0}^{k}(-1)^j\binom{k}{j}(k-j)^m \quad [s := k-j] \\
&= \sum_{s=k}^{0}(-1)^{k-s}\binom{k}{k-s}s^m \quad [\text{za } s=0,\ s^m=0] \\
&= \sum_{s=1}^{k}(-1)^{k-s}\binom{k}{k-s}s^m \quad \left[\text{prema } \binom{n}{r} = \binom{n}{n-r}\right] \\
&= \sum_{s=1}^{k}(-1)^{k-s}\binom{k}{s}s^m \overset{[j:=s]}{=} \sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^m.
\end{aligned}$$

Izjednačavajući ovu formulu s (3.3), dobivamo traženu formulu (3.2). □

Specijalno, iz Teorema 3.1 dobivamo:

$$\begin{aligned}
&\text{za } k=2: \quad |\mathcal{P}(\mathcal{A};2)| = \tfrac{1}{2}(2^m-2) = 2^{m-1}-1, \\
&\text{za } k=3: \quad |\mathcal{P}(\mathcal{A};3)| = \tfrac{1}{2}\left(1-2^m+3^{m-1}\right).
\end{aligned}$$

Primjer 3.1. *Broj svih k-particija skupa $\mathcal{A}$ koje zadovoljavaju Definiciju 3.1 može biti ogroman. Za $m = 5, 10, 50, 1200, 10^6$ i $k = 2, 3, 4, 5, 6, 8, 10$ približni broj svih k-particija skupa $\mathcal{A}$ prikazan je u Tablici 3.1.*

$\approx \lvert\mathcal{P}(\mathcal{A}; k)\rvert$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=8$	$k=10$
$m=5$	15	25	10	1	–	–	–
$m=10$	511	9330	34105	42525	22827	750	1
$m=50$	10^{15}	10^{23}	10^{29}	10^{33}	10^{36}	10^{41}	10^{44}
$m=1200$	10^{361}	10^{572}	10^{721}	10^{837}	10^{931}	10^{1079}	10^{1193}
$m=10^6$	10^{301030}	10^{477120}	10^{602058}	10^{698968}	10^{778148}	10^{903085}	10^{10^6}

Tablica 3.1: Približni broj k-particija u ovisnosti o broju m elemenata skupa $\mathcal{A}$ i broju k klastera

Primjer 3.2. *Zadan je skup $\mathcal{A} \subset \mathbb{R}^2$ prikazan na Slici 4.6a, str. 78, koji sadržava $m = 1200$ elemenata. U Tablici 3.1 može se vidjeti približan broj svih njegovih k-particija s $k = 2, 3, 4, 5, 6, 8$ i 10 klastera.*

Ako uvedemo kriterij da je bolja particija ona čiji su klasteri kompaktniji i bolje razdvojeni, onda bismo mogli postaviti pitanje globalno optimalne (najbolje) particije.

3.1 Optimalna k-particija

Kad bismo uveli neku kvazimetričku funkciju $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, onda bismo mjeru kompaktnosti i dobre razdvojenosti klastera u nekoj particiji Π s k klastera $\pi_1, \ldots, \pi_k$ mogli definirati na sljedeći način:

1. U svakom klasteru π_j odredimo centar $c_j \in \operatorname{argmin}_{x \in \mathbb{R}^n} \sum_{a^i \in \pi_j} d(x, a^i)$,
2. Za svaki klaster π_j odredimo ukupno „rasipanje” (suma udaljenosti točaka klastera π_j do centra c_j) $\mathcal{F}(\pi_j) = \sum_{a^i \in \pi_j} d(c_j, a^i)$,
3. Zbroj svih vrijednosti $\mathcal{F}(\pi_j)$ po svim klasterima daje mjeru kompaktnosti i dobre razdvojenosti klastera u particiji i predstavlja funkciju cilja u ovom optimizacijskom problemu (vidi (3.5)).

Na Slici 4.6, str. 78, možemo vidjeti po jednu particiju s $k = 2, 3, 4, 5, 6, 7$ i 8 klastera i odgovarajuće vrijednosti LS-funkcije cilja. Primijetimo da se

povećanjem broja klastera smanjuje vrijednost funkcije cilja. Primjerice, na Slici 4.6c prikazana je jedna od mnogobrojnih (vidi Tablicu 3.1) 3-particija skupa $\mathcal{A}$. Na njoj funkcija cilja $\mathcal{F}_{LS}$ postiže vrijednost 19860. Opravdano je postaviti pitanje je li to i najbolja 3-particija, tj. može li se pronaći neka druga 3-particija s nižom vrijednosti funkcije cilja?

Općenito bismo mogli postaviti barem nekoliko sljedećih pitanja:

1. Je li navedena funkcija cilja najprikladnija za ovaj primjer?
2. Koliki je najprikladniji broj klastera u particiji?
3. Imaju li particije prikazane na Slici 4.6, str. 78 najniže vrijednosti funkcije cilja od svih mogućih particija s tim brojem klastera?

Iz navedenog primjera vidi se da odgovori na postavljena pitanja neće biti jednostavni. Pitanje izbora funkcije cilja kao i pitanje najprikladnijeg broja klastera u particiji zadire u prethodnu statističku analizu podataka. U ovom ćemo udžbeniku za definiranje funkcije cilja uglavnom koristiti LS-kvazimetričku funkciju i ℓ_1-metričku funkciju. Izborom najprikladnije particije sa sferičnim klasterima bavit ćemo se u točki 5, str. 87, izborom najprikladnije particije s elipsoidnim klasterima u točki 6.5, str. 114, a izborom najprikladnije fuzzy-particije u točki 7.4.

Odmah treba reći da problem traženja optimalne particije spada u NP-teške probleme [59] nekonveksne optimizacije općenito nediferencijabilne funkcije više varijabli, koja najčešće posjeduje značajan broj stacionarnih točaka. Traženje optimalne particije općenito neće biti moguće provesti pretraživanjem čitavog skupa $\mathcal{P}(\mathcal{A};k)$. U ovom udžbeniku bavit ćemo se traženjem optimalne particije sa sferičnim klasterima u točki 4, str. 67, traženjem optimalne particije s elipsoidnim klasterima u točki 6.4, str. 108 i traženjem optimalne fuzzy-particije u točki 7, str. 117.

Općenito, ako je $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $\mathbb{R}_+ = [0, +\infty\rangle$ neka kvazimetrička funkcija (vidi točku 2, str. 9), onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar

$$c_j \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{a\in\pi_j} d(x,a). \tag{3.4}$$

Kvaliteta particije određena je vrijednošću funkcije cilja $\mathcal{F}\colon \mathcal{P}(\mathcal{A};k) \to \mathbb{R}_+$, koja se obično definira kao suma udaljenosti do centra klastera po svim klasterima. U tom smislu, globalno optimalnom k-particijom (k-GOPart) smatramo rješenje sljedećeg globalno optimizacijskog problema (GOP):

$$\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi), \qquad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} d(c_j,a). \tag{3.5}$$

Teorem 3.2. *Povećanjem broja klastera u particiji vrijednost funkcije cilja $\mathcal{F}$ se ne povećava.*

Dokaz teorema vidi na str. 65.

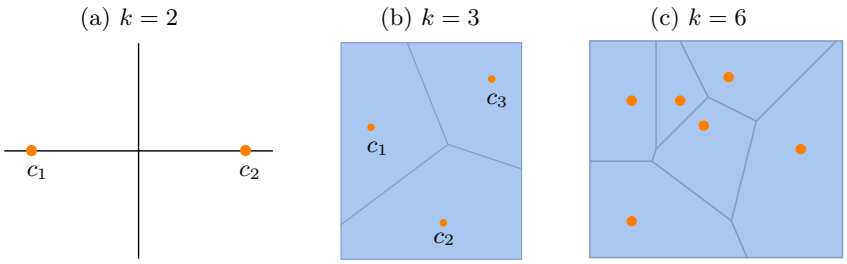
3.1.1 Princip minimalnih udaljenosti i Voronoijev dijagram

Princip minimalnih udaljenosti (vidi Algoritam 3.1, str. 39 ili (3.39), str. 58) usko je povezan s tzv. Voronoijevim dijagramom ili Dirichletovom teselacijom (vidi primjerice [3, 66, 116]).

Neka je d obična euklidska udaljenost u ravnini. Pogledajmo najprije slučaj $k = 2$ klastera s centrima c_1, c_2 u ravnini $\mathbb{R}^2$. Svi elementi $a \in \mathcal{A} \subset \mathbb{R}^2$ koji leže na simetrali $\sigma(c_1, c_2)$ segmenta $\overline{c_1c_2}$ jednako su udaljeni od centara $c_1, c_2 \in \mathbb{R}^2$. Simetrala $\sigma(c_1, c_2)$ okomita je na segment $\overline{c_1c_2}$ dijeli ravninu $\mathbb{R}^2$ na dvije poluravnine – Voronoijeva područja:

$$VR(c_1) = \{x \in \mathbb{R}^2 \colon d(c_1, x) < d(c_2, x)\},$$
$$VR(c_2) = \{x \in \mathbb{R}^2 \colon d(c_1, x) > d(c_2, x)\}.$$

Simetrala σ predstavlja Voronoijev dijagram skupa centara $\{c_1, c_2\}$ (vidi Sliku 3.2a).



Slika 3.2: Princip minimalnih udaljenosti i Voronoijev dijagram

U slučaju $k = 3$ klastera s centrima c_1, c_2, c_3 u ravnini $\mathbb{R}^2$ simetrala $\sigma(c_1, c_2)$ segmenta $\overline{c_1c_2}$ definira dvije poluravnine $M(c_1, c_2)$, $M(c_2, c_1)$, simetrala $\sigma(c_1, c_3)$ segmenta $\overline{c_1c_3}$ definira poluravnine $M(c_1, c_3)$, $M(c_3, c_1)$, a simetrala $\sigma(c_2, c_3)$ segmenta $\overline{c_2c_3}$ definira dvije poluravnine $M(c_2, c_3)$, $M(c_3, c_2)$. Voronoijeva područja s centrima c_1, c_2, c_3 definirana su s:

$$VR(c_1) = M(c_1, c_2) \cap M(c_1, c_3),$$
$$VR(c_2) = M(c_2, c_1) \cap M(c_2, c_3),$$
$$VR(c_3) = M(c_3, c_1) \cap M(c_3, c_2),$$

a Voronoijev dijagram centara c_1, c_2, c_3 (vidi Sliku 3.2b) s:

$$V(c_1,c_2,c_3) = (\overline{VR(c_1)} \cap \overline{VR(c_2)}) \cup (\overline{VR(c_1)} \cap \overline{VR(c_3)}) \cup (\overline{VR(c_2)} \cap \overline{VR(c_3)})$$

Općenito, u slučaju k klastera s centrima $c_1, \ldots, c_k$, Voronoijeva područja definirana su kao:

$$VR(c_j) = \bigcap_{s \neq j} M(c_j, c_s), \quad j = 1, \ldots, k,$$

dok je Voronoijev dijagram definiran kao unija presjeka zatvarača svih parova Voronoijevih područja:

$$V(c_1, \ldots, c_k) = \bigcup_{s \neq j} \overline{VR(c_j)} \cap \overline{VR(c_s)}.$$

Primijetite da se klaster $\pi(c_j)$ dobiven principom minimalnih udaljenosti (3.39) nalazi u Voronoijevom području $VR(c_j)$ ograničenom Voronoijevim dijagramom.

Slično se definira i Voronoijevo područje i Voronoijev dijagram za $k > 3$ (vidi Sliku 3.2c dobivenu korištenjem programskog sustava *Mathematica*).

Zadatak 3.1. *Definirajte i nacrtajte Voronoijeve dijagrame za slučaj ℓ_1 i ℓ_∞ metričkih funkcija.*

Zadatak 3.2. *Odredite Voronoijev dijagram u slučaju $k = 3$ razmatrajući trokutu $\triangle(c_1, c_2, c_3)$ opisanu kružnicu. Može li se ovakvo razmišljanje primijeniti i u slučaju $k > 3$?*

3.1.2 k-means algoritam I

Ne postoji metoda koja bi uspješno riješila GOP (3.5). Ipak, postoji dobro poznati k-means algoritam koji daje lokalno optimalno rješenje koje jako ovisi o izboru početne aproksimacije. Za izabranu početnu particiju $\Pi^{(0)}$, k-means algoritam u konačno mnogo koraka pronalazi lokalno optimalnu particiju. Algoritam se obično zadaje u dva koraka koji se iterativno sukcesivno ponavljaju, a završava kada više nema razlike između prethodne i aktualne particije.

Algoritam 3.1. *[k-means algoritam I]*

Korak A: *Pridruživanje (assignment step). Za dani skup točaka $z_1, \ldots, z_k \in \mathbb{R}^n$ principom minimalnih udaljenosti odrediti klastere π_j, $j = 1, \ldots, k$, particije $\Pi = \{\pi_1, \ldots, \pi_k\}$,*

$$\pi_j := \pi_j(z_j) = \{a \in \mathcal{A} \colon d(z_j, a) \leq d(z_s, a)\, \forall s = 1, \ldots, k\}.$$

Korak B: *Korekcija (update step). Za danu particiju* $\Pi = \{\pi_1, \ldots, \pi_k\}$ *odrediti centre klastera* $c_j \in \operatorname{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a)$, $j = 1, \ldots, k$ *i izračunati vrijednost funkcije cilja* $\mathcal{F}(\Pi)$ *prema* (3.5)*;*

Staviti $z_j = c_j$, $j = 1, \ldots, k$*;*

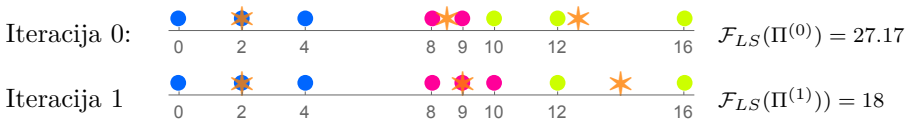
Primjedba 3.1. *U Koraku A može se dogoditi da se neki element* $a \in \mathcal{A}$ *pojavi na granici između dva ili više klastera. Odluka o svrstavanju tog elementa u neki klaster može u značajnoj mjeri predodrediti daljni tijek iterativnog procesa (vidi [92]). Jedan primjer ovakve situacije pojavljuje se kod problema optimalnog definiranja izbornih jedinica u nekoj zemlji (vidi Primjer 1.8, str. 6). Pri tome se gotovo uvijek pojavljuje potreba dijeljenja glasača nekog grada u dvije ili više izbornih jedinica (u Republici Hrvatskoj to je slučaj s gradom Zagrebom). Ovakav problem razmatrat ćemo kod fuzzy grupiranja podataka u točki 7, str. 117.*

Kod običnog grupiranja podataka podatak koji se pojavi na granici dva ili više klastera najčešće se konvencionalno smješta u prvi po redu od ovih klastera.

Primjer 3.3. *Primjenom k-means algoritma pronađimo LS-optimalnu 3-particiju skupa* $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$ *krenuvši od početne particije* $\Pi^{(0)} = \{\{0, 2, 4\}, \{8, 9\}, \{10, 12, 16\}\}$.

Iteracija	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$
0	{0, 2, 4}	{8, 9}	{10, 12, 16}	2	8.5	12.67	27.17
1	{0, 2, 4}	{8, 9, 10}	{12, 16}	2	9	14	18
2	{0, 2, 4}	{8, 9, 10}	{12, 16}	2	9	14	18

Tablica 3.2: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$. Plavo su označeni rezultati Koraka A:, a crveno rezultati Koraka B:



Slika 3.3: Traženje LS-optimalne 3-particije skupa $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$

Zadatak 3.3. *Primjenom k-means algoritma pronađite* ℓ_1*-optimalnu 3-particiju skupa* $\mathcal{A}$ *iz Primjera 3.3 krenuvši od iste početne particije.*

Sljedeći teorem pokazuje da niz funkcijskih vrijednosti dobiven k-means algoritmom ima svojstvo monotonog pada (vidi također Teorem 4.1, str.4.1).

Teorem 3.3. *Neka je $\mathcal{A} \subset \mathbb{R}^n$ skup, $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ kvazimetrička funkcija i $\mathcal{F}$ funkcija cilja zadana s* (3.5)*. Primjenom k-means algoritma funkcija cilja $\mathcal{F}$ neće se povećati.*

Dokaz. Neka je $\Pi^{(t)} = \{\pi_1^{(t)}, \ldots, \pi_k^{(t)}\}$ particija s centrima $c^{(t)} = \{c_1^{(t)}, \ldots, c_k^{(t)}\}$ i $\mathcal{F}(\Pi^{(t)})$ odgovarajuća vrijednost funkcije cilja.

Ako na skup $\mathcal{A}$ primijenimo `Korak A` (princip minimalnih udaljenosti s centrima $c^{(t)}$), dobivamo novu particiju $\Pi^{(t+1)} = \{\pi_1^{(t+1)}, \ldots, \pi_k^{(t+1)}\}$ za koju vrijedi:

$$\mathcal{F}(\Pi^{(t)}) = \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t)}} d(c_j^{(t)}, a) \overset{(A)}{\geq} \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t+1)}} d(c_j^{(t)}, a).$$

Nadalje, ako u svakom klasteru $\pi_j^{(t+1)}$ primijenimo `Korak B` (odredimo novi reprezentant $c_j^{(t+1)}$), dobivamo:

$$\mathcal{F}(\Pi^{(t)}) = \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t)}} d(c_j^{(t)}, a) \overset{(A)}{\geq} \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t+1)}} d(c_j^{(t)}, a) \overset{(B)}{\geq} \sum_{j=1}^{k} \sum_{a \in \pi_j^{(t+1)}} d(c_j^{(t+1)}, a).$$

Dakle, $\mathcal{F}(\Pi^{(t)}) \geq \mathcal{F}(\Pi^{(t+1)})$. □

Primjer 3.4. *Treba pronaći LS-optimalnu 2-particiju skupa $\mathcal{A} = \{0, 2, 3\}$ primjenom k-means algoritma uz početnu particiju $\Pi^{(0)} = \{\{0, 2\}, \{3\}\}$.*

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$
1	{0,2}	{3}	1	3	2
2	{0,2}	{3}	1	3	2

Tablica 3.3: Traženje LS-optimalne 2-particije skupa $\mathcal{A} = \{0, 2, 3\}$

Kao što se vidi iz Tablice 3.3, k-means algoritam uz primjenu LS-kvazimetričke funkcije ne može pronaći bolju particiju od početne. Međutim, bolja particija u ovom je slučaju particija $\Pi^\star = \{\{0\}, \{2, 3\}\}$ jer je $\mathcal{F}_{LS}(\Pi^\star) = 0.5$. Na ovom jednostavnom primjeru pokazano je da k-means algoritam daje lokalno optimalnu particiju. Izborom neke druge početne particije možda bismo dobili k-GOPart. Pokušajte!

Uz spomenuti nedostatak k-means algoritma da jako ovisi o početnoj particiji i da kao rezultat daje neku lokalno optimalnu particiju kao u Primjeru 3.4, treba spomenuti još jedan nedostatak: tijekom iterativnog postupka može se dogoditi da neki od klastera postanu prazni skupovi, tj. može se dogoditi da se broj klastera u particiji smanji.

3.2 Grupiranje podataka s jednim obilježjem

Neka je $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}$ skup realnih brojeva koje treba grupirati u k klastera $\pi_1, \ldots, \pi_k$, u skladu s Definicijom 3.1, str. 33. Primjerice, dane u godini možemo grupirati u tri klastera prema prosječnoj dnevnoj temperaturi izraženoj u °C: klaster hladnih dana, klaster dana s umjerenom temperaturom i klaster toplih dana. Svaki element $a \in \mathcal{A}$ temeljem tog obilježja reprezentirat ćemo jednim realnim brojem kojeg ćemo također označavati s a. Zato ćemo nadalje pretpostavljati da je skup $\mathcal{A} = \{a^1, \ldots, a^m\}$ multiskup podataka-realnih brojeva, tj. neki elementi u $\mathcal{A}$ mogu se pojaviti više puta.

Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar c_j na sljedeći način:

$$c_j \in \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} d(x, a), \qquad j = 1, \ldots, k. \tag{3.6}$$

Nadalje, ako na skupu svih particija $\mathcal{P}(\mathcal{A}; k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a), \tag{3.7}$$

onda optimalnu k-`GOPart` tražimo rješavanjem sljedećeg optimizacijskog problema:

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi). \tag{3.8}$$

Primijetite da na taj način k-`GOPart` ima svojstvo da je suma „rasipanja" (suma odstupanja) elemenata klastera oko njegovog centra minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i međusobnu razdvojenost (separiranost) klastera.

Primjedba 3.2. *Broj svih k-particija skupa $\mathcal{A}$ s m elemenata može biti jako velik (vidi Tablicu 3.1). Međutim, u slučaju podataka s jednim obilježjem ($\mathcal{A} \subset \mathbb{R}$), očigledno je da se optimalna particija može očekivati između particija čiji se klasteri međusobno nastavljaju jedan na drugoga. To znači*

da se svi elementi klastera π_2 nalaze desno od klastera π_1, svi elementi klastera π_3 nalaze se desno od klastera π_2, itd. (vidi [80], str. 161). Broj svih takvih particija znatno je manji i daje ga sljedeća propozicija (vidi također i Tablicu 3.4).

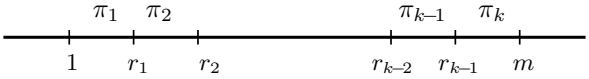
$\binom{m-1}{k-1}$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=8$	$k=10$
$m=10$	9	36	84	126	126	36	1
$m=30$	29	406	3 654	23 751	118 755	1 560 780	10 015 005
$m=50$	49	1 176	18 424	211 876	1 906 884	85 900 584	2 054 455 634

Tablica 3.4: Broj k-particija skupa $\mathcal{A} \subset \mathbb{R}$ čiji se klasteri nastavljaju jedan na drugoga

Propozicija 3.1. *Neka je $\mathcal{A} = \{a^i \in \mathbb{R} : i = 1, \ldots, m\}$. Broj svih k-particija skupa $\mathcal{A}$ čiji se klasteri $\pi_1, \ldots, \pi_k$ međusobno nastavljaju jedan na drugi iznosi*

$$\binom{m-1}{k-1}. \tag{3.9}$$

Dokaz. Zbog jednostavnosti, a bez smanjenja općenitosti, pretpostavimo da je $\mathcal{A} = \{1, \ldots, m\}$.



Očigledno najmanji element klastera π_1 mora biti $1 \in \mathcal{A}$, a najveći element klastera π_k mora biti $m \in \mathcal{A}$. S $r_1, \ldots, r_{k-1}$ označimo redom najveće elemente klastera $\pi_1, \ldots, \pi_{k-1}$. Za njih vrijedi $1 \leq r_1 < r_2 < \cdots < r_{k-1} < m$. Primijetite da ako bi bilo $r_{k-1} = m$, klaster π_k bio bi prazan pa to više ne bi bila k-particija. Zato se pitanje broja svih k-particija skupa $\mathcal{A}$ čiji se klasteri $\pi_1, \ldots, \pi_k$ međusobno nastavljaju jedan na drugi svodi na pitanje broja elemenata skupa

$$S = \{(r_1, \ldots, r_{k-1}) \in \mathcal{A}^{k-1} : 1 \leq r_1 < r_2 < \cdots < r_{k-1} < m\},$$

odnosno broja svih podskupova skupa $\{1, \ldots, m-1\}$ s $(k-1)$ članova. To je broj svih $(k-1)$-kombinacija skupa od $m-1$ elemenata. □

3.2.1 Primjena LS-kvazimetričke funkcije

Neka je $\mathcal{A} \subset \mathbb{R}$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija. Ako je $d_{LS}\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, $d_{LS}(x, y) = (x - y)^2$, LS-kvazimetrička funkcija, centri klastera $\pi_1, \ldots, \pi_k$ nazivaju se centroidi i određuju se na sljedeći način:

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} (x - a)^2 = \frac{1}{|\pi_j|} \sum_{a \in \pi_j} a, \qquad j = 1, \ldots, k, \tag{3.10}$$

a funkcija cilja (3.7) definirana je s

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} (c_j - a)^2. \tag{3.11}$$

Primjer 3.5. *Zadan je skup $\mathcal{A} = \{2, 4, 8, 10, 16\}$. Treba pronaći sve njegove 3-particije koje zadovoljavaju Definiciju 3.1 i koje se međusobno nastavljaju jedna na drugu. Za njih treba odrediti pripadne centroide i vrijednosti funkcije cilja $\mathcal{F}_{LS}$.*

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
{2}	{4}	{8,10,16}	2	4	11.33	0+0+34.67=34.67	36+16+33.33=85.33
{2}	{4,8}	{10,16}	2	6	13	0+8+18=26	36+8+50=94
{2}	{4,8,10}	{16}	2	7.33	16	0+18.67+0=18.67	36+1.33+64=101.33
{2,4}	{8}	{10,16}	3	8	13	2+0+18=20	50+0+50=100
{2,4}	{8,10}	{16}	3	9	16	2+2+0=4	50+2+64=116
{2,4,8}	{10}	{16}	4.67	10	16	18.67+0+0=18.67	33.33+4+64=101.33

Tablica 3.5: Sve 3-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$

Prema Stirlingovoj formuli (3.2), str. 33, broj svih 3-particija skupa $\mathcal{A}$ je 25. Međutim, broj 3-particija istog skupa koje se nastavljaju jedna na drugu prema (3.9) iznosi samo $\binom{5-1}{3-1} = \frac{4!}{2! \cdot 2!} = 6$ (vidi Tablicu 3.5). Kao što se vidi iz Tablice 3.5, LS-optimalna 3-particija u ovom slučaju je $\Pi^\star = \{\{2, 4\}, \{8, 10\}, \{16\}\}$ jer na njoj funkcija cilja $\mathcal{F}_{LS}$ zadana s (3.11) postiže najmanju vrijednost (globalni minimum). Dakle, particija $\Pi^\star$ je `LS 3-GOPart`.

Zadatak 3.4. *Zadan je skup $\mathcal{A} = \{1, 4, 5, 8, 10, 12, 15\}$. Koliko ovaj skup ima 3-particija, a koliko 3-particija čiji se klasteri međusobno nastavljaju? Ispišite sve 3-particije skupa $\mathcal{A}$ čiji se klasteri međusobno nastavljaju i među njima pronađite LS-optimalnu 3-particiju.*

Rješenje: Broj svih particija je 301, a broj svih particija čiji se klasteri međusobno nastavljaju je 15. LS-optimalna 3-particija je $\Pi^\star = \{\{1, 4, 5\}, \{8, 10\}, \{12, 15\}\}$. $\mathcal{F}(\Pi^\star) = \frac{91}{6} = 15.1667$.

3.2.2 Dualni problem

Sljedeća lema pokazuje da je u slučaju primjene LS-kvazimetričke funkcije „rasipanje" skupa $\mathcal{A}$ oko njegovog centra c jednako zbroju „rasipanja" klastera π_j, $j = 1, \ldots, k$ oko njihovih centara c_j, $j = 1, \ldots, k$ i težinske sume kvadrata odstupanja centra c od centara c_j, pri čemu su težine određene veličinom skupova π_j.

Lema 3.2. *Neka je $\mathcal{A} = \{a^1, \ldots, a^m\}$ skup podataka, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija s klasterima $\pi_1, \ldots, \pi_k$. Neka je nadalje*

$$c = \frac{1}{m}\sum_{i=1}^{m} a^i, \qquad c_j = \frac{1}{|\pi_j|}\sum_{a\in\pi_j} a, \quad j = 1, \ldots, k. \tag{3.12}$$

Tada vrijedi:

$$\sum_{i=1}^{m}(c - a^i)^2 = \mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi), \tag{3.13}$$

gdje je:

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k}\sum_{a\in\pi_j}(c_j - a)^2, \tag{3.14}$$

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k}|\pi_j|(c_j - c)^2. \tag{3.15}$$

Dokaz. Primijetimo najprije da za c_j vrijedi $\sum_{a^i\in\pi_j}(c_j - a^i) = 0$. Koristeći ovaj identitet za proizvoljni $x \in \mathbb{R}$ dobivamo:

$$\begin{aligned}\sum_{a^i\in\pi_j}(x - a^i)^2 &= \sum_{a^i\in\pi_j}((x - c_j) + (c_j - a^i))^2 \\ &= \sum_{a^i\in\pi_j}(x - c_j)^2 + 2\sum_{a^i\in\pi_j}(x - c_j)(c_j - a^i) + \sum_{a^i\in\pi_j}(c_j - a^i)^2 \\ &= |\pi_j|(x - c_j)^2 + \sum_{a^i\in\pi_j}(c_j - a^i)^2,\end{aligned}$$

odnosno:

$$\sum_{a^i\in\pi_j}(x - a^i)^2 = \sum_{a^i\in\pi_j}(c_j - a^i)^2 + |\pi_j|(c_j - x)^2, \quad j = 1, \ldots, k. \tag{3.16}$$

Ako u (3.16) umjesto x stavimo $c = \frac{1}{m}\sum_{i=1}^{m} a^i$ i zbrojimo sve jednakosti, dobivamo (3.13). □

U izrazu (3.13) prirodno se pojavila funkcija cilja $\mathcal{F}_{LS}$. Izraz (3.13) pokazuje da se ukupno rasipanje elemenata skupa $\mathcal{A}$ oko njegovog centroida c može prikazati kao zbroj dviju funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$.

Specijalno, LS-optimalna 3-particija skupa $\mathcal{A}$ iz Primjera 3.5 postiže se na particiji $\Pi^\star = \{\{2,4\},\{8,10\},\{16\}\}$, za koju je $\mathcal{F}(\Pi^\star) = 4$ (vidi Tablicu 3.5). Postavlja se pitanje: što je u tom slučaju $\mathcal{G}(\Pi^\star)$?

Kako bismo odgovorili na to pitanje, najprije dopunimo Tablicu 3.5 vrijednostima koje funkcija $\mathcal{G}$ prima na pojedinim particijama (plavi dio tablice). Primijetite da je zbroj $\mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi)$ uvijek konstantan i jednak $\sum_{i=1}^{m}(c-a^i)^2 = 120$, što je u skladu s (3.13), a da se najveća vrijednost funkcije $\mathcal{G}$ postiže baš na LS-optimalnoj 3-particiji $\Pi^\star$, na kojoj je funkcija cilja F_{LS} primila najmanju vrijednost.

Je li to slučajno?

Kako bismo odgovorili na ovo pitanje, pokušajmo najprije riješiti sljedeći zadatak, gdje se razmatra sličan problem. Za rješavanje ovog zadatka bit će nam potrebno predznanje osnova matematičke analize (vidi primjerice [47]).

Primjer 3.6. *Neka su $\varphi, \psi \in C^2(\mathbb{R})$ dvije funkcije za koje vrijedi $\varphi(x) + \psi(x) = \kappa$, $\kappa \in \mathbb{R}$. Funkcija φ postiže lokalni minimum u točki $x_0 \in \mathbb{R}$ onda i samo onda ako funkcija ψ u točki $x_0 \in \mathbb{R}$ postiže lokalni maksimum i vrijedi:*

$$\min_{x\in\mathbb{R}} \varphi(x) = \kappa - \max_{x\in\mathbb{R}} \psi(x), \qquad \textit{odnosno} \qquad \varphi(x_0) = \kappa - \psi(x_0).$$

Ako je $\varphi'(x_0) = 0$, onda je $\psi'(x_0) = 0$, i obratno. Također, ako je $\varphi''(x_0) > 0$, onda je $\psi''(x_0) < 0$, i obratno. Zato vrijedi

- $x_0 \in \operatorname{argmin}_{x\in\mathbb{R}} \varphi(x)$ onda i samo onda ako je $x_0 \in \operatorname{argmax}_{x\in\mathbb{R}} \psi(x)$;
- $\min_{x\in\mathbb{R}} \varphi(x) = \kappa - \max_{x\in\mathbb{R}} \psi(x)$, odnosno $\varphi(x_0) = \kappa - \psi(x_0)$.

Provjerite imaju li funkcije $\varphi(x) = x^2 - 1$ i $\psi(x) = -x^2 + 3$ navedena svojstva. Nacrtajte njihove grafove u jednom koordinatnom sustavu. Pokušajte sami konstruirati još jedan primjer para funkcija φ, ψ koje zadovoljavaju navedena svojstva.

Iz Leme 3.2 neposredno slijedi tvrdnja sljedećeg teorema [103].

Teorem 3.4. *Uz oznake kao u Lemi 3.2 postoji* $\Pi^\star \in \mathcal{P}(\mathcal{A};k)$ *takva da je*

$$(i)\quad \Pi^\star \in \operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi),$$

$$(ii)\quad \min_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \mathcal{F}_{LS}(\Pi^\star) \quad \& \quad \max_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi) = \mathcal{G}(\Pi^\star),$$

pri čemu je $\mathcal{G}(\Pi^\star) = \sum_{i=1}^{m}(c-a^i)^2 - \mathcal{F}_{LS}(\Pi^\star)$.

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije $\mathcal{F}_{LS}$ zadane s (3.11), možemo maksimizirati funkciju $\mathcal{G}$

$$\operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j|(c_j - c)^2. \tag{3.17}$$

Optimizacijski problem (3.17) zovemo dualni problem u odnosu na optimizacijski problem $\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi)$.

Možemo reći da LS-optimalna particija ima svojstvo da je suma „rasipanja elemenata klastera" (suma LS-udaljenosti elemenata klastera do svog centroida) minimalna, a da su pri tome centroidi klastera međusobno maksimalno razdvojeni. Na taj način postižemo najbolju unutrašnju kompaktnost i međusobnu razdvojenost (separiranost) klastera.

3.2.3 Princip najmanjih apsolutnih odstupanja

Neka je $\mathcal{A} \subset \mathbb{R}$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija. Ako je $d_1 \colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$, $d_1(x,y) = |x-y|$, ℓ_1-metrička funkcija, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s:

$$c_j = \operatorname{med}(\pi_j) \in \operatorname{Med}(\pi_j) = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{a\in\pi_j} |x-a| = \operatorname{med}(\pi_j), \quad j = 1, \ldots, k, \tag{3.18}$$

a funkcija cilja (3.7) s

$$\mathcal{F}_1(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} |c_j - a|. \tag{3.19}$$

Ako pri tome iskoristimo (3.20) iz Zadatka 3.6, onda za izračunavanje funkcije cilja (3.19) nije potrebno poznavati centre klastera (3.18), što može značajno ubrzati proces izračunavanja.

Primjer 3.7. *Zadan je skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *kao u Primjeru 3.5, str. 44. Treba pronaći sve njegove tročlane particije koje zadovoljavaju Definiciju 3.1 i čiji se klasteri nastavljaju jedan na drugoga.*

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}_1(\Pi)$
{2}	{4}	{8,10,16}	2	4	10	0+0+8=8
{2}	{4,8}	{10,16}	2	6	13	0+4+6=10
{2}	{4,8,10}	{16}	2	8	16	0+6+0=6
{2,4}	{8}	{10,16}	3	8	13	2+0+6=8
{2,4}	{8,10}	{16}	3	9	16	2+2+0=4
{2,4,8}	{10}	{16}	4	10	16	6+0+0=6

Tablica 3.6: Particije skupa $\mathcal{A}$ čiji se klasteri međusobno nastavljaju

Za ove particije odredimo pripadne centre i vrijednosti funkcije cilja $\mathcal{F}_1$ uz primjenu ℓ_1-metričke funkcije, a zatim pronađimo globalno ℓ_1-optimalnu 3-particiju.

Broj svih tročlanih particija čiji se klasteri međusobno nastavljaju je $\binom{m-1}{k-1} = 6$, a kao što se vidi iz Tablice 3.6, ℓ_1-optimalna 3-particija je $\{\{2,4\},\{8,10\},\{16\}\}$ jer na njoj funkcija cilja $\mathcal{F}_1$ zadana s (3.19) postiže najmanju vrijednost (globalni minimum). Dakle, particija $\Pi^\star$ je ℓ_1-GOPart.

Zadatak 3.5. *Između svih particija skupa $\mathcal{A} = \{1,4,5,8,10,12,15\}$ iz Zadatka 3.4, str. 44, pronađite ℓ_1-optimalnu 3-particiju.*

Zadatak 3.6. *Neka je $\mathcal{A} = \{a^1,\ldots,a^m\}$ konačan rastući niz realnih brojeva. Pokažite da vrijedi:*

$$\sum_{i=1}^{m} |a^i - \mathrm{med}(\mathcal{A})| = \sum_{i=1}^{\lceil \frac{m}{2} \rceil} \left(a^{m-i+1} - a^i\right), \tag{3.20}$$

gdje je[1] *$\lceil x \rceil$ jednak x ako je x cijeli broj, a $\lceil x \rceil$ je najmanji cijeli broj veći od x ako x nije cijeli broj. Primjerice, $\lceil 20 \rceil = 20$, ali $\lceil 20.3 \rceil = 21$.*

3.2.4 Grupiranje podataka s težinama

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a^1,\ldots,a^m\} \subset \mathbb{R}$ na pravcu, pri čemu je svakom podatku a^i pridružena odgovarajuća težina $w_i > 0$. Primjerice u Primjeru 3.8 [86, str. 30], u kojemu se analizira problem pojave visokog vodostaja rijeke Drave na mjernom mjestu Donji Miholjac, težine

[1] U programskom sustavu *Mathematica* veličina $\lceil x \rceil$ dobiva se kao `Ceiling[x]`, a veličina $\lfloor x \rfloor$ kao `Floor[x]`.

podataka su veličine vodostaja. Funkcija cilja (3.7) u slučaju težinskih podataka postaje:

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} w_i d(c_j, a^i), \tag{3.21}$$

gdje je:

$$c_j \in \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a^i \in \pi_j} w_i d(x, a^i), \quad j = 1, \ldots, k. \tag{3.22}$$

Specijalno, kod primjene LS-kvazimetričke funkcije centri c_j klastera π_j težinske su aritmetičke sredine podataka iz klastera π_j:

$$c_j = \frac{1}{\kappa^j} \sum_{a^i \in \pi_j} w_i a^i, \qquad \kappa^j = \sum_{a^i \in \pi_j} w_i, \tag{3.23}$$

a kod primjene ℓ_1-metričke funkcije centri c_j klastera π_j težinski su medijani podataka koji pripadaju klasteru π_j [77, 109]:

$$c_j = \operatorname*{med}_{a^i \in \pi_j} (w_i, a^i) \in \operatorname{Med}(w, \mathcal{A}). \tag{3.24}$$

Primjer 3.8. *Promatrajmo ponovno skup* $\mathcal{A} = \{1, 4, 5, 8, 10, 12, 15\}$ *iz Zadatka 3.4, str. 44. Svim podacima, osim posljednjeg, pridružimo težinu* 1*, a posljednjem podatku pridružimo težinu* 3*. Sada LS-optimalna* 3*-particija postaje* $\Pi^\star = \{\{1, 4, 5\}, \{8, 10, 12\}, \{15\}\}$ *s centroidima:* $\frac{10}{3}, 10, 15$ *i vrijednosti funkcije cilja* $\mathcal{F}(\Pi^\star) = \frac{50}{3} = 16.667$*.*

U slučaju primjene ℓ_1-metričke funkcije za određivanje centara klastera treba znati izračunati težinski medijan podataka. Kao što smo naveli u točki 2.1.3, str. 15, to može biti složen postupak. Ako su težine cijeli brojevi, problem se može svesti na određivanje običnog medijana podataka (vidi Primjer 2.4, str. 19). Ako težine nisu cijeli brojevi, množenjem nekim brojem i zaokruživanjem one se mogu svesti na cijele brojeve.

Zadatak 3.7. *Pronađite* ℓ_1*-optimalnu* 3*-particiju skupa* $\mathcal{A}$ *iz prethodnog primjera u slučaju kada su sve težine jednake* 1 *i u slučaju kada su podacima pridružene težine kao u prethodnom primjeru.*

Zadatak 3.8. *Napišite formule za centroid skupa* $\mathcal{A}$ *i funkcije cilja* $\mathcal{F}$ *i* $\mathcal{G}$ *za slučaj skupa podataka* $\mathcal{A}$ *s težinama* $w_1, \ldots, w_m > 0$*.*

Rješenje: $\mathcal{G}(\Pi) = \sum_{j=1}^{k} \big(\sum_{\pi_j} w_s\big)(c_j - c)^2$.

3.3 Grupiranje podataka s dva ili više obilježja

Neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i)^T \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup, koji u smislu Definicije 3.1 treba grupirati u $1 \leq k \leq m$ nepraznih disjunktnih klastera. Primjerice, skup $\mathcal{A} \subset \mathbb{R}^2$ iz Primjera 3.2, str. 36, ima dva obilježja (apscise i ordinate točaka), a možemo ga grupirati u $2, 3, 4, 5, 6, 7, 8$ ili više klastera (vidi Sliku 4.6, str. 78).

Neka je $\Pi \in \mathcal{P}(\mathcal{A}; k)$ neka particija skupa $\mathcal{A}$. Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar c_j na sljedeći način:

$$c_j \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k. \tag{3.25}$$

Nadalje, potpuno analogno kao u prethodnom slučaju, ako na skupu svih particija $\mathcal{P}(\mathcal{A}; k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a), \tag{3.26}$$

onda optimalnu k-particiju tražimo rješavanjem sljedećeg `GOP`:

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A}; k)} \mathcal{F}(\Pi). \tag{3.27}$$

Primijetite da na taj način optimalna k-particija ima svojstvo da je suma „rasipanja" (suma d-udaljenosti elemenata klastera do svog centra) minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost klastera.

3.3.1 Princip najmanjih kvadrata

Neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i)^T \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova particija. Ako je $d_{LS}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d_{LS}(a, b) = \|a - b\|_2^2$, LS-kvazimetrička funkcija, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ nazivaju se centroidi i određuju se na sljedeći način:

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} \|x - a\|_2^2 = \tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a = \Big(\tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a_1, \ldots, \tfrac{1}{|\pi_j|} \sum_{a \in \pi_j} a_n \Big)^T, \tag{3.28}$$
$$j = 1, \ldots, k,$$

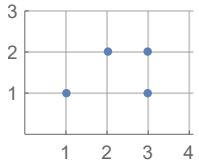
pri čemu $\sum_{a \in \pi_j} a_1$ označava sumu prvih komponenti svih elemenata klastera π_j, a $\sum_{a \in \pi_j} a_n$ sumu n-tih komponenti svih elemenata klastera π_j. Funkcija

cilja (3.26) u ovom slučaju zadana je s

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|c_j - a\|_2^2. \tag{3.29}$$

Primjer 3.9. *Neka je* $\mathcal{A} = \{a^1 = (1,1)^T, a^2 = (3,1)^T, a^3 = (3,2)^T, a^4 = (2,2)^T\}$ *skup točaka u ravnini. Broj svih njegovih* 2*-particija je* $\mathcal{P}(\mathcal{A};2) = 2^{4-1} - 1 = 7$*, a prikazane su u Tablici 3.7. Između svih* 2*-particija skupa* $\mathcal{A}$ *potražimo LS-optimalnu.*

Skup $\mathcal{A}$ sastoji se od elemenata s dva obilježja (koordinate), pa se jednostavnije može zapisati kao $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, 4\}$ i grafički prikazati u ravnini (vidi Sliku 3.4).



Slika 3.4: Skup $\mathcal{A} \subset \mathbb{R}^2$

Prema (3.2), skup $\mathcal{A}$ ima 7 različitih 2-particija. Neka je $\Pi = \{\pi_1, \pi_2\}$ bilo koja od njih. Centroidi njenih klastera zadani su s:

$$c_1 = \frac{1}{|\pi_1|} \sum_{a \in \pi_1} a, \qquad c_2 = \frac{1}{|\pi_2|} \sum_{a \in \pi_2} a,$$

a odgovarajuća LS-funkcija cilja je:

$$\mathcal{F}_{LS}(\Pi) = \sum_{a \in \pi_1} \|c_1 - a\|_2^2 + \sum_{a \in \pi_2} \|c_2 - a\|_2^2.$$

U ovom slučaju vrijednost funkcije $\mathcal{F}_{LS}$ predstavlja sumu „sume kvadrata udaljenosti" točaka klastera π_1 do njegovog centroida c_1 i sumu „sume kvadrata udaljenosti" točaka klastera π_2 do njegovog centroida c_2.

π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}(\Pi)$	$\mathcal{G}(\Pi)$
$\{(1,1)^T\}$	$\{(2,2)^T, (3,1^T, (3,2)^T\}$	$(1,1)^T$	$(2.67, 1.67)^T$	0+1.33=1.33	1.82+0.60=2.42
$\{(3,1)^T\}$	$\{(1,1)^T, (2,2)^T, (3,2)^T\}$	$(3,1)^T$	$(2.,1.67)^T$	0+2.67=2.67	0.81+0.27=1.08
$\{(3,2)^T\}$	$\{(1,1)^T, (2,2)^T, (3,1)^T\}$	$(3,2)^T$	$(2.,1.3)^T$	0+2.67=2.67	0.81+0.27=1.08
$\{(2,2)^T\}$	$\{(1,1)^T, (3,1)^T, (3,2)^T\}$	$(2,2)^T$	$(2.3,1.3)^T$	0+3.33=3.33	0.31+0.10=0.42
$\{(1,1)^T, (3,1)^T\}$	$\{(2,2)^T, (3,2)^T\}$	$(2,1)^T$	$(2.5,2.)^T$	2+0.5=2.5	0.625+0.625=1.25
$\{(1,1)^T, (3,2)^T\}$	$\{(2,2)^T, (3,1)^T\}$	$(2,1.5)^T$	$(2.5,1.5)^T$	2.5+1.=3.5	0.125+0.125=0.25
$\{(1,1)^T, (2,2)^T\}$	$\{(3,1)^T, (3,2)^T\}$	$(1.5,1.5)^T$	$(3.,1.5)^T$	1+0.5=1.5	1.125+1.125=2.25

Tablica 3.7: Particije, centri i funkcije cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$ iz Primjera 3.9

U Tablici 3.7 navedene su sve particije s centroidima odgovarajućih klastera i vrijednostima funkcije cilja $\mathcal{F}_{LS}$. Kao što se vidi iz Tablice 3.7, particija $\{\{(1,1)^T\}, \{(2,2)^T, (3,1)^T, (3,2)^T\}\}$ je LS-optimalna jer na njoj funkcija cilja $\mathcal{F}_{LS}$ postiže globalni minimum (vidi također Sliku 3.4).

3.3.2 Dualni problem

Sljedeća lema pokazuje da je u slučaju primjene LS-kvazimetričke funkcije „rasipanje" skupa $\mathcal{A}$ oko njegovog centra c jednako zbroju „rasipanja" klastera π_j, $j = 1, \ldots, k$ oko njihovih centara c_j, $j = 1, \ldots, k$ i težinske sume kvadrata odstupanja centra c od centara c_j, pri čemu su težine određene veličinom skupova π_j.

Lema 3.3. *Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup podataka, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija s klasterima $\pi_1, \ldots, \pi_k$. Neka su nadalje*

$$c = \frac{1}{m} \sum_{i=1}^{m} a^i, \qquad c_j = \frac{1}{|\pi_j|} \sum_{a^i \in \pi_j} a^i, \quad j = 1, \ldots, k \tag{3.30}$$

redom centroid čitavog skupa $\mathcal{A}$ i centroidi klastera $\pi_1, \ldots, \pi_k$. Tada vrijedi

$$\sum_{i=1}^{m} \|c - a^i\|^2 = \mathcal{F}_{LS}(\Pi) + \mathcal{G}(\Pi), \tag{3.31}$$

gdje je:

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} \|c_j - a^i\|^2, \tag{3.32}$$

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j| \|c_j - c\|^2. \tag{3.33}$$

Dokaz. Primijetimo najprije da za c_j vrijedi svojstvo aritmetičke sredine:

$$\sum_{a^i \in \pi_j} (c_j - a^i) = 0. \tag{3.34}$$

Za proizvoljni $x \in \mathbb{R}^n$ računamo:

$$\begin{aligned}\sum_{a^i \in \pi_j} \|x - a^i\|^2 &= \sum_{a^i \in \pi_j} \|(x - c_j) + (c_j - a^i)\|^2 \\ &= \sum_{a^i \in \pi_j} \|x - c_j\|^2 + 2 \sum_{a^i \in \pi_j} \langle x - c_j, c_j - a^i \rangle + \sum_{a^i \in \pi_j} \|c_j - a^i\|^2.\end{aligned}$$

Kako je $\sum\limits_{a^i \in \pi_j} \langle x - c_j, c_j - a^i \rangle = \langle x - c_j, \sum\limits_{a^i \in \pi_j} (c_j - a^i) \rangle \overset{(3.34)}{=} 0$, iz prethodne jednakosti dobivamo:

$$\sum_{a^i \in \pi_j} \|x - a^i\|^2 = \sum_{a^i \in \pi_j} \|c_j - a^i\|^2 + |\pi_j| \|c_j - x\|^2, \quad j = 1, \ldots, k. \tag{3.35}$$

Ako u (3.35) umjesto x stavimo $c = \frac{1}{m} \sum\limits_{i=1}^{m} a^i$ i zbrojimo sve jednakosti, dobivamo (3.31). □

U izrazu (3.31) prirodno se pojavila funkcija cilja $\mathcal{F}_{LS}$. Izraz (3.31) pokazuje da se ukupno rasipanje elemenata skupa $\mathcal{A}$ oko njegovog centroida c može prikazati kao zbroj dviju funkcija cilja $\mathcal{F}_{LS}$ i $\mathcal{G}$.

Analogno, kao u točki 3.2.2, str. 45, korištenjem Leme 3.3 može se pokazati da vrijedi sljedeći teorem [27, 103].

Teorem 3.5. *Uz oznake kao u Lemi 3.3 postoji* $\Pi^\star \in \mathcal{P}(\mathcal{A}; k)$*, takav da je*

$$(i) \quad \Pi^\star \in \operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \operatorname*{argmax}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi),$$

$$(ii) \quad \min_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi) = \mathcal{F}_{LS}(\Pi^\star) \quad \& \quad \max_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi) = \mathcal{G}(\Pi^\star),$$

pri čemu je $\mathcal{G}(\Pi^\star) = \sum\limits_{i=1}^{m} \|c - a^i\|^2 - \mathcal{F}_{LS}(\Pi^\star)$.

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije $\mathcal{F}_{LS}$ zadane s (3.32), možemo rješavati problem maksimuma za funkciju $\mathcal{G}$

$$\operatorname*{argmax}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = \sum_{j=1}^{k} |\pi_j| \|c_j - c\|^2. \tag{3.36}$$

Optimizacijski problem (3.36) zovemo dualni problem u odnosu na optimizacijski problem $\operatorname{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}_{LS}(\Pi)$.

Možemo reći da LS-optimalna particija ima svojstvo da je suma „rasipanja elemenata klastra" (suma LS-udaljenosti elemenata klastera do svog centroida) minimalna, a da su pri tome centroidi klastera međusobno maksimalno razdvojeni. Na taj način postižemo najbolju unutrašnju kompaktnost i međusobnu razdvojenost (separiranost) klastera.

Primjer 3.10. *Kod Primjera 3.9, str. 51, može se razmatrati i odgovarajući dualni problem.*

Specijalno, u ovom slučaju jednakost (3.31) glasi:

$$\sum_{i=1}^{m} \|c - a^i\|_2^2 = \Big(\sum_{a\in\pi_1} \|c_1 - a\|_2^2 + \sum_{a\in\pi_2} \|c_2 - a\|_2^2 \Big) + \big(m_1\|c_1 - c\|_2^2 + m_2\|c_2 - c\|_2^2\big),$$

a dualni optimizacijski problem (3.36) postaje:

$$\operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi), \qquad \mathcal{G}(\Pi) = m_1\|c_1 - c\|_2^2 + m_2\|c_2 - c\|_2^2.$$

Za svaku 2-particiju u Tablici 3.7, str. 52, plavom bojom prikazane su vrijednosti dualne funkcije cilja $\mathcal{G}$. Kao što se vidi, funkcija $\mathcal{G}$ prima maksimalnu vrijednost na 2-particiji $\{\{(1,1)^T\},\{(2,2)^T,(3,1)^T,(3,2)^T\}\}$ na kojoj je funkcija $\mathcal{F}_{LS}$ zadana s (3.29) primila minimalnu vrijednost.

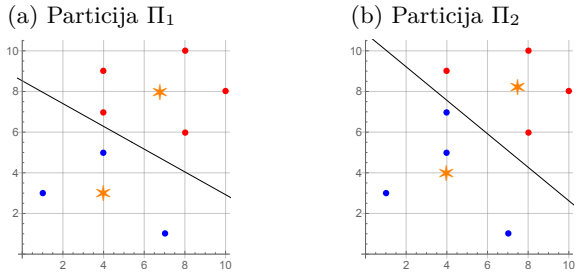
Primjer 3.11. *Skup* $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1,\ldots,8\} \subset \mathbb{R}^2$ *zadan je s:*

i	1	2	3	4	5	6	7	8
x_i	1	4	4	4	7	8	8	10
y_i	3	5	7	9	1	6	10	8

Uz primjenu LS-kvazimetričke funkcije za 2-particije:

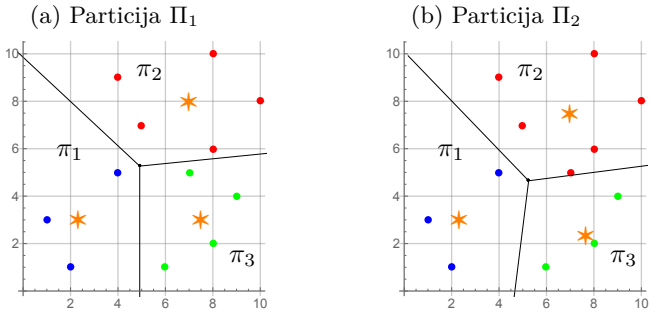
$$\Pi_1 = \{\{a^1, a^2, a^5\}, \{a^3, a^4, a^6, a^7, a^8\}\},$$
$$\Pi_2 = \{\{a^1, a^2, a^3, a^5\}, \{a^4, a^6, a^7, a^8\}\},$$

prikazane na Slici 3.5, čiji su klasteri označeni plavom odnosno crvenom bojom, treba odrediti centroide i odgovarajuće vrijednosti funkcija cilja $\mathcal{F}_{LS}$ *i* $\mathcal{G}$ *te na osnovi toga ustanoviti koja je particija bliža optimalnoj.*



Slika 3.5: Dvije particije skupa $\mathcal{A}$ iz Primjera 3.11

Za particiju Π_1 dobivamo: $c_1 = (4,3)^T$, $c_2 = (6.8,8)^T$, $\mathcal{F}_{LS} = 26 + 38.8 = 64.8$ i $\mathcal{G} = 61.575$, a za particiju Π_2: $c_1 = (4,4)^T$, $c_2 = (7.5,8.25)^T$, $\mathcal{F}_{LS} = 38 + 27.75 = 65.75$ i $\mathcal{G} = 60.625$. Dakle, 2-particija Π_1 bliža je LS-optimalnoj. Primjenom *Mathematica*-modula `WKMeans[]` provjerite je li to ujedno i globalno LS-optimalna 2-particija. Primijetite (formula (3.2)) da u ovom slučaju ukupno postoji $2^7 - 1 = 127$ različitih 2-particija.



Slika 3.6: Usporedba dviju particija iz Zadatka 3.9

Zadatak 3.9. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, m\}$ *prikazan na Slici 3.6, gdje je:*

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	2	4	4	5	6	7	8	8	8	9	10
y_i	3	1	5	9	7	1	5	2	6	10	4	8

Treba ustanoviti na kojoj od dviju 3*-particija prikazanih na Slici 3.6 LS-funkcija cilja* $\mathcal{F}_{LS}$ *zadana s* (3.29) *prima manju vrijednost.*

Rješenje:

$$\Pi_1 = \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^9, a^{10}, a^{12}\}, \{a^6, a^7, a^8, a^{11}\}\} \quad \ldots \quad \text{Slika 3.6a}$$
$$\Pi_2 = \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^7, a^9, a^{10}, a^{12}\}, \{a^6, a^8, a^{11}\}\} \quad \ldots \quad \text{Slika 3.6b}$$
$$\Pi_1 : c_1 = (2.33, 3)^T,\ c_2 = (7, 8)^T,\ c_3 = (7.5, 3)^T;$$
$$\mathcal{F}_{LS} = 12.67 + 34 + 15 = 61.67; \quad \mathcal{G} = 127.25,$$
$$\Pi_2 : c_1 = (2.33, 3)^T,\ c_2 = (7, 7.5)^T,\ c_3 = (7.67, 2.33)^T;$$
$$\mathcal{F}_{LS} = 12.67{+}41.5{+}9.33{=}63.5; \quad \mathcal{G} = 125.42.$$

Dakle, manja vrijednost LS-funkcije cilja $\mathcal{F}_{LS}$ (i veća vrijednost dualne funkcije $\mathcal{G}$) postiže se na 3-particiji Π_1 pa nju smatramo bliže LS-optimalnoj. Primjenom *Mathematica*-modula `WKMeans[]` uz izbor različitih početnih particija pokušajte pronaći bolju 3-particiju.

3.3.3 Princip najmanjih apsolutnih odstupanja

Neka je $\mathcal{A} \subset \mathbb{R}^n$ skup, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka njegova k-particija. Ako je $d_1 \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $d_1(x, y) = \|x - y\|_1$, ℓ_1-metrička funkcija, centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$\begin{aligned} c_j &= \operatorname{med}(\pi_j) = \big(\operatorname*{med}_{a\in\pi_j} a_1, \ldots, \operatorname*{med}_{a\in\pi_j} a_n\big)^T \\ &\in \operatorname{Med}(\pi_j) = \big(\operatorname*{Med}_{a\in\pi_j} a_1, \ldots, \operatorname*{Med}_{a\in\pi_j} a_n\big)^T = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{a\in\pi_j} \|x - a\|_1 \end{aligned} \tag{3.37}$$

pri čemu $\operatorname*{med}_{a\in\pi_j} a_1$ označava medijan prvih komponenti svih elemenata klastera π_j, a $\operatorname*{med}_{a\in\pi_j} a_n$ medijan n-tih komponenti svih elemenata klastera π_j. ℓ_1-funkcija cilja (3.26) u ovom je slučaju zadana s

$$\mathcal{F}_1(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} \|c_j - a\|_1 \tag{3.38}$$

Zadatak 3.10. *Pokažite da je uz primjenu ℓ_1-metričke funkcije globalno optimalna 2-particija iz Primjera 3.9:* $\{\{(1,1)^T, (3,2)^T\}, \{(2,2)^T, (3,1)^T\}\}$ *s centrima klastera* $c_1 = (2, 1.5)^T$, $c_2 = (2.5, 1.5)^T$ *i vrijednosti funkcije cilja* $\mathcal{F}_1 = 3 + 2 = 5$.

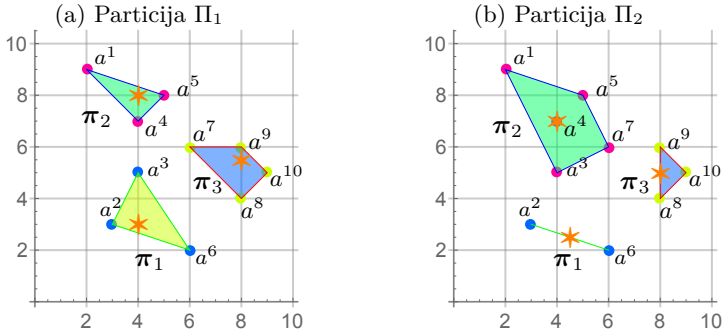
Zadatak 3.11. *Na particije iz Zadatka 3.9, str. 55, primijenite princip najmanjih apsolutnih odstupanja.*

Primjer 3.12. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 10\}$, *gdje je:*

i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	4	4	5	6	6	8	8	9
y_i	9	3	5	7	8	2	6	4	6	5

Treba ustanoviti na kojoj od dviju niže navedenih 3*-particija* ℓ_1*-funkcija cilja* (3.38) *prima manju vrijednost.*

$$\Pi_1 = \{\{a^2, a^3, a^6\}, \{a^1, a^4, a^5\}, \{a^7, a^8, a^9, a^{10}\}\} \quad \ldots \quad \textit{Slika 3.7a,}$$
$$\Pi_2 = \{\{a^2, a^6\}, \{a^1, a^3, a^4, a^5, a^7\}, \{a^8, a^9, a^{10}\}\} \quad \ldots \quad \textit{Slika 3.7b.}$$



Slika 3.7: Usporedba dviju particija

Niže su izračunati ℓ_1-centri pojedinih klastera u objema particijama i vrijednost funkcije cilja na objema particijama. Vidi se da je Π_1 „bolja" particija jer se na njoj postiže niža vrijednost funkcije cilja.

	c_1	c_2	c_3	$\mathcal{F}_1$
Π_1	$(4,3)^T$	$(4,8)^T$	$(8,5.5)^T$	(1+2+3)+(3+1+1)+(2.5+1.5+.5+1.5) =17
Π_2	$(4.5,2.5)^T$	$(4,7)^T$	$(8,5)^T$	(2+2)+(4+2+0+2+3)+(1+1+1) =18

Primjedba 3.3. *Kao što smo u točki 3.2.4, str. 48, razmatrali problem grupiranja jednodimenzionalnih težinskih podataka, slično bismo mogli postupiti i u slučaju grupiranja težinskih dvodimenzionalnih i višedimenzionalnih podataka.*

3.4 Funkcija cilja $F(c_1, \ldots, c_k) = \sum_{i=1}^{m} \min_{1 \le j \le k} d(c_j, a^i)$

Funkcija cilja iz GOP (3.5) nije prikladna za primjenu standardnih optimizacijskih metoda jer je nezavisna varijabla particija. Zato ćemo GOP (3.5) preformulirati tako da funkcija cilja postane obična funkcija više realnih varijabli. Pregled najpopularnijih metoda za traženje optimalnih particija može se vidjeti u [106].

Kao što je naznačeno u k-means algoritmu na str. 39, za dani skup centara $c_1, \ldots, c_k \in \mathbb{R}^n$, skup $\mathcal{A}$ razdijelit ćemo u k klastera $\pi(c_1), \ldots, \pi(c_k)$[2] tako da u klaster π_j dođu oni elementi skupa $\mathcal{A}$ koji su najbliži centru c_j, tj. tako da za svaki $a^i \in \mathcal{A}$ bude

$$a^i \in \pi_j(c_j) \iff d(c_j, a^i) \le d(c_s, a^i), \quad \forall s = 1, \ldots, k. \tag{3.39}$$

Pri tome treba voditi računa da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru. Ovaj princip, koji nazivamo principom minimalnih udaljenosti, daje particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ s klasterima $\pi_1, \ldots, \pi_k$.

Zbog toga se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći GOP (vidi također [103]):

$$\operatorname*{argmin}_{c \in \mathbb{R}^{n \times k}} F(c), \qquad F(c) = \sum_{i=1}^{m} \min_{1 \le j \le k} d(c_j, a^i), \tag{3.40}$$

gdje je $c \in \mathbb{R}^{n \times k}$ konkatenacija vektora $c_1, \ldots, c_k$. Funkcija F je nenegativna, simetrična, nediferencijabilna, nekonveksna, ali Lipschitz neprekidna funkcija.

Sljedeći teorem pokazuje da je funkcija F zadana s (3.40) Lipschitz neprekidna u slučaju primjene LS-kvazimetričke funkcije. U [81] slično je pokazano da je ova funkcija Lipschitz neprekidna i u slučaju primjene ℓ_1-metričke funkcije. Ovo je važno svojstvo funkcije F jer omogućava primjenu globalno optimizacijskog algoritma DIRECT [38, 46, 69, 85].

Teorem 3.6. *Zadan je skup* $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\} \subset \Delta$, *gdje je* $\Delta = \{x \in \mathbb{R}^n : \alpha_i \le x_i \le \beta_i\}$ *i* $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n$. *Funkcija* $F : \Delta^k \to \mathbb{R}_+$,

$$F(c) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} \|c_j - a^i\|^2,$$

je Lipschitz neprekidna na Δ^k.

[2] Primijetite da klaster $\pi(c_j)$ ovisi i o susjednim klasterima, a oznaka $\pi(c_j)$ ukazuje na to da je klaster $\pi(c_j)$ pridružen centru c_j.

U svrhu dokaza ovog teorema minimizirajuću funkciju F do na proizvoljni $\epsilon > 0$ aproksimirat ćemo diferencijabilnom (glatkom) funkcijom F_ϵ. Prije toga bit će potrebna niže navedene teorijska priprema (vidi [50]).

Lema 3.4. *Funkcija* $\psi : \mathbb{R}^n \to \mathbb{R}$, $\psi(x) = \ln(e^{x_1} + \cdots + e^{x_n})$ *je konveksna funkcija.*

Dokaz. Treba pokazati da za sve $x, y \in \mathbb{R}^n$ vrijedi:

$$\psi(\lambda x + (1-\lambda)y) \leq \lambda\psi(x) + (1-\lambda)\psi(y), \quad \forall \lambda \in [0,1], \tag{3.41}$$

odnosno

$$\psi(\alpha x + \beta y) \leq \alpha\psi(x) + \beta\psi(y), \tag{3.42}$$

gdje su $\alpha, \beta > 0$, takvi da je $\alpha + \beta = 1$.

Definirajmo $p := \frac{1}{\alpha}$, $q := \frac{1}{\beta}$ za koje vrijedi $\frac{1}{p} + \frac{1}{q} = 1$. Primijetite da zbog $\alpha + \beta = 1$ jedan od brojeva α, β mora biti manji od 1, iz čega slijedi da jedan od brojeva p, q mora biti veći od 1. Neka su $x = (x_1, \ldots, x_n)^T$, $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Na vektore

$$a = (e^{\alpha x_1}, \ldots, e^{\alpha x_n})^T, \quad b = (e^{\beta y_1}, \ldots, e^{\beta y_n})^T \in \mathbb{R}^n,$$

primijenimo Hölderovu nejednakost[3] (vidi [53]) i dobivamo:

$$|\langle a, b\rangle| \leq \Big(\sum_{i=1}^n (e^{\alpha x_i})^p\Big)^{1/p} \Big(\sum_{i=1}^n (e^{\beta y_i})^q\Big)^{1/q},$$

odnosno

$$\sum_{i=1}^n e^{\alpha x_i + \beta y_i} \leq \Big(\sum_{i=1}^n e^{x_i}\Big)^{\alpha} \Big(\sum_{i=1}^n e^{y_i}\Big)^{\beta}.$$

Odavde logaritmiranjem neposredno slijedi tražena nejednakost (3.42). □

Korolar 3.1. *Neka je* $A \in \mathbb{R}^{n\times n}$ *kvadratna matrica,* $b \in \mathbb{R}^n$ *vektor i funkcija* $\psi : \mathbb{R}^n \to \mathbb{R}$, *zadana s* $\psi(x) = \ln(e^{x_1} + \cdots + e^{x_n})$. *Tada je funkcija* $\Phi(x) = \psi(Ax + b)$ *konveksna.*

[3] Za vektore $a, b \in \mathbb{R}^n$ i realne brojeve p, q, takve da je $\frac{1}{p} + \frac{1}{q} = 1$, $p > 1$, vrijedi Hölderova nejednakost $\sum_{i=1}^n |a_i b_i| \leq \|a\|_p \|b\|_q$, odnosno

$$\sum_{i=1}^n |a_i b_i| \leq \Big(\sum_{i=1}^n |a_i|^p\Big)^{1/p} \Big(\sum_{i=1}^n |b_i|^q\Big)^{1/q}.$$

Specijalno, za $p = q = 2$ ovo postaje poznata Cauchy-Schwarz-Buniakowsky nejednakost (vidi [89]).

Dokaz. Slično kao u dokazu Leme 3.4 dovoljno je pokazati da za proizvoljne $x, y \in \mathbb{R}^n$ i $\alpha, \beta > 0$, takve da je $\alpha + \beta = 1$ vrijedi:

$$\Phi(\alpha x + \beta y) \leq \alpha\Phi(x) + \beta\Phi(y).$$

Kako je

$$\begin{aligned}\Phi(\alpha x + \beta y) =& \psi(A(\alpha x + \beta y) + b) = \psi(\alpha Ax + \beta Ay + b)\\ =& \psi(\alpha Ax + \alpha b + \beta Ay + \beta b - (\alpha + \beta)b + b)\\ =& \psi(\alpha(Ax + b) + \beta(Ay + b))\\ \leq& \alpha\Phi(x) + \beta\Phi(y),\end{aligned}$$

slijedi traženo. □

Zadatak 3.12. *Pokažite da je* $\psi : \mathbb{R}^n_+ \to \mathbb{R}$, $\psi(x) = \ln(\frac{1}{x_1} + \cdots + \frac{1}{x_n})$ *konveksna funkcija.*

Lema 3.5. *Za svaki* $\epsilon > 0$ *funkcija* $\psi_\epsilon \colon \mathbb{R} \to \mathbb{R}_+$,

$$\psi_\epsilon(x) = \epsilon \ln\left(e^{-\frac{x}{\epsilon}} + e^{\frac{x}{\epsilon}}\right) = \epsilon \ln\left(2 \operatorname{ch} \tfrac{x}{\epsilon}\right). \tag{3.43}$$

konveksna je funkcija klase $C^\infty(\mathbb{R})$ *za koju vrijedi:*

$$0 < \psi_\epsilon(x) - |x| \leq \epsilon \ln 2, \quad \forall x \in \mathbb{R}, \tag{3.44}$$

$$\psi'_\epsilon(x) = \operatorname{th} \tfrac{x}{\epsilon}, \qquad \psi''_\epsilon(x) = \frac{1}{\epsilon \operatorname{ch}^2 \frac{x}{\epsilon}}, \qquad \operatorname*{argmin}_{x \in \mathbb{R}} \psi_\epsilon(x) = 0, \tag{3.45}$$

pri čemu jednakost u (3.44) vrijedi onda i samo onda ako je $x = 0$.

a) Funkcije $x \mapsto |x|$ i $x \mapsto \psi_\epsilon(x)$ b) Funkcija $u \mapsto \frac{2 \operatorname{ch} u}{e^{|u|}}$

Slika 3.8: Glatka aproksimacija funkcije $x \mapsto |x|$

Dokaz. Ako stavimo $n = 2, x_1 = -\frac{x}{\epsilon}, x_2 = \frac{x}{\epsilon}$, konveksnost funkcije ψ_ϵ direktno slijedi iz *Leme 3.4*. U svrhu dokaza tvrdnje (3.44) primijetimo da:

$$\begin{aligned}\psi_\epsilon(x) - |x| &= \epsilon \ln\left(2 \operatorname{ch} \tfrac{x}{\epsilon}\right) - \epsilon\tfrac{|x|}{\epsilon}\\ &= \epsilon\Big(\ln\left(2 \operatorname{ch} \tfrac{x}{\epsilon}\right) - \ln \exp \tfrac{|x|}{\epsilon}\Big) = \epsilon \ln \frac{2 \operatorname{ch} \frac{x}{\epsilon}}{\exp \frac{|x|}{\epsilon}}.\end{aligned}$$

Budući da za svaki $u \in \mathbb{R}$ vrijedi (vidi Zadatak 3.13): $1 < \frac{2 \operatorname{ch} u}{\exp |u|} \leq 2$, zbog monotonosti logaritamske funkcije iz prethodne jednakosti slijedi (3.44):

$$\epsilon \ln 1 < \psi_\epsilon(x) - |x| \leq \epsilon \ln 2.$$

Formule u (3.45) dobivaju se izravno. □

Zadatak 3.13. *Dokažite da za svaki $u \in \mathbb{R}$ vrijedi (vidi Sliku 3.8b)*

$$1 < \frac{2 \operatorname{ch} u}{e^{|u|}} \leq 2.$$

Sukladno (3.44), primijetite da funkciju $x \mapsto |x|$ za $x \in \mathbb{R}$ možemo aproksimirati funkcijom ψ_ϵ (vidi Sliku 3.8a).

Općenito, nediferencijabilnu funkciju $f\colon \mathbb{R}^k \to \mathbb{R}$, $f(z) = \max\limits_{j=1,\ldots,k} z_j$ možemo aproksimirati diferencijabilnom funkcijom

$$\psi_\epsilon(z) = \psi_\epsilon(z_1, \ldots, z_k) = \epsilon \ln \sum_{j=1}^{k} \exp\left(\tfrac{z_j}{\epsilon}\right). \tag{3.46}$$

Naime,

$$\begin{aligned}
\psi_\epsilon(z) - f(z) &= \epsilon \ln \sum_{j=1}^{k} \exp\left(\tfrac{z_j}{\epsilon}\right) - \epsilon \frac{\max\limits_{i=1,\ldots,k} z_i}{\epsilon} \\
&= \epsilon \Big(\ln \sum_{j=1}^{k} \exp\left(\tfrac{z_j}{\epsilon}\right) - \ln \exp \frac{\max z_i}{\epsilon} \Big) \\
&= \epsilon \ln \frac{\sum\limits_{j=1}^{k} \exp\left(\tfrac{z_j}{\epsilon}\right)}{\exp \frac{\max z_i}{\epsilon}} = \epsilon \ln \sum_{j=1}^{k} \exp \tfrac{z_j - \max z_i}{\epsilon} \leq \epsilon \ln \sum_{j=1}^{k} e^0 = \epsilon \ln k.
\end{aligned}$$

Nadalje, kako je $\min\limits_{j=1,\ldots,k} z_j = -\max\limits_{j=1,\ldots,k} (-z_j)$, ovaj rezultat možemo iskoristiti da funkciju $F(c_1, \ldots, c_k) = \sum\limits_{i=1}^{m} \min\limits_{1 \leq j \leq k} d(c_j, a^i)$ aproksimiramo funkcijom

$$F_\epsilon(c_1, \ldots, c_k) = -\epsilon \sum_{i=1}^{m} \ln \sum_{j=1}^{k} \exp\left(-\tfrac{d(c_j, a^i)}{\epsilon}\right). \tag{3.47}$$

Sada smo u mogućnosti provesti dokaz Teorema 3.6, str. 58.

Dokaz. (Teorema 3.6, str. 58) Ako sukladno (3.47) definiramo pomoćnu funkciju $F_\varepsilon : \Delta^k \to \mathbb{R}_+$,

$$F_\varepsilon(u) = -\varepsilon \sum_{i=1}^{m} \ln \sum_{j=1}^{k} \exp\left(-\tfrac{\|c_j - a^i\|^2}{\varepsilon}\right),$$

onda prema [50], imamo

$$0 \leq F(u) - F_\varepsilon(u) \leq \varepsilon m \ln k,$$

i zbog toga

$$\begin{aligned} |F(u) - F(v)| &= |(F(u) - F_\varepsilon(u)) + (F_\varepsilon(v) - F(v)) + (F_\varepsilon(u) - F_\varepsilon(v))| \\ &\leq |F(u) - F_\varepsilon(u)| + |F_\varepsilon(v) - F(v)| + |F_\varepsilon(u) - F_\varepsilon(v)| \\ &\leq 2\varepsilon m \ln k + |F_\varepsilon(u) - F_\varepsilon(v)|. \end{aligned} \tag{3.48}$$

Kako je

$$\frac{\partial F_\varepsilon(x)}{\partial x_p} = 2 \sum_{i=1}^{m} \frac{(x_p - a^i) \exp\left(-\frac{\|x_p - a^i\|^2}{\varepsilon}\right)}{\sum_{j=1}^{k} \exp\left(-\frac{\|x_j - a^i\|^2}{\varepsilon}\right)},$$

slijedi

$$\begin{aligned} \left\| \tfrac{\partial F_\varepsilon(x)}{\partial x_p} \right\| &\leq 2 \sum_{i=1}^{m} \|x_p - a^i\| \leq 2 \sum_{i=1}^{m} \max_{j=1,\ldots,m} \|a^i - a^j\| \\ &\leq 2\, m \max_{i,j \in \{1,\ldots,m\}} \|a^i - a^j\|,\ p = 1, \ldots, k, \end{aligned}$$

tj. gradijent $\nabla F_\varepsilon(x)$ je neprekidan i ograničen na Δ^k. Koristeći Lagrangeov teorem o srednjoj vrijednosti za funkciju F_ϵ na Δ^k, zaključujemo da postoji $L > 0$ (neovisan o ε) takav da je

$$|F_\varepsilon(u) - F_\varepsilon(v)| \leq L\|u - v\|, \quad u, v \in \Delta^k.$$

Konačno, ako $\varepsilon \to 0^+$, iz (3.48) slijedi $|F(u) - F(v)| \leq L\|u - v\|$. □

Sljedeća lema i korolar daju vezu između funkcije cilja $\mathcal{F}$ zadane s (3.5) i funkcije cilja F zadane s (3.40).

Lema 3.6. *Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ konačan skup u $\mathbb{R}^n$, $z_1, \ldots, z_k \in \mathbb{R}^n$ skup međusobno različitih točaka, a $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$*

kvazimetrička funkcija. Neka je nadalje, $\Pi = \{\pi_1(z_1), \ldots, \pi_k(z_k)\}$ *particija čiji su klasteri dobiveni principom minimalnih udaljenosti a* $c_j \in \operatorname{argmin}_{x\in\mathbb{R}^n} \sum_{a\in\pi_j} d(x,a)$, $j = 1, \ldots, k$ *centri tih klastera. Tada vrijedi:*

$$F(z_1, \ldots, z_k) \overset{(\star)}{\geq} \mathcal{F}(\Pi) \overset{(\star\star)}{\geq} F(c_1, \ldots, c_k), \tag{3.49}$$

pri čemu jednakost u $(\star)$ *i* $(\star\star)$ *vrijedi onda i samo onda ako je* $z_j = c_j$ *za svaki* $j = 1, \ldots, k$.

Dokaz. U svrhu dokaza nejednakosti $(\star)$ sumu $\sum_{i=1}^{m}$ rastavit ćemo na k suma $\sum_{j=1}^{k} \sum_{a\in\pi_j}$.

$$\begin{aligned}
F(z_1, \ldots, z_k) &= \sum_{i=1}^{m} \min\{d(z_1, a^i), \ldots, d(z_k, a^i)\} \\
&= \sum_{j=1}^{k} \sum_{a^i\in\pi_j} \min\{d(z_1, a^i), \ldots, d(z_k, a^i)\} \\
&= \sum_{j=1}^{k} \sum_{a^i\in\pi_j} d(z_j, a^i) \\
&\overset{(\star)}{\geq} \sum_{j=1}^{k} \sum_{a^i\in\pi_j} d(c_j, a^i) = \mathcal{F}(\{\pi_1, \ldots, \pi_k\}).
\end{aligned}$$

U svrhu dokaza jednakosti $(\star\star)$ najprije primijetite da za svaki $a \in \pi_j$ vrijedi:

$$d(c_j, a) \geq \min\{d(c_1, a), \ldots, d(c_k, a)\}.$$

Zato vrijedi:

$$\begin{aligned}
\mathcal{F}(\{\pi_1, \ldots, \pi_k\}) &= \sum_{j=1}^{k} \sum_{a^i\in\pi_j} d(c_j, a^i) \\
&\geq \sum_{j=1}^{k} \sum_{a^i\in\pi_j} \min\{d(c_1, a^i), \ldots, d(c_k, a^i)\} \\
&= \sum_{i=1}^{m} \min\{d(c_1, a^i), \ldots, d(c_k, a^i)\} = F(c_1, \ldots, c_k),
\end{aligned}$$

iz čega slijedi tražena nejednakost. □

Teorem 3.7. *Neka je* $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\} \subset \mathbb{R}^n$. *Tada:*

(i) $c^\star = (c_1^\star, \ldots, c_k^\star)^T \in \underset{c_1,\ldots,c_k \in \mathbb{R}^n}{\operatorname{argmin}} F(c_1, \ldots, c_k)$ *onda i samo onda ako* $\Pi^\star = \{\pi_1^\star(c_1^\star), \ldots, \pi_k^\star(c_k^\star)\} \in \underset{\Pi \in \mathcal{P}(\mathcal{A};k)}{\operatorname{argmin}} \mathcal{F}(\Pi)$,

(ii) $\min\limits_{c_1,\ldots,c_k \in \mathbb{R}^n} F(c_1, \ldots, c_k) = \min\limits_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi)$.

Dokaz. (a) Neka je $c^\star = (c_1^\star, \ldots, c_k^\star)^T \in \underset{c_1,\ldots,c_k \in \mathbb{R}^n}{\operatorname{argmin}} F(c_1, \ldots, c_k)$. S $\pi_j^\star$ označimo odgovarajuće klastere dobivene principom minimalnih udaljenosti i $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$. Prema Lemi 3.6 je

$$F(c^\star) = \mathcal{F}(\Pi^\star). \tag{3.50}$$

Tvrdimo da je

$$\Pi^\star \in \underset{\Pi \in \mathcal{P}(\mathcal{A};k)}{\operatorname{argmin}} \mathcal{F}(\Pi). \tag{3.51}$$

Naime, ksd bi postojala particija $\mathcal{N}^\star = \{\nu_1^\star, \ldots, \nu_k^\star\} \in \mathcal{P}(\mathcal{A}; k)$ s centrima klastera $\zeta^\star = (\zeta_1^\star, \ldots, \zeta_k^\star)^T$ takva da je $\mathcal{F}(\mathcal{N}^\star) < \mathcal{F}(\Pi^\star)$, onda bi bilo

$$F(\zeta^\star) \overset{(Lema\ 3.6)}{=} \mathcal{F}(\mathcal{N}^\star) < \mathcal{F}(\Pi^\star) \overset{(Lema\ 3.6)}{=} F(c^\star),$$

a to nije moguće jer je $c^\star \in \underset{c \in \mathbb{R}^{n\times k}}{\operatorname{argmin}} F(c)$.

(b) Neka je $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\} \in \underset{\Pi \in \mathcal{P}(\mathcal{A};k)}{\operatorname{argmin}} \mathcal{F}(\Pi)$. S $c^\star = (c_1^\star, \ldots, c_k^\star)^T$ označimo centre klastera $\pi_1^\star, \ldots, \pi_k^\star$. Prema Lemi 3.6 je

$$F(c^\star) = \mathcal{F}(\Pi^\star). \tag{3.52}$$

Tvrdimo da je

$$c^\star \in \underset{c \in \mathbb{R}^{n\times k}}{\operatorname{argmin}} F(c). \tag{3.53}$$

Naime, kad bi postojao $\zeta^\star = (\zeta_1^\star, \ldots, \zeta_k^\star)^T$ takav da je $F(\zeta^\star) < F(c^\star)$, onda bi za particiju $\mathcal{N}^\star(\zeta^\star)$ vrijedilo

$$\mathcal{F}(\Pi^\star) \overset{(Lema\ 3.6)}{=} F(c^\star) > F(\zeta^\star) \overset{(Lema\ 3.6)}{=} \mathcal{F}(\mathcal{N}^\star),$$

a to nije moguće jer je $\Pi^\star \in \underset{\Pi \in \mathcal{P}(\mathcal{A};k)}{\operatorname{argmin}} \mathcal{F}(\Pi)$. □

Primjer 3.13. *Zadan je skup* $\mathcal{A} = \{1, 3, 4, 8\}$ *s* $m = 4$ *podatka. U Tablici 3.8 navedene su neke vrijednosti funkcija cilja* $\mathcal{F}_{LS}$ *i* F_{LS} *iz kojih se potvrđuje tvrdnja Leme 3.6 i Teorema 3.7. Jednakost u* $(\star)$ *postiže se na optimalnoj particiji pri čemu se* z_1, z_2 *podudaraju s centrima klastera (četvrti redak).*

	z_1	z_2	$F_{LS}(z_1, z_2)$	π_1	π_2	c_1	c_2	$\mathcal{F}_{LS}$	$F_{LS}(c_1, c_2)$
1.	1	4	17	{1}	{3,4,8}	1	5	14	14
2.	1	5	14	{1,3}	{4,8}	2	6	10	10
3.	3	7	6	{1,3,4}	{8}	$\frac{8}{3}$	8	$\frac{14}{3}$	$\frac{14}{3}$
4.	$\frac{8}{3}$	8	$\frac{14}{3}$	{1,3,4}	{8}	$\frac{8}{3}$	8	$\frac{14}{3}$	$\frac{14}{3}$

Tablica 3.8: Usporedba vrijednosti funkcije cilja $\mathcal{F}_{LS}$ i F_{LS} za $\mathcal{A} = \{1, 3, 4, 8\}$

Primjer 3.14. *Zadan je skup* $\mathcal{A} = \{16, 11, 2, 9, 2, 8, 15, 19, 8, 17\}$ *s* $m = 10$ *podataka. U Tablici 3.9 navedene su neke vrijednosti funkcija cilja* $\mathcal{F}_1$ *i* F_1 *iz kojih se potvrđuje tvrdnja Leme 3.6 i Teorema 3.7. Posebno obratite pažnju na treći redak koji pokazuje strogu nejednakost u* $(\star\star)$.

	z_1	z_2	$F_1(z_1, z_2)$	π_1	π_2	(c_1, c_2)	$\mathcal{F}_1$	$F_1(c_1, c_2)$
1.	2	6	55	{2,2}	{8,8,9,11,15,16,17,19}	{2,13}	31	31
2.	2	13	31	{2,2}	{8,8,9,11,15,16,17,19}	{2,13}	31	31
3.	3	15	29	{2,2,8,8,9}	{11,15,16,17,19}	{8,16}	23	21
4.	6	16	25	{2,2,8,8,9,11}	{15,16,17,19}	$\{8, \frac{33}{2}\}$	21	21
5.	8	16	21	{2,2,8,8,9,11}	{15,16,17,19}	$\{8, \frac{33}{2}\}$	21	21

Tablica 3.9: Usporedba vrijednosti funkcije cilja $\mathcal{F}_1$ i F_1

Zadatak 3.14. *Sličnu provjeru kao u Primjeru 3.13 provedite uz primjenu* ℓ_1*-metričke funkcije. Također, sličnu provjeru kao u Primjeru 3.14 provedite za slučaj LS-kvazimetričke funkcije.*

Koristeći Teorem 3.7 sada smo u mogućnosti provesti dokaz Teorema 3.2, str. 38, prema kojemu se povećanjem broja klastera u particiji ne povećava vrijednost funkcije cilja $\mathcal{F}$.

Dokaz. (Teorema 3.2, str. 38) Neka su $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_{k-1})^T$ centri optimalne $(k-1)$-particije $\Pi^{(k-1)}$, a $c^\star = (c_1^\star, \ldots, c_k^\star)^T$ centri optimalne k-particije

$\Pi^{(k)}$. Izaberimo $\zeta \in \mathbb{R}^n \setminus \{\hat{c}_1, \ldots, \hat{c}_{k-1}\}$ i označimo

$$\delta_{k-1}^i := \min_{1 \le s \le k-1} d(\hat{c}_s, a^i), \quad i = 1, \ldots, m.$$

Vrijedi:

$$\begin{aligned}
\mathcal{F}(\Pi^{(k-1)}) &\overset{(Teorem\ 3.7)}{=} F(\hat{c}) = \sum_{i=1}^{m} \min\{d(\hat{c}_1, a^i), \ldots, d(\hat{c}_{k-1}, a^i)\} = \sum_{i=1}^{m} \delta_{k-1}^i \\
&\geq \sum_{i=1}^{m} \min\{\delta_{k-1}^i, d(\zeta, a^i)\} \\
&\overset{(opt.part.\ \Pi^{(k)})}{\geq} \sum_{i=1}^{m} \min\{d(c_1^\star, a^i), \ldots, d(c_k^\star, a^i)\} \\
&= F(c^\star) \overset{(Teorem\ 3.7)}{=} \mathcal{F}(\Pi^{(k)}),
\end{aligned}$$

što potvrđuje da povećanjem broja klastera optimalne particije vrijednost funkcije cilja ne raste. □

Primjedba 3.4. *U dokazu je implicitno pokazano da i funkcija F ima svojstvo monotonosti.*

Lema 3.6 i Teorem 3.7 motivacija su za uvođenje sljedeće definicije.

Definicija 3.2. *Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ konačan skup u $\mathbb{R}^n$, $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ kvazimetrička funkcija i $\hat{\Pi} = \{\hat{\pi}_1, \ldots, \hat{\pi}_k\}$ particija čiji su klaster-centri $\hat{c}_1, \ldots, \hat{c}_k$ komponente vektora na kojemu se postiže lokalni minimum funkcije F. Particiju $\hat{\Pi}$ zvat ćemo* lokalno optimalnom k-particijom `(LOPart)` *skupa $\mathcal{A}$ ako vrijedi:*

$$\mathcal{F}(\hat{\Pi}) = F(\hat{c}_1, \ldots, \hat{c}_k). \tag{3.54}$$

Poglavlje 4

Traženje optimalne particije

Ako je $\mathcal{A} \subset \mathbb{R}^n$ konačan skup, onda je k-`GOPart` rješenje `GOP` (3.5), str. 37 ili ekvivalentnog problema (3.40), str. 58. Ne postoji metoda kojom bi se pouzdano pronašla k-`GOPart`. Primjena globalno optimizacijskog algoritma `DIRECT` na rješavanje problema (3.40) bila bi numerički vrlo neučinkovita jer minimizacijska funkcija obično ima veliki broj nezavisnih varijabli složenih u k vektora, a budući da je funkcija cilja simetrična u tih k vektora, algoritam `DIRECT` tražio bi svih $k!$ rješenja.

U slučaju podataka s jednim obilježjem postoje određene modifikacije i prilagođavanja algoritma `DIRECT` koja daju k-`GOPart` (vidi primjerice [38, 68–70, 85, 100]).

U općem slučaju algoritam `DIRECT` možemo upotrijebiti samo za traženje dobre početne aproksimacije, a onda primijeniti neku od poznatih lokalno optimizacijskih metoda (vidi primjerice [93, 94]). Najpoznatija lokalno optimizacijska metoda koja se u tu svrhu koristi je k-means algoritam kao i njegove brojne modifikacije. Na taj način možemo pronaći tek neku `LOPart`.

Kao početnu aproksimaciju za k-means algoritam u literaturi se predlažu i razne varijante inkrementalnog algoritma (vidi primjerice [6, 10, 62, 98]).

4.1 Transformacija podataka

Promatramo skup podataka $\mathcal{A} = \{a^i = (a^i_1, \ldots, a^i_n)^T : i = 1, \ldots, m\} \subset \Delta \subset \mathbb{R}^n$, gdje je $\Delta = [\alpha_1, \beta_1] \times \cdots \times [\alpha_n, \beta_n]$. Ako komponente $a^i_1, \ldots, a^i_n$ podataka $a^i \in \mathcal{A}$ nisu jednakog ranga, tj. ako se duljine intervala $[\alpha_j, \beta_j]$ međusobno značajnije razlikuju, treba ih normalizirati. To znači da skup $\mathcal{A}$ treba transformirati u skup $\mathcal{B} = \{\mathcal{T}(a^i) \in [0,1]^n : a^i \in \mathcal{A}\} \subset [0,1]^n$

primjenom preslikavanja $\mathcal{T}\colon \Delta \to [0,1]^n$ (vidi [38, 97], gdje je

$$\mathcal{T}(x) = D(x-\alpha), \qquad D = \operatorname{diag}\left(\tfrac{1}{\beta_1-\alpha_1}, \ldots, \tfrac{1}{\beta_n-\alpha_n}\right). \tag{4.1}$$

Nakon završenog postupka grupiranja skupa $\mathcal{B}$, dobivene rezultate treba vratiti nazad u skup Δ primjenom preslikavanja $\mathcal{T}^{-1}\colon [0,1]^n \to \Delta$, gdje je:

$$\mathcal{T}^{-1}(x) = D^{-1}x + \alpha. \tag{4.2}$$

Zato nadalje možemo pretpostavljati da je cijeli skup podataka $\mathcal{A}$ sadržan u hiperkocki $[0,1]^n$.

4.2 *k*-means algoritam II

4.2.1 Zapis funkcije cilja $\mathcal{F}$ pomoću matrice pripadnosti

Neka je $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ kvazimetrička funkcija, $\mathcal{A} \subset \mathbb{R}^n$, $|\mathcal{A}| = m$, konačan skup i $\Pi = \{\pi_1, \ldots, \pi_k\}$ k-particija s centrima klastera $c_j \in \underset{c\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{a\in\pi_j} d(c,a)$, $j = 1, \ldots, k$. Tada se funkcija cilja $\mathcal{F}$ iz (3.40) definira s:

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} d(c_j, a).$$

Funkcija $\mathcal{F}$ može se zapisati na drugi način, tj. uvođenjem tzv. matrice pripadnosti $U \in \{0,1\}^{m\times k}$ s elementima:

$$u_{ij} = \begin{cases} 1, & \text{if } a^i \in \pi_j \\ 0, & \text{if } a^i \notin \pi_j \end{cases}, \qquad i = 1, \ldots m, \quad j = 1, \ldots, k. \tag{4.3}$$

Uvjet (4.3) znači da svaki element $a^i \in \mathcal{A}$ mora pripasti točno jednom klasteru, a funkciju $\mathcal{F}$ možemo zapisati kao [13, 14, 106]:

$$\mathcal{F}(\Pi) = \Phi(c, U) = \sum_{j=1}^{k} \sum_{i=1}^{m} u_{ij} d(c_j, a^i), \tag{4.4}$$

gdje je $\Phi\colon \mathbb{R}^{n\times k} \times \{0,1\}^{m\times k} \to \mathbb{R}_+$.

Primjer 4.1. *Jedna 3-optimalna LS-particija skupa* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *je:*

$$\Pi = \{\{2,4\}, \{8,10\}, \{16\}\}.$$

Odgovarajuća matrica pripadnosti je:

$$U = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Primjedba 4.1. *Primijetite da za elemente matrice U vrijedi:*

$$\sum_{j=1}^{k} u_{ij} = 1, \quad i = 1, \ldots, m, \tag{4.5}$$

$$1 \leq \sum_{i=1}^{m} u_{ij} \leq m, \quad j = 1, \ldots, k. \tag{4.6}$$

Jednakosti (4.5) *znače da svaki element $a^i \in \mathcal{A}$ mora pripasti barem jednom klasteru, a nejednakosti* (4.6) *znače da svaki klaster ima najmanje jedan, a najviše m elemenata.*

Zadatak 4.1. *Koliko ima matrica $U \in \{0,1\}^{m\times k}$ koje zadovoljavaju uvjete* (4.5) *i* (4.6)*? U kakvoj je vezi broj svih ovakvih matrica s brojem* (3.2)*, str. 33 svih k-particija skupa $\mathcal{A}$ s m elemenata?*

U cilju optimizacije izračunavanja, funkcija (4.4) piše se u obliku:

$$\Phi(c, U) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} d(c_j, a^i), \tag{4.7}$$

a traženje globalno optimalne k-particije može se zapisati kao sljedeći `GOP`:

$$\underset{c\in\mathbb{R}^{n\times k},\, U\in\{0,1\}^{m\times k}}{\operatorname{argmin}} \Phi(c, U). \tag{4.8}$$

4.2.2 Standardni *k*-means algoritam

Promatramo globalni optimizacijski problem (4.8). Optimizacijska funkcija Φ je nelinearna nekonveksna funkcija s ogromnim brojem varijabli: $n \times k + m \times k$. Primjerice, funkcija Φ iz Primjera 4.1 sa samo $m = 5$ podataka ima $1 \times 3 + 5 \times 3 = 18$ nezavisnih varijabli. Samo u vrlo specijalnom slučaju kada je broj atributa $n = 1$ i kada broj klastera k nije prevelik, može se provesti direktna minimizacija nekom od specijaliziranih globalno optimizacijskih metoda za simetričnu Lipschitz neprekidnu funkciju (vidi [38, 69, 85] koje daju dovoljno točno rješenje.

U općem slučaju ne postoji optimizacijska metoda kojom bi se riješio GOP (4.8). Umjesto toga, u literaturi se mogu pronaći brojne metode koje pronalaze stacionarnu točku ili u najboljem slučaju, lokalni minimum. Na taj način dobiva se particija za koju obično ne znamo koliko je blizu optimalnoj.

Najpopularniji algoritam ovakvog tipa dobro je poznati *k-means algoritam* koji smo već ranije spominjali na str. 39. Sada ćemo konstruirati k-means algoritam koristeći zapis (4.7) funkcije cilja Φ pomoću matrice pripadnosti uz korištenje opće optimizacijske metode (vidi primjerice [67]): počevši od nekog $c^{(0)} \in \mathbb{R}^{k\times n}$ sukcesivno ponavljamo sljedeća dva koraka:

Korak A: Uz fiksni $c^{(0)}$ definiramo klastere $\pi_1, \ldots, \pi_k$ principom minimalnih udaljenosti (3.39). Na taj način s (4.3) određena je matrica $U^{(1)}$;

Korak B: Uz fiksnu matricu $U^{(1)}$ pronađemo optimalni $c^{(1)}$ rješavanjem k optimizacijskih problema:

$$c_1^{(1)} \in \operatorname*{argmin}_{c_1\in\mathbb{R}^n} \sum_{i=1}^{m} u_{i1}^{(1)} d(c_1, a^i), \ldots, c_k^{(1)} \in \operatorname*{argmin}_{c_k\in\mathbb{R}^n} \sum_{i=1}^{m} u_{ik}^{(1)} d(c_k, a^i),$$

i izračunamo $\Phi(c^{(1)}, U^{(1)})$ prema (4.7);

Na taj način sukcesivnom primjenom Koraka A i Koraka B dobivamo niz $(c^{(1)}, U^{(1)}), (c^{(2)}, U^{(2)}), \ldots$. Budući da za izbor matrice U imamo konačno mnogo mogućnosti, algoritam se u nekom momentu počinje ponavljati i tada smo dobili lokalno optimalnu particiju. Po konstrukciji, niz odgovarajućih funkcijskih vrijednosti $\Phi(c^{(1)}, U^{(1)}) \geq \Phi(c^{(2)}, U^{(2)}) \geq \cdots$ monotono je padajući i konačan.

Zbog toga se k-means algoritam formalno zapisuje s dva koraka koji se sukcesivno izmjenjuju. Algoritam se može zaustaviti kada relativna vrijednost funkcije cilja padne ispod nekog unaprijed zadanog pozitivnog realnog broja $\epsilon_{\text{KM}} > 0$:

$$\frac{\Phi_{r+1}-\Phi_r}{\Phi_r} < \epsilon_{\text{KM}} > 0. \tag{4.9}$$

Algoritam 4.1. [k-means algoritam II]

Korak A: *Pridruživanje (assignment step). Za dani skup $z_1, \ldots, z_k \in \mathbb{R}^n$ međusobno različitih točaka principom minimalnih udaljenosti odrediti klastere $\pi_1, \ldots, \pi_k$ (odnosno matricu pripadnosti $U \in \{0,1\}^{m\times k}$),*

Korak B: *Korekcija (update step). Za danu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ (odnosno matricu pripadnosti $U \in \{0,1\}^{m\times k}$) odrediti odgovarajuće klaster-centre $c_1, \ldots, c_k \in \mathbb{R}^n$;*

Staviti $z_j := c_j$, $j = 1, \ldots, k$.

Dobivene rezultate sumirajmo u sljedećem teoremu (usporedite s Teoremom 3.3, str. 41).

Teorem 4.1. *k-means algoritam 4.1 u konačno (T) koraka pronalazi* `LOPart`*, a pri tome je niz funkcijskih vrijednosti $(\mathcal{F}(\Pi^{(t)}))$, $t = 0, 1, \ldots, T$ monotono padajući.*

Dokaz. Već ranije zaključili smo da je niz funkcijskih vrijednosti $(\mathcal{F}(\Pi^{(t)}))$ dobiven k-means algoritmom monotono padajući te da nakon konačno mnogo koraka postaje stacionaran. Prema Lemi 3.6 i Teoremu 3.7, str. 62 za neki $T > 0$ vrijedi $\mathcal{F}(\Pi^{(T)}) = F(c_1, \ldots, c_k)$, gdje su c_j centri klastera particije $\Pi^{(T)}$. Zbog navedenog svojstva, sukladno Definiciji 3.2, str. 66, particija $\Pi^{(T)}$ je k-`LOPart`. □

Kao što smo već ranije naveli na str. 39, rješenje dobiveno k-means algoritmom jako ovisi o izboru početne aproksimacije (početnih centara ili početne particije) pa za način izbora početne aproksimacije u literaturi postoje brojne heurističke metode (vidi primjerice [5, 10, 52, 98]).

Primjedba 4.2. *Poznato je da se prilikom izvođenja k-means algoritma može dogoditi:*

- *da rješenje bude samo lokalno optimalna particija,*
- *da neki od klastera postanu prazni skupovi (vidi Primjer 3.4, str. 41),*
- *da se neki element $a \in \mathcal{A}$ može pojaviti na granici dva ili više klastera (vidi Primjer 4.4 i detaljnije u [92]).*

Primjer 4.2. *Neka je $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$. Potražimo optimalnu 2-particiju uz primjenu LS-kvazimetričke funkcije. Uz početnu particiju $\Pi^{(1)} = \{\{1, 2, 3\}, \{8, 9, 10, 25\}\}$ k-means algoritam kao 2-optimalnu particiju prepoznaje baš tu particiju (vidi Sliku 4.1a).*

Uz početnu particiju $\Pi^{(1)} = \{\{1, 2, 3, 8\}, \{9, 10, 25\}\}$ k-means algoritam kao 2-optimalnu particiju prepoznaje particiju $\Pi^{\star}$ prikazanu na Slici 4.1b.

Mathematica-programom `K-Means-IniPar-IniCen`[1] *lako se može provjeriti da je $\Pi^{\star}$ globalno optimalna 2-particija, a da je $\Pi^{(1)}$ jedina lokalno optimalna 2-particija.*

Uz vrijednosti kriterijskih funkcija cilja u niže navedenoj tablici prikazane su i odgovarajuće vrijednosti `CH` *i* `DB` *indeksa (vidi točku 5, str. 87).*

[1] `https://www.mathos.unios.hr/images/homepages/scitowsk/K-Means-IniPar-IniCen.nb`

Primijetite da se vrijednosti funkcija cilja F_{LS} i $\mathcal{F}_{LS}$ podudaraju na optimalnim particijama, a veličina `CH` *(odnosno* `DB`*) indeksa veća je (odnosno manja) na globalno optimalnoj particiji.*

Particije	$\mathcal{G}$	$\mathcal{F}_{LS}$	F_{LS}	`CH`	`DB`
$\Pi^{(1)}$	207.429	196	196	5.29	0.71
$\Pi^{\star}$	325.929	77.5	77.5	21.03	0.18

Slika 4.1: Lokalno i globalno optimalne 2-particije skupa $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$

Zadatak 4.2. *Za skup iz Primjera 4.2 korištenjem istog Mathematica programa odredite 2-optimalne particije uz primjenu ℓ_1-metričke funkcije. Postoje li i u ovom slučaju različite lokalno i globalno optimalne particije?*

Primjer 4.3. *Zadan je skup $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, 8\}$, gdje je:*

i	1	2	3	4	5	6	7	8
x_i	2	3	5	6	7	8	9	10
y_i	3	6	8	5	7	1	5	3

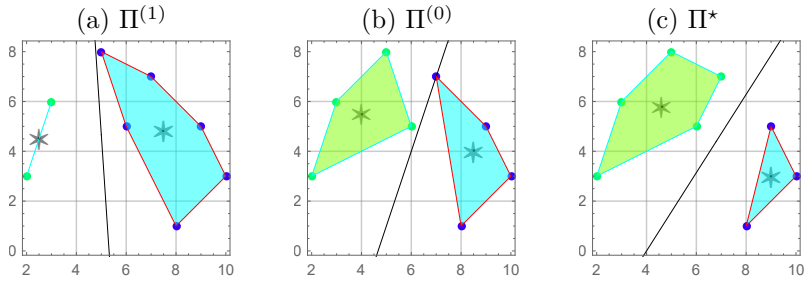
Potražimo optimalnu 2-particiju uz primjenu LS-kvazimetričke funkcije. Početnu particiju $\Pi^{(1)}$ prikazanu na Slici 4.2a k-means algoritam prepoznaje baš kao 2-optimalnu particiju.

Uz početnu particiju $\Pi^{(0)}$ prikazanu na Slici 4.2b, k-means algoritam kao 2-optimalnu particiju prepoznaje particiju $\Pi^{\star}$ prikazanu na Slici 4.2c.

Mathematica-programom `K-Means-IniPar-IniCen` *lako se može provjeriti da je $\Pi^{\star}$ globalno optimalna 2-particija, a da je $\Pi^{(1)}$ jedina lokalno optimalna 2-particija.*

Uz vrijednosti kriterijskih funkcija cilja, u niže navedenoj tablici prikazane su i odgovarajuće vrijednosti `CH` *i* `DB` *indeksa. Primijetite da se vrijednosti funkcija cilja F_{LS} i $\mathcal{F}_{LS}$ podudaraju na optimalnim particijama, a veličina* `CH` *(odnosno* `DB`*) indeksa veća je (odnosno manja) na globalno optimalnoj particiji.*

Particije	$\mathcal{G}$	$\mathcal{F}_{LS}$	F_{LS}	`CH`	`DB`
$\Pi^{(1)}$	37.67	55.33	55.33	4.08	.89
$\Pi^{\star}$	51	42	42	7.28	.83

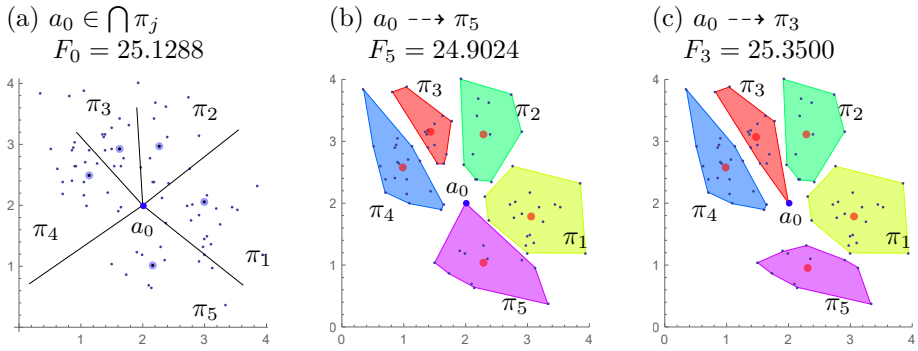


Slika 4.2: Lokalno ($\Pi^{(1)}$) i globalno ($\Pi^{\star}$) optimalna 2-particije skupa $\mathcal{A} \subset \mathbb{R}^2$

Zadatak 4.3. *Za skup iz Primjera 4.3 korištenjem istog Mathematica programa odredite* 2*-optimalne particije uz primjenu* ℓ_1*-metričke funkcije. Postoje li i u ovom slučaju različite lokalno i globalno optimalne particije?*

Sljedeći primjer pokazuje kako samo jedna točka može utjecati na dobivanje različitih lokalno optimalnih k-particija.

Primjer 4.4. *[92] Skup podataka* $\mathcal{A}$ *definiran je na sljedeći način. Najprije je određena točka* $a_0 \in \mathbb{R}^2$ *i pet različitih točaka* $z_1, \ldots, z_5 \in \mathbb{R}^2$ *slučajno odabranih na kružnici sa središtem u* $a_0 \in \mathbb{R}^2$*. U okolini svake točke* z_j *generirane su pseudoslučajne točke iz binormalne distribucije s očekivanjem* $0 \in \mathbb{R}^2$ *i jediničnom kovarijacijskom matricom.*



Slika 4.3: Traženje optimalne pozicije točke a_0

Pomoću točaka $z_1, \ldots, z_5$ *primjenom principa minimalnih udaljenosti definirani su klasteri* $\pi_j = \pi(z_j)$, $j = 1, \ldots, 5$*. Pri tome točka* a_0 *leži na zajedničkom rubu svih pet klastera (Slika 4.3a).*

Na Slici. 4.3b prikazana je lokalno optimalna particija, kod koje je točka a_0 *pridružena klasteru* π_5*. Slično, na Slici. 4.3c prikazana je druga lokalno optimalna particija, pri čemu je točka* a_0 *pridružena klasteru* π_3*.*

4.2.3 *k*-means algoritam s višestrukim pokretanjem

Problem izbora dobre početne aproksimacije prilikom traženja optimalne k-particije može se pokušati zaobići tako da se standardni k-means algoritam pokrene više puta s novim slučajno izabranim početnim centrima, i da se zadrži trenutno najbolja particija (vidi [52]). Ovaj princip opisan je u niže navedenom Algoritmu 1.

Algoritam 1 : (`k-means algoritam s višestrukim pokretanjem`)

Input: $\mathcal{A} \subset \Delta \subset \mathbb{R}^n$ {Skup podataka}; $k \geq 2$, $It > 1$;

1: Odredi $c^{(0)} \in \Delta^k$ slučajno;
2: Primijeni k-means algoritam na skup $\mathcal{A}$ s početnim centrom $c^{(0)}$, rješenje označi s $\hat{c} = \hat{c}^{(0)}$ i stavi $F_0 = F(\hat{c})$;
3: **for** $i = 1$ to It **do**
4: Odredi $c^{(i)} \in \Delta^k$ slučajno;
5: Primijeni k-means algoritam na skup $\mathcal{A}$, s početnim centrom $c^{(i)}$, rješenje označi s $\hat{c}^{(i)}$ i stavi $F_1 = F(\hat{c}^{(i)})$;
6: **if** $F_1 \leq F_0$ **then**
7: Stavi $\hat{c} = \hat{c}^{(i)}$ i stavi $F_0 = F_1$;
8: **end if**
9: **end for**

Output: $\{\hat{c}, F(\hat{c})\}$.

Primjedba 4.3. *Primijetite da linija 5 uključuje mogućnost da k-means algoritam smanji broj klastera. U tom slučaju, vrijednost funkcije F poraste, pa ta particija više nije kandidat za optimalnu k-particiju.*

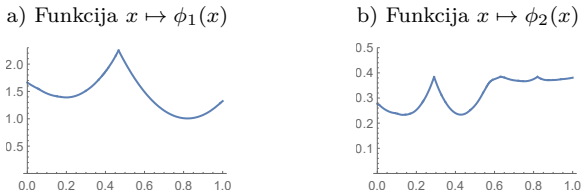
4.3 Inkrementalni algoritam

Pretpostavimo sada da najprikladniji broj klastera nije unaprijed poznat. Na temelju metode razvijene u radovima [5–7, 9–11, 62, 98], konstruirat ćemo inkrementalni algoritam koji će sukcesivno tražiti optimalne particije s $2, 3, \ldots$ klastera. Kako bismo ustanovili koja je od ovih particija najprikladnija u točki 5, str. 87, primijenit ćemo više različitih indeksa.

Primjer 4.5. *Zadana su tri centra $c_1^\star = 0.2$, $c_2^\star = 0.4$, $c_3^\star = 0.8$ i u njihovoj okolini po 10 normalno distribuiranih slučajnih brojeva, koji čine skup $\mathcal{A}$ (vidi Sliku 4.5a, odnosno Sliku 4.5d). Primjenom inkrementalnog algoritma pokušat ćemo rekonstruirati centre od kojih je nastao skup $\mathcal{A}$.*

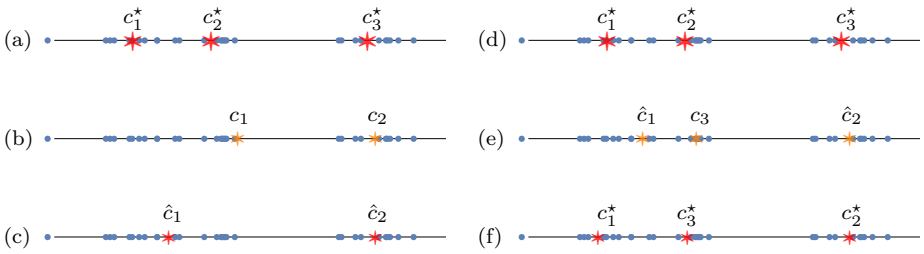
Najprije izaberimo $c_1 = \text{mean}(\mathcal{A})$, odredimo $\phi_1 = \sum_{a \in \mathcal{A}} (c_1 - a)^2$ i definirajmo funkciju (vidi Sliku 4.4a):

$$\phi_1(x) = \sum_{i=1}^{m} \min\{(c_1 - a^i)^2, (x - a^i)^2\}.$$



Slika 4.4: Traženje prvog i drugog centra

Primjenom *Mathematica*-modula `NMinimize[]` riješimo problem globalne optimizacije $\underset{x \in [0,1]}{\text{argmin}}\, \phi_1(x)$ i rješenje označimo s c_2 (Slika 4.5b).



Slika 4.5: Traženje sljedećeg centra

Nakon toga, na skup $\mathcal{A}$ primijenimo k-means algoritam s početnim centrima c_1, c_2. Tako dobivamo optimalnu 2-particiju skupa $\mathcal{A}$ s centrima $\hat{c}_1, \hat{c}_2$ (Slika 4.5c). Odredimo $\phi_2 = \phi(\hat{c}_1, \hat{c}_2)$.

U cilju pronalaženja optimalne 3-particije skupa $\mathcal{A}$ definirajmo pomoćnu funkciju (vidi Sliku 4.4b):

$$\phi_2(x) = \sum_{i=1}^{m} \min\{(\hat{c}_1 - a^i)^2, (\hat{c}_2 - a^i)^2, (x - a^i)^2\}.$$

Primjenom *Mathematica*-modula `NMinimize[]` riješimo problem globalne optimizacije $\underset{x \in [0,1]}{\text{argmin}}\, \phi_2(x)$ i rješenje označimo s c_3 (Slika 4.5e).

Nakon toga, na skup $\mathcal{A}$ primijenimo k-means algoritam s početnim centrima $(\hat{c}_1, \hat{c}_2, c_3)$. Tako dobivamo optimalnu 3-particiju skupa $\mathcal{A}$ s centrima $c_1^\star, c_2^\star, c_3^\star$ (koji

se u ovom slučaju vrlo dobro podudaraju s originalnim centrima - vidi Sliku 4.5f). Odredimo $\phi_3 = \phi(c_1^\star, c_2^\star, c_3^\star)$.

Odluku o zaustavljannju iterativnog procesa donosimo na osnovi kriterija (vidi primjerice [10]) $\frac{\phi_k - \phi_{k-1}}{\phi_1} < \epsilon$ za neki maleni $\epsilon > 0$ (recimo 0.05). U našem je primjeru $\frac{\phi_3 - \phi_2}{\phi_1} = 0.04$.

Zadatak 4.4. *Provedite sljedeći korak, odredite ϕ_4 i izračunajte $\frac{\phi_4 - \phi_3}{\phi_1}$.*

Konstruirajte i provedite inkrementalni algoritam na skupu $\mathcal{A}$ iz Primjera 4.5 uz primjenu ℓ_1-metričke funkcije.

Općenito, neka je $\mathcal{A} \subset \mathbb{R}^n$. Inkrementalni algoritam započinje izborom početnog centra $c_1 \in \mathbb{R}^n$. Primjerice, to može biti centroid ili medijan skupa $\mathcal{A}$. Sljedeći centar c_2 dobit ćemo kao rješenje `GOP` za funkciju $\Phi\colon \mathbb{R}^n \to \mathbb{R}$:

$$c_2 \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \Phi(x), \quad \Phi(x) := \sum_{i=1}^{m} \min\{\|c_1 - a^i\|_2^2, \|x - a^i\|_2^2\}. \tag{4.10}$$

Nakon toga, na centre c_1, c_2 primijenimo k-means algoritam i time dobivamo centre $c_1^\star, c_2^\star$ lokalno optimalne 2-particije $\Pi^{(2)}$.

Općenito, poznavajući k centara $c_1, \ldots, c_k$, aproksimaciju sljedećeg centra odredit ćemo rješavanjem `GOP`:

$$c_{k+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \Phi(x), \quad \Phi(x) := \sum_{i=1}^{m} \min\{\delta_k^i, \|x - a^i\|_2^2\}, \tag{4.11}$$

gdje je $\delta_k^i = \min\limits_{1 \le s \le k} \|c_s - a^i\|_2$. Nakon toga, primjenom k-means algoritma dobivamo centre $\{c_1^\star, \ldots, c_k^\star, c_{k+1}^\star\}$ optimalne $(k+1)$-particije $\Pi^{(k+1)}$.

U Koraku 1 i Koraku 4 Algoritma 2 dovoljno je provesti samo nekoliko iteracija (recimo 10) optimizacijskog algoritma `DIRECT` i tako dobiti dovoljno dobru početnu aproksimaciju za k-means algoritam koji nakon toga daje `LOPart`.

Algoritam 2 : `Inkrementalni algoritam`

Input: $\mathcal{A} \subset \mathbb{R}^n$ {Skup podataka}; $\epsilon > 0$;

1: Izabrati $c_1 \in \mathbb{R}^n$, izračunati $F_1 = F(c_1)$, riješiti `GOP` (4.10) i rješenje označiti s c_2;

2: Primijeniti k-means algoritam na centre $\hat{c}_1, \hat{c}_2$, rješenje označiti s $c^\star = \{c_1^\star, c_2^\star\}$, a odgovarajuću vrijednost funkcije cilja s $F_2 := F(c^\star)$;

3: Staviti $c_1 = c_1^\star$; $c_2 = c_2^\star$; $k = 2$; $F_k =: F_2$;

4: Za centre $c_1, \ldots, c_k$ riješi `GOP` $\underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \Phi(x)$, gdje je:

$$\Phi(x) := \sum_{i=1}^{m} \min\{\delta_k^i, \|x - a^i\|_2^2\}, \quad \delta_k^i = \min_{1 \le s \le k} \|c_s - a^i\|_2^2, \tag{4.12}$$

i rješenje označiti s c_{k+1};

5: Primijeniti k-means algoritam na centre $c_1, \ldots, c_k, c_{k+1}$, rješenje označiti s $c^\star = \{c_1^\star, \ldots, c_k^\star, c_{k+1}^\star\}$, a odgovarajuću vrijednost funkcije cilja s $F_{k+1} := F(c^\star)$;

6: **if** $\frac{1}{F_1}(F_k - F_{k+1}) > \epsilon$ **then**

7: $c_1 = c_1^\star, \ldots, c_{k+1} = c_{k+1}^\star$; $F_k = F_{k+1}$; $k = k+1$ i prijeći na Korak 4;

8: **end if**

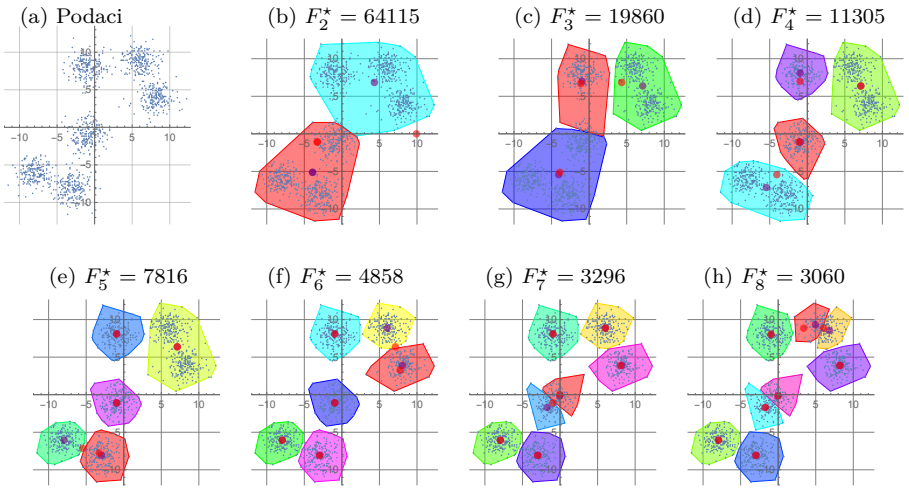
Output: $\{c_1^\star, \ldots, c_k^\star, c_{k+1}^\star\}$

Primjedba 4.4. *Primijetite da smo Algoritam 2 slično mogli pokrenuti i s više od jednog početnog centra.*

Primjer 4.6. *Skup podataka $\mathcal{A} \subset \mathbb{R}^2$ definirat ćemo na sljedeći način:*

```
In[1]:= SeedRandom[1213];
        c={{-8,-6},{-3,-8},{-1,-1},{8,4},{6,9},{-1,8}}; m=200;
        kov = 1.5 {{1,0},{0,1}};
        podaci = Table[RandomReal[
                    MultinormalDistribution[c[[i]], kov], m], {i,Length[c]}
                 ]];
        A = Flatten[podaci, 1];
```

Skup $\mathcal{A}$ sadrži $m = 1200$ podataka prikazanih na Slici 4.6a. Na ovaj skup primijenili smo inkrementalni algoritam i odredili `LOPart` s $2, \ldots, 8$ klastera, što je prikazano na Slici 4.6. U zaglavlju slike navedene su vrijednosti odgovarajuće funkcije cilja. Naravno, ostaje otvoreno pitanje koja od dobivenih optimalnih particija ima najprikladniji broj klastera.



Slika 4.6: Inkrementalni algoritam

4.4 Aglomerativni hijerarhijski algoritmi

Jedna mogućnost traženja optimalne particije su tzv. aglomerativni hijerarhijski algoritmi. Ovi algoritmi najviše se primjenjuju u području društvenih znanosti, biologije, medicine, arheologije, ali i u računarskim znanostima [49, 61, 104, 106].

4.4.1 Uvod i motivacija

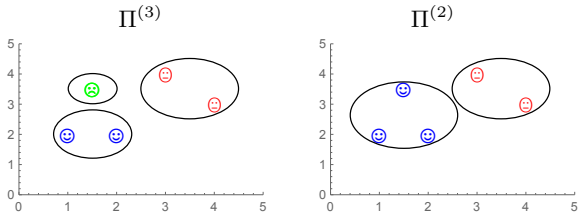
Osnovna ideja aglomerativnih hijerarhijskih algoritama sastoji se u tome da se polazeći od poznate particije $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ sastavljene od $1 < k \leq m$ klastera konstruira particija $\Pi^{(r)}$ s r klastera tako da se barem dva klastera particije $\Pi^{(k)}$ spoje u jedan $(r < k)$ ili tako da se jedan klaster particije $\Pi^{(k)}$ rastavi u barem dva klastera $(r > k)$. U tom smislu uvodimo sljedeću definiciju.

Definicija 4.1. *Kažemo da je particija* $\Pi^{(k)}$ *ugniježdena (engl.:* nested*) u particiju* $\Pi^{(r)}$ *i pišemo* $\Pi^{(k)} \sqsubset \Pi^{(r)}$ *ako vrijedi:*

(i) $r < k$,

(ii) svaki klaster iz $\Pi^{(k)}$ *podskup je nekog klastera iz* $\Pi^{(r)}$.

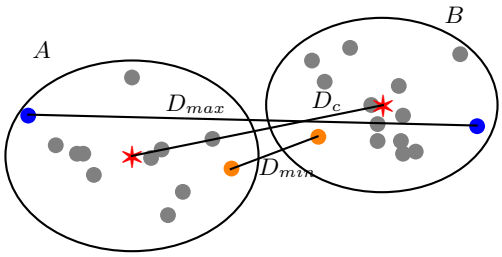
Na Slici 4.7 prikazana je particija $\Pi^{(3)}$ koja je ugniježđena u particiju $\Pi^{(2)}$. Nadalje ćemo razmatrati samo aglomerativne hijerarhijske algoritme koji u svakom koraku povezuju najviše po dva klastera promatrane



Slika 4.7: Particija $\Pi^{(3)}$ ugniježđena je u particiju $\Pi^{(2)}$ ($\Pi^{(3)} \sqsubset \Pi^{(2)}$)

k-particije $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$. Ta dva klastera odabrat ćemo tako da razmotrimo sve moguće parove klastera. Ukupan broj ovih parova jednak je broju svih kombinacija bez ponavljanja od k elemenata drugog razreda:

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}.$$



Slika 4.8: Različite mjere udaljenosti skupova A i B

Ako uvedemo neku kvazimetričku funkciju $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ (kao u točki 2, str. 9), onda kao mjeru sličnosti (odnosno različitosti) dva klastera možemo promatrati različite mjere njihove međusobne udaljenosti [50, 105, 106]. Općenito, udaljenost skupa A do skupa B možemo definirati na sljedeći način:

$$D_c(A,B) = d(c_A, c_B), \quad \text{[udaljenost centara } c_A, c_B \text{ skupova]} \tag{4.13}$$

$$D_{min}(A,B) = \min_{a\in A, b\in B} d(a,b) \quad \text{[minimalna udaljenost]} \tag{4.14}$$

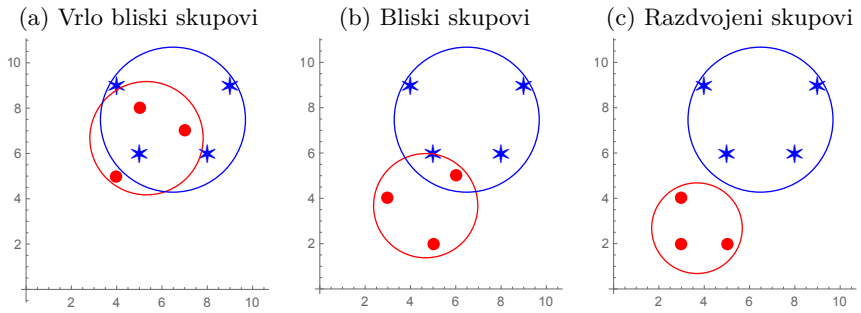
$$D_{max}(A,B) = \max_{a\in A, b\in B} d(a,b) \quad \text{[maksimalna udaljenost]} \tag{4.15}$$

$$D_{avg}(A,B) = \frac{1}{|A||B|} \sum_{a\in A} \sum_{b\in B} d(a,b) \quad \text{[prosječna udaljenost]} \tag{4.16}$$

$$\mathtt{HD}(A,B) = \max\{\max_{a\in A} \min_{b\in B} d(a,b), \max_{b\in B} \min_{a\in A} d(a,b)\}$$
$$\text{[Hausdorffova udaljenost]} \tag{4.17}$$

Zadatak 4.5. *Što je Hausdorffova udaljenost skupova A i B na Slici 4.8?*

Primjer 4.7. *Na Slici 4.9 prikazan je skup A (3 crvene točkice) i skup B (4 plave zvjezdice) u različitim međusobnim položajima, a u Tablici 4.1 navedene su odgovarajuće D_c, D_{min} i Hausdorffove* `HD` *udaljenosti za LS-kvazimetričku i ℓ_1-metričku funkciju za navedena tri slučaja.*



Slika 4.9: Udaljenost skupa A (crvene točke) od skupa B (plave zvjezdice)

	Vrlo bliski skupovi	Bliski skupovi	Razdvojeni skupovi
D_c uz ℓ_1-metričku funkciju	2	5	9
D_c uz LS-kvazimetričku funkciju	2.05	18	31.4
D_{min} uz ℓ_1-metričku funkciju	2	2	4
D_{min} uz LS-kvazimetričku funkciju	2	2	8
`HD` uz ℓ_1-metričku funkciju	4	7	11
`HD` uz LS-kvazimetričku funkciju	8	25	61

Tablica 4.1: Različite mjere udaljenosti dva skupa

U nastavku teksta detaljnije ćemo razmotriti primjenu mjere sličnosti (udaljenosti) klastera definiranu s (4.13) koristeći ranije spomenute kvazimetričke funkcije.

Neka je $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ zadani skup. Aglomerativni hijerarhijski algoritam može započeti od neke particije $\Pi^{(\mu)}$ koja sadržava μ klastera ($1 < \mu \leq m$) i završiti s particijom $\Pi^{(k)}$ koja sadržava k klastera ($1 \leq k < \mu \leq m$). Mi ćemo algoritam najčešće pokretati od particije s m klastera

$$\Pi^{(m)} = \{\{a^1\}, \{a^2\}, \ldots, \{a^m\}\},$$

tj. od particije u kojoj je svaki element skupa $\mathcal{A}$ za sebe poseban klaster. Udaljenost između klastera definirat ćemo kao udaljenost njihovih centara uz primjenu LS-kvazimetričke ili ℓ_1-metričke funkcija.

Primijetite da je vrijednost kriterijske funkcije cilja $\mathcal{F}$ na particiji $\Pi^{(m)}$ jednaka nuli. U prvom koraku biramo dva najsličnija klastera, tj. dva najbliža elementa skupa $\mathcal{A}$. Njih ćemo spojiti u jedan klaster. Primijetite da se sukladno Teoremu 3.2, str. 38, na taj način povećava vrijednost kriterijske funkcije cilja $\mathcal{F}$.

Algoritam se može zaustaviti na particiji s unaprijed zadanim brojem klastera, a moguće je razmotriti i problem određivanja particije s najprikladnijim brojem klastera [49, 110], što ćemo detaljnije razmatrati u Poglavlju 5.

Algoritam 3 (Agglomerative Nesting (AGNES))

Input: $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}, \quad 1 < k < m, \quad \mu = 0;$

1: Definirati početnu particiju $\Pi^{(m)} = \{\pi_1, \ldots, \pi_m\}$, $\pi_j = \{a^j\}$;
2: Za particiju $\Pi^{(m-\mu)}$ konstruirati matricu sličnosti $R_{m-\mu} \in \mathbb{R}^{(m-\mu)\times(m-\mu)}$, $r_{ij} = D(\pi_i, \pi_j)$;
3: Riješiti optimizacijski problem $\{i_0, j_0\} \subseteq \underset{1<i<j\leq m}{\operatorname{argmin}}\ r_{i,j}$;
4: Konstruirati novu particiju $\Pi^{(m-\mu-1)} = (\Pi^{(m-\mu)} \setminus \{\pi_{i_0}, \pi_{j_0}\}) \bigcup \{\pi_{i_0} \bigcup \pi_{j_0}\}$;
5: **if** $\mu < m - k$, **then**
6: $\quad \mu := \mu + 1$ i prijeći na Korak 2;
7: **else**
8: $\quad$ STOP;
9: **end if**

Output: $\{\Pi^{(k)}\}$.

U `Koraku 3` traži se pozicija $\{i_0, j_0\}$ najmanjeg elementa u gornjem trokutu matrice sličnosti $R_{m-\mu}$. U `Koraku 4` konstruira se nova particija tako da se klasteri π_{i_0}, π_{j_0} spoje u jedan klaster. U `Koraku 5` provjerava se kriterij zaustavljanja algoritma.

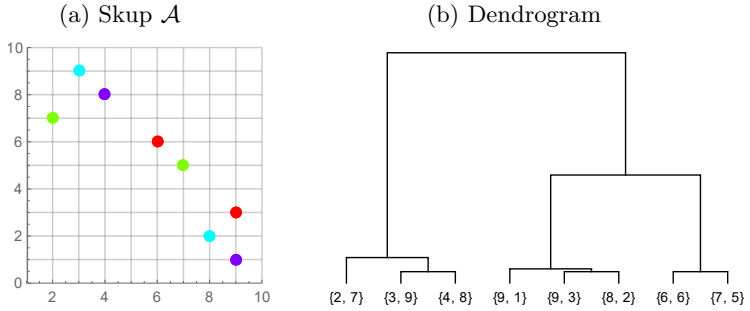
Koristan ilustrativni prikaz Algoritma 3 može se napraviti pomoću tzv. dendrograma, koji pokazuje svaki korak algoritma i daje razinu sličnosti. Algoritam ćemo ilustrirati na sljedećem jednostavnom primjeru.

Primjer 4.8. *Na skupu* $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, 8\}$ *zadanim s:*

i	1	2	3	4	5	6	7	8
x_i	2	3	4	6	7	8	9	9
y_i	7	9	8	6	5	2	1	3

(vidi Sliku 4.10a), polazeći od particije $\Pi^{(m)} = \{\{a^1\}, \ldots, \{a^m\}\}$ *pokrenut ćemo Algoritam* ***AGNES*** *koristeći mjeru sličnosti* (4.13) *uz primjenu* ℓ_1*-metričke funkcije (vidi Sliku 4.8):*

$$D_1(A, B) = \|c_A - c_B\|_1, \quad c_A = \operatorname*{med}_{a \in A} a \quad c_B = \operatorname*{med}_{b \in B} b.$$



Slika 4.10: Skup $\mathcal{A}$ i odgovarajući dendrogram uz primjenu ℓ_1-metričke funkcije

Sve particije skupa $\mathcal{A}$ mogu se dobiti programskim sustavom *Mathematica* (vidi Sliku 4.10b).

```
In[1]:= Needs["HierarchicalClustering`"]
In[2]:= DendrogramPlot[A, Linkage->"Median", HighlightLevel->2, LeafLabels->(# &)]
```

Pokazat ćemo kako se to može dobiti primjenom Algoritam `AGNES`[2]. U prvom prolazu kroz Algoritam `AGNES` krećemo od 8-particije $\Pi^{(8)}$ i najprije odredimo centre klastera (u ovom slučaju to su sami elementi skupa $\mathcal{A}$) i gornji trokut matrice sličnosti R_8. U njoj potražimo najmanji element (ako ima više jednakih najmanjih elemenata, izaberemo bilo koji), koji određuje dva najsličnija klastera. U našem slučaju izabrani su jednočlani klasteri

$$\{(3, 9)^T\}, \quad \{(4, 8)^T\}.$$

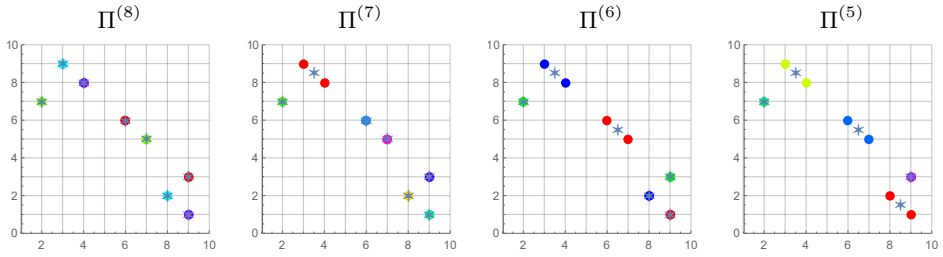
Oni se izbacuju iz particije, a dodaje se novi dvočlani klaster $\{(3, 9)^T, (4, 8)^T\}$. Tako dobivamo optimalnu 7-particiju $\Pi^{(7)}$ prikazanu na Slici 4.11.

[2] `https://www.mathos.unios.hr/images/homepages/scitowsk/Agglomerative_Nesting.nb`

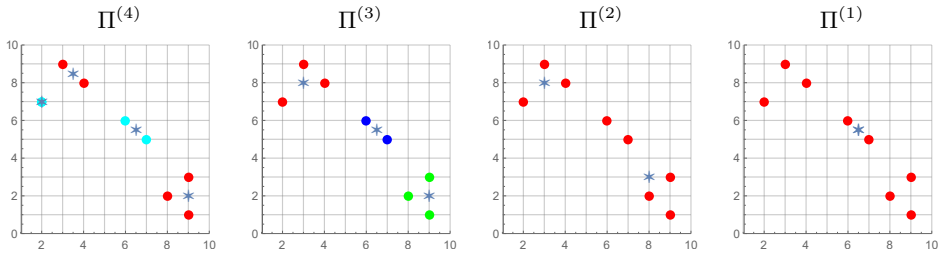
Ponavljajući postupak redom dobivamo ostale optimalne particije: $\Pi^{(6)}, \Pi^{(5)}, \Pi^{(4)}, \Pi^{(3)}, \Pi^{(2)}, \Pi^{(1)}$.

$$R_8 = \begin{bmatrix} 3 & 3 & 5 & 7 & 11 & 13 & 11 \\ & 2 & 6 & 8 & 12 & 14 & 12 \\ & & 4 & 6 & 10 & 12 & 10 \\ & & & 2 & 6 & 8 & 6 \\ & & & & 4 & 6 & 4 \\ & & & & & 2 & 2 \\ & & & & & & 2 \end{bmatrix}; \quad R_7 = \begin{bmatrix} 5 & 7 & 11 & 13 & 11 & 3 \\ & 2 & 6 & 8 & 6 & 5 \\ & & 4 & 6 & 4 & 7 \\ & & & 2 & 2 & 11 \\ & & & & 2 & 13 \\ & & & & & 11 \end{bmatrix}; \quad R_6 = \begin{bmatrix} 11 & 13 & 11 & 3 & 6 \\ & 2 & 2 & 11 & 5 \\ & & 2 & 13 & 7 \\ & & & 11 & 5 \\ & & & & 6 \end{bmatrix};$$

$$R_5 = \begin{bmatrix} 11 & 3 & 6 & 12 \\ & 11 & 5 & 2 \\ & & 6 & 12 \\ & & & 6 \end{bmatrix}; \quad R_4 = \begin{bmatrix} 3 & 6 & 12 \\ & 6 & 12 \\ & & 6 \end{bmatrix}; \quad R_3 = \begin{bmatrix} 6 & 6 \\ & 12 \end{bmatrix};$$



Slika 4.11: Optimalne k-particije. Zvjezdice označavaju centre klastera.



Slika 4.12: Optimalne k-particije. Zvjezdice označavaju centre klastera.

Zadatak 4.6. *Na skupu $\mathcal{A}$ iz Primjera 4.8 polazeći od particije $\Pi^{(m)} = \{\{a^1\}, \ldots, \{a^m\}\}$ provedite Algoritam* ***AGNES*** *koristeći mjeru sličnosti* (4.13) *uz primjenu LS-kvazimetričke funkcije.*

Zadatak 4.7. *Na skupu $\mathcal{A}$ iz Primjera 4.8 polazeći od particije $\Pi^{(m)} = \{\{a^1\}, \ldots, \{a^m\}\}$ provedite Algoritam* ***AGNES*** *koristeći mjeru sličnosti* (4.14) *uz primjenu ℓ_1-metričke funkcije.*

Zadatak 4.8. *Na skupu $\mathcal{A}$ iz Primjera 4.8 polazeći od particije $\Pi^{(m)} = \{\{a^1\}, \ldots, \{a^m\}\}$ provedite Algoritam* `AGNES` *koristeći mjeru sličnosti* (4.14) *uz primjenu LS-kvazimetričke funkcije.*

4.4.2 Primjena principa najmanjih kvadrata

Neka je $\Pi^{(k)} = \{\pi_1, \ldots, \pi_k\}$ neka k-particija konačnog skupa $\mathcal{A} \subset \mathbb{R}^n$. Uz primjenu LS-kvazimetričke funkcije, sličnost (udaljenost) klastera $\pi_r, \pi_s \in \Pi^{(k)}$ definirat ćemo kao udaljenost njihovih centroida:

$$D_{LS}(A, B) = \|c_r - c_s\|_2^2, \qquad c_r = \operatorname{mean} \pi_r, \quad c_s = \operatorname{mean} \pi_s. \tag{4.18}$$

Tvrdnja sljedećeg zadatka pomoći će nam kod dokaza Teorema 4.2 koji daje eksplicitne formule za centroid unije dva klastera i za vrijednost funkcije cilja na toj uniji.

Zadatak 4.9. *Slično kako kod dokaza Leme 3.2, str. 45, dokažite da za skup $A = \{a^i \in \mathbb{R}^n : i = 1, \ldots, p\}$ i njegov centroid $c_A = \operatorname{mean} A$ vrijedi:*

$$\sum_{i=1}^{p} (a^i - c_A) = 0, \tag{4.19}$$

$$\sum_{i=1}^{p} \|a^i - x\|_2^2 = \sum_{i=1}^{p} \|a^i - c_A\|_2^2 + p\|x - c_A\|_2^2, \quad \forall x \in \mathbb{R}^n. \tag{4.20}$$

Teorem 4.2. *Ako je skup $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ sastavljen od dva disjunktna neprazna klastera $\mathcal{A} = A \cup B$,*

$$A = \{a^1, \ldots, a^p\}, \quad |A| = p, \quad c_A = \tfrac{1}{p} \sum_{i=1}^{p} a^i;$$

$$B = \{b^1, \ldots, b^q\}, \quad |B| = q, \quad c_B = \tfrac{1}{q} \sum_{j=1}^{q} b^j,$$

tada je centroid skupa $\mathcal{A} = A \cup B$ zadan s:

$$c = c(\mathcal{A}) = \tfrac{p}{p+q} c_A + \tfrac{q}{p+q} c_B, \tag{4.21}$$

i vrijedi:

$$\mathcal{F}_{LS}(A \cup B) = \mathcal{F}_{LS}(A) + \mathcal{F}_{LS}(B) + p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2, \tag{4.22}$$

gdje je $\mathcal{F}_{LS}$ LS-funkcija cilja (vidi točku 3, str. 33).

Dokaz. Jednakost (4.21) neposredno slijedi iz:

$$c = c(A \cup B) = \tfrac{1}{p+q}\Big(\sum_{i=1}^{p} a^i + \sum_{j=1}^{q} b^j\Big) = \tfrac{p}{p+q}\tfrac{1}{p}\sum_{i=1}^{p} a^i + \tfrac{q}{p+q}\tfrac{1}{q}\sum_{j=1}^{q} b^j.$$

Korištenjem Zadatka 4.9 dobivamo:

$$\begin{aligned}\mathcal{F}_{LS}(A \cup B) =& \sum_{i=1}^{p} \|a^i - c\|_2^2 + \sum_{j=1}^{q} \|b^j - c\|_2^2 \\ =& \sum_{i=1}^{p} \|a^i - c_A\|_2^2 + p\|c - c_A\|_2^2 + \sum_{j=1}^{q} \|b^j - c_B\|_2^2 + q\|c - c_B\|_2^2 \\ =& \mathcal{F}_{LS}(A) + \mathcal{F}_{LS}(B) + p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2.\end{aligned}$$

□

Izraz $\Delta := p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2$ iz (4.22) može se pojednostaviti kao u sljedećem korolaru.

Korolar 4.1. *Ako je skup* $\mathcal{A} = \{a^i \in \mathbb{R}^n : i = 1, \ldots, m\}$ *s centroidom* $c = \frac{1}{m}\sum_{i=1}^{m} a^i$ *unija od dva klastera* $\mathcal{A} = A \cup B$,

$$A = \{a^1, \ldots, a^p\}, \quad |A| = p, \quad c_A = \tfrac{1}{p}\sum_{i=1}^{p} a^i;$$

$$B = \{b^1, \ldots, b^q\}, \quad |B| = q, \quad c_B = \tfrac{1}{q}\sum_{j=1}^{q} b^j,$$

tada je

$$\Delta := p\|c_A - c\|_2^2 + q\|c_B - c\|_2^2 = \tfrac{pq}{p+q}\|c_A - c_B\|_2^2. \tag{4.23}$$

Dokaz. Kako je prema (4.21)

$$\begin{aligned}p\|c_A - c\|_2^2 &= p\|\tfrac{p}{p+q}c_A + \tfrac{q}{p+q}c_B - c_A\|_2^2 = \tfrac{pq^2}{(p+q)^2}\|c_A - c_B\|_2^2, \\ q\|c_B - c\|_2^2 &= q\|\tfrac{p}{p+q}c_A + \tfrac{q}{p+q}c_B - c_B\|_2^2 = \tfrac{p^2q}{(p+q)^2}\|c_A - c_B\|_2^2,\end{aligned}$$

vrijedi:

$$\Delta = \tfrac{pq^2}{(p+q)^2}\|c_A - c_B\|_2^2 + \tfrac{p^2q}{(p+q)^2}\|c_A - c_B\|_2^2,$$

iz čega neposredno slijedi (4.23). □

Zbog toga se kao mjera sličnosti dva klastera A i B umjesto (4.18) može koristiti tzv. Wardova udaljenost [115]:

$$D_W(A, B) = \frac{|A||B|}{|A|+|B|} \|c_A - c_B\|_2^2, \tag{4.24}$$

gdje je $|A|$ broj elemenata klastera A, a $|B|$ broj elemenata klastera B. Kao što se može vidjeti iz Teorema 6.1 i Korolara 4.1, ako kao mjeru sličnosti dva klastera koristimo Wardovu udaljenost (4.24), onda će spajanjem klastera A i B vrijednost funkcije cilja porasti upravo za $D_W(A, B)$. Može se pokazati [106] da ovaj izbor osigurava najmanji mogući porast funkcije cilja $\mathcal{F}_{LS}$.

Poglavlje 5

Indeksi

5.1 Izbor particije s najprikladnijim brojem klastera

Možemo postaviti sljedeće pitanje:

> U koliko bi klastera bilo najprihvatljivije grupirati promatrani skup podataka $\mathcal{A}$, odnosno kako za promatrani skup podataka $\mathcal{A}$ odabrati particiju s najprikladnijim brojem klastera (engl.: Most Appropriate Partition (`MAPart`))?

Odgovor na ovo pitanje jedan je od najsloženijih problema klaster analize. O tome postoji brojna stručna literatura [14, 20, 25, 33, 49, 50, 78, 80, 90, 103, 106, 110], a obično se rješava ispitivanjem različitih pokazatelja koje jednostavno nazivamo indeksi.

U nekim jednostavnim slučajevima broj klastera u particiji određen je samom prirodom problema. Primjerice, prirodno je studente grupirati u $k = 5$ klastera prema postignutom uspjehu na studiju, ali nije lako dati odgovor na pitanje u koliko bi klastera bilo najprihvatljivije grupirati skup svih kukaca ili u koliko skupina treba grupirati privredne subjekte neke administrativne jedinice.

Ako broj klastera u koji treba grupirati skup $\mathcal{A}$ nije unaprijed poznat, prirodno bi bilo potražiti particiju koja se sastoji od klastera koji su interno što kompaktniji, a eksterno što bolje međusobno razdvojeni. Za takvu particiju reći ćemo da ima najprikladniji broj klastera – to će biti `MAPart`.

Razmotrit ćemo primjenu samo nekoliko najpoznatijih indeksa, koji pretpostavljaju korištenje LS-kvazimetričke funkcije.

5.1.1 Calinski–Harabasz indeks

Ovaj indeks predložili su T. Calinski i J. Harabasz u svom radu *„A dendrite method for cluster analysis"* objavljenom 1974. godine u časopisu Communications in Statistics. Nakon toga Calinski–Harabasz (CH) indeks doživio je brojna usavršavanja i prilagođavanja (vidi primjerice [14, 90, 110]).

CH-indeks definira se tako da interno kompaktnija particija čiji su klasteri dobro međusobno razdvojeni ima veću CH vrijednost.

Najprije primijetimo (vidi Teorem 3.2, str. 38) da vrijednost funkcije cilja ne raste povećanjem broja klastera u particiji, tj. da je niz funkcijskih vrijednosti na optimalnim particijama monotono padajući.

Ako prilikom određivanja optimalne k-particije $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ koristimo LS-kvazimetričku funkciju, onda odgovarajuću funkciju cilja $\mathcal{F}$ možemo zapisati kao:

$$\mathcal{F}_{LS}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|c_j - a\|_2^2. \tag{5.1}$$

Vrijednost funkcija $\mathcal{F}_{LS}$ na optimalnoj particiji $\Pi^\star$ pokazuje ukupno „rasipanje" elemenata svih klastera $\pi_1^\star, \ldots, \pi_k^\star$ te particije do njihovih centroida $c_1^\star, \ldots, c_k^\star$. Kao što smo ranije primijetili, što je vrijednost funkcije $\mathcal{F}_{LS}$ manja, time je „rasipanje" manje, što znači da su klasteri interno kompaktniji.

Zato ćemo pretpostaviti da je CH-indeks optimalne particije $\Pi^\star$ obrnuto proporcionalan vrijednosti funkcije cilja $\mathcal{F}_{LS}(\Pi^\star)$.

Kao što smo primijetili ranije u Poglavlju 3.2.2, str. 45, odnosno str. 52, prilikom traženja optimalne particije, osim minimizacije funkcije $\mathcal{F}_{LS}$, možemo potražiti maksimum odgovarajuće dualne funkcije:

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j \|c_j - c\|_2^2, \quad m_j = |\pi_j|, \tag{5.2}$$

pri čemu je $c = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \|x - a^i\|_2^2 = \frac{1}{m} \sum_{i=1}^{m} a^i$ centroid čitavog skupa $\mathcal{A}$.

Vrijednost funkcija $\mathcal{G}$ na particiji $\Pi^\star$ pokazuje ukupnu težinsku razdvojenost centroida $c_1^\star, \ldots, c_k^\star$ klastera $\pi_1^\star, \ldots, \pi_k^\star$. Što je vrijednost funkcije $\mathcal{G}$ veća, time su i LS-udaljenosti centroida $c_j^\star$ do centroida čitavog skupa $c^\star$ veće, pri čemu udaljenosti ponderiramo s brojem elemenata u pojedinom klasteru. To znači da su i centroidi $c_j^\star$ međusobno maksimalno moguće razdvojeni.

Zato ćemo pretpostaviti da je CH-indeks optimalne particije $\Pi^\star$ proporcionalan vrijednosti funkcije cilja $\mathcal{G}(\Pi^\star)$.

Uzevši u obzir još i statističke razloge povezane s brojem m elemenata skupa $\mathcal{A}$ i brojem k klastera u particiji $\Pi^\star$, mjera interne kompaktnosti i eksterne razdvojenosti klastera optimalne particije $\Pi^\star$ u slučaju primjene LS-kvazimetričke funkcije definirana je brojem [20, 110]:

$$\texttt{CH}(k) = \frac{\frac{1}{k-1}\mathcal{G}(\Pi^\star)}{\frac{1}{m-k}\mathcal{F}_{LS}(\Pi^\star)}, \tag{5.3}$$

koji nazivamo `CH`-indeks particije $\Pi^\star$. Particiju s najvećim `CH`-indeksom smatramo `MAPart`.

Primjer 5.1. *Promatrajmo skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *iz Primjera 3.5, str. 44. Dobivenu LS-optimalnu* 3*-particiju* $\Pi_3^\star = \{\{2, 4\}, \{8, 10\}, \{16\}\}$ *usporedimo s LS-optimalnom* 2*-particijom. Optimalne particije možemo potražiti pomoću k-means algoritma ili pomoću Inkrementalnog algoritma.*

(a) LS-optimalna 2-particija (b) LS-optimalna 3-particija

Slika 5.1: Izbor particije s prikladnijim brojem klastera

LS-optimalna 2-particija je $\Pi_2^\star = \{\{2, 4\}, \{8, 10, 16\}\}$ (Slika 5.1 a) za koju je

$$\mathcal{F}_{LS}(\Pi_2^\star) = 36.67, \quad \mathcal{G}(\Pi_2^\star) = 83.33.$$

Za LS-optimalnu 3-particiju $\Pi_3^\star$ (Slika 5.1 b) dobili smo

$$\mathcal{F}_{LS}(\Pi_3^\star) = 4, \quad \mathcal{G}(\Pi_3^\star) = 116.$$

Primijetite da sukladno teoriji vrijedi $\mathcal{F}_{LS}(\Pi_2^\star) \geq \mathcal{F}_{LS}(\Pi_3^\star)$ i također $\mathcal{G}(\Pi_2^\star) \leq \mathcal{G}(\Pi_3^\star)$.

Odgovarajući `CH`-indeksi su

$$\texttt{CH}(2) = \frac{83.33/1}{36.67/3} = 6.82, \qquad \texttt{CH}(3) = \frac{116/2}{4/2} = 29.$$

Budući da je $\texttt{CH}(3) > \texttt{CH}(2)$, particija $\Pi_3^\star$ particija je s prihvatljivijim (prikladnijim) brojem klastera.

Sljedeća propozicija pokazuje da se najveća vrijednost `CH`-indeksa k-particije postiže na globalno optimalnoj k-particiji.

Propozicija 5.1. *Neka je* $\mathcal{A} \subset \mathbb{R}^n$ *konačan skup i neka su* Π_1 *i* Π_2 *dvije različite* k*-particije skupa* $\mathcal{A}$*. Tada vrijedi:*

$$\mathit{CH}(\Pi_1) \geq \mathit{CH}(\Pi_2) \quad \Leftrightarrow \quad \mathcal{F}_{LS}(\Pi_1) \leq \mathcal{F}_{LS}(\Pi_2). \tag{5.4}$$

Dokaz. Neka je $m = |\mathcal{A}|$ i $c = \text{mean}(\mathcal{A})$. Označimo $\kappa := \sum_{i=1}^{m} \|c - a^i\|_2^2$. Tada prema (3.31), str. 52, vrijedi:

$$\mathcal{G}(\Pi_1) = \kappa - \mathcal{F}_{LS}(\Pi_1) \quad \text{i} \quad \mathcal{G}(\Pi_2) = \kappa - \mathcal{F}_{LS}(\Pi_2).$$

Zato vrijedi:

$$\begin{aligned} \texttt{CH}(\Pi_1) \geq \texttt{CH}(\Pi_2) &\Leftrightarrow \tfrac{m-k}{k-1} \frac{\mathcal{G}(\Pi_1)}{\mathcal{F}_{LS}(\Pi_1)} \geq \tfrac{m-k}{k-1} \frac{\mathcal{G}(\Pi_2)}{\mathcal{F}_{LS}(\Pi_2)} \\ &\Leftrightarrow \frac{\kappa - \mathcal{F}_{LS}(\Pi_1)}{\mathcal{F}_{LS}(\Pi_1)} \geq \frac{\kappa - \mathcal{F}_{LS}(\Pi_2)}{\mathcal{F}_{LS}(\Pi_2)} \\ &\Leftrightarrow \tfrac{\kappa}{\mathcal{F}_{LS}(\Pi_1)} - 1 \geq \tfrac{\kappa}{\mathcal{F}_{LS}(\Pi_2)} - 1 \Leftrightarrow \mathcal{F}_{LS}(\Pi_1) \leq \mathcal{F}_{LS}(\Pi_2). \square \end{aligned}$$

5.1.2 Davies–Bouldin indeks

Ovaj indeks predložili su D. Davies i D. Bouldin u svom radu *„A cluster separation measure"* objavljenom 1979. godine u časopisu IEEE Transactions on Pattern Analysis and Machine Intelligence. I ovaj indeks kasnije je doživio brojne prilagodbe i usavršavanja (vidi primjerice [14, 90, 110, 111, 113]).

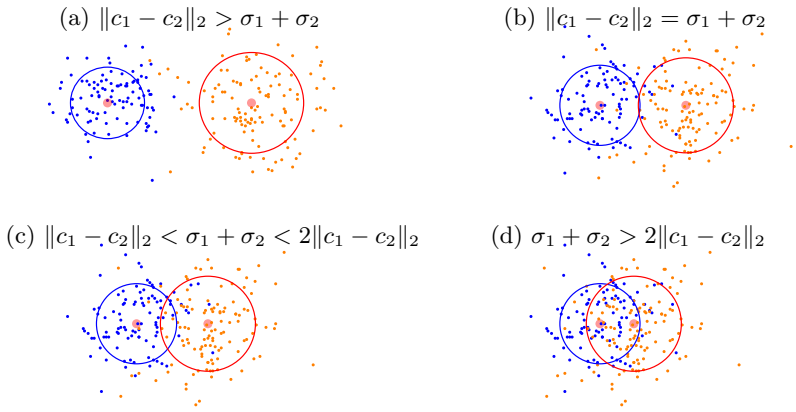
DB-indeks definira se tako da interno kompaktnija particija čiji su klasteri međusobno bolje razdvojeni ima manju DB vrijednost.

Niže navedeni koncept preuzet je iz [111]. Neka je $c \in \mathbb{R}^2$ točka u ravnini oko koje je primjenom Gaussove normalne distribucije s varijancom σ^2 generirano m slučajnih točaka a^i. Ovaj skup točaka čini *sferičan skup podataka* kojeg ćemo označiti s $\mathcal{A}$.

Iz statističke literature [12] poznato je da se u krugu $K(c, \sigma)$ sa središtem u točki c i radijusom σ (standardna devijacija) nalazi oko 68% točaka skupa $\mathcal{A}$. Ovaj krug zvat ćemo glavni krug skupa podataka $\mathcal{A}$.

Pretpostavimo da su za dvije različite točke $c_1, c_2 \in \mathbb{R}^2$ i dvije različite varijance σ_1^2, σ_2^2 na prethodno opisan način generirana dva sferična skupa podataka $\mathcal{A}_1, \mathcal{A}_2$ i da su $K_1(c_1, \sigma_1)$, $K_2(c_2, \sigma_2)$ njihovi odgovarajući glavni krugovi. Radijus σ_1 prvog kruga standardna je devijacija skupa $\mathcal{A}_1$, a radijus drugog kruga σ_2 standardna je devijacija skupa $\mathcal{A}_2$.

Mogući odnosi skupova $\mathcal{A}_1$ i $\mathcal{A}_2$ s obzirom na međusobni položaj njihovih glavnih krugova $K_1(c_1, \sigma_1)$ i $K_2(c_2, \sigma_2)$ prikazani su na Slici 5.2. Na Slici 5.2 a prikazani su skupovi $\mathcal{A}_1$ i $\mathcal{A}_2$ čiji se glavni krugovi ne sijeku i za koje vrijedi $\|c_1 - c_2\|_2 > \sigma_1 + \sigma_2$, na Slici 5.2 b prikazani su skupovi $\mathcal{A}_1$ i $\mathcal{A}_2$ čiji se glavni krugovi dodiruju i za koje vrijedi $\|c_1 - c_2\|_2 = \sigma_1 + \sigma_2$, itd.



Slika 5.2: Međusobno različiti odnosi dva sferična skupa podataka

Dakle, možemo reći da se glavni krugovi $K_1(c_1, \sigma_1)$, $K_2(c_2, \sigma_2)$ skupova $\mathcal{A}_1, \mathcal{A}_2$ presijecaju (imaju neprazan presjek) ako vrijedi [111]:

$$\|c_1 - c_2\|_2 \leq \sigma_1 + \sigma_2, \tag{5.5}$$

odnosno možemo reći da su glavni krugovi skupova $\mathcal{A}_1, \mathcal{A}_2$ razdvojeni ako vrijedi:

$$\frac{\sigma_1 + \sigma_2}{\|c_1 - c_2\|_2} < 1.$$

Promatrajmo sada optimalnu particiju $\Pi^\star$ skupa $\mathcal{A}$ s klasterima $\pi_1^\star, \ldots, \pi_k^\star$ i njihovim centroidima $c_1^\star, \ldots, c_k^\star$. Uočimo jedan od klastera $\pi_j^\star$ i razmotrimo njegov odnos prema ostalim klasterima. Primijetite da je veličinom

$$D_j := \max_{s \neq j} \frac{\sigma_j + \sigma_s}{\|c_j^\star - c_s^\star\|_2}, \quad \sigma_j^2 := \frac{1}{|\pi_j^\star|} \sum_{a \in \pi_j^\star} \|c_j^\star - a\|_2^2, \tag{5.6}$$

zadano najveće moguće preklapanje klastera $\pi_j^\star$ s nekim drugim klasterom. Veličina

$$\frac{1}{k}(D_1 + \cdots + D_k) \tag{5.7}$$

prosjek je brojeva (5.6), a predstavlja još jednu mjeru interne kompaktnosti i eksterne razdvojenosti klastera u particiji. Jasno je da što je broj (5.7) manji, klasteri su kompaktniji i bolje razdvojeni. Zato se DB-indeks optimalne particije $\Pi^\star$ skupa $\mathcal{A}$ s klasterima $\pi_1^\star, \ldots, \pi_k^\star$ i njihovim centroidima $c_1^\star, \ldots, c_k^\star$ definira na sljedeći način [14, 25, 80, 90, 110, 111]

$$\mathtt{DB}(k) = \frac{1}{k} \sum_{j=1}^{k} \max_{s \neq j} \frac{\sigma_j + \sigma_s}{\|c_j^\star - c_s^\star\|_2}, \quad \sigma_j^2 = \frac{1}{|\pi_j^\star|} \sum_{a \in \pi_j^\star} \|c_j^\star - a\|_2^2. \tag{5.8}$$

Particiju s najmanjim DB-indeksom smatramo `MAPart`.

Primjer 5.2. *Promatrajmo ponovo skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *iz Primjera 5.1. Primjenom DB-indeksa odredimo particiju s najprikladnijim brojem klastera.*

LS-optimalna 2-particija je $\Pi_2^\star = \{\{2, 4\}, \{8, 10, 16\}\}$. Centri njenih klastera su: $c_1^\star = 3$, $c_2^\star = 11.33$, a odgovarajuće standardne devijacije: $\sigma_1 = 1$, $\sigma_2 = 3.4$. Odgovarajući DB-indeks je

$$\mathtt{DB}(2) = \frac{1}{2}\left(\frac{\sigma_1+\sigma_2}{\|c_1^\star - c_2^\star\|_2} + \frac{\sigma_2+\sigma_1}{\|c_2^\star - c_1^\star\|_2}\right) = 0.58788.$$

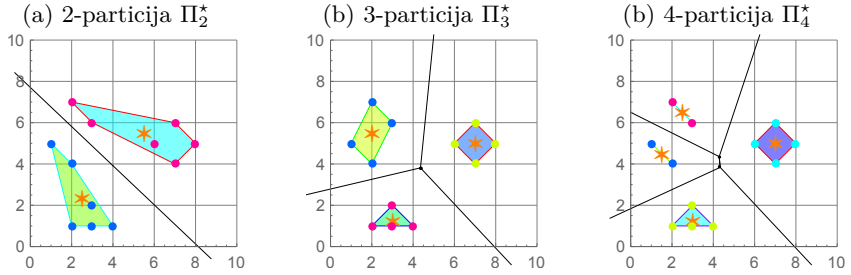
LS-optimalna 3-particija je $\Pi_3^\star = \{\{2, 4\}, \{8, 10\}, \{16\}\}$. Centri njenih klastera su: $c_1^\star = 3$, $c_2^\star = 9$, $c_3^\star = 16$, a odgovarajuće standardne devijacije: $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 0$. Odgovarajući DB-indeks je

$$\mathtt{DB}(3) = \frac{1}{3}\left(\max\left\{\frac{\sigma_1+\sigma_2}{\|c_1^\star - c_2^\star\|_2}, \frac{\sigma_1+\sigma_3}{\|c_1^\star - c_3^\star\|_2}\right\} + \max\left\{\frac{\sigma_2+\sigma_1}{\|c_2^\star - c_1^\star\|_2}, \frac{\sigma_2+\sigma_3}{\|c_2^\star - c_3^\star\|_2}\right\} + \max\left\{\frac{\sigma_3+\sigma_1}{\|c_3^\star - c_1^\star\|_2}, \frac{\sigma_3+\sigma_2}{\|c_3^\star - c_2^\star\|_2}\right\}\right) = 0.26984.$$

Budući da je DB(3)<DB(2), particija $\Pi_3^\star$ je particija s prihvatljivijim (prikladnijim) brojem klastera što je u suglasju s ranije dobivenim zaključkom pomoću CH-indeksa.

Primjer 5.3. *Potražimo particiju skupa* $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, 12\}$ *s najprikladnijim brojem klastera, pri čemu je*

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	2	3	2	2	3	4	3	6	8	7	7
y_i	5	4	6	7	1	2	1	1	5	5	4	6



Slika 5.3: Izbor LS-optimalne particije s najprikladnijim brojem klastera

Za optimalne LS-particije $\Pi_2^\star, \ldots, \Pi_6^\star$ s $2, 3, \ldots, 6$ klastera izračunat ćemo vrijednost funkcije cilja $\mathcal{F}_{LS}$, vrijednost CH-indeksa i vrijednost DB-indeksa. Na Slici 5.3 prikazane su LS-optimalna 2-particija, 3-particija i 4-particija, a na Slici 5.4 grafovi koji prikazuju vrijednost spomenutih indeksa za LS-optimalne particije.

(a) Calinski–Harabasz indeks (b) Davies–Bouldin indeks

Slika 5.4: Izbor particije s najprikladnijim brojem klastera

Kao što se može vidjeti, CH-indeks prima najveću, a DB-indeks najmanju vrijednost na istoj particiji. To znači da s visokom sigurnošću možemo tvrditi da je particija s najprikladnijim brojem klastera upravo 3-particija, što je i vizualno očekivano (vidi Sliku 5.3).

Zadatak 5.1. *Neka je $\mathcal{A} \subset \mathbb{R}^n$ konačan skup i neka su Π_1 i Π_2 dvije različite k-particije skupa $\mathcal{A}$. Vrijedi li:*

$$DB(\Pi_1) \leq DB(\Pi_2) \quad \Leftrightarrow \quad \mathcal{F}_{LS}(\Pi_1) \leq \mathcal{F}_{LS}(\Pi_2)? \tag{5.9}$$

Primjedba 5.1. *U Primjeru 4.2 za skup s jednim obilježjem izračunate su vrijednosti CH i DB-indeksa za lokalno i globalno optimalnu particiju. Slično je urađeno u Primjeru 4.3 za skup s dva obilježja. Pokazuje se da je veličina CH (odnosno DB) indeksa veća (odnosno manja) na globalno optimalnoj particiji.*

Primjer 5.4. *Ako na skupu $\mathcal{A}$ iz Primjera 4.8 (vidi Sliku 4.10a, str. 81) polazeći od particije $\Pi^{(m)} = \{\{a^1\}, \ldots, \{a^m\}\}$ provedemo Algoritam AGNES koristeći mjeru sličnosti* (4.13) *uz primjenu LS-kvazimetričke funkcije*

$$D_2(A, B) = \|c_A - c_B\|_2^2, \quad c_A = \operatorname*{mean}_{a \in A} a, \quad c_B = \operatorname*{mean}_{b \in B} b,$$

dobit ćemo sličan rezultat kao u Primjeru 4.8, ali u ovom slučaju možemo odrediti i odgovarajuće vrijednosti CH i DB-indeksa (vidi Tablicu 5.1).

Indeksi	$\Pi^{(2)}$	$\Pi^{(3)}$	$\Pi^{(4)}$	$\Pi^{(5)}$	$\Pi^{(6)}$	$\Pi^{(7)}$
CH	17.76	33.65	30.34	26.97	21.78	18.31
DB	0.50	0.42	0.36	0.38	0.25	0.18

Tablica 5.1: CH i DB-indeksi LS-optimalnih particija

Dok CH-indeks jasno pokazuje da je $\Pi^{(3)}$ particija s najprikladnijim brojem klastera, DB-indeks ne daje jasan odgovor.

Općenito, treba reći da navedeni indeksi daju prihvatljive zaključke ako su podaci takvi da se uklapaju u pretpostavke na osnovi kojih su indeksi konstruirani. Pri tome, kao što pokazuju i prethodno navedeni primjeri, zaključivanje o najprikladnijem broju klastera na skupu s relativno malenim brojem podataka neće biti dovoljno pouzdano. Štoviše, zaključivanje o najprikladnijem broju klastera u particiji temeljem navedenih indeksa nije uvijek jednoznačno.

5.1.3 Kriterij širine siluete

Kriterij širine siluete (engl. *Silhouette Width Criterion* (`SWC`)) vrlo je popularan u klaster analizi i primjenama [49, 106, 110]. Za k-`LOPart` $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ `SWC` se definira na sljedeći način: Za svaki $a^i \in \mathcal{A} \cap \pi_r^\star$ računaju se brojevi

$$\alpha_{ir} = \frac{1}{|\pi_r^\star|} \sum_{b \in \pi_r^\star} d(a^i, b), \quad \beta_{ir} = \min_{q \neq r} \frac{1}{|\pi_q^\star|} \sum_{b \in \pi_q^\star} d(a^i, b), \tag{5.10}$$

a odgovarajuci `SWC` indeks definiran je s

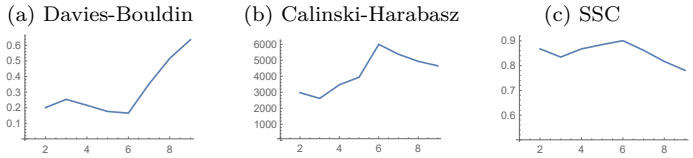
$$\texttt{SWC}(k) = \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_{ir} - \alpha_{ir}}{\max\{\alpha_{ir}, \beta_{ir}\}}. \tag{5.11}$$

Kompaktniji i bolje separirani klasteri u particiji daju veći `SWC` broj. Particiju s najvećim `SWC` indeksom smatramo `MAPart`.

Zbog složene numeričkog postupka `SWC`-indeksa, obično se koristi Pojednostavljeni kriterij širine siluete (engl. *Simplified Silhouette Width Criterion* (`SSC`)) koji umjesto prosječne vrijednosti (5.10) koristi udaljenost od elementa $a_i \in \mathcal{A} \cap \pi_r^\star$ do centara $c_1^\star, \ldots, c_k^\star$:

$$\alpha_{ir} = d(a_i, c_r^\star),\ \beta_{ir} = \min_{q \neq r} d(a^i, c_q^\star), \quad \texttt{SSC}(k) = \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_{ir} - \alpha_{ir}}{\max\{\alpha_{ir}, \beta_{ir}\}}. \tag{5.12}$$

Prema ovom kriteriju, particiju s najvećim `SSC`-indeksom smatramo `MAPart`.



Slika 5.5: Indeksi

Primjer 5.5. *Promatrajmo ponovo skup $\mathcal{A}$ iz Primjera 4.6, str. 77. Pokušajmo odrediti najprikladniji broj klastera primjenom prethodno spomenutih* `CH, DB` *i* `SSC`*-indeksa.*

Vrijednosti svih indeksa za optimalne particije $\Pi_2^\star, \ldots, \Pi_8^\star$ navedene su u Tablici 5.2 (indeksi najprikladnije particije označeni su bold), a grafički su prikazani na Slici 5.5. Sva tri indeksa ukazuju na to da je $\Pi_6^\star$ particija s najprihvatljivijem brojem klastera.

	$(k=2)$	$(k=3)$	$(k=4)$	$(k=5)$	$(k=6)$	$(k=7)$	$(k=8)$
DB	0.201	0.253	0.215	0.176	**0.165**	0.352	0.516
CH	2979	2621	3471	3946	**6001**	5384	4940
SSWC	0.868	0.835	0.867	0.885	**0.901**	0.862	0.817

Tablica 5.2: Vrijednosti indeksa k-optimalnih particije iz Primjera 5.5

5.2 Usporedba particija

Problem usporedbe dviju različitih particija $\Pi^{(1)} = \{\pi_1^{(1)}, \ldots, \pi_k^{(1)}\}$ i $\Pi^{(2)} = \{\pi_1^{(2)}, \ldots, \pi_\ell^{(2)}\}$ pojavljuje se u različitim situacijama kao primjerice ako želimo usporediti particiju dobivenu nekom metodom s originalnom particijom na kojoj smo testirali metodu. Pokazat ćemo kako se to može napraviti primjenom tzv. Rand ili Jaccard indeksa ovih particija ili tako da odredimo udaljenost skupa centara klastera originalne particije i skupa centara klastera izračunate particije.

5.2.1 Rand indeks dviju particija

Pojam Rand indeksa uvest ćemo prema [43]. Neka su $\Pi^{(1)} = \{\pi_1^{(1)}, \ldots, \pi_k^{(1)}\}$ i $\Pi^{(2)} = \{\pi_1^{(2)}, \ldots, \pi_\ell^{(2)}\}$ dvije particije skupa $\mathcal{A}$ s razdvojenim klasterima.

Definicija 5.1. *Neka je* $\mathcal{C} = \{(a^i, a^j) \in \mathcal{A} \times \mathcal{A} \colon 1 \leq i < j \leq m\} \subset \mathbb{R}^{n \times n}$. *Kažemo da su elementi para* $(a, b) \in \mathcal{C}$ *spareni (paired) u* $\Pi^{(1)}$ *ako pripadaju istom klasteru u* $\Pi^{(1)}$, *tj. ako postoji* $\pi_s^{(1)} \in \Pi^{(1)}$, *tako da je* $a, b \in \pi_s^{(1)}$. *Analogno se definira i sparenost u particiji* $\Pi^{(2)}$.

Primijetite da je skup $\mathcal{C}$ skup svih kombinacija bez ponavljanja drugog razreda od m elemenata skupa $\mathcal{A}$ pa je $|\mathcal{C}| = \binom{m}{2} = \frac{m(m-1)}{2}$.

Definicija 5.2. *Definiramo sljedeće podskupove skupa* $\mathcal{C}$*:*

$\mathcal{C}_1$: *svi parovi* $(a, b) \in \mathcal{C}$ *spareni u* $\Pi^{(1)}$ *i spareni u* $\Pi^{(2)}$,

$\mathcal{C}_2$: *svi parovi* $(a, b) \in \mathcal{C}$ *koji su spareni u* $\Pi^{(1)}$, *ali nisu spareni u* $\Pi^{(2)}$,

$\mathcal{C}_3$: *svi parovi* $(a, b) \in \mathcal{C}$ *koji nisu spareni u* $\Pi^{(1)}$, *ali su spareni u* $\Pi^{(2)}$,

$\mathcal{C}_4$: *svi parovi* $(a, b) \in \mathcal{C}$ *koji nisu spareni ni u* $\Pi^{(1)}$ *ni u* $\Pi^{(2)}$.

Primjedba 5.2. *Skup* $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$ *jedna je particija skupa* $\mathcal{C}$. *Ako uvedemo oznake:* $a := |\mathcal{C}_1|$, $b := |\mathcal{C}_2|$, $c := |\mathcal{C}_3|$, $d := |\mathcal{C}_4|$, *onda vrijedi:*

$$\frac{m(m-1)}{2} = |\mathcal{C}| = a + b + c + d.$$

Kažemo da su parovi iz skupa $\mathcal{C}_1 \cup \mathcal{C}_4$ *sukladni parovi (concordant pairs) jer imaju isti status u obje particije, dok za parove iz skupa* $\mathcal{C}_2 \cup \mathcal{C}_3$ *kažemo da nisu sukladni jer nemaju isti status u obje particije.*

Rand indeks definira se na sljedeći način:

$$R(\Pi^{(1)}, \Pi^{(2)}) := \frac{a + d}{a + b + c + d}. \tag{5.13}$$

Očigledno je $R(\Pi^{(1)}, \Pi^{(2)}) \in [0, 1]$, a particije su sličnije što je vrijednost Rand indeksa bliža 1. Primijetite da bi se u ovom slučaju Rand indeks mogao izračunati kao $R(\Pi^{(1)}, \Pi^{(2)}) = \frac{a+d}{|\mathcal{C}|}$, gdje je $|\mathcal{C}| = \binom{m}{2} = \frac{m(m-1)}{2}$.

U cilju izbjegavanja simetričnosti stukture kod Rand indeksa, definira se Jaccard indeks:

$$J(\Pi^{(1)}, \Pi^{(2)}) := \frac{a}{a + b + c} \tag{5.14}$$

Zadatak 5.2. *Pokažite da je* $D_R(\Pi^{(1)}, \Pi^{(2)}) = 1 - R(\Pi^{(1)}, \Pi^{(2)})$ *metrika na* $\mathcal{C}$.

U svrhu usporedbe particija $\Pi^{(1)}$ i $\Pi^{(2)}$ može se definirati i tzv. matrica prijelaza[1]:

$$S(\Pi^{(1)}, \Pi^{(2)}) = (s_{ij}), \tag{5.15}$$

gdje je s_{ij} broj elemenata klastera $\pi_i^{(1)}$ koji se pojavljuju u klasteru $\pi_j^{(2)}$ (vidi [106]). Ova matrica na finiji način pokazuje stupanj podudaranja particija $\Pi^{(1)}$ i $\Pi^{(2)}$.

Zadatak 5.3. *Neka je $\mathcal{A} \subset \mathbb{R}^n$. Ako je $U^{(1)}$ matrica pripadnosti k-particije $\Pi^{(1)}$, a $U^{(2)}$ matrica pripadnosti ℓ-particije $\Pi^{(2)}$, pokažite da se matrica prijelaza može dobiti kao*

$$S(\Pi^{(1)}, \Pi^{(2)}) = \left(U^{(1)}\right)^T \cdot U^{(2)}.$$

Primjer 5.6. *Neka je $\mathcal{A} = \{a^i : i = 1, \ldots, 9\}$. Promatramo dvije particije:*

$$\begin{aligned}\Pi^{(1)} &= \{\{a^1, a^2, a^3\}, \{a^4, a^5, a^6\}, \{a^7, a^8, a^9\}\},\\ \Pi^{(2)} &= \{\{a^1, a^2\}, \{a^3, a^4, a^5, a^6\}, \{a^7, a^8\}, \{a^9\}\}.\end{aligned}$$

Napišimo pripadne matrice pripadnosti (Membership Matrices). Tada se lako vidi da je:

$$\begin{aligned}a = 5 = |\mathcal{C}_1| &= |\{(a^1, a^2), (a^4, a^5), (a^4, a^6), (a^5, a^6), (a^7, a^8)\}|\\ b = 4 = |\mathcal{C}_2| &= |\{(a^1, a^3), (a^2, a^3), (a^7, a^9), (a^8, a^9)\}|\\ c = 3 = |\mathcal{C}_3| &= |\{(a^3, a^4), (a^3, a^5), (a^3, a^6)\}|\\ d = 24 = |\mathcal{C}_4| &= |\{(a^1, a^4), (a^1, a^5), (a^1, a^6), (a^1, a^7), (a^1, a^8), (a^1, a^9), (a^2, a^4), (a^2, a^5), (a^2, a^6),\\ &\quad (a^2, a^7), (a^2, a^8), (a^2, a^9), (a^3, a^7), (a^3, a^8), (a^3, a^9), (a^4, a^7), (a^4, a^8), (a^4, a^9),\\ &\quad (a^5, a^7), (a^5, a^8), (a^5, a^9), (a^6, a^7), (a^6, a^8), (a^6, a^9), \}|\\ &\quad R(\Pi^{(1)}, \Pi^{(2)}) = \tfrac{29}{36} = 0.805556.\end{aligned}$$

Primjer 5.7. *Promatramo podatke $\mathcal{A}$ iz Primjera 4.6, str. 77. Vrijednosti Rand odnosno Jaccard indeksa između originalne 6-particije Π i k-*`LOPart` *$\Pi^\star(2), \ldots, \Pi^\star(8)$ dobivenih inkrementalnim algoritmom i prikazanih na Slici 4.6, str. 78 navedeni su u Tablici 5.3.*

[1] Bez obzira što se u engleskoj literaturi može naći pod nazivom „confusion matrix", mislimo da je primjereniji naziv „matrica prijelaza".

Indeksi	$\Pi^\star(2)$	$\Pi^\star(3)$	$\Pi^\star(4)$	$\Pi^\star(5)$	$\Pi^\star(6)$	$\Pi^\star(7)$	$\Pi^\star(8)$
`Rand`	0.111	0.238	0.482	0.661	0.924	0.918	0.853
`Jaccard`	0.111	0.206	0.393	0.561	0.880	0.870	0.778

Tablica 5.3: `Rand` i `Jaccard` indeksi za k-`LOPart`, $k = 2, \ldots, 8$

Kao što se i očekivalo, najveće vrijednosti `Rand` i `Jaccard` indeksa dobivaju se za 6-`LOPart` $\Pi^\star(6)$. Pri tome matrica prijelaza (5.15) pokazuje vrlo dobro prepoznavanje elemenata klastera

$$S_6 = \begin{bmatrix} 197 & 3 & 0 & 0 & 0 & 0 \\ 5 & 195 & 0 & 0 & 0 & 0 \\ 2 & 8 & 190 & 0 & 0 & 0 \\ 0 & 0 & 1 & 178 & 21 & 0 \\ 0 & 0 & 0 & 0 & 200 & 0 \\ 0 & 0 & 0 & 0 & 0 & 200 \end{bmatrix}$$

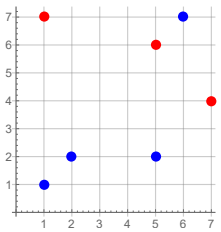
5.2.2 Primjena Hausdorffove udaljenosti

Dvije particije $\Pi^{(1)} = \{\pi_1^{(1)}, \ldots, \pi_k^{(1)}\}$ i $\Pi^{(2)} = \{\pi_1^{(2)}, \ldots, \pi_\ell^{(2)}\}$ možemo usporediti i tako da odredimo udaljenost skupa centara $c_1, \ldots, c_k$ klastera particije $\Pi^{(1)}$ i skupa centara $z_1, \ldots, z_\ell$ klastera particije $\Pi^{(2)}$. U tu svrhu iskoristit ćemo poznatu Hausdorffovu metričku funkciju [90] koju ćemo primijeniti na skupove centara.

Definicija 5.3. *Hausdorffova udaljenost dvaju skupova $\mathcal{A}$ and $\mathcal{B}$ definira se kao*

$$d_H(\mathcal{A}, \mathcal{B}) = \max\{\max_{x\in\mathcal{A}} \min_{y\in\mathcal{B}} d(x, y), \max_{y\in\mathcal{B}} \min_{x\in\mathcal{A}} d(x, y)\}. \tag{5.16}$$

Primjer 5.8. *Zadana su dva skupa točaka $A = \{(4, 9)^T, (5, 6)^T, (8, 6)^T, (9, 9)^T\}$ i $B = \{(3, 4)^T, (5, 2)^T, (6, 5)^T\}$ koja su prikazana plavim i crvenim točkama na Slici 5.6. Hausdorffova udaljenost između njih je $d_H(A, B) = 6$.*



Slika 5.6: Hausdorffova udaljenost dvaju skupova točaka

Poglavlje 6

Mahalanobis grupiranje podataka

U prethodnom poglavlju razmatrali smo problem grupiranja podataka u sferične klastere. Korak prema realnim problemima napravit ćemo u ovom poglavlju promatrajući problem grupiranja podataka u elipsoidalne klastere. Priroda problema ili geometrijski razlozi često ukazuju na potrebu traženja particije s elipsoidnim klasterima (vidi Primjer 1.1, str. 1.1, Primjer 1.4, str. 1.4, Primjer 1.9, str. 1.9).

Inicijalno, problem ćemo promatrati u ravnini, gdje ključnu ulogu ima TLS-pravac. Nakon toga bit će lako napraviti poopćenje na $\mathbb{R}^n$.

6.1 TLS-pravac u ravnini

Razmotrimo poznati problem određivanja najboljeg pravca u ravnini u smislu potpunih najmanjih kvadrata (*Total Least Squares* (TLS))[1] (vidi [22, 48, 64]). Za dani skup podataka $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ treba odrediti parametre pravca

$$ax + by + c = 0, \tag{6.1}$$

tako da suma kvadrata ortogonalnih udaljenosti točaka a^i do pravca (6.1) bude minimalna. Kako bi jednadžba (6.1) određivala neki pravac, na parametre a, b, c postavlja se zahtjev $a^2 + b^2 \neq 0$, koji osigurava da barem jedan od parametara a ili b bude različit od nule. Ako jednadžbu (6.1) podijelimo

[1] U znanstvenoj literaturi ovaj problem može se pronaći još i korištenjem ključne riječi „Errors-in-variables".

s $\sqrt{a^2+b^2}$ i uvedemo oznake: $\alpha := \frac{a}{\sqrt{a^2+b^2}}$, $\beta := \frac{b}{\sqrt{a^2+b^2}}$, $\gamma := \frac{c}{\sqrt{a^2+b^2}}$, zahtjev $a^2+b^2 \neq 0$ prelazi u $\alpha^2+\beta^2=1$, a problem određivanja optimalnih parametara traženog pravca možemo postaviti kao sljedeći GOP u $\mathbb{R}^3$:

$$\underset{\alpha,\beta,\gamma\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{m}(\alpha x_i+\beta y_i+\gamma)^2, \quad \text{uz uvjet} \quad \alpha^2+\beta^2=1. \tag{6.2}$$

Sljedeća lema omogućit će nam redukciju dimenzije ovog GOP na $\mathbb{R}^2$.

Lema 6.1. *Zadan je skup podataka $\mathcal{A}=\{a^i=(x_i,y_i)^T\in\mathbb{R}^2 : i=1,\ldots,m\}$ s težinama $w_i>0$. TLS-pravac prolazi centroidom podataka $(\bar{x},\bar{y})^T$, gdje je:*

$$\bar{x}=\tfrac{1}{W}\sum_{i=1}^{m} w_i x_i, \qquad \bar{y}=\tfrac{1}{W}\sum_{i=1}^{m} w_i y_i, \qquad W=\sum_{i=1}^{m} w_i.$$

Dokaz. Primijetimo najprije da ako normalni pravac $\alpha x+\beta y+\gamma=0$, $\alpha^2+\beta^2=1$, prolazi centroidom podataka $(\bar{x},\bar{y})^T$, onda njegova jednadžba glasi:

$$\alpha(x-\bar{x})+\beta(y-\bar{y})=0, \quad \text{uz uvjet} \quad \alpha^2+\beta^2=1. \tag{6.3}$$

Koristeći svojstvo (2.6), str.11 i linearnost aritmetičke sredine, dobivamo:

$$\begin{aligned}&\sum_{i=1}^{m} w_i(\alpha x_i+\beta y_i-(-\gamma))^2\\ \geq &\sum_{i=1}^{m} w_i(\alpha x_i+\beta y_i-(\alpha\bar{x}+\beta\bar{y}))^2=\sum_{i=1}^{m} w_i\left(\alpha(x_i-\bar{x})+\beta(y_i-\bar{y})\right)^2,\end{aligned}$$

pri čemu jednakost vrijedi onda i samo onda ako je $\gamma=-\alpha\bar{x}-\beta\bar{y}$. To znači da između svih normalnih pravaca $\alpha x_i+\beta y_i+\gamma=0$, $\alpha^2+\beta^2=1$, pravac (6.3) koji prolazi centroidom podataka ima najmanju moguću težinsku sumu kvadrata ortogonalnih odstupanja. □

Sukladno *Lemi 6.1*, TLS-pravac tražit ćemo u obliku (6.3) tako da umjesto rješavanja GOP (6.2), rješavamo GOP

$$\begin{aligned}&\underset{\alpha,\beta\in\mathbb{R}}{\operatorname{argmin}} F(\alpha,\beta), \quad \text{uz uvjet} \quad \alpha^2+\beta^2=1, \quad \text{gdje je}\\ &F(\alpha,\beta)=\sum_{i=1}^{m} w_i\left[\alpha(x_i-\bar{x})+\beta(y_i-\bar{y})\right]^2.\end{aligned} \tag{6.4}$$

Uz oznake:

$$B := \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_m - \bar{x} & y_m - \bar{y} \end{bmatrix}, \quad D = \operatorname{diag}(w_1, \ldots, w_m), \quad t = \begin{bmatrix} \alpha \\ \beta \end{bmatrix},$$

funkcija F može se zapisati u obliku:

$$F(\alpha, \beta) = \|\sqrt{D}\, Bt\|_2^2, \quad \|t\|_2 = 1. \tag{6.5}$$

Naime, $\|\sqrt{D}\, Bt\|_2^2 = (\sqrt{D}\, Bt)^T(\sqrt{D}\, Bt) = t^T B^T DBt$, a kako je

$$B^T DB = \begin{bmatrix} x_1 - \bar{x} & \cdots & x_m - \bar{x} \\ y_1 - \bar{y} & \cdots & y_m - \bar{y} \end{bmatrix} \cdot \begin{bmatrix} w_1(x_1 - \bar{x}) & w_1(y_1 - \bar{y}) \\ \cdots & \cdots \\ w_m(x_m - \bar{x}) & w_m(y_m - \bar{y}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^m w_i(x_i - \bar{x})^2 & \sum_{i=1}^m w_i(x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^m w_i(x_i - \bar{x})(y_i - \bar{y}) & \sum_{i=1}^m w_i(y_i - \bar{y})^2 \end{bmatrix},$$

dobivamo

$$[\alpha, \beta](B^T DB) \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = [\alpha, \beta] \begin{bmatrix} \alpha \sum_{i=1}^m w_i(x_i - \bar{x})^2 + \beta \sum_{i=1}^m w_i(x_i - \bar{x})(y_i - \bar{y}) \\ \alpha \sum_{i=1}^m w_i(x_i - \bar{x})(y_i - \bar{y}) + \beta \sum_{i=1}^m w_i(y_i - \bar{y})^2 \end{bmatrix}$$

$$= \alpha^2 \sum_{i=1}^m w_i(x_i - \bar{x})^2 + 2\alpha\beta \sum_{i=1}^m w_i(x_i - \bar{x})(y_i - \bar{y}) + \beta^2 \sum_{i=1}^m w_i(y_i - \bar{y})^2$$

$$= \sum_{i=1}^m w_i \big[\alpha(x_i - \bar{x}) + \beta(y_i - \bar{y})\big]^2 = F(\alpha, \beta).$$

Zadatak 6.1. *Pokažite da je matrica $B^T DB$ pozitivno definitna onda i samo onda ako točke $(x_i, y_i)^T$, $i = 1, \ldots, m$ ne leže na pravcu.*

Sukladno [64], vrijedi sljedeći teorem. Potrebni korišteni pojmovi mogu se pronaći u [22, 108].

Teorem 6.1. *Funkcija F zadan u* (6.4) *postiže svoj globalni minimum na jediničnom svojstvenom vektoru $t = (\alpha, \beta)^T$ koji odgovara manjoj svojstvenoj vrijednosti simetrične pozitivno definitne matrice $B^T DB$.*

Dokaz. Primijetimo da je $B^T DB \in \mathbb{R}^{2\times 2}$ simetrična matrica. Prema teoremu o dijagonalizaciji simetrične matrice [108], postoji ortogonalna matrica V i dijagonalna matrica $\Delta = \operatorname{diag}(\lambda_1, \lambda_2)$ takva da je $B^T DB = V\Delta V^T$.

Pri tome su λ_1 i λ_2 pozitivne svojstvene vrijednosti, a stupci matrice V odgovarajući svojstveni vektori matrice B^TDB.

Za proizvoljni jedinični vektor $t \in \mathbb{R}^2$ vrijedi:

$$\|\sqrt{D}Bt\|_2^2 = (\sqrt{D}Bt)^T(\sqrt{D}Bt) = t^TB^TDBt = t^TV\Delta V^Tt = s^T\Delta s,$$

gdje je $s = V^Tt$. Pri tome, zbog ortogonalnosti matrice V^T, vrijedi $\|s\|_2 = 1$. Zato je

$$\|\sqrt{D}Bt\|_2^2 = s^T\Delta s = \sum_{i=1}^{2} \lambda_i s_i^2 \geq \lambda_{\min}(B^TDB).$$

Posljednja nejednakost posljedica je činjenice da se minimum konveksne kombinacije postiže na najmanjem broju. Nije teško vidjeti da se jednakost postiže upravo kada je t jednak jediničnom svojstvenom vektoru koji je pridružen najmanjoj (manjoj) svojstvenoj vrijednosti matrice B^TDB. □

Zadatak 6.2. *Odredite eksplicitne formule za svojstvene vrijednosti matrice* B^TDB.

Primjer 6.1. *Zadani su podaci*

w_i	1	1	1	1	1
x_i	1	2	3	4	5
y_i	1	3	4	2	3

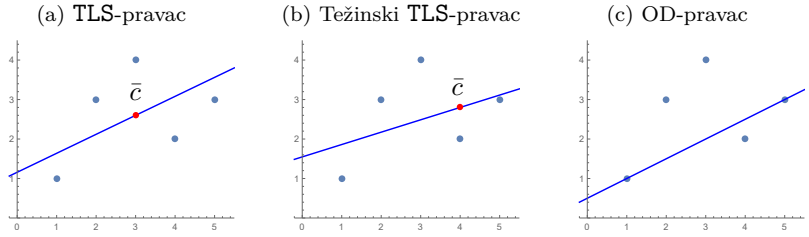
Centroid podataka je u točki $\bar{c} = (3, \frac{13}{5})^T$, a matrice B, D i B^TDB u ovom slučaju su:

$$B = \begin{bmatrix} -2 & -8/5 \\ -1 & 2/5 \\ 0 & 7/5 \\ 1 & -3/5 \\ 2 & 2/5 \end{bmatrix}, \quad D = \mathrm{diag}(1,1,1,1,1), \quad B^TDB = \begin{bmatrix} 10 & 3 \\ 3 & 26/5 \end{bmatrix}.$$

Svojstvene vrijednosti matrice B^TDB su $\lambda_1 = 11.442$, $\lambda_2 = 3.758$, a jedinični svojstveni vektor koji odgovara manjoj svojstvenoj vrijednosti je $t = (-0.43, 0.90)^T$. Zato jednadžba najboljeg TLS-pravca glasi $-0.433\,(x-3) + 0.901\,(y - \frac{13}{5}) = 0$, odnosno $-0.433x + 0.901y + 3.643 = 0$. Graf ovog pravca prikazan je na Slici 6.1a.

Ako težinu $w_5 = 1$ zamijenimo s $w_5 = 6$, centroid podataka postaje točka $\bar{c} = (4, \frac{14}{5})^T$, a matrice B, D i B^TDB u ovom slučaju su:

$$B = \begin{bmatrix} -3 & -9/5 \\ -2 & 1/5 \\ -1 & 6/5 \\ 0 & -4/5 \\ 1 & 1/5 \end{bmatrix}, \quad D = \mathrm{diag}(1,1,1,1,6), \quad B^TDB = \begin{bmatrix} 20 & 5 \\ 5 & 28/5 \end{bmatrix}.$$



Slika 6.1: Najbolji TLS i OD pravac

Svojstvene vrijednosti matrice B^TDB postaju $\lambda_1 = 21.566$ i $\lambda_2 = 4.034$, a jedinični svojstveni vektor koji odgovara manjoj svojstvenoj vrijednosti je $t = (-0.299, 0.954)^T$. Zato jednadžba TLS-pravca glasi $-0.299\,(x-4)+0.954\,(y-\frac{14}{5}) = 0$, odnosno $-0.299x+0.954y+3.867 = 0$. Graf ovog pravca prikazan je na Slici 6.1b.

$\star \quad \star \quad \star \quad \star \quad \star$

Matrica $\Sigma = \frac{1}{W}B^TDB$, uz neke uvjete na skup podataka $\mathcal{A}$ (vidi Lemu 6.2), pozitivno je definitna simetrična matrica. Njezine svojstvene vrijednosti su realni brojevi, a odgovarajući svojstveni vektori međusobno su okomiti. U smjeru svojstvenog vektora, koji odgovara većoj svojstvenoj vrijednosti, usmjeren je TLS-pravac. Jedinični svojstveni vektor koji odgovara manjoj svojstvenoj vrijednosti ove matrice je jedinični vektor normale na TLS-pravac $\vec{n}_0 = \xi\vec{i} + \eta\vec{j}$, a budući da TLS-pravac mora proći centroidom $(\bar{x}, \bar{y})^T$ skupa $\mathcal{A}$, njegova jednadžba glasi:

$$\xi(x - \bar{x}) + \eta(x - \bar{y}) = 0.$$

U statističkoj literaturi matrica Σ naziva se *kovarijacijska matrica* (engl.: covariance matrix) slučajnih varijabli X, Y, a smjerovi svojstvenih vektora nazivaju se glavni smjerovi (engl.: principal components). U smjeru svojstvenog vektora, koji odgovara većoj svojstvenoj vrijednosti, varijanca podataka je veća, a u smjeru svojstvenog vektora, koji odgovara manjoj svojstvenoj vrijednosti, varijanca podataka je manja.

6.2 Mahalanobis kvazimetrička funkcija u ravnini

Neka je $\mathcal{S}\colon X_0 \to X_0$ linearni operator kontrakcije/dilatacije kojemu u bazi $e = \{e_1, e_2\}$ pripada dijagonalna matrica

$$\Sigma(e) = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}, \quad \alpha, \beta > 0.$$

Operator $\mathcal{S}$ jediničnu kružnicu (uz primjenu l_2-udaljenosti) sa središtem u ishodištu

$$K = \{x \in \mathbb{R}^2 \colon \|x\|_2^2 = 1\} = \{(x_1, x_2)^T \in \mathbb{R}^2 \colon x_1^2 + x_2^2 = 1\}$$

preslikava na elipsu s poluosima α, β:

$$E = \{\xi = (\xi_1, \xi_2)^T \in \mathbb{R}^2 \colon \tfrac{\xi_1^2}{\alpha^2} + \tfrac{\xi_2^2}{\beta^2} = 1\}.$$

Naime, neki vektor $x = x_1 e_1 + x_2 e_2 \in K$ operator $\mathcal{S}$ preslikava u vektor $\xi = \xi_1 e_1 + \xi_2 e_2$ pa jednakost $\mathcal{S}(x) = \xi$ daje:

$$\alpha x_1 e_1 + \beta x_2 e_2 =: \xi_1 e_1 + \xi_2 e_2.$$

Zato je veza između komponenti vektora x i ξ zadana s:

$$x_1 = \tfrac{\xi_1}{\alpha}, \quad x_2 = \tfrac{\xi_2}{\beta},$$

pa kružnica K prelazi u elipsu E.

Definirajmo takvu kvazimetričku funkciju $d_m \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ primjenom koje će točke na elipsi E biti jednako udaljene od ishodišta O, odnosno primjenom koje će elipsa E postati jedinična kružnica u prostoru snabdjevenom kvazimetričkom funkcijom d_m. To ćemo postići tako da udaljenosti u smjeru svojstvenih vektora uzimamo obrnuto proporcionalno veličini odgovarajuće svojstvene vrijednosti operatora $\mathcal{S}$. Upravo takvo djelovanje ima inverzni linearni operator $\mathcal{S}^{-1}$.

Zato kvazimetričku funkciju $d_m \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ definiramo s:

$$d_m(x, y; \Sigma) := \|\Sigma^{-1/2}(x - y)\|_2^2 = (x - y)^T \Sigma^{-1} (x - y). \tag{6.6}$$

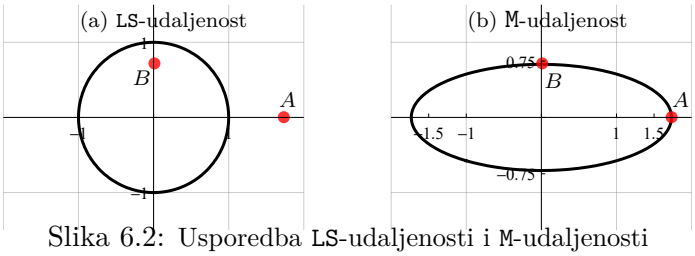
Kvazimetrička funkcija d_m naziva se Mahalanobis kvazimetrička funkcija, a udaljenost $d_m(x, y; \Sigma)$ naziva se Mahalanobis udaljenost (M-udaljenost) točaka $x, y \in \mathbb{R}^2$. Na taj način elipsa E postaje jedinična Mahalanobis kružnica (M-kružnica):

$$M(O; \Sigma) = \{x \in \mathbb{R}^2 \colon d_m(O, x; \Sigma) = 1\}.$$

Primjer 6.2. *Linearni operator $\mathcal{S} \colon X_0 \to X_0$ u bazi $e = \{e_1, e_2\}$ zadan je matricom $\Sigma(e) = \begin{bmatrix} 3 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$. Jedinična M-kružnica u ovom slučaju je:*

$$\tfrac{\xi^2}{(\sqrt{3})^2} + \tfrac{\eta^2}{(1/\sqrt{2})^2} = 1,$$

a geometrijski predstavlja elipsu s glavnom poluosi $\sqrt{3}$ i sporednom poluosi $\frac{1}{\sqrt{2}}$. Primijetite da točke $A = (\sqrt{3}, 0)$ i $B = (0, \frac{1}{\sqrt{2}})$ leže na jediničnoj M-kružnici.



Slika 6.2: Usporedba LS-udaljenosti i M-udaljenosti

6.3 Mahalanobis udaljenost inducirana skupom točaka iz ravnine

Zbog jednostavnosti, problem promatramo na $\mathbb{R}^2$. Za dani skup podataka $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ s težinama $w_1, \ldots, w_m > 0$ potražit ćemo glavne smjerove izduženja. U tu svrhu promatramo kovarijacijsku matricu

$$\Sigma = \frac{1}{W} B^T D B = \begin{bmatrix} \frac{1}{W}\sum_{i=1}^{m} w_i (x_i - \bar{x})^2 & \frac{1}{W}\sum_{i=1}^{m} w_i (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{W}\sum_{i=1}^{m} w_i (x_i - \bar{x})(y_i - \bar{y}) & \frac{1}{W}\sum_{i=1}^{m} w_i (y_i - \bar{y})^2 \end{bmatrix}, \tag{6.7}$$

gdje je $\bar{a} = (\bar{x}, \bar{y})^T$ centroid skupa $\mathcal{A}$

$$\bar{x} = \frac{1}{W}\sum_{i=1}^{m} w_i x_i, \quad \bar{y} = \frac{1}{W}\sum_{i=1}^{m} w_i y_i, \quad W = \sum_{i=1}^{m} w_i.$$

Kao što smo već primijetili, smjer TLS-pravca, a onda i glavnog smjera podataka (first principal component) u smjeru je svojstvenog vektora koji pripada većoj svojstvenoj vrijednosti matrice Σ. Sporedni smjer (second principal component) uzima se okomito na prvi, dakle, u smjeru svojstvenog vektora koji odgovara manjoj svojstvenoj vrijednosti iste matrice. Zbog toga ćemo glavne smjerove u skupu $\mathcal{A}$ tražiti u smjeru svojstvenih vektora matrice Σ.

Primjedba 6.1. *Kovarijacijsku matricu Σ skupa podataka $\mathcal{A}$ s centroidom $\bar{a} = \text{mean}(\mathcal{A})$ možemo zapisati pomoću Kroneckerovog produkta:*

$$\Sigma = \frac{1}{W}\sum_{i=1}^{m} w_i (\bar{a} - a^i)(\bar{a} - a^i)^T. \tag{6.8}$$

Sljedeća lema pokazuje da će kovarijacijska matrica Σ zadana s (6.7) biti pozitivno definitna ako skup podataka $\mathcal{A} \subset \mathbb{R}^2$ ne leži na pravcu kroz centroid (usporedi sa Zadatkom 6.1, str. 101).

Lema 6.2. *Neka je* $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 : i = 1, \ldots, m\}$ *skup točaka u ravnini s centroidom* $\bar{a} = (\bar{x}, \bar{y})^T$.

Tada je kovarijacijska matrica Σ *zadana s* (6.7) *simetrična pozitivno definitna matrica onda i samo onda ako su vektori* $(x_1 - \bar{x}, \ldots, x_m - \bar{x})^T$, $(y_1 - \bar{y}, \ldots, y_m - \bar{y})^T$ *linearno nezavisni.*

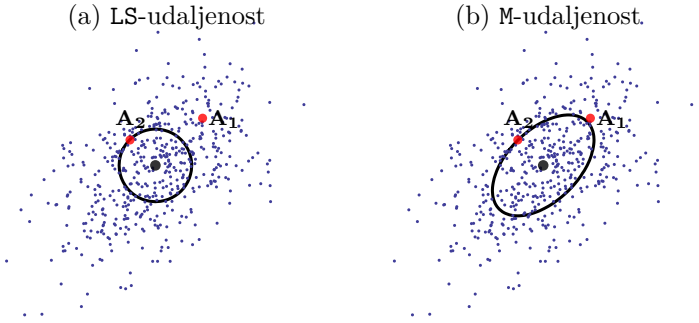
Dokaz. Simetričnost matrice Σ je očigledna, a pozitivna definitnost slijedi iz Cauchy-Schwarz-Buniakowskyjeve nejednakosti. □

Zadatak 6.3. *Neka su* $u, v \in \mathbb{R}$ *proizvoljni realni brojevi i* $a = (x_1, \ldots, x_m)^T$, $b = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$ *vektori.*

a) Ako su vektori $a, b \in \mathbb{R}^m$ *linearno nezavisni (zavisni), moraju li i vektori* $(x_1 - u, \ldots, x_m - u)^T$, $(y_1 - v, \ldots, y_m - v)^T \in \mathbb{R}^m$ *biti linearno nezavisni (zavisni) ?*

b) Ako su vektori $(x_1 - u, \ldots, x_m - u)^T$, $(y_1 - v, \ldots, y_m - v)^T \in \mathbb{R}^m$ *linearno nezavisni (zavisni), moraju li i vektori* $a, b \in \mathbb{R}^m$ *biti linearno nezavisni (zavisni) ?*

Primjer 6.3. *Svojstvene vrijednosti simetrične matrice* $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ *su* $\lambda_1 = 3$, $\lambda_2 = 1$, *a odgovarajući jedinični svojstveni vektori* $u_1 = \frac{\sqrt{2}}{2}(1, 1)^T$, $u_2 = \frac{\sqrt{2}}{2}(-1, 1)^T$. *Pomoću ove matrice u okolini ishodišta* $O = (0, 0)^T$ *generirat ćemo skup* $\mathcal{A}$ *s* $m = 500$ *slučajnih točaka naredbom (vidi Sliku 6.3)*



Slika 6.3: Skup točaka generiran kovarijacijskom matricom S

```
In[1]:= SeedRandom[23];
        m=500; O={0,0}; cov={{2,1},{1,2}};
        RandomReal[MultinormalDistribution[O, cov],m];
```

Centroid $\bar{a} = (\bar{x}, \bar{y})^T$ i kovarijacijska matrica prema (6.7) je

$$\bar{a} = \begin{bmatrix} 0.015 \\ -0.051 \end{bmatrix}, \qquad Cov = \begin{bmatrix} 1.960 & 0.975 \\ 0.975 & 1.933 \end{bmatrix}.$$

Svojstvene vrijednosti matrice Cov su $\lambda_1 = 2.921$, $\lambda_2 = 0.972$, a odgovarajući svojstveni vektori $u_1 = (-0.712, -0.702)^T$ (glavni smjer - first principal component) i $u_2 = (0.702, -0.712)^T$ (sporedni smjer - second principal component).

Izaberimo točku $A_1 = (1.3, 1.3)^T$ na glavnom smjeru. Njena LS-udaljenost do ishodišta O je $d_{LS}(A_1, O) = 3.38$, a njena M-udaljenost do ishodišta je $d_m(A_1, O) = 1.16$ (Slika 6.3a). Točku $A_2 = (-.7, .7)^T$ izaberimo na sporednom smjeru. Njena LS-udaljenost do ishodišta je $d_{LS}(A_2, O) = 0.98$, a njena M-udaljenost do ishodišta je $d_m(A_2, O) = 1.008$ (Slika 6.3b).

Kao što se može vidjeti u ravnini snabdjevenoj LS-kvazimetričkom funkcijom točka A_1 daleko je od jedinične kružnice, a točka A_2 blizu jedinične kružnice. U ravnini snabdjevenoj M-kvazimetričkom funkcijom obje točke A_1, A_2 blizu su jedinične M-kružnice.

Teorem 6.2. *Neka je $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 : i = 1, \ldots, m\}$ skup točaka u ravnini s odgovarajućim težinama $w_i > 0$, a $\Sigma \in \mathbb{R}^{2\times 2}$ simetrična pozitivno definitna matrica. Tada se centroid skupa $\mathcal{A}$ podudara s M-centroidom skupa $\mathcal{A}$.*

Dokaz. U skladu s Definicijom 2.2, M-centroid skupa $\mathcal{A}$ je vektor

$$c_m^\star = \operatorname*{argmin}_{x\in\mathbb{R}^2} \sum_{i=1}^m w_i d_m(x, a^i) = \operatorname*{argmin}_{x\in\mathbb{R}^2} \sum_{i=1}^m w_i (x - a^i)^T \Sigma^{-1} (x - a^i),$$

na kojemu se postiže minimum funkcije

$$F_m(x) = \sum_{i=1}^m w_i d_m(x, a^i) = \sum_{i=1}^m w_i (x - a^i)^T \Sigma^{-1} (x - a^i).$$

Stacionarna točka $c_m^\star$ funkcije F_m određena je s:

$$F_m'(c_m^\star) = 2 \sum_{i=1}^m w_i \Sigma^{-1} (c_m^\star - a^i) = 0.$$

Kako je Hessian $H_{F_m} = 2W\Sigma^{-1} > 0$ funkcije F_m pozitivno definitan, na vektoru

$$c_m^\star := \tfrac{1}{W} \sum_{i=1}^m w_i a^i, \quad W = \sum_{i=1}^m w_i,$$

postiže se globalni minimum funkcije F_m. Dakle, M-centroid skupa $\mathcal{A}$ podudara se s običnim centroidom skupa $\mathcal{A}$. □

6.3.1 Mahalanobis udaljenost inducirana skupom točaka iz $\mathbb{R}^n$

Općenito, neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i)^T \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup podataka s težinama $w_1, \ldots, w_m > 0$. Tada vrijedi:

(i) Centroid skupa $\mathcal{A}$ zadan je s:

$$\bar{a} = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{m} w_i d_m(x, a^i) = \tfrac{1}{W} \sum_{i=1}^{m} w_i a^i, \qquad W = \sum_{i=1}^{m} w_i.$$

(ii) Neka je

$$B = \begin{bmatrix} \bar{a}_1 - a_1^1 & \cdots & \bar{a}_n - a_n^1 \\ \vdots & \ddots & \vdots \\ \bar{a}_1 - a_1^m & \cdots & \bar{a}_n - a_n^m \end{bmatrix}, \qquad D = \operatorname{diag}(w_1, \ldots, w_m).$$

Tada je kovarijacijska matrica $\Sigma = \frac{1}{W} B^T D B$ zadana s:

$$\Sigma = \tfrac{1}{W} \begin{bmatrix} \sum w_i(\bar{a}_1 - a_1^i)^2 & \cdots & \sum w_i(\bar{a}_1 - a_1^1)(\bar{a}_n - a_n^i) \\ \sum w_i(\bar{a}_2 - a_2^i)(\bar{a}_1 - a_1^i) & \cdots & \sum w_i(\bar{a}_2 - a_2^1)(\bar{a}_n - a_n^i) \\ \vdots & \ddots & \vdots \\ \sum w_i(\bar{a}_n - a_n^i)(\bar{a}_1 - a_1^i) & \cdots & \sum w_i(\bar{a}_n - a_n^i)^2 \end{bmatrix},$$

što možemo pisati pomoću Kroneckerovog produkta (vidi također Primjedbu 6.1, str. 105):

$$\Sigma = \tfrac{1}{W} \sum_{i=1}^{m} w_i (\bar{a} - a^i)(\bar{a} - a^i)^T.$$

(iii) Mahalanobis kvazimetrička funkcija definira se s: $d_m : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$,

$$d_m(x, y; \Sigma) = (x - y)^T \Sigma^{-1} (x - y).$$

6.4 Metode za traženje optimalne particije s elipsoidnim klasterima

Promatramo skup $\mathcal{A} \subset \Delta \subset \mathbb{R}^n$ s $m = |\mathcal{A}|$ elemenata, koji treba grupirati u k klastera. Za traženje optimalnih particija skupa $\mathcal{A}$ s elipsoidnim

klasterima konstruirat ćemo odgovarajuće modifikacije k-means algoritma i Inkrementalnog algoritma.

Neka je $\Pi = \{\pi_1, \ldots, \pi_k\}$ particija skupa $\mathcal{A}$ gdje je za svaki klaster π_j određen njegov centroid $c_j = \text{mean}(\pi_j)$ i pripadna kovarijacijska matrica $\Sigma_j = \frac{1}{|\pi_j|} \sum_{a \in \pi_j} (c_j - a)(c_j - a)^T$. Ako se svojstvene vrijednosti kovarijacijske matrice međusobno bitnije razlikuju, to je svakako razlog da particije skupa $\mathcal{A}$ potražimo u formi elipsoidnih klastera.

U cilju osiguranja monotonog pada vrijednosti funkcije cilja kod implementacije Mahalanobis k-means Algoritma 6.1 definirat ćemo (vidi [62, 103]) *normaliziranu Mahalanobis kvazimetričku funkciju* $d_M \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$:

$$d_M(u, v; \Sigma) := \sqrt[n]{\det \Sigma}\,(u - v)^T \Sigma^{-1} (u - v) = \|u - v\|_\Sigma^2. \tag{6.9}$$

Definicija 6.1. *Neka je $\Sigma \in \mathbb{R}^{n \times n}$ pozitivno definitna simetrična matrica i $u \in \mathbb{R}^n$.*

- *Skup $M(u; \Sigma) = \{x \in \mathbb{R}^n \colon (x - u)^T \Sigma^{-1} (x - u) = 1\}$ zovemo Mahalanobis kružnica (M-kružnica),*
- *Skup $M_N(u; \Sigma) = \{x \in \mathbb{R}^n \colon \sqrt[n]{\det \Sigma}(x - u)^T \Sigma^{-1} (x - u) = 1\}$ zovemo normalizirana Mahalanobis kružnica (M_N-kružnica).*

Lema 6.3. *Neka je $\Sigma \in \mathbb{R}^{n \times n}$ pozitivno definitna simetrična matrica sa svojstvenim vrijednostima $\lambda_1 \geq \cdots \geq \lambda_n > 0$. Tada:*

(i) M-kružnica $M(u; \Sigma)$ je hiperelipsoid s centrom u točki $u \in \mathbb{R}^n$ i glavnim poluosima duljine $\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_n} > 0$ u smjeru svojstvenih vektora matrice Σ,

(ii) Normalizirana M-kružnica $M_N(u; \Sigma)$ je hiperelipsoid s centrom u točki $u \in \mathbb{R}^n$ i glavnim poluosima duljine $\frac{\sqrt{\lambda_1}}{\sqrt[n]{\det \Sigma}} \geq \cdots \geq \frac{\sqrt{\lambda_n}}{\sqrt[n]{\det \Sigma}} > 0$ u smjeru svojstvenih vektora matrice Σ. Normalizirajući faktor $\sqrt[n]{\det \Sigma}$ osigurava konstantnost volumena hiperelipsoida: $\prod_{i=1}^{n} \frac{\sqrt{\lambda_i}}{\sqrt[n]{\det \Sigma}} = 1$.

Dokaz. Dokaz tvrdnje (i) slijedi iz prethodnih razmatranja (vidi također [57]).

U svrhu dokaza tvrdnje (ii) primijetite da jednadžba M_N-kružnice može biti zapisana kao

$$(x - u)^T \left(\frac{1}{\sqrt[n]{\det \Sigma}} \Sigma\right)^{-1} (x - u) = 1,$$

iz čega korištenjem (i) za svojstvene vrijednosti $\frac{\lambda_i}{\sqrt[n]{\det \Sigma}}$ dobivamo traženu tvrdnju (ii). □

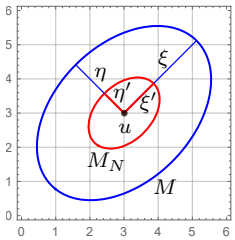
Primjer 6.4. *Neka je*

$$\Sigma = \begin{bmatrix} \frac{13}{2} & \frac{5}{2} \\ \frac{5}{2} & \frac{11}{2} \end{bmatrix} = Q \begin{bmatrix} 3^2 & 0 \\ 0 & 2^2 \end{bmatrix} Q^T, \quad \textit{gdje je} \quad Q = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

dekompozicija na svojstvene vrijednosti pozitivno definitne simetrične matrice $\Sigma \in \mathbb{R}^{2\times 2}$.

M-kružnica $M((3,3)^T;\Sigma)$ *elipsa je s centrom u točki* $u = (3,3)^T$ *i poluosima duljine* $\xi = 3$ *i* $\eta = 2$ *u smjeru svojstvenih vektora (stupci ortogonalne matrice* Q*) (plava elipsa na Slici 6.4).*

Budući da je $\sqrt[4]{\det S} = \sqrt{6}$, *normalizirana M-kružnica* $M_N((3,3)^T;S)$ *elipsa je s centrom u točki* $u = (3,3)^T$ *i poluosima duljine* $\xi' = \frac{3}{\sqrt{6}}$ *i* $\eta' = \frac{2}{\sqrt{6}}$ *u smjeru svojstvenih vektora (stupci ortogonalne matrice* Q*) (crvene elipse na Slici 6.4). Primijetite da je* $\xi' \cdot \eta' = 1$.



Slika 6.4: M-kružnica i normalizirana M_N-kružnica

Funkciju cilja tada također možemo definirati na jedan od sljedeća dva načina:

$$\mathcal{F}_M(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} d_M^{(j)}(c_j, a, \Sigma_j); \tag{6.10}$$

$$F_M(c_1,\ldots,c_k) = \sum_{i=1}^{m} \min_{1\le j\le k} d_M^{(j)}(c_j, a^i, \Sigma_j). \tag{6.11}$$

Ako je kovarijacijska matrica $\Sigma_j \approx I$ (sve svojstvene vrijednosti su ≈ 1), onda klaster π_j ima približno sferičan oblik, a ako je $\lambda_{max}^{(j)} \gg \lambda_{min}^{(j)}$, onda klaster π_j ima vrlo izdužen elipsoidni oblik u smjeru svojstvenog vektora

koji pripada najvećoj svojstvenoj vrijednosti. Kako treba postupiti kada je $\text{cond}\,(S_j) = \lambda^{(j)}_{max}/\lambda^{(j)}_{min}$ veliki broj (primjerice, 10^{20}) može se vidjeti u [4].

Za traženje optimalne k-particije s elipsoidnim klasterima pokazat ćemo Mahalanobis k-means algoritam i Mahalanobis inkrementalni algoritam (vidi [62]).

6.4.1 Mahalanobis k-means algoritam

Mahalanobis k-means algoritam započet će inicijalizacijom u `Koraku~0`, a nakon toga će se sukcesivno izmjenjivati `Korak~A` i `Korak~B`.

Algoritam 6.1. [Mahalanobis k-means algoritam]

Korak 0: *Za dani skup* $z_1, \ldots, z_k \in \mathbb{R}^n$ *međusobno različitih točaka za svaki* $j = 1, \ldots, k$ *odrediti:*
Klastere: π_j; *[princip minimalnih udaljenosti];*
Centroide: $c_j := \text{mean}[\pi_j]$;
Kovarijacijske matrice: $\Sigma_j := \frac{1}{m} \sum\limits_{a \in \pi_j} (c_j - a)(c_j - a)^T$;
M-kvazimetričke funkcije: $d^{(j)}_M(x, y, \Sigma_j) := \sqrt[n]{\det \Sigma_j}(x - y)^T \Sigma_j^{-1}(x - y)$.

Korak A: *Pridruživanje.* *Za dane centroide* $c_j \in \mathbb{R}^n$, *kovarijacijske matrice* Σ_j *i M-kvazimetričke funkcije:* $d^{(j)}_M(x, y, \Sigma_j)$ *principom minimalnih udaljenosti odrediti nove klastere*

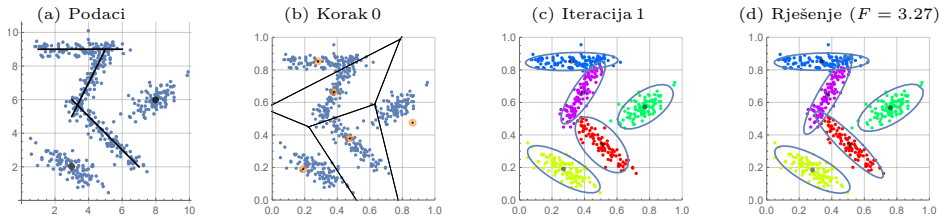
$$\pi_j = \{a^i \in \mathcal{A} \colon d^{(j)}_M(c_j, a^i, S_j) \leq d^{(s)}_M(c_s, a^i, S_s),\ \forall s \in J\},\ j = 1, \ldots, k;$$

Korak B: *Korekcija.* *Za danu particiju* $\Pi = \{\pi_1, \ldots, \pi_k\}$ *odrediti:*
Centroide: $c_j := \text{mean}[\pi_j]$;
Kovarijacijske matrice: $\Sigma_j := \frac{1}{m} \sum\limits_{a \in \pi_j} (c_j - a)(c_j - a)^T$;
M-kvazimetričke funkcije: $d^{(j)}_M(x, y, \Sigma_j) := \sqrt[n]{\det \Sigma_j}(x - y)^T \Sigma_j^{-1}(x - y)$.

Budući da niz funkcijskih vrijednosti dobiven u svakoj iteraciji monotono opada (vidi [103]), kriterij zaustavljanja Algoritma 6.1 može biti kao u standardnom slučaju (4.9). Analogno, kao u Teoremu 4.1, str.71 može se pokazati da Algoritma 6.1 daje `LOPart`.

Primjer 6.5. *Sintetički skup podataka konstruirat ćemo slično kao u [106]. Izaberimo dvije točke* $C_1 = (3, 2)^T$, $C_2 = (8, 6)^T \in \Delta = [0, 10]^2$. *U okolini točke* C_1, *odnosno* C_2, *generiramo po* 100 *slučajnih točaka iz binormalne distibucije s očekivanjem* $0 \in \mathbb{R}^2$ *i kovarijacijskom matricom* $\Sigma_1 = \frac{1}{2}\begin{bmatrix} 2 & -1 \\ -1 & .9 \end{bmatrix}$, *odnosno kovarijacijskom matricom* $\Sigma_2 = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$. *Osim toga izaberimo tri segmenta u ravnini* $\ell_1 = [(1, 9)^T, (6, 9)^T]$, $\ell_2 = [(3, 5)^T, (5, 9)^T]$,

$\ell_3 = [(3,6)^T, (7,2)^T]$ i u okolini svakog od njih generiramo po 100 *normalno distribuiranih slučajnih točaka (vidi Sliku 6.5a). Nakon transformacije podataka na kvadrat* $[0,1]^2$ *dobivamo particiju* $\Pi^{(0)} = \{\pi_1, \ldots, \pi_5\}$ *i skup* $\mathcal{A} = \cup \pi_j \subset [0,1]^2$ *s* $m = 500$ *točaka.*



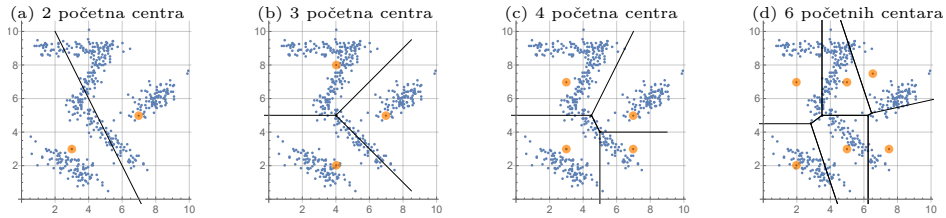
Slika 6.5: Mahalanobis k-means algoritam

Korištenjem Mahalanobis k-means Algoritma 6.1 potražit ćemo optimalnu 5-particiju. Najprije sukladno `Koraku-0` izaberemo 5 inicijalnih centara (narančaste točke na Slici 6.5b) i principom minimalnih udaljenosti odredimo inicijalne klastere $\pi_1, \ldots, \pi_5$ (Voronoijev dijagram na Slici 6.5b), njihove centroide, pripadne kovarijacijske matrice i M-kvazimetričke funkcije. Prva iteracija Algoritma 6.1 prikazana je na Slici 6.5c, a M-optimalna 5-particija $\Pi^\star$ dobivena nakon 4 iteracije na Slici 6.5d. Elipse na slikama obuhvaćaju 95% točaka pojedinog klastera.

Rand indeks (0.863), Jaccard indeks (0.803) kao i matrica prijelaza

$$S(\Pi^0, \Pi^\star) = \begin{bmatrix} 100 & 0 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 & 0 \\ 0 & 0 & 97 & 3 & 0 \\ 0 & 0 & 12 & 87 & 1 \\ 0 & 1 & 0 & 13 & 86 \end{bmatrix}$$

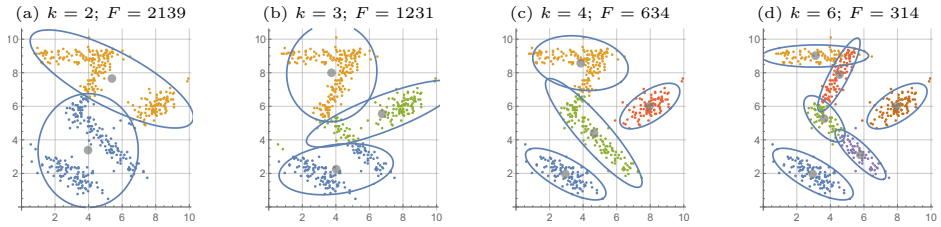
definirani u točki 5.2.1, str.95 pokazuju visoki stupanj podudaranja originalne i izračunate particije jer je njezina struktura skoro dijagonalna.



Slika 6.6: Izbor početnih centara

Primjer 6.6. *Mahalanobis k-means Algoritam 6.1 mogli smo pokrenuti i s manje ili više početnih centara. Na Slici 6.6 prikazani su Voronoijevi dijagrami skupa $\mathcal{A}$ iz Primjera 6.5 za dva, tri, četiri i šest izabranih početnih*

centara, a na Slici 6.7 za te centre prikazane su dobivene Mahalanobis k-`LOPart` *za* $k = 2, 3, 4, 6$. *Elipse na slikama obuhvaćaju* 95% *točaka pojedinog klastera.*



Slika 6.7: Rezultati Mahalanobis k-means algoritma

Može se postaviti pitanje nalazi li se između ovih Mahalanobis k-`LOPart` također i Mahalanobis `MAPart` i koja je to particija.

Primjedba 6.2. *Može se pokazati da se Mahalanobis k-means Algoritam 6.1 značajno podudara s* Generalized Mixture Decomposition Algorithmic Scheme *[106] kao specijalan slučaj* Expectation Maximization *algoritma (vidi [118, str. 31]), ali je učinkovitost Algoritma 6.1, mjerena potrebnim* `CPU`*-vremenom, značajno bolja.*

Zadatak 6.4. *Konstruirajte Algoritam 6.1 za podatake s težinama* $w_i > 0$.

6.4.2 Mahalanobis inkrementalni algoritam

Drugi algoritam za traženje Mahalanobis k-`LOPart` koji ćemo navesti je `Mahalanobis inkrementalni algoritam`. Neka je $\mathcal{A} \subset \mathbb{R}^n$. Inkrementalni algoritam započinje izborom početnog centra $c_1 \in \mathbb{R}^n$. Primjerice, to može biti centroid skupa $\mathcal{A}$. Sljedeći centar c_2 dobit ćemo kao rješenje `GOP` za funkciju $\Phi\colon \mathbb{R}^n \to \mathbb{R}$:

$$c_2 \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \Phi(x), \quad \Phi(x) := \sum_{i=1}^{m} \min\{\|c_1 - a^i\|^2, \|x - a^i\|^2\}.$$

Nakon toga, na centre c_1, c_2 primijenimo Mahalanobis k-means Algoritam 6.1 i time dobivamo centre $c_1^\star, c_2^\star$ 2-`LOPart` $\Pi^{(2)}$.

Općenito, poznavajući k centara $c_1^\star, \ldots, c_k^\star$, sljedeći centar c_{k+1} potražit ćemo kao rješenje `GOP` za funkciju $\Phi\colon \mathbb{R}^n \to \mathbb{R}$:

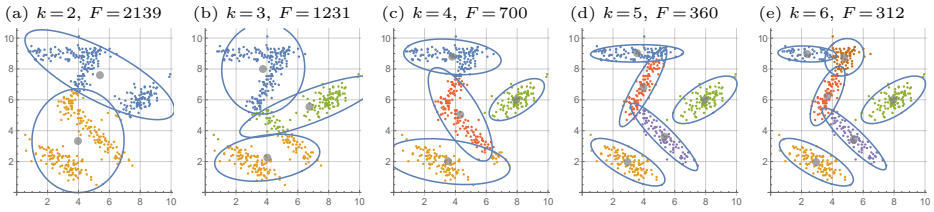
$$c_{k+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \Phi(x), \quad \Phi(x) := \sum_{i=1}^{m} \min\{\delta_k^i, \|x - a^i\|^2\}, \tag{6.12}$$

gdje je $\delta_k^i = \min_{1\le s\le k} \|c_s^\star - a^i\|^2$. Ovdje možemo primijeniti nekoliko (primjerice 10) iteracija algoritma `DIRECT`.

Nakon toga, primjenom Mahalanobis k-means Algoritma 6.1, dobivamo centre $(c_1^\star, \ldots, c_{k+1}^\star)$ $(k+1)$-`LOPart` $\Pi^{(k+1)}$.

Primjedba 6.3. *Primijetite da smo Algoritam 2 slično mogli pokrenuti i s više od jednog početnog centra.*

Primjer 6.7. *Mahalanobis inkrementalni algoritam pokrenut ćemo na skupu podataka $\mathcal{A}$ iz Primjera 6.5 počevši s centroidom skupa $\mathcal{A}$.*



Slika 6.8: Inkrementalni Mahalanobis algoritam

Na Slici 6.8 prikazani su dobiveni Mahalanobis k-`LOPart` za $k = 2, 3, 4, 5, 6$ s odgovarajućim vrijednostima funkcije cilja. Elipse na slikama obuhvaćaju 95% točaka pojedinog klastera.

Primijetite da za svaki k nisu dobivene ni iste particije niti iste vrijednosti funkcije cilja kao kod primjene k-means algoritma u Primjeru 6.6. Također se može postaviti pitanje nalazi li se između ovih Mahalanobis k-particija i Mahalanobis `MAPart` i koja je to.

6.5 Izbor particije s najprikladnijim brojem elipsoidnih klastera

Slično kao i u slučaju sferičnih klastera (vidi točku 5, str.87) možemo postaviti sljedeće pitanje:

> U koliko bi elipsoidnih klastera bilo najprihvatljivije grupirati promatrani skup podataka $\mathcal{A}$, odnosno kako za promatrani skup podataka $\mathcal{A}$ odabrati Mahalanobis particiju s najprikladnijim brojem klastera (Mahalanobis `MAPart`)?

U tu svrhu definirat ćemo poopćene indekse navedene u točki 5, str.87 (vidi [62]). Neka je $\Pi = \{\pi_1, \ldots, \pi_k\}$ particija skupa $\mathcal{A} \subset \Delta$ s $m = |\mathcal{A}|$ elemenata, gdje je za svaki klaster π_j određen njegov centroid $c_j = \text{mean}(\pi_j)$, pri-

padna kovarijacijska matrica $\Sigma_j = \frac{1}{|\pi_j|} \sum_{a \in \pi_j} (c_j - a)(c_j - a)^T$ i M-kvazimetrička funkcija $d_M^{(j)}$ zadana s (6.9).

- Mahalanobis Calinski-Harabasz indeks definirat ćemo kao:

$$\texttt{MCH[k]} = \frac{\frac{1}{k-1} \sum_{j=1}^{k} d_M^{(j)}(c, c_j, \Sigma_j)}{\frac{1}{m-k} \sum_{j=1}^{k} \sum_{a \in \pi_j} d_M^{(j)}(c_j, a, \Sigma_j)}, \qquad c = \text{mean}[\mathcal{A}]; \tag{6.13}$$

- Mahalanobis Davies-Bouldin indeks definirat ćemo kao:

$$\texttt{MDB[k]} = \frac{1}{k} \sum_{j=1}^{k} \max_{s \neq j} \frac{V(\pi_j) + V(\pi_s)}{d_M^{(j)}(c_j, c_s, \Sigma_j + \Sigma_s)}, \; V(\pi_j) = \sum_{a_s \in \pi_j} d_M^{(j)}(c_j, a^s, \Sigma_j); \tag{6.14}$$

- Mahalanobis pojednostavljeni kriterij širine silhouette (MSSC) definirat ćemo tako da najprije za svaki $a^i \in \mathcal{A} \cap \pi_r$ izračunamo brojeve:

$$\alpha_{ir} = d_M^{(r)}(c_r^\star, a^i, S_r^\star), \quad \beta_{ir} = \min_{s \neq r} d_M^{(s)}(c_s^\star, a^i, S_s^\star), \quad s_i = \frac{\beta_{ir} - \alpha_{ir}}{\max\{\alpha_{ir}, \beta_{ir}\}},$$

  a nakon toga MSSC-indeks definiramo kao prosjek:

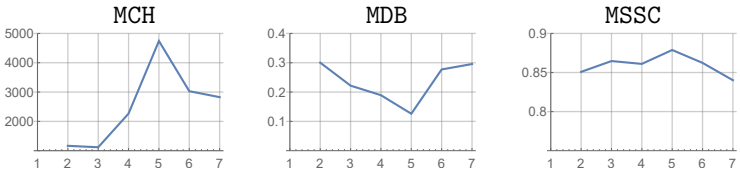
$$\texttt{MSSC[k]} = \frac{1}{m} \sum_{i=1}^{m} s_i. \tag{6.15}$$

Particiju s najvećim MCH-indeksom, odnosno particiju s najmanjim MDB-indeksom, odnosno particiju s najvećim MSSC-indeksom zvat ćemo *Mahalanobis particija s najprihvatljivijim brojem klastera* (Mahalanobis MAPart).

Indeks	$\Pi^{(2)}$	$\Pi^{(3)}$	$\Pi^{(4)}$	$\Pi^{(5)}$	$\Pi^{(6)}$
MCH	1166	1117	2266	**4742**	3029
MDB	0.300	0.222	0.189	**0.126**	0.277
MSSC	0.851	0.865	0.861	**0.879**	0.862

Tablica 6.1: Vrijednosti Mahalanobis indeksa na particijama iz Primjera 6.7

Primjer 6.8. *Za particije iz Primjera 6.7 s podacima definiranim u Primjeru 6.5 u Tablici 6.1 navedene su vrijednosti* MCH, MDB *i* MSSC*-indeksa. Svi indeksi ukazuju na* 5*-particiju kao najprihvatljiviju particiju (Mahalanobis* MAPart*).*



Slika 6.9: Sva tri Mahalanobis indeksa ukazuju na 5-particiju iz Primjera 6.7 kao Mahalanobis `MAPart`

Na Slici 6.9 prikazani su odgovarajući grafovi korištenih indeksa iz kojih se također lijepo vidi da svi indeksi ukazuju na 5-particiju kao Mahalanobis `MAPart`.

Poglavlje 7

Fuzzy grupiranje podataka

Ako se može očekivati da neki od elemenata skupa $\mathcal{A}$ prirodno mogu djelomično pripadati u dva ili više klastera, treba primijeniti fuzzy grupiranje (*neizrazito grupiranje*). U znanstvenoj literaturi fuzzy grupiranje smatra se jednim od oblika tzv. mekog grupiranja podataka (engl.: *soft clustering*). Fuzzy grupiranje primjenjuje se kod analize slike i signala, medicinske dijagnostike, tomografije, astronomije, kod prepoznavanja govora, u znanosti o okolišu, itd. (vidi primjerice [14, 99, 106]). Kao što smo već spomenuli u t. 3, str. 40, ovakav problem također se pojavljuje i prilikom definiranja izbornih jedinica u nekoj zemlji zbog potrebe smještanja dijela glasača nekog područja ili grada u dvije ili više izbornih jedinica (u Republici Hrvatskoj to je slučaj s gradom Zagrebom).

Ako pretpostavimo da elementi skupa podataka $\mathcal{A}$ mogu djelomično pripadati različitim klasterima, onda elementi u_{ij} matrice pripadnosti U u funkciji (4.7) trebaju biti $u_{ij} \in [0, 1]$. Prema [13, 14, 106], stupanj pripadnosti elementa a^i u klasteru π_j određen je funkcijom pripadnosti $q \mapsto u_{ij}^q$, $q > 1$. Veličina q naziva se *fuzzifier*, a funkcija cilja (4.7) uz poznavanje kvazimetričke funkcije $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ postaje:

$$\Phi(c, U) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij}^q(c)\, d(c_j, a^i), \tag{7.1}$$

uz uvjete:

$$\sum_{j=1}^{k} u_{ij} = 1, \quad i = 1, \ldots, m, \tag{7.2}$$

$$1 \leq \sum_{i=1}^{m} u_{ij} \leq m, \quad j = 1, \ldots, k. \tag{7.3}$$

Primijetite da se praktično može dogoditi da svi elementi skupa $\mathcal{A}$ pripadnu jednom klasteru, ali s manjim ili većim stupnjem pripadnosti. To znači da granice klastera nisu jasno određene pa se u literaturi ovaj pristup često naziva *meko grupiranje* (soft clustering).

7.1 Fuzzy c-means algoritam

Algoritam za traženje fuzzy-`LOPart` analogan k-means algoritmu koji smo analizirali u točki 4.2, str. 68 naziva se *Fuzzy c-means algoritam*. Kao što smo u točki 3.1.2, str. 39 pokazali, za konstrukciju algoritma potrebno je sukcesivno provoditi sljedeća dva koraka.

Korak A: Za fiksni $c^{(0)}$ principom minimalnih udaljenosti treba odrediti klastere $\pi_1, \ldots, \pi_k$ i na taj način definirati matricu pripadnosti $U^{(1)}$.

Korak B: Za fiksnu matricu pripadnosti $U^{(1)}$ treba pronaći optimalne centre $c_1^{(1)}, \ldots, c_k^{(1)}$ rješavanjem k optimizacijskih problema:

$$c_1^{(1)} \in \operatorname*{argmin}_{c_1 \in \mathbb{R}^n} \sum_{i=1}^{m} u_{i1}^{(1)} d(c_1, a^i), \ldots, c_k^{(1)} \in \operatorname*{argmin}_{c_k \in \mathbb{R}^n} \sum_{i=1}^{m} u_{ik}^{(1)} d(c_k, a^i),$$

i izračunati vrijednost funkcije cilja $\Phi(c^{(1)}, U^{(1)})$ prema (7.1).

7.1.1 Određivanja matrice pripadnosti

U cilju određivanja matrice pripadnosti u `Koraku A FCM`-algoritma treba odrediti funkcije pripadnosti u_{ij}. U tu svrhu definirajmo Lagrangeovu funkciju [106]

$$\mathcal{J}(c, U, \lambda) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij}^q(c) d(c_j, a^i) + \sum_{i=1}^{m} \lambda_i \Big(\sum_{j=1}^{k} u_{ij}(c) - 1 \Big). \tag{7.4}$$

Ako parcijalnu derivaciju funkcije $\mathcal{J}$ po u_{rs}

$$\frac{\partial \mathcal{J}(c, U, \lambda)}{\partial u_{rs}} = q u_{rs}^{q-1} d(c_s, a^r) - \lambda_r \tag{7.5}$$

izjednačimo s nulom, dobivamo:

$$u_{rs} = \Big(\frac{\lambda_r}{q d(c_s, a^r)} \Big)^{\frac{1}{q-1}}, \quad s = 1, \ldots, k. \tag{7.6}$$

Uvrštavajući (7.6) u (7.2) dobivamo:

$$\sum_{j=1}^{k} \Big(\frac{\lambda_r}{q d(c_j, a^r)} \Big)^{\frac{1}{q-1}} = 1,$$

iz čega dobivamo parametar λ_r:

$$\lambda_r = \frac{q}{\Big(\sum\limits_{j=1}^{k} \big(\frac{1}{d(c_j,a^r)} \big)^{\frac{1}{q-1}} \Big)^{q-1}}, \tag{7.7}$$

a uvrštavajući (7.7) u (7.6) konačno dobivamo

$$u_{rs} = \frac{1}{\sum\limits_{j=1}^{k} \Big(\frac{d(c_s,a^r)}{d(c_j,a^r)} \Big)^{1/(q-1)}}. \tag{7.8}$$

Primjedba 7.1. *Preciznije, funkcija pripadnosti* $u_{ij} \colon \mathcal{A} \to [0,1]$ *definira se kao (vidi [42]):*

$$u_{ij}(c) = \begin{cases} \dfrac{1}{\sum\limits_{s=1}^{k} \left(\frac{d(c_j,a^i)}{d(c_s,a^i)} \right)^{1/(q-1)}} = \dfrac{d(c_j,a^i)^{1/(1-q)}}{\sum\limits_{s=1}^{k} d(c_s,a^i)^{1/(1-q)}}, & \text{ako } I_i = \emptyset \\ \dfrac{1}{|I_i|}, & \text{ako } I_i \neq \emptyset \ \& \ j \in I_i \\ 0, & \text{ako } I_i \neq \emptyset \ \& \ j \notin I_i \end{cases}, \tag{7.9}$$

gdje je $I_i = \{j \in \{1, \ldots, k\} \colon c_j = a^i\}$.

Traženje optimalne vrijednosti parametra q *za koji funkcija cilja* (7.1) *uz uvjet* (7.2) *postiže svoj globalni minimum složeni je problem globalne optimizacije (vidi primjerice [83]). Zbog toga se u primijenjenim istraživanjima najčešće koristi* $q \in [1.5, 2.5]$.

7.1.2 Određivanja centara klastera

Poznavajući matricu pripadnosti, centar c_j klastera π_j dobit ćemo tako da parcijalnu derivaciju funkcije $\mathcal{J}$ po c_j izjednačimo s nulom:

$$\frac{\partial \mathcal{J}(c, U, \lambda)}{\partial c_j} = \sum_{i=1}^{m} u_{ij}^q(c) \frac{\partial d(c_j, a^i)}{\partial c_j} = 0. \tag{7.10}$$

Specijalno, za LS-kvazimetričku funkciju $d(c_j, a^i) := (c_j - a^i)^T (c_j - a^i)$, (7.10) postaje:

$$2 \sum_{i=1}^{m} u_{ij}^q(c)(c_j - a^i) = 0,$$

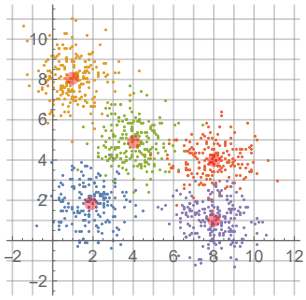
iz čega dobivamo centar c_j klastera π_j

$$c_j = \frac{1}{\sum_{i=1}^{m} u_{ij}^q(c)} \sum_{i=1}^{m} u_{ij}^q(c) a^i. \tag{7.11}$$

Slično, u slučaju ℓ_1-kvazimetričke funkcije $d(c_j, a^i) := |c_j - a^i|$ odgovarajući centar težinski je medijan [77]

$$c_j = \operatorname*{med}_{i=1,\ldots,m} (u_{ij}^q(c), a^i). \tag{7.12}$$

Primjer 7.1. *Kao u [99], zadane su točke* $(2,2)^T, (1,8)^T, (4,5)^T, (8,4)^T$, $(8,1)^T \in [0,10]^2$. *U okolini svake točke generirano je po* 100 *točaka iz Gaussove binormalne distribucije. Na taj način određena je originalna* 5*-particija* $\Pi = \{\pi_1, \ldots, \pi_5\}$ *i skup* $\mathcal{A} = \cup \pi_j$ *s* 500 *točaka.*



Slika 7.1: Originalna 5-particija Π

Bez obzira na izbor 5 različitih početnih centara iz $[0,10]^2$ `FCM`-algoritam pronalazi optimalnu 5 particiju $\Pi^\star$ vrlo sličnu originalnoj, što se vidi po postotnoj strukturi matrice prijelaza:

$$S(\Pi, \Pi^\star)(\%) = \begin{bmatrix} 79.9 & 4.4 & 8.1 & 3.5 & 4.2 \\ 3.7 & 81.5 & 10.6 & 2.5 & 1.7 \\ 9.7 & 6.7 & 69.9 & 9.1 & 4.5 \\ 2.7 & 1.6 & 8.6 & 74.3 & 12.8 \\ 3.2 & 1.3 & 5.6 & 12.2 & 77.7 \end{bmatrix}.$$

7.2 Gustafson-Kessel fuzzy c-means algoritam

Za traženje fuzzy optimalne particije s unaprijed poznatim brojem klastera elipsoidnog oblika konstruiran je dobro poznati *Gustafson-Kessel fuzzy c-means* (`GK`) algoritam (vidi [40, 106]).

Slično kao u slučaju običnog `FCM`-algoritma opisanog u točki 7.1 i Gustafson-Kessel fuzzy c-means algoritam provodi se sukcesivnom primjenom spomenuta dva koraka.

U `Koraku A`, uz poznavanje međusobno različitih točaka $c_1 \ldots, c_k \in \mathbb{R}^n$, korištenjem (7.8), odnosno (7.9), treba odrediti matricu pripadnosti $U \in [0,1]^{k\times m}$ sa svojstvima (7.2)-(7.3), a nakon toga kovarijacijske matrice:

$$\Sigma_j = \frac{1}{\sum\limits_{i=1}^{m} u_{ij}^q(c)} \sum_{i=1}^{m} u_{ij}^q(c)(c_j - a^i)(c_j - a^i)^T, \quad j = 1, \ldots, k, \tag{7.13}$$

i Mahalanobis kvazimetričke funkcije:

$$d_M^{(j)}(x, y; \Sigma_j) := \sqrt[n]{\det \Sigma_j}\,(x-y)^T \Sigma_j^{-1}(x-y), \quad j = 1, \ldots, k. \tag{7.14}$$

U `Koraku B`, uz poznavanje matrice pripadnosti $U \in [0,1]^{k\times m}$ sa svojstvima (7.2)-(7.3), korištenjem (7.11) treba definirati centre $c_1, \ldots, c_k \in \mathbb{R}^n$.

`GK`-algoritam započinje izborom početne aproksimacije $c_1^{(0)}, \ldots, c_k^{(0)} \in \mathbb{R}^n$, a nakon toga, sljedeća dva koraka sukcesivno se izmjenjuju[1].

Korak A Za dani skup međusobno različitih točaka $c_1 \ldots, c_k \in \mathbb{R}^n$ treba odrediti odgovarajuću matricu pripadnosti $U \in [0,1]^{m\times k}$, kovarijacijske matrice Σ_j i Mahalanobis kvazimetričke funkcije $d_M^{(j)}(x, y; \Sigma_j)$:

$$u_{ij} = \Big(\sum_{s=1}^{k} \Big(\tfrac{d_M(c_j, a^i; S_j)}{d_M(c_s, a^i; S_s)} \Big)^{1/(q-1)} \Big)^{-1},$$

$$\Sigma_j = \big(\sum_{i=1}^{m} u_{ij}^q(c) \big)^{-1} \sum_{i=1}^{m} u_{ij}^q(c)(c_j - a^i)(c_j - a^i)^T, \quad j = 1, \ldots, k;$$

$$d_M^{(j)}(x, y; \Sigma_j) = \sqrt[n]{\det \Sigma_j}\,(x-y)^T \Sigma_j^{-1}(x-y), \quad j = 1, \ldots, k.$$

Korak B Za danu matricu pripadnosti $U \in [0,1]^{m\times k}$ treba definirati odgovarajuće centre $c_1, \ldots, c_k \in \mathbb{R}^n$

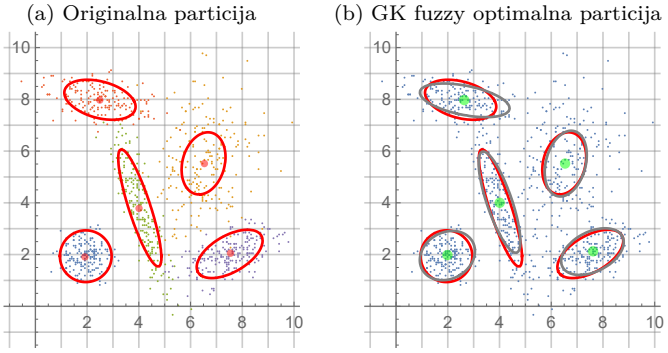
$$c_j = \frac{1}{\sum\limits_{i=1}^{m} u_{ij}^q(c)} \sum_{i=1}^{m} u_{ij}^q(c) a^i, \quad j = 1, \ldots, k.$$

Alogoritam se zaustavlja kada je $\|U^{(j)} - U^{(j-1)}\| < \epsilon$ za neki unaprijed zadani maleni $\epsilon > 0$.

[1] Gustafson-Kessel c-means algoritam i odgovarajući MATLAB kod dostupni su u [4].

Primjer 7.2. *Skup podataka $\mathcal{A} \subset [0,10]^2$ definiran je na sljedeći način. Najprije odaberimo pet različitih točaka C_j, pet parova glavnih poluosi $\{\xi_j, \eta_j\}$ i pet kutova rotacije ϑ_j. Na osnovi ovih podataka definirajmo kovarijacijske matrice Σ_j kao u Primjeru 6.4, str. 110. Nadalje, svakoj točki C_j pridružimo slučajni cijeli broj $m_j \in [160, 240]$. Skup podataka $\mathcal{A}$ tada generiramo korištenjem pet Gaussvih multinormalnih slučajnih generatora $\mathcal{N}(C_j, S_j)$.*

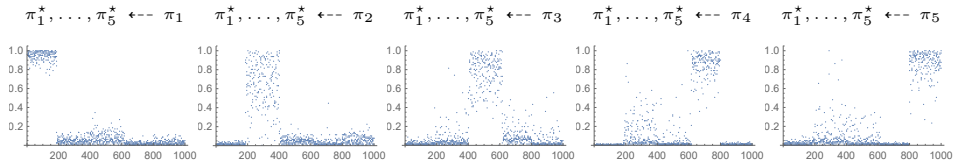
Na taj način, dobivena je originalna particija $\Pi = \{\pi_1, \ldots, \pi_5\}$ s $m = \sum_{j=1}^{5} m_j = 1000$ podataka (vidi Sliku 7.2a). Crvene točke na slici označavaju centre c_j klastera π_j, a crvene elipse su odgovarajuće M_N-kružnice.



Slika 7.2: Originalna particija Π i GK fuzzy optimalna particija $\Pi^\star$

Optimalnu particiju $\Pi^\star$ skupa $\mathcal{A}$ potražit ćemo primjenom `GK`-algoritma. Na Slici 7.2b prikazani su centri $c_j^\star$ (zelene točke), originalne M_N-kružnice (crvene elipse) i rekonstruirane M_N-kružnice (sive elipse).

Kao ilustraciju kvalitete rekonstrukcije pogledajmo Sliku 7.3 na kojoj je prikazano rasipanje elemenata klastera originalne particije Π po klasterima particije $\Pi^\star$.



Slika 7.3: Rasipanje elemenata klastera originalne particije Π po klasterima particije $\Pi^\star$

Matricu prijelaza $S(\Pi, \Pi^\star)$ možemo odrediti primjenom Zadatka 5.3, str. 97. Ona kvalitetnije pokazuje mjeru rasipanje elemenata klastera originalne particije Π

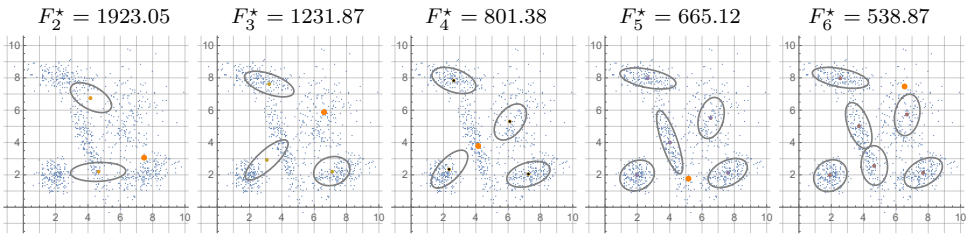
po klasterima particije $\Pi^\star$

$$S(\Pi, \Pi^\star)(\%) = \begin{bmatrix} 94.6 & 1.4 & 1.8 & 0.4 & 1.9 \\ 4.8 & 65.4 & 7.0 & 12.4 & 10.3 \\ 7.6 & 5.2 & 77.4 & 4.3 & 5.6 \\ 1.8 & 2.6 & 9.4 & 85.5 & 0.8 \\ 2.8 & 5.8 & 4.3 & 0.9 & 86.2 \end{bmatrix}.$$

7.3 Fuzzy inkrementalni algoritam

Pretpostavimo sada da broj klastera u particiji nije unaprijed poznat. Slično kao što smo u točki 4.3, str. 74 konstruirali Inkrementalni algoritam 2 za sferične klastere, a u točki 6.4.2, str. 113 Mahalanobis inkrementalni algoritam za elipsoidne klastere, možemo konstruirati i fuzzy varijantu ovih algoritama. Konstrukcija algoritma uglavnom se podudara s Inkrementalnim algoritmom 2 s tim da u slučaju primjene `LS`-kvazimetričke funkcije u `Koraku 5` primijenimo `FCM`-algoritam, a u slučaju primjene Mahalanobis kvazimetričke funkcije u `Koraku 5` primijenimo `GK`-algoritam. Na taj način dobit ćemo fuzzy-optimalne particije s $2, 3, \ldots$ klastera.

Primjer 7.3. *Za skup podataka $\mathcal{A}$ iz Primjera 7.2 i $\epsilon = .05$ provest ćemo Algoritam 2 počevši s centrom $\hat{c}_1 = \text{mean}(\mathcal{A}) = (4.6, 4.2)$.*



Slika 7.4: Fuzzy inkrementalni algoritam

Nakon što smo odredili centar $\hat{c}_2$, primjenom nekoliko iteracija `DIRECT`-algoritma i `GK`-algoritma dobivamo dva centra $c_1^\star, c_2^\star$ i njihove M_N-circles prikazane na Slici 7.4a. Na istoj slici crvenom točkom označen je i novi (treći) centar $\hat{c}_3$ dobiven primjenom `DIRECT` algoritma. Primjenom `GK`-algoritma dobivamo fuzzy optimalnu 3-particiju prikazanu na Slici 7.4b, itd. Povećanjem broja klastera vrijednost funkcije cilja opada, kao što je naznačeno na Slici 7.4, a za postizanje kriterija zaustavljanja algoritma bilo je potrebno šest iteracija.

Naravno, još ostaje problem izbora najprikladnije particije.

7.4 Izbor particije s najprikladnijim brojem klastera

Kako bismo mogli zaključiti koja je od dobivenih particija najprihvatljivija, koristit ćemo uobičajeni postupak primjene različitih indeksa. Neki indeksi nastali su prilagođavanjem već korištenih indeksa, a neki su originalno konstruirani za slučaj traženja najprihvatljivije fuzzy optimale particije (vidi [14, 110, 117]).

Neka je $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ k-optimalna fuzzy particija s matricom pripadnosti $U^\star$ i odgovarajućim klaster-centrima $c_1^\star, \ldots, c_k^\star$. Za slučaj fuzzy sferičnih klastera spomenimo sljedeće indekse:

- Prema Shannonovoj teoriji informacija [101], Klasifikacijska entropija definira se kao:
$$\texttt{CE}(k) = -\tfrac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} u_{ij}^\star \ln u_{ij}^\star, \tag{7.15}$$
a najmanja CE-vrijednost indicira najprihvatljiviju particiju [117];

- Xie-Beni fuzzy indeks definira se kao (vidi primjerice [99, 117]):
$$\texttt{XB}(k) = \frac{1}{m \min\limits_{i\neq j}\|c_i^\star - c_j^\star\|^2}\sum_{i=1}^{m}\sum_{j=1}^{k} u_{ij}^{\star q}\|c_j^\star - a^i\|^2, \tag{7.16}$$
a najmanja XB-vrijednost indicira najprihvatljiviju particiju;

- Davies – Bouldin fuzzy indeks definira se kao (vidi primjerice [99, 121])
$$\texttt{FDB}(k) = \tfrac{1}{k}\sum_{j=1}^{k}\max_{s\neq j}\frac{V(\pi_j^\star)+V(\pi_s^\star)}{\|c_j^\star - c_s^\star\|^2}, \quad V(\pi_j^\star) = \frac{1}{\sum\limits_{i=1}^{m} u_{ij}^{\star q}}\sum_{i=1}^{m} u_{ij}^{\star q}\|c_j^\star - a^i\|^2, \tag{7.17}$$
a najmanja FDB-vrijednost indicira najprihvatljiviju particiju;

- Chalinski-Harabasz fuzzy indeks može se definirati kao (vidi primjerice [99, 117, 121]):
$$\texttt{FCH}(k) = \frac{\sum\limits_{j=1}^{k}\kappa_j^\star\|c_A - c_j^\star\|^2}{\Phi(c^\star, U^\star)/(m-k)}, \quad c_A = \tfrac{1}{m}\sum_{i=1}^{m} a^i, \quad \kappa_j^\star = \sum_{i=1}^{m} u_{ij}^\star, \tag{7.18}$$
a najveća FCH-vrijednost indicira najprihvatljiviju particiju;

U slučaju fuzzy elipsoidnih klastera Klasifikacijska entropija ostaje nepromijenjena, a drugi indeksi trebaju se prilagoditi korištenjem kovarijacijskih matrica $\Sigma_j^\star$ definiranih s (7.13) i Mahalanobis kvazimetričkih funkcija definiranih s (7.14) (vidi primjerice [14, 99, 117, 121]).

- Xie-Beni fuzzy indeks definira se kao:

$$\texttt{XB}(k) = \frac{1}{m \min\limits_{i\neq j} d_M(c_i^\star, c_j^\star; \Sigma_i^\star + \Sigma_j^\star)} \sum_{i=1}^{m}\sum_{j=1}^{k} u_{ij}^{\star q} d_M(c_j^\star, a^i, \Sigma_j^\star), \tag{7.19}$$

  a najmanja XB-vrijednost indicira najprihvatljiviju particiju;

- Davies – Bouldin fuzzy indeks definira se kao:

$$\texttt{FDB}(k) = \frac{1}{k}\sum_{j=1}^{k} \max_{s\neq j} \frac{V(\pi_j^\star) + V(\pi_s^\star)}{d_M(c_j^\star, c_s^\star; \Sigma_j^\star + \Sigma_s^\star)},\ V(\pi_j^\star) = \frac{1}{\sum\limits_{i=1}^{m} u_{ij}^{\star q}} \sum_{i=1}^{m} u_{ij}^{\star q} d_M(c_j^\star, a^i; \Sigma_j^\star), \tag{7.20}$$

  a najmanja FDB-vrijednost indicira najprihvatljiviju particiju;

- Hipervolumni fuzzy indeks definira se kao:

$$\texttt{FHV}(k) = \sum_{j=1}^{k} \sqrt{\det \Sigma_j^\star}, \tag{7.21}$$

  a najmanja FHV-vrijednost indicira najprihvatljiviju particiju;

- Chalinski-Harabasz fuzzy indeks može se definirati kao:

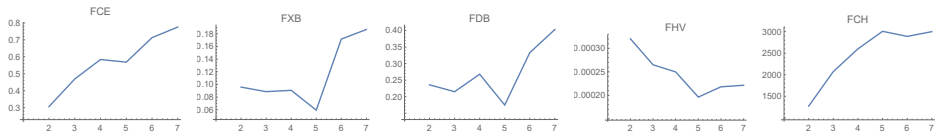
$$\texttt{FCH}(k) = \frac{\sum\limits_{j=1}^{k} \kappa_j^\star d_M(c_A, c_j^\star, \Sigma_j^\star + \Sigma)}{F_M(c^\star, U^\star)/(m-k)}, \quad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(c_A - a^i)(c_A - a^i)^T, \tag{7.22}$$

  gdje je $c_A = \frac{1}{m}\sum\limits_{i=1}^{m} a^i$, $\kappa_j^\star = \sum\limits_{i=1}^{m} u_{ij}^\star$, a najveća FCH-vrijednost indicira najprihvatljiviju particiju.

Primjer 7.4. *Za skup podataka $\mathcal{A}$ iz Primjera 7.2 i za fuzzy optimalne particije s $2, \ldots, 6$ klastera navedenih u Primjeru 7.3 izračunat ćemo odgovarajuće vrijednosti spomenutih indeksa. U Tablici 7.1 i na Slici 7.5 vidi se da je po svim indeksima 5-particija najprihvatljivija fuzzy particija.*

	$(k=2)$	$(k=3)$	$(k=4)$	$(k=5)$	$(k=6)$
FCE	**0.306**	0.469	0.584	**0.569**	0.713
FXB	0.096	0.089	0.091	**0.059**	0.172
FDB	0.236	0.215	0.268	**0.175**	0.333
FHV	0.00032	0.00026	0.00025	**0.0002**	0.00022
FCH	1272	2073	2600	**3007**	2889

Tablica 7.1: Vrijednosti indeksa optimalnih fuzzy particija iz Primjera 7.3. Vrijednosti indeksa najprihvatljivije particije posebno su istaknute



Slika 7.5: Vrijednosti indeksa optimalnih fuzzy particija iz Primjera 7.3

7.4.1 Fuzzy varijanta Rand indeksa

Neka su $\Pi^{(1)} = \{\pi_1^{(1)}, \ldots, \pi_k^{(1)}\}$ i $\Pi^{(2)} = \{\pi_1^{(2)}, \ldots, \pi_\ell^{(2)}\}$ dvije fuzzy particije skupa $\mathcal{A}$. Ako promatramo matricu pripadnosti U u slučaju tvrdog grupiranja (vidi primjerice Primjer 5.6), vidimo da je neki par (a^r, a^s) sparen (paired) ako je skalarni produkt $U(r) \cdot U(s)$ pripadnih redaka matrice U jednak 1, a nije sparen ako je $U(r) \cdot U(s) = 0$. Vodeći se tom analogijom i u slučaju fuzzy grupiranja za matricu U, čiji su elementi u_{rs} zadani s (7.6), može se definirati *fuzzy Rand indeks* (5.13), gdje je (vidi [21, 36, 43]):

$$
\begin{aligned}
a &= \sum_{(a^r,a^s)\in\mathcal{C}} \psi^{(1)}(a^r,a^s)\psi^{(2)}(a^r,a^s),\\
b &= \sum_{(a^r,a^s)\in\mathcal{C}} \psi^{(1)}(a^r,a^s)(1-\psi^{(2)}(a^r,a^s)),\\
c &= \sum_{(a^r,a^s)\in\mathcal{C}} (1-\psi^{(1)}(a^r,a^s))\psi^{(2)}(a^r,a^s),\\
d &= \sum_{(a^r,a^s)\in\mathcal{C}} (1-\psi^{(1)}(a^r,a^s))(1-\psi^{(2)}(a^r,a^s)),
\end{aligned}
$$

gdje je:

$$
\psi^{(\kappa)}(a^r,a^s) = \sum_{(a^r,a^s)\in\mathcal{C}} U^{(\kappa)}(a^r)\, U^{(\kappa)}(a^s), \quad \kappa = 1,2.
$$

Ovdje $U^{(\kappa)}(a^r)$ predstavlja redak matrice pripadnosti $U^{(\kappa)}$ particije $\Pi^{(\kappa)}$ pridružen elementu $a^r \in \mathcal{A}$. Preciznije, ako s $U^{(\kappa)}(r)$ označimo r-ti redak

matrice pripadnosti $U^{(\kappa)}$ koji pokazuje "rasipanje" elementa a^r po klasterima particije $\Pi^{(\kappa)}$, onda možemo pisati:

$$\psi^{(\kappa)}(a^r, a^s) = \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} U^{(\kappa)}(r) \cdot U^{(\kappa)}(s), \quad \kappa = 1, 2, \tag{7.23}$$

gdje je $U^{(\kappa)}(r) \cdot U^{(\kappa)}(s)$ uobičajeni skalarni produkt redaka $U^{(\kappa)}(r)$ i $U^{(\kappa)}(s)$ matrice $U^{(\kappa)}$.

Naravno, sve je moguće primijeniti i na Jaccard index (5.14).

Zadatak 7.1. *Na jednostavnom primjeru pokazite da u fuzzy slučaju funkcija D_R iz Zadatka 5.2 nije metrika (vidi primjerice [43]).*

Primjer 7.5. *Za originalnu* Π *i* 5*-optimalnu fuzzy particiju* $\Pi^\star$ *iz Primjera 7.2 fuzzy Rand indeks (*0.869*) i fuzzy Jaccard indeks (*0.508*) ukazuju na to da je dobivena* 5*-optimalna particija vrlo slična originalnoj.*

Poglavlje 8

Prepoznavanje geometrijskih objekata u ravnini

U ovom poglavlju promatrat ćemo problem prepoznavanja više unaprijed nepoznatih istovrsnih geometrijskih objekata u ravnini kao što su:

- više pravaca u ravnini:

$$p_j(u_j, v_j, z_j): \quad u_j x + v_j y + z_j = 0, \quad j = 1, \ldots, k, \tag{8.1}$$

- više kružnica u ravnini:

$$K_j(p_j, q_j, r_j): \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} p_j \\ q_j \end{bmatrix} + r_j \begin{bmatrix} \cos t \\ \sin t \end{bmatrix}, \; t \in [0, 2\pi], \quad j = 1, \ldots, k, \tag{8.2}$$

  gdje su $S_j = (p_j, q_j)$ središta, a $r_j > 0$ radijusi tih kružnica,

- više elipsi u ravnini:

$$E_j(p_j, q_j, \xi_j, \eta_j, \vartheta_j): \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} p_j \\ q_j \end{bmatrix} + U(\vartheta) \begin{bmatrix} \xi_j \cos t \\ \eta_j \sin t \end{bmatrix}, \; t \in [0, 2\pi], \tag{8.3}$$

  za $j = 1, \ldots, k$ gdje su $S_j = (p_j, q_j)^T$ centri, $\xi_j, \eta_j > 0$ poluosi i ϑ_j kutevi zakreta tih elipsi, a $U(\vartheta) = \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix}$ pripadna matrica rotacije, itd.

Primijetite da svaki pravac iz (8.1) ima 3 parametra, svaka kružnica iz (8.2) ima 3 parametra, svaka elipsa iz (8.3) ima 5 parametara, itd. U literaturi se također pojavlju i problemi s više generaliziranih kružnica [107], više segmenata [19], više kružnih lukova, itd.

Neka je $\Delta = [a,b] \times [c,d] \subset \mathbb{R}^2$, $a < b, c < d$ pravokutnik u ravnini, a $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \Delta \colon i = 1, \ldots, m\}$ skup podataka koji potječe od više unaprijed nepoznatih istovrsnih geometrijskih objekata u ravnini koje treba prepoznati ili rekonstruirati. Nepoznati su geometrijski objekti, odnosno njihovi parametri, ali i njihov broj.

Nadalje, pretpostavljamo da podaci koji dolaze od nekog geometrijskog objekta zadovoljavaju „svojstvo homogenosti", tj. pretpostavljamo da su podaci pretežno „homogeno rasuti" oko tog geometrijskog objekta. U tom smislu uvedimo sljedeću definiciju.

Definicija 8.1. *Neka je $\pi(\gamma)$ klaster koji potječe od geometrijskog objekta γ duljine $|\gamma|$ u pravokutniku Δ. Broj $\rho(\pi) = \frac{|\pi(\gamma)|}{|\gamma|}$ zovemo* lokalna gustoća *klastera π.*

8.1 Broj geometrijskih objekata unaprijed je poznat

Promatrajmo problem prepoznavanja poznatog broja k istovrsnih geometrijskih objekata u ravnini od kojih svaki ima po n parametara, a koje je na temelju skupa podataka $\mathcal{A}$ potrebno rekonstruirati ili detektirati. Ovaj problem skraćeno ćemo označiti kao `MGD`-problem (*Multiple Geometric Object Detection problem*).

Geometrijski objekt označimo s $\gamma_j(t_j)$, gdje je $t_j \in \mathbb{R}^n$ vektor parametara. Rješavanje ovog problema promatrat ćemo kao jedan problem grupiranja gdje su centri klastera spomenuti geometrijski objekti (`G`-klaster-centri).

Najprije treba dobro definirati udaljnost $\mathfrak{D}$ točke $a \in \mathcal{A}$ do geometrijskog objekta γ_j. Problem traženja optimalne k-particije $\Pi = \{\pi_1, \ldots, \pi_k\}$, čiji su klaster-centri geometrijski objekti $\gamma_1(t_1), \ldots, \gamma_k(t_k)$, možemo postaviti na jedan od sljedeća dva načina:

$$\operatorname*{argmin}_{\Pi} \mathcal{F}(\Pi), \quad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} \mathfrak{D}(a^i, \gamma_j(t_j)), \tag{8.4}$$

$$\operatorname*{argmin}_{t \in \mathbb{R}^{kn}} F(t), \quad F(t) = \sum_{i=1}^{m} \min_{1 \le j \le k} \mathfrak{D}(a^i, \gamma_j(t_j)), \quad t = (t_1, \ldots, t_k)^T. \tag{8.5}$$

8.1.1 Metoda za traženje k-LOPart s G-klaster-centrima

Za navedene primjere funkcija cilja definirana u (8.5) je Lipschitz-neprekidna funkcija [93], ali ipak direktno rješavanje GOP (8.5) primjenom globalno optimizacijskog algoritma DIRECT nije prihvatljivo jer se radi o simetričnoj funkciji s nk varijabli složenih u k vektor-parametara $t_1, \ldots, t_k \in \mathbb{R}^n$, a koja zbog toga ima barem $k!$ različitih točaka u kojima postiže globalni minimum. Naime, algoritam DIRECT tražio bi sve točke globalnog minimuma.

Zato se za rješavanje GOP (8.5) može primijeniti sljedeća metoda [93] koju ukratko možemo opisati u dva koraka:

Korak 1: Primjenom algoritma DIRECT pronaći dovoljno dobru početnu aproksimaciju G-klaster-centara ili primjenom algoritma DBSCAN pronaći dovoljno dobru početnu particiju;

Korak 2: Primjenom k-means algoritma modificiranog tako da centri klastera budu naši geometrijski objekti pronaći optimalno rješenje.

Naravno, ovakvim pristupom možemo očekivati pronalaženje k-LOPart. Pored navedenog pristupa, k-LOPart možemo potražiti primjenom k-means algoritma uz odgovarajući izbor početne aproksimacije ili primjenom nekih drugih pristupa (vidi primjerice točku 8.10, str. 170).

8.1.2 Traženje početne aproksimacije

Početnu aproksimaciju za MGD problem (8.5) potražit ćemo tako da promatramo problem traženja optimalne particije čiji su klaster-centri neki jednostavniji geometrijski objekti $\tilde{\gamma}(\tilde{t}_j)$ koji nalikuju našim geometrijskim objektima i po mogućnosti imaju manji broj $\tilde{n} \leq n$ parametara. GOP bi za ovu situaciju trebao biti što jednostavniji kako bi algoritam DIRECT brzo mogao dati prihvatljivo rješenje.

Kako bismo mogli primijeniti algorithm DIRECT na GOP (8.5), funkciju cilja F treba transformirati u funkciju $f\colon [0,1]^{k\tilde{n}} \to \mathbb{R}$, $f(x) = (F \circ T^{-1})(x)$, gdje je: $T\colon [\alpha_1, \beta_1]^k \times \cdots \times [\alpha_{\tilde{n}}, \beta_{\tilde{n}}]^k \to [0,1]^{k\tilde{n}}$, a $[\alpha_j, \beta_j]$ segment u kojemu se može očekivati vrijednost parametra $\tilde{t}_j$. Nakon manjeg broja iteracija algoritma DIRECT (recimo 10-20) dobivamo vektor $\hat{x}$, a početna aproksimacija za GOP (8.5) tada je zadana s $T^{-1}(\hat{x})$. Na taj način pokušat ćemo barem dobro pozicionirati tražene geometrijske objekte.

8.1.3 Modifikacija *k*-means algoritma za G-klaster centre

Ako raspolažemo dobrom početnom aproksimacijom (početni G-klaster-centri ili početna particija), za traženje k-GOPart možemo iskoristiti dobro poznati k-means algoritam modificiran za slučaj G-klaster-centara. Algoritam se slično kao u točki 4.2.2, str. 69 može opisati u dva koraka koji se sukcesivno ponavljaju.

Algoritam 8.1. *(Modifikacija k-means algoritma za G-klaster centre (KCG))*

Korak A: *(Pridruživanje) Za dani skup međusobno različitih geometrijskih objekata $\gamma_1(t_1), \ldots, \gamma_k(t_k)$, skup podataka $\mathcal{A}$ treba razdijeliti na k disjunktnih nepraznih klastera $\pi_1, \ldots, \pi_k$ korištenjem principa minimalnih udaljenosti za $j = 1, \ldots, k$*

$$\pi_j := \{a \in \mathcal{A} \colon \mathfrak{D}(a, \gamma_j(t_j)) \leq \mathfrak{D}(a, \gamma_s(t_s)),\ \forall s \neq j\}; \tag{8.6}$$

Korak B: *(Korekcija) Za danu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$, treba odrediti odgovarajuće G-klaster-centre $\hat{\gamma}_j(\hat{t}_j)$, $j = 1, \ldots, k$ kao rješenja sljedećih* GOP*:*

$$\hat{\gamma}_j(\hat{t}_j) \in \operatorname*{argmin}_{t \in \mathbb{R}^n} \sum_{a \in \pi_j} \mathfrak{D}(\gamma_j(t), a), \quad \forall j = 1, \ldots, k. \tag{8.7}$$

Slično kao u slučaju običnih sferičnih ili elipsoidalnih klastera (vidi Teorem 4.1, str. 71) može se pokazati da vrijedi sljedeći teorem.

Teorem 8.1. *KCG-algoritam u konačno mnogo koraka pronalazi lokalno optimalnu particiju, a pri tome je niz funkcijskih vrijednosti definiranih s* (8.4) *monotono padajući.*

Budući da niz funkcijskih vrijednosti (F_n) monotono opada, KCG-algoritam možemo zaustaviti kada je

$$\frac{F_{n-1} - F_n}{F_n} < \epsilon_{\text{KCG}}, \tag{8.8}$$

za neki maleni $\epsilon_{\text{KCG}} > 0$ (primjerice .005).

8.2 Broj geometrijskih objekata nije unaprijed poznat

8.2.1 Inkrementalni algoritam za G-klaster-centre

Ako broj klastera u particiji nije unaprijed poznat, možemo potražiti optimalne particije s $k = 1, 2, \ldots, k_{max}$ klastera i nakon toga u cilju prepoznavanja najprihvatljivije particije pokušati primijeniti neki od poznatih indeksa prilagođenih za slučaj G-klaster-centara. U tu svrhu možemo na odgovarajući način modificirati poznati inkrementalni algoritam (vidi primjerice [8, 98, 113]).

Inkrementalni algoritam za rješavanje MGD problema može započeti određivanjem najboljeg reprezentanta skupa $\mathcal{A}$

$$\hat{\gamma}_1(\hat{t}_1) \in \operatorname*{argmin}_{t\in\mathbb{R}^n} \sum_{a\in\mathcal{A}} \mathfrak{D}(a, \gamma(t)), \tag{8.9}$$

ali može započeti i s nekim drugim početnim objektom.

Nadalje, ako su poznati G-klaster-centri $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$, sljedeći centar $\hat{\gamma}_{k+1}$ odredit ćemo rješavanjem sljedećeg GOP

$$\operatorname*{argmin}_{t\in\mathbb{R}^n} \Phi(t), \quad \Phi(t) = \sum_{i=1}^{m} \min\{\delta_k^{(i)}, \mathfrak{D}(a^i, \gamma(t))\}, \tag{8.10}$$

gdje je: $\delta_k^{(i)} = \min\{\mathfrak{D}(a^i, \hat{\gamma}_1), \ldots, \mathfrak{D}(a^i, \hat{\gamma}_k)\}$.

Nakon toga, primjenom KCG-algoritma dobivamo optimirane G-klaster centre $\gamma_1^\star, \ldots, \gamma_k^\star, \gamma_{k+1}^\star$ lokalno optimalne $(k+1)$-particije $\Pi^{(k+1)}$. Budući da KCG-algoritam daje k-LOPart (vidi Teorem 8.1, str. 132), na taj način inkrementalni algoritam generira niz k-LOPart.

U općem slučaju, inkrementalni algoritam možemo zaustaviti (vidi [8]) kada je za neki k relativna vrijednost funkcije cilja (8.4) manja od nekog unaprijed zadanog broja $\epsilon_{\text{B}} > 0$ (primjerice, .005):

$$\frac{F_k - F_{k-1}}{F_1} < \epsilon_{\text{B}}. \tag{8.11}$$

Motivacija za ovakav kriterij zaustavljanja procesa leži u činjenici da povećanjem broja klastera u particiji vrijednost funkcije cilja opada (Teorem 3.2, str. 38).

Primjedba 8.1. *Primijetite da smo opisani algoritam analogno mogli pokrenuti i s više od jednog početnog geometrijskog objekta.*

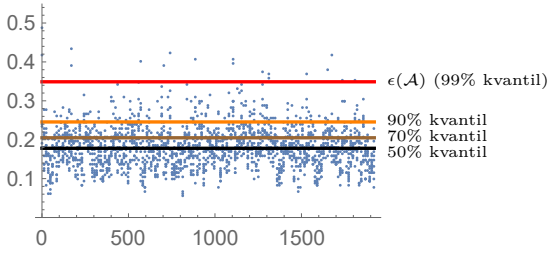
Pokazalo se da se u slučaju rješavanja MGD problema, kriterij zaustavljanja inkrementalnog algoritma može bolje definirati korištenjem poznate metode koja se koristi u algoritmu DBSCAN (vidi primjerice [32, 95]) za procjenu lokalne gustoće skupa podataka $\mathcal{A}$.

Definicija 8.2. *Neka je* $MinPts > 2$ *i neka je za svaki* $a \in \mathcal{A}$ *određen radijus* $\epsilon_a > 0$ *najmanjeg kruga sa središtem u* a *koji sadrži barem* $MinPts$ *elemenata skupa* $\mathcal{A}$. *Tada* ϵ-*gustoću skupa* $\mathcal{A}$ *definiramo kao* 99% *kvantil skupa* $\mathcal{R}(\mathcal{A}) = \{\epsilon_a : a \in \mathcal{A}\}$ *i označavamo s* $\epsilon(\mathcal{A})$. *Pri tome pod* p% *kvantilom* [čitaj: p postotnim kvantilom] *skupa* $\mathcal{R}(\mathcal{A})$ *podrazumijevamo broj* $q > 0$ *sa svojstvom da je* p% *elemenata skupa* $\mathcal{R}(\mathcal{A})$ *manje od* q.

Primjedba 8.2. *Parametar* $MinPts$ *možemo definirati kao* $MinPts = \lfloor \log |\mathcal{A}| \rfloor$ *(vidi [95]).*

Primijetite također da za skoro sve točke $a \in \mathcal{A}$ *odgovarajući krug s centrom u* a *i radijusom* $\epsilon(\mathcal{A})$ *sadrži barem* $MinPts$ *elemenata skupa* $\mathcal{A}$.

Primjer 8.1. *Za skup podataka* $\mathcal{A}$ *prikazan na Slici 8.11a odredit ćemo* ϵ-*gustoću* $\epsilon(\mathcal{A})$. *Parametar* $MinPts$ *određen je s* $MinPts = \lfloor \log |\mathcal{A}| \rfloor = 6$. *Na Slici 8.1 prikazan je skup točaka* $\{(i, \epsilon_{a^i}) : i = 1, \ldots, |\mathcal{A}|\}$ *i pravci koji definiraju* 50%, 70%, 90% *i* 99% *kvantil skupa* $\mathcal{R}(\mathcal{A})$.



Slika 8.1: ϵ-gustoća skupa $\mathcal{A}$ prikazanog na Slici 8.11a: $\epsilon(\mathcal{A}) = 0.349$

Prijeđimo na definiranje kriterija zaustavljanja inkrementalnog algoritma. Neka je u k-tom koraku inkrementalnog algoritma dobivena k-LOPart $\Pi = \{\pi_1, \ldots, \pi_k\}$. Za svaki klaster $\pi_j \in \Pi$ s G-klaster-centrom γ_j definirajmo skup

$$D_j := \{\mathfrak{D}_1(a, \gamma_j) : a \in \pi_j\},$$

gdje je $\mathfrak{D}_1(a, \gamma_j)$ ortogonalna udaljenost točke a do G-klaster-centra γ_j. Primjerice, za slučaj pravaca ova udaljenost dana je s (8.26) u Lemi 8.1, str. 140, za slučaj kružnica s (8.60), str. 161, a za slučaj elipsi s (8.77), str. 169. S

QD$[\pi_j]$ označimo 90% kvantil skupa D_j i definirajmo

$$\texttt{QD}[\Pi] = \max_{\pi \in \Pi} \texttt{QD}[\pi]. \tag{8.12}$$

Očekujemo da je k-particija Π blizu k-LOPart ako vrijedi:

$$\texttt{QD}[\Pi] < \epsilon(\mathcal{A}). \tag{8.13}$$

Zato kriterij zaustavljanja (8.11) inkrementalnog algoritma možemo zamijeniti boljim uvjetom (8.13).

Primijetite da će uvjet (8.13) biti lakše ispuniti ako QD$[\pi_j]$ definiramo kao $p\%$ ($p < 90$) kvantil skupa D_j. Na taj način možemo utjecati na broj geometrijskih objekata koji će se pojaviti u izlaznim rezultatima inkrementalnog algoritma.

8.3 Traženje MAPart i prepoznavanje geometrijskih objekata

Inkrementalni algoritam, ali i neki drugi pristupi (vidi primjerice točku 8.10, str. 170), u pravilu proizvode više k-LOPart s različitim brojem klastera. Najvažnije pitanje kod rješavanja MGD problema je sljedeće: postoji li među dobivenim k-GOPart i ona čiji se G-klaster-centri podudaraju s originalnim geometrijskim objektima (na osnovi kojih su nastali podaci) i ako postoji, kako je identificirati? Ovu k-LOPart zvat ćemo najprihvatljivija particija (*Most Appropriate Partition* (MAPart)). Posebnu pažnju posvetit ćemo upravo definiranju kriterija za prepoznavanje MAPart.

8.3.1 Modifikacija klasičnih indeksa

Budući da MGD problem rješavamo koristeći grupiranje podataka, spomenuti problem prepoznavanja MAPart mogli bismo pokušati riješiti modifikacijom nekih od poznatih indeksa iz točke 5, str. 87.

Navedimo modifikaciju CH-indeksa i DB-indeksa. Neka je $m = |\mathcal{A}|$, $\Pi^\star = \{\pi_1^\star, \ldots, \pi_\kappa^\star\}$ dana κ-GOPart s G-klaster-centrima $\gamma_1^\star, \ldots, \gamma_\kappa^\star$ i neka

je $\gamma^\star$ reprezentant cijelog skupa $\mathcal{A}$. Tada je (vidi [37, 90])

$$\texttt{CHG}(\kappa) = \frac{\frac{1}{\kappa-1}\sum_{j=1}^{\kappa}|\pi_j^\star| d_H(\gamma_j^\star,\gamma^\star)}{\frac{1}{m-\kappa}\sum_{j=1}^{\kappa}\sum_{a\in\pi_j^\star}\mathfrak{D}(a,\gamma_j^\star)}, \tag{8.14}$$

$$\texttt{DBG}(\kappa) = \frac{1}{\kappa}\sum_{j=1}^{\kappa}\max_{s\neq j}\frac{\sigma_j+\sigma_s}{d_H(\gamma_j^\star,\gamma_s^\star)}, \quad \text{gdje je} \quad \sigma_j^2 = \frac{1}{|\pi_j^\star|}\sum_{a\in\pi_j^\star}\mathfrak{D}(a,\gamma_j^\star), \tag{8.15}$$

gdje je $\mathfrak{D}$ kvazimetrička funkcija kojom je definirana udaljenost točke $a \in \mathcal{A}$ do krivulje $\gamma_j^\star$, $d_H(\gamma_j^\star,\gamma^\star)$ udaljenost G-klaster-centra $\gamma_j^\star$ do reprezentanta $\gamma^\star$ čitavog skupa $\mathcal{A}$, a $d_H(\gamma_j^\star,\gamma_s^\star)$ udaljenost dva G-klaster-centra.

Primijetite da u nekim slučajevima neće biti jednostavno definirati udaljenost dviju krivulja. U takvim slučajevima na svakoj krivulji odredimo diskretni skup točaka, a udaljenost između krivulja definiramo kao Hausdorffovu udaljenost pripadnih skupova. Iz tog razloga koristimo oznaku d_H.

8.3.2 Novi indeks za MGD probleme

Budući da su originalni indeksi primarno konstruirani za sferične ili elipsoidalne klastere, pokazalo se da je u ovom slučaju ipak bolje iskoristiti specijalnu strukturu MGD problema.

U tom smislu, u [91] definiran je novi indeks specijaliziran za MGD problem. Kratko ćemo ga opisati. Za svaki klaster $\pi_j^\star$ sukladno Definiciji 8.1 najprije treba odrediti lokalnu gustoću $\rho_j = \frac{|\pi_j^\star|}{|\gamma_j^\star|}$. Primijetite da će klaster, čiji je G-klaster-centar blizak nekoj od krivulja od koje potječu podaci, imati relativno veću gustoću.

Nadalje, za skup $\rho = \{\rho_1,\ldots,\rho_\kappa\}$ određuje se varijanca

$$\texttt{Var}(\rho) = \frac{1}{\kappa-1}\sum_{j=1}^{\kappa}(\rho_j-\bar{\rho})^2, \quad \bar{\rho} = \frac{1}{\kappa}\sum_{j=1}^{\kappa}\rho_j, \quad \kappa \geq 2, \tag{8.16}$$

i definira Geometrical Objects-indeks (GO) za κ-LOPart s G-klaster-centrima

$$\texttt{GO}(\kappa) := \begin{cases} \texttt{Var}(\rho_1,\ldots,\rho_\kappa), & \text{ako } \kappa \geq 2, \\ F, & \text{ako } \kappa = 1, \end{cases} \tag{8.17}$$

gdje je F vrijednost funkcije cilja iz (8.5). Manja vrijednost GO-indeksa pripada prikladnijoj particiji.

Budući da je inkrementalni algoritam lokalnog karaktera, u pravilu će proizvesti više krivulja γ nego što se može očekivati obzirom na konfiguraciju skupa $\mathcal{A}$ (kao što se može vidjeti na primjerima pravaca na Slici 8.11, str. 151), ali se između njih ne moraju pojaviti i oni od kojih su nastali podaci. To znači da se između dobivenih particija često ne nalazi i `MAPart` pa korištenjem modifikacije klasičnih indeksa ne možemo uvijek očekivati detekciju `MAPart`.

8.3.3 Novi pristup

U nastavku ćemo pokazati pristup koji koristi specifičnost promatranog problema i koji se u primjenama pokazao efikasnijim.

Neka je $\Pi = \{\pi_1, \ldots, \pi_k\}$ originalna k-particija skupa $\mathcal{A}$. Budući da za svaki $a \in \mathcal{A}$ odgovarajući kružić radijusa $\epsilon(\mathcal{A})$ sadrži barem $MinPts$ elemenata skupa $\mathcal{A}$, donja granica lokalne gustoće za svaki klaster $\pi \in \Pi$ može se procijeniti s (vidi Primjedbu 8.2):

$$\rho(\pi) \geq \tfrac{MinPts}{2\epsilon(\mathcal{A})}. \tag{8.18}$$

Zbog toga, za svaki klaster $\pi^\star \in \Pi^\star$ za koji ne vrijedi (8.18), pripadne `G`-klaster-centre treba ispustiti. Pomoću preostalih `G`-klaster-centara $\gamma_1^\star, \ldots, \gamma_r^\star$ definiramo rezidual (ostatak):

$$\mathcal{R}(r) = \mathcal{A} \setminus \bigcup_{j=1}^{r} \Gamma_j, \quad \Gamma_j = \{a \in \mathcal{A} \colon \mathfrak{D}_1(a, \gamma_j^\star) < \epsilon(\mathcal{A})\}. \tag{8.19}$$

Ako je $|\mathcal{R}(r)| \leq MinPts$, r-`LOPart` dobivena primjenom `KCG`-algoritma na `G`-klaster-centre $\gamma_1^\star, \ldots, \gamma_r^\star$ je `MAPart` (vidi Primjer 8.5, Primjer 8.13 i Primjer 8.14).

Ako je $|\mathcal{R}(r)| > MinPts$, r-`LOPart` dobivena primjenom `KCG`-algoritma na `G`-klaster-centre $\gamma_1^\star, \ldots, \gamma_r^\star$ nije `MAPart` (vidi Primjer 8.6).

Sada smo u mogućnosti definirati nužne uvjete da lokalno optimalna particija r-`LOPart` $\Pi^\star = \{\pi_1^\star, \ldots, \pi_r^\star\}$ bude ujedno i `MAPart`:

$\mathbf{A}_0$: $\texttt{QD}[\Pi^\star] < \epsilon(\mathcal{A})$,

$\mathbf{A}_1$: $\rho(\pi_j^\star) \geq \tfrac{MinPts}{2\epsilon(\mathcal{A})}$ za sve klastere $\pi_j^\star \in \Pi^\star$,

$\mathbf{A}_2$: $|\mathcal{R}(r)| \leq MinPts$.

8.4 Pravac kao reprezentant skupa podataka iz $\mathbb{R}^2$

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 : i = 1, \ldots, m\} \subset \Delta$ koji potječe od jednog unaprijed nepoznatog pravca p zadanog u eksplicitnom obliku $y = \alpha x + \beta$ i da su pogreške u izmjerenim vrijednostima nezavisne varijable x_i zanemarive, a pogreške u izmjerenim vrijednostima zavisne varijable y_i normalno distribuirane nezavisne slučajne varijable s očekivanjem 0 i varijancom σ^2, tj. da je za svaki $i = 1, \ldots, m$,

$$y_i = \alpha x_i + \beta + \epsilon_i, \quad \epsilon_i \in \mathcal{N}(0, \sigma^2).$$

Tada parametre α, β pravca p možemo potražiti rješavanjem `GOP`

$$\underset{\alpha,\beta\in\mathbb{R}}{\operatorname{argmin}} F(\alpha, \beta), \qquad F(\alpha, \beta) = \sum_{i=1}^{m} (y_i - \alpha x_i - \beta)^2. \tag{8.20}$$

Ovo je jednostavni linearni problem najmanjih kvadrata [77, 84]. Pripadni sustav *normalnih jednadžbi* možemo riješiti primjenom QR-dekompozicije ili dekompozicijom na singularne vrijednosti (vidi [84, 108]).

Ako se značajne pogreške mogu očekivati u izmjerenim vrijednostima nezavisne varijable x_i i u izmjerenim vrijednostima zavisne varijable y_i, tj. ako je za svaki $i = 1, \ldots, m$,

$$y_i = \alpha(x_i + \delta_i) + \beta + \epsilon_i, \quad \delta_i, \epsilon_i \in \mathcal{N}(0, \sigma^2),$$

onda pravac p možemo potražiti rješavanjem sljedećeg `TLS`-problema:

$$\underset{\alpha,\beta\in\mathbb{R},\, \delta\in\mathbb{R}^m}{\operatorname{argmin}} T(\alpha, \beta, \delta), \quad T(\alpha, \beta, \delta) = \sum_{i=1}^{m} \left((y_i - \alpha(x_i + \delta_i) - \beta)^2 + \delta_i^2\right), \tag{8.21}$$

gdje je $\delta = (\delta_1, \ldots, \delta_m)^T \in \mathbb{R}^m$ (vidi [48]). `GOP` (8.21) nelinearni je globalno optimizacijski problem s $m + 2$ varijable. Postoje specijalne, ali numerički zahtjevne metode za rješavanje ovakvih problema (vidi primjerice [15]). U ovom slučaju radi se o najjednostavnijem `TLS`-problemu koji smo već razmatrali u t 6.1, str. 99.

U sljedećoj točki navest ćemo neke osnovne činjenice i svojstva vezana uz pravac u ravnini koje će nam pomoći kod rješavanja problema prepoznavanja više pravaca u ravnini.

8.4.1 Pravac u ravnini

Općenito, pravac u ravnini zadajemo u implicitnom obliku:

$$\alpha x + \beta y + \gamma = 0, \quad \alpha^2 + \beta^2 \neq 0. \tag{8.22}$$

Skup svih pravaca u ravnini označit ćemo s $\mathcal{L}$. Dokažimo najprije sljedeće pomoćne rezultate.

Propozicija 8.1. *Neka je $(O;(\vec{i},\vec{j}))$ pravokutni koordinatni sustav u ravnini u kojemu je jednadžbom (8.22) zadan pravac p u toj ravnini. Tada:*

(i) $\vec{n}=\alpha\vec{i}+\beta\vec{j}$ *vektor je normale na pravac p,*

(ii) Vektori $\vec{u}_1=\beta\vec{i}-\alpha\vec{j}$, *odnosno* $\vec{u}_2=-\beta\vec{i}+\alpha\vec{j}$, *određuju smjer pravca p,*

(iii) Ako je $P_0=(x_0,y_0)^T$ *točka na pravcu p, onda je* $\gamma=-\alpha x_0-\beta y_0$.

Dokaz. (i) Neka su $P_1=(x_1,y_1)^T$, $P_2=(x_2,y_2)^T$ dvije različite točke koje leže na pravcu p. To znači da je

$$\alpha x_s+\beta y_s+\gamma=0,\quad s=1,2.$$

Oduzimanjem ovih jednadžbi dobivamo:

$$\alpha(x_2-x_1)+\beta(y_2-y_1)=0, \tag{8.23}$$

što pokazuje da su vektori $\vec{n}=\alpha\vec{i}+\beta\vec{j}$ i $\overrightarrow{P_1P_2}$ okomiti, tj. da je $\vec{n}$ normala na pravac p.

(ii) Direktnom provjerom može se vidjeti da su vektori $\vec{u}_1,\vec{u}_2$ okomiti na vektor $\vec{n}$ pa time određuju smjer pravca p.

(iii) Ako je $P_0=(x_0,y_0)^T$ točka na pravcu p, onda vrijedi $\alpha(x-x_0)+\beta(y-y_0)=0$, odnosno $\alpha x+\beta y+\gamma=0$, gdje je $\gamma=-\alpha x_0-\beta y_0$. □

8.4.2 Normalna jednadžba pravca u ravnini

Nadalje, bez smanjenja općenitosti možemo pretpostaviti da je pravac (8.22) zadan uz uvjet $\alpha^2+\beta^2=1$. Naime, u protivnom dijeljenjem jednadžbe (8.22) s $\sqrt{\alpha^2+\beta^2}$ dobivamo

$$\hat{\alpha}x+\hat{\beta}y+\hat{\gamma}=0,\quad \hat{\alpha}^2+\hat{\beta}^2=1. \tag{8.24}$$

Jednadžbu (8.24) zvat ćemo normalna jednadžba pravca u ravnini. Analogno Propoziciji 8.1 vrijedi:

Propozicija 8.2. *Neka je $(O;(\vec{i},\vec{j}))$ pravokutni koordinatni sustav u ravnini u kojemu je normalnom jednadžbom (8.24) zadan pravac p u toj ravnini. Tada:*

(i) $\vec{n}_0 = \hat{\alpha}\vec{i} + \hat{\beta}\vec{j}$ *jedinični je vektor normale na pravac* p,

(ii) Jedinični vektori $\vec{u}_1 = \hat{\beta}\vec{i} - \hat{\alpha}\vec{j}$, *odnosno* $\vec{u}_2 = -\hat{\beta}\vec{i} + \hat{\alpha}\vec{j}$, *određuju smjer pravca* p,

(iii) Ako je $P_0 = (x_0, y_0)^T$ *točka na pravcu* p *zadanom s* (8.24), *onda je* $\hat{\gamma} = -\hat{\alpha}x_0 - \hat{\beta}y_0$, *a jednadžba* (8.24) *postaje*

$$\hat{\alpha}(x - x_0) + \hat{\beta}(y - y_0) = 0, \quad \hat{\alpha}^2 + \hat{\beta}^2 = 1. \tag{8.25}$$

Pri tome parametar $\hat{\gamma}$ *do na predznak određuje udaljenost ishodišta* O *do pravca* p.

Dokaz Propozicije 8.2 može se provesti slično dokazu Propozicije 8.1. Posebno dokažimo samo tvrdnju (iii). Ako je $P_0 = (x_0, y_0)^T$ točka na pravcu p, onda je $\hat{\gamma} = -\hat{\alpha}x_0 - \hat{\beta}y_0$. Ako s $\vec{r}_0$ označimo radij-vektor točke P_0, onda je $\hat{\gamma}$ negativna vrijednost skalarnog produkta vektora $\vec{r}_0$ i $\vec{n}_0$, tj. $\hat{\gamma} = -\langle \vec{r}_0, \vec{n}_0 \rangle$, odakle slijedi tvrdnja (iii) Propozicije 8.2.

Sljedeća lema daje eksplicitne formule za udaljenost točke do pravca i za ortogonalnu projekciju točke na pravac zadan s (8.24).

Lema 8.1. *Neka je* $(O; (\vec{i}, \vec{j}))$ *pravokutni koordinatni sustav u ravnini, neka je* $\alpha x + \beta y + \gamma = 0$, $\alpha^2 + \beta^2 = 1$ *normalna jednadžba pravca* p *u ravnini i neka je* $a^i = (x_i, y_i)^T \in \mathbb{R}^2$ *proizvoljna točka u ravnini. Tada vrijedi:*

(i) Udaljenost točke a^i *do pravca* p *zadana je formulom:*

$$d(a^i, p) = |\alpha x_i + \beta y_i + \gamma|, \tag{8.26}$$

(ii) Ortogonalna projekcija a^i_p *točke* a^i *na pravac* p *zadana je radij vektorom:*

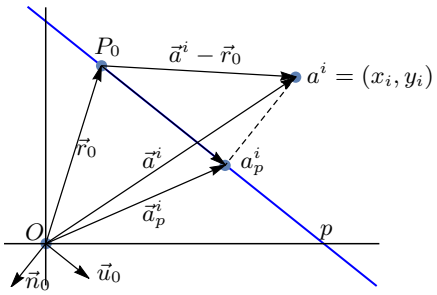
$$\vec{a}^i_p = \langle \vec{a}^i, \vec{u}_0 \rangle \vec{u}_0 - \gamma \vec{n}_0, \tag{8.27}$$

gdje je $\vec{a}^i = x_i\vec{i} + y_i\vec{j}$ *radij vektor točke* a^i, $\vec{n}_0 = \alpha\vec{i} + \beta\vec{j}$ *jedinični vektor normale, a* $\vec{u}_0 = \beta\vec{i} - \alpha\vec{j}$ *jedinični vektor u smjeru pravca* p.

Dokaz. (i) Neka je $P_0 = (x_0, y_0)^T$ točka na pravcu p i

$$(\overrightarrow{P_0a^i})_{\vec{n}} = \langle \overrightarrow{P_0a^i} \cdot \vec{n}_0 \rangle \vec{n}_0$$

ortogonalna projekcije vektora $\overrightarrow{P_0a^i} = (x_i - x_0)\vec{i} + (y_i - y_0)\vec{j}$ na normalu pravca p zadanu jediničnim vektorom $\vec{n}_0 = \alpha\vec{i} + \beta\vec{j}$ (vidi Sliku 8.2).



Slika 8.2: Udaljenost točke do pravca i projekcija točke na pravac

Udaljenost točke a^i do pravca p tada je:

$$\begin{aligned} d(a^i, p) =& \|\langle \overrightarrow{P_0 a^i} \cdot \vec{n}_0 \rangle \vec{n}_0\| = |\langle \overrightarrow{P_0 a^i} \cdot \vec{n}_0 \rangle| \\ =& |\alpha(x_i - x_0) + \beta(y_i - y_0)| = |\alpha x_i + \beta y_i - \alpha x_0 - \beta y_0|. \end{aligned}$$

Kako je prema Propoziciji 8.2, $\gamma = -\alpha x_0 - \beta y_0$, slijedi tražena tvrdnja.

(ii) Ako s $\vec{r}_0$ označimo radij vektor točke P_0, onda je $\overrightarrow{P_0 a^i} = \vec{a}^i - \vec{r}_0$, a radij-vektor projekcije a^i_p točke a^i na pravac p zadan je s (vidi Sliku 8.2):

$$\begin{aligned} \vec{a}^i_p =& \vec{r}_0 + \langle \vec{a}^i - \vec{r}_0, \vec{u}_0 \rangle \vec{u}_0 \\ =& \langle \vec{a}^i, \vec{u}_0 \rangle \vec{u}_0 + \vec{r}_0 - \langle \vec{r}_0, \vec{u}_0 \rangle \vec{u}_0. \end{aligned} \tag{8.28}$$

Kako su $\vec{u}_0, \vec{n}_0$ dva međusobno okomita jedinična vektora, vrijedi:

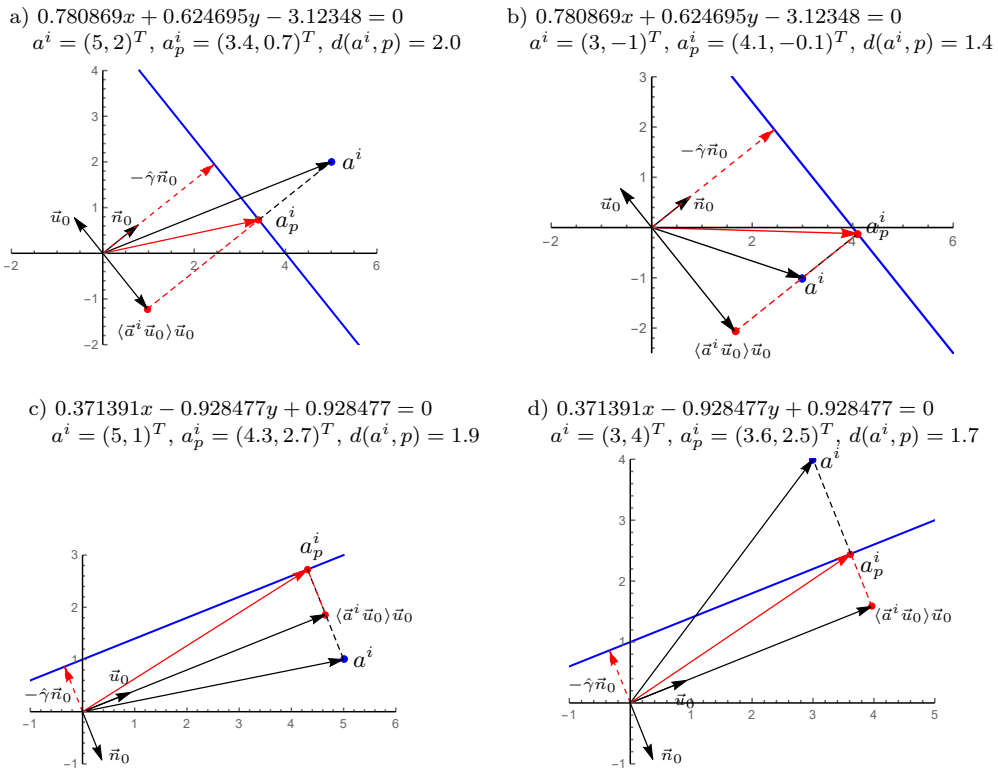
$$\vec{r}_0 = \langle \vec{r}_0, \vec{u}_0 \rangle \vec{u}_0 + \langle \vec{r}_0, \vec{n}_0 \rangle \vec{n}_0,$$

iz čega slijedi $\vec{r}_0 - \langle \vec{r}_0, \vec{u}_0 \rangle \vec{u}_0 = \langle \vec{r}_0, \vec{n}_0 \rangle \vec{n}_0$, što uvršteno u (8.28) daje

$$\vec{a}^i_p = \langle \vec{a}^i, \vec{u}_0 \rangle \vec{u}_0 + \langle \vec{a}^i, \vec{n}_0 \rangle \vec{n}_0. \tag{8.29}$$

Kako je prema Propoziciji 8.2, $\gamma = -\langle \vec{r}_0, \vec{n}_0 \rangle$, iz (8.29) slijedi (8.27). □

Primjer 8.2. *Na Slici 8.3 prikazano je nekoliko karakterističnih slučajeva za koje je određena ortogonalna projekcija i udaljenost točke a^i do pravca p.*



Slika 8.3: Ortogonalna projekcija i udaljenost točke a^i do pravca p

Zadatak 8.1. *Nađite formulu za ortogonalnu projekciju točke $a^i \in \mathbb{R}^2$ na pravac p zadan u eksplicitnom obliku $y = kx + l$ i dokažite da je udaljenost točke a^i do tog pravca zadana s:*

$$d(a^i, p) = \frac{|kx_i + l - y_i|}{\sqrt{k^2+1}}.$$

Udaljenost dvaju pravaca u ravnini može se definirati na više načina [113]. Jednu mogućnost navodimo u sljedećoj definiciji.

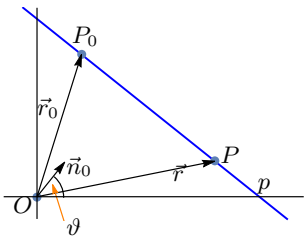
Definicija 8.3. *Neka su p_1: $\alpha_1 x + \beta_1 y + \gamma_1 = 0$, p_2: $\alpha_2 x + \beta_2 y + \gamma_2 = 0$ jednadžbe dvaju pravaca zadanih u normalnom obliku, a $\vec{n}_1 = \alpha_1 \vec{i} + \beta_1 \vec{j}$ i $\vec{n}_2 = \alpha_2 \vec{i} + \beta_2 \vec{j}$ odgovarajući jedinični vektori normala. Tada je njihova međusobna udaljenost određena kvazimetričkom funkcijom $d\colon \mathcal{L}^2 \to [0,1]$,*

$$d(p_1, p_2) = \begin{cases} 1 - |\langle \vec{n}_1, \vec{n}_2 \rangle|, & \text{ako } |\langle \vec{n}_1, \vec{n}_2 \rangle| < 1, \\ \frac{|\gamma_1 - \gamma_2|}{1 + |\gamma_1 - \gamma_2|}, & \text{ako } |\langle \vec{n}_1, \vec{n}_2 \rangle| = 1. \end{cases} \tag{8.30}$$

Ako pravci nisu paralelni ($|\langle \vec{n}_1, \vec{n}_2 \rangle| < 1$), udaljenost između njih određuje se pomoću kuta između njihovih normala, a ako su pravci paralelni ($|\langle \vec{n}_1, \vec{n}_2 \rangle| = 1$), udaljenost između njih definira se pomoću njihovih slobodnih koeficijenata γ_1, γ_2. Tu se koristi svojstvo (iii) iz Propozicije 8.2.

8.4.3 Hesseov normalni oblik jednadžbe pravca

Neka je $(O; (\vec{i}, \vec{j}))$ pravokutni koordinatni sustav u ravnini. Pravac p u ravnini možemo zadati jednom njegovom točkom P_0 i jediničnim vektorom normale $\vec{n}_0 = \cos\vartheta\vec{i} + \sin\vartheta\vec{j}$ na pravac p *koji ima smjer od ishodišta* O *prema pravcu* p, a s pozitivnim smjerom osi x zatvara kut ϑ (vidi Slika 8.4).



Slika 8.4: Hesseov normalni oblik jednadžbe pravca u ravnini

S $\vec{r} = x\vec{i} + y\vec{j}$ označimo radij vektor proizvoljne točke $P = (x, y)^T$ koja leži na pravcu p, a s $\vec{r}_0$ radij vektor točke P_0. Vektor $\overrightarrow{P_0P} = \vec{r} - \vec{r}_0$ okomit je na normalu $\vec{n}_0$, pa vrijedi:

$$\langle \vec{r} - \vec{r}_0, \vec{n}_0 \rangle = 0. \tag{8.31}$$

Tako dobivamo *Hesseov normalni oblik jednadžbe pravca*:

$$x\cos\vartheta + y\sin\vartheta - \delta = 0, \tag{8.32}$$

gdje je:

$$\delta := \langle \vec{r}_0, \vec{n}_0 \rangle = (r_0)_{n_0} > 0$$

udaljenost ishodišta O do pravca p.

Uočite vezu između Hesseovog normalnog oblika (8.32) i normalnog oblika (8.24).

Udaljenost točke $a^i = (x_i, y_i)^T$ do pravca p određena je s:

$$d(a^i, p) = |x_i\cos\vartheta + y_i\sin\vartheta - \delta|. \tag{8.33}$$

Ako je poznata neka točka $P_0 = (x_0, y_0)^T$ koja leži na pravcu p, onda (8.32) postaje:

$$(x - x_0)\cos\vartheta + (y - y_0)\sin\vartheta = 0, \tag{8.34}$$

a udaljenost (8.33) točke a^i do pravca p postaje:

$$d(a^i, p) = |(x_i - x_0)\cos\vartheta + (y_i - y_0)\sin\vartheta - \delta|. \tag{8.35}$$

8.4.4 Traženje TLS-pravca u Hesseovom normalnom obliku

Umjesto rješavanja problema (6.4), str. 6.4 TLS-pravac možemo tražiti u Hesseovom normalnom obliku tako da rješavamo sljedeći GOP:

$$\underset{\vartheta\in[0,\pi]}{\operatorname{argmin}}\,\Phi(\vartheta), \qquad \Phi(\vartheta) = \sum_{i=1}^{m} w_i \left[(x_i - \bar{x})\cos\vartheta + (y_i - \bar{y})\sin\vartheta\right]^2. \tag{8.36}$$

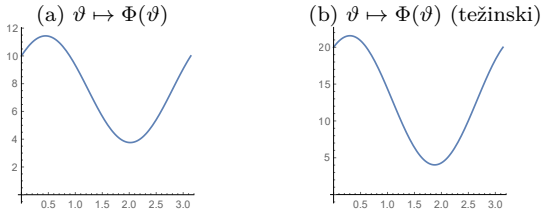
Primijetite da koristeći zapis (6.5) funkciju Φ iz (8.36) možemo zapisati kao:

$$\Phi(\vartheta) = \|\sqrt{D}B[\cos\vartheta, \sin\vartheta]^T\|_2^2. \tag{8.37}$$

Na taj način problem traženja najboljeg TLS-pravca sveli smo na rješavanje GOP samo s jednom nezavisnom varijablom $\vartheta \in [0, \pi]$, što će nam biti posebno važno kod konstrukcije inkrementalnog algoritma za rješavanje MLD problema u točki 8.5.3, str. 148.

Primjer 8.3. *Potražimo* TLS*-pravac u Hesseovom normalnom obliku za podatke iz Primjera 6.1.*

Minimizirajuća funkcija Φ zadana s (8.36) za ovaj primjer prikazana je na Slici 8.5a i postiže svoj minimum za $\vartheta^\star = 2.019$, čime je određen TLS-pravac: $(x-3)\cos\vartheta^\star + (x - \frac{13}{5})\sin\vartheta^\star = 0$. Primijetite da smo na ovaj način dobili traženi TLS-pravac u normalnom obliku: $-0.433x + 0.901y - 1.044 = 0$.



Slika 8.5: Minimizirajuća funkcija (8.36)

U slučaju težinskih podataka s $w_5 = 6$ minimizirajuća funkcija Φ prikazana je na Slici 8.5b i postiže svoj minimum za $\vartheta^\star = 1.8743$, čime je određen TLS-pravac: $(x-4)\cos\vartheta^\star + (x - \frac{14}{5})\sin\vartheta^\star = 0$, odnosno: $-0.299x + 0.954y - 1.477 = 0$.

8.4.5 OD-pravac

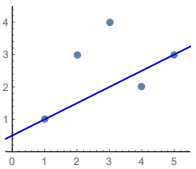
Pravac kao reprezentant skupa podataka $\mathcal{A} \subset \mathbb{R}^2$ može se odrediti također i kao najbolji Orthogonal Distance (OD) pravac: [77]

$$\underset{\alpha,\beta,\gamma\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{m} \frac{|\alpha x_i+\beta y_i+\gamma|}{\sqrt{\alpha^2+\beta^2}}, \tag{8.38}$$

ili u Hesseovom normalnom obliku rješavajući GOP:

$$\underset{\vartheta\in[0,\pi],\,\delta\in[0,M]}{\operatorname{argmin}} \sum_{i=1}^{m} |x_i\cos\vartheta + y_i\sin\vartheta - \delta|, \tag{8.39}$$

gdje je $M = \max\limits_{i=1,\ldots,m} \|a^i\|_2$.



Slika 8.6: Najbolji OD-pravac iz Primjera 6.1, str. 102

U ovom slučaju ne vrijedi lema analogna Lemi 6.1, str. 100 pa se ne može smanjiti broj nezavisni varijabli. Međutim, može se pokazati [77] da najbolji OD-pravac prolazi kroz barem dvije točke podataka (vidi Sliku 8.6).

8.5 Prepoznavanje više pravaca u ravnini

Za problem prepoznavanja više pravaca u ravnini (*Multiple Line Detection Problem* (MLD)) u literaturi se može pronaći više pristupa. Najpoznatija je primjena Houghovih transformacija [31, 34, 63]. U [102] prvi je put primijenjen „center-based clustering pristup". Specijalno, u radu [8] primjenjuje se inkrementalni algoritam, a u [111] jedna kombinacija grupiranja i TLS-pristupa u cilju rješavanja problema prepoznavanja redova zasijanja u poljoprivredi (vidi također [51, 112]). Detaljnije ćemo razmotriti metode koje se zasnivaju na grupiranju podataka.

Ako s $\mathfrak{U} = \{p = (\xi,\eta,\zeta)^T \in \mathbb{R}^3 \colon \xi^2+\eta^2 = 1\}$ označimo skup vektora-parametara, onda se svaki pravac $\ell \in \mathcal{L}$ može zapisati u obliku $\ell(\xi,\eta,\zeta) \equiv \xi x + \eta y + \zeta = 0$, $(\xi,\eta,\zeta)^T \in \mathfrak{U}$. Očigledno, postoji bijekcija između skupova $\mathcal{L}$ i $\mathfrak{U}$.

Ako udaljenost točke $a^i = (x_i, y_i)^T \in \mathcal{A}$ do pravca $\ell \in \mathcal{L}$ kojemu je pridružen vektor-parametara $p = (\xi, \eta, \zeta)^T \in \mathfrak{U}$, definiramo koristeći (8.26) iz Leme 8.1, str. 140

$$\mathfrak{D}(a^i, \ell(p)) = (\xi\, x_i + \eta\, y_i + \zeta)^2, \quad \ell(p) \equiv \xi x + \eta y + \zeta = 0, \tag{8.40}$$

onda se skup $\mathcal{A} \subset \mathbb{R}^n$ može grupirati u k klastera $\pi_1, \ldots, \pi_k$ čiji su centri pravci $\ell_j(p_j)$ s vektorima parametara p_j određenim s:

$$p_j \in \operatorname*{argmin}_{p \in \mathfrak{U}} \sum_{a^i \in \pi_j} \mathfrak{D}(a^i, \ell(p)). \tag{8.41}$$

Zato ćemo klaster s pravcem-centrom $\ell_j(p_j)$ označavati s $\pi(p_j)$ ili jednostavno s π_j.

Primijetite da je rješenje `GOP` (8.41) `TLS`-pravac skupa π_j (vidi točku 6.1, str. 99). Zato se traženje globalno optimalne k-particije (k-`GOPart`) skupa $\mathcal{A}$ može promatrati kao traženje rješenja sljedećeg `GOP` (vidi [8, 102, 113]):

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi), \qquad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} \mathfrak{D}(a^i, \ell_j), \tag{8.42}$$

gdje je $\mathcal{P}(\mathcal{A}; k)$ skup svih k-particija skupa $\mathcal{A}$.

Poznato je (vidi [98, 103]) da se problem traženja k-`GOPart` (8.42) može zamijeniti rješavanjem sljedećeg `GOP`:

$$\operatorname*{argmin}_{p_j \in \mathfrak{U}} F(p_1 \ldots, p_k), \quad F(p_1 \ldots, p_k) = \sum_{i=1}^{m} \min_{1 \le j \le k} \mathfrak{D}(a^i, \ell_j(p_j)), \tag{8.43}$$

jer se njihova rješenja podudaraju.

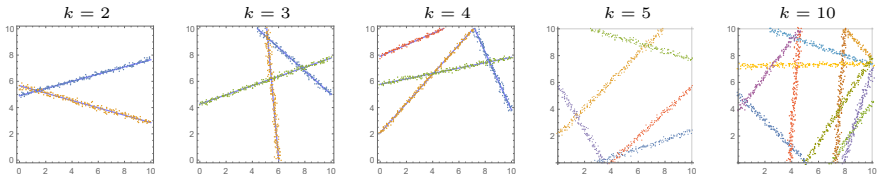
Općenito, funkcija iz (8.43) nekonveksna je i nediferencijabilna, ali slično kao u [93], može se pokazati da je Lipschitz-neprekidna.

8.5.1 Generiranje podataka koji potječu od više pravaca u ravnini

Razmatrani problem detekcije više pravaca u ravnini ilustrirat ćemo na umjetnom skupu podataka $\mathcal{A} \subset \Delta = [0, 10] \times [0, 10]$ koji potječe od $k \le 10$ pravaca u ravnini.

Najprije izaberemo broj $k \in \mathcal{U}(2, 10)$. Nakon toga k puta slučajno izaberemo dvije točke $A_j, B_j \in \partial\Delta$ na rubu pravokutnika Δ koje ne leže na istoj stranici pravokutnika Δ, ali tako da za njihovu Hausdorffovu udaljenost vrijedi $d_H(\{A_r, B_r\}, \{A_s, B_s\}) \ge 1$ za $r \ne s$.

Parovima točaka (A_j, B_j), $j = 1, \ldots, k$ određeno je k pravaca koji sijeku pravokutnik $[0, 10] \times [0, 10]$ (vidi Sliku 8.7).



Slika 8.7: Pet izabranih primjera s $2, 3, 4, 5$ i 10 pravaca

Kako bismo osigurali homogenost podataka u okolini pravaca vidljivih u pravokutniku Δ, na svakom pravcu ℓ_j generirat ćemo približno jednak broj uniformno distribuiranih točaka lokalne gustoće $\rho = \frac{|\pi_j|}{|\ell_j|} = 21$, gdje je $|\ell_j|$ duljina vidljivog dijela pravca ℓ_j u pravokutniku Δ, a $|\pi_j|$ broj generiranih točaka. Nakon toga, svakoj tako izabranoj točki na pravcu dodajemo slučajnu pogrešku iz binormalne distribucije s očekivanjem $0 \in \mathbb{R}^2$ i kovarijacijskom matricom $\sigma^2 I$, $\sigma^2 = .01$, gdje je I jedinična matrica drugog reda. Tako smo definirali skup podataka $\mathcal{A}$ i njegovu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$, gdje je π_j skup podataka koji potječe od pravca ℓ_j (vidi Sliku 8.7). Primijetite da pravac ℓ_j ne mora biti centar klastera π_j. Na taj način za svaki par (k, σ^2) definirat ćemo po 100 skupova ovakvih pravaca i pripadnih podataka na kojima će se testirati naša metoda.

8.5.2 Algoritam k-najbližih pravaca (KCL)

Najpoznatija metoda za traženje k-GOPart u slučaju poznatog broja k pravaca je KCG-algoritam (točka 8.1, str. 132) pri čemu su u ovom slučaju G-klaster-centri pravci pa ćemo i algoritam zvati *Algoritam k-najbližih pravaca* (k-closest line (KCL)) (vidi također [8, 102, 113]).

KCL-algoritam može započeti početnim centar-pravcima ili početnom particijom. Posebnu pozornost treba posvetiti izboru početne aproksimacije za KCL-algoritam.

Jedna mogućnost izbora početne aproksimacije u slučaju poznatog broja k, kao što je navedeno u točki 8.1.1, str. 131, je primjenom algoritma DIRECT pronaći dovoljno dobre početne centar-pravce. Pri tome te pravce nije dobro tražiti u eksplicitnom ili implicitnom zapisu jer bi u tom slučaju bilo teško odrediti prihvatljivu konačnu domenu parametara koja je potrebna za algoritam DIRECT.

Zato ćemo početne pravce tražiti u Hesseovom normalnom obliku

$$p_j(\vartheta_j, \delta_j) \equiv x \cos \vartheta_j + y \sin \vartheta_j - \delta_j, \quad j = 1, \ldots, k.$$

Kako je u ovom slučaju LS-udaljenost točke $a^i = (x_i, y_i)^T \in \mathcal{A}$ do pravca

$\ell_j(\vartheta_j, \delta_j)$ zadana s:

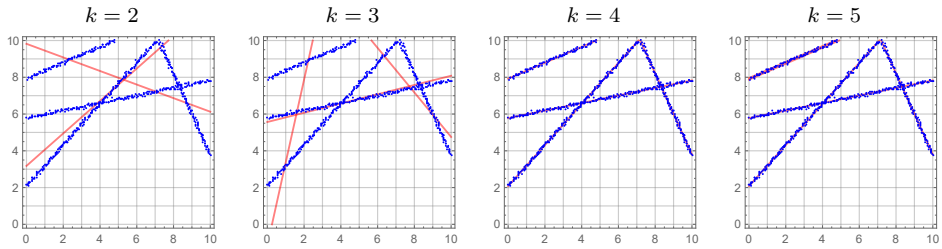
$$\mathfrak{D}(p_j(\vartheta_j, \delta_j), a^i) = (x\cos\vartheta_j + y\sin\vartheta_j - \delta_j)^2,$$

početnu aproksimaciju potražit ćemo rješavanjem `GOP` (usporedi (8.43))

$$\underset{\substack{\boldsymbol{\vartheta}\in[0,\pi]^k,\\ \boldsymbol{\delta}\in[0,M]^k}}{\operatorname{argmin}} G(\boldsymbol{\vartheta}, \boldsymbol{\delta}), \quad G(\boldsymbol{\vartheta}, \boldsymbol{\delta}) = \sum_{i=1}^{m} \min_{1\le j\le k} (x_i\cos\vartheta_j + y_i\sin\vartheta_j - \delta_j)^2 \tag{8.44}$$

uz primjenu algoritma `DIRECT`. Domena parametra ϑ_j je $[0,\pi]$, a domenu parametra δ_j možemo definirati s $[0, M]$, $M := \max_{i=1,\dots,m} \|a^i\|_2$. Rješenje ovog problema je k pravaca $\ell_j(\hat{\vartheta}_j, \hat{\delta}_j) \equiv \cos\hat{\vartheta}_j x + \sin\hat{\vartheta}_j y - \hat{\delta}_j = 0$, koje ćemo iskoristiti kao početnu aproksimaciju u `KCL`-algoritmu.

Primjer 8.4. *Za primjer s 4 pravca prikazana na Slici 8.7 pomoću algoritma `DIRECT` pronađena je početna aproksimacija s $k = 2, 3, 4, 5$ pravaca u Hesseovom normalnom obliku. S ovim aproksimacijama KCL algoritam dao je rezultate prikazane na Slici 8.8. Ostaje pitanje kako između ovih rješenja detektirati particiju s najprihvatljivijim brojem klastera.*



Slika 8.8: Optimalne k-particije ($k = 2, 3, 4, 5$) dobivene `KCL`-algoritmom uz početnu aproksimaciju dobivenu primjenom 20 iteracija algoritma `DIRECT` na rješavanje `GOP` (8.44)

8.5.3 Inkrementalni algoritam

Ako broj klastera u particiji nije unaprijed poznat, možemo potražiti optimalne particije s $k = 1, 2, \dots, k_{max}$ klastera i nakon toga u cilju prepoznavanja najprihvatljivije particije pokušati primijeniti neki od poznatih indeksa prilagođenih za slučaj pravaca-centara. U tu svrhu u literaturi se također može pronaći nekoliko modifikacija dobro poznatog inkrementalnog algoritma prilagođenog za slučaj centara-pravaca (vidi primjerice [8, 102, 113]).

Inkrementalni algoritam za rješavanje MLD problema započinje određivanjem najboljeg TLS-pravca $\tilde{\ell}_1$ skupa $\mathcal{A}$ sukladno točki 6.1, str. 99.

Nadalje, ako su pravci $\tilde{\ell}_1, \ldots, \tilde{\ell}_k$ poznati, sljedeći pravac $\tilde{\ell}_{k+1}$ odredit ćemo rješavanjem sljedećeg GOP:

$$\underset{\mathbf{p}\in\mathfrak{U}}{\operatorname{argmin}}\, \phi(\mathbf{p}), \quad \phi(\mathbf{p}) = \sum_{i=1}^{m} \min\{d_k^{(i)}, \mathfrak{D}(a^i, \ell(\mathbf{p}))\}, \tag{8.45}$$

gdje je: $d_k^{(i)} = \min\{\mathfrak{D}(a^i, \tilde{\ell}_1), \ldots, \mathfrak{D}(a^i, \tilde{\ell}_k)\}$.

Slično kao u točki 8.5.2, str. 147 zbog zahtjeva algoritma DIRECT za konačnom domenom parametara, sljedeći pravac $\tilde{\ell}_{k+1}$ tražit ćemo u Hesseovom normalnom obliku rješavajući sljedeći GOP:

$$\underset{\substack{\vartheta\in[0,\pi],\\ \delta\in[0,M]}}{\operatorname{argmin}}\, \Phi(\vartheta, \delta), \quad \Phi(\vartheta, \delta) = \sum_{i=1}^{m} \min\{d_k^{(i)}, \mathfrak{D}_H(a^i, \ell(\vartheta, \delta))\}, \tag{8.46}$$

gdje je: $\mathfrak{D}_H(a^i, \ell(\vartheta, \delta))\} = (x_i \cos\vartheta + y_i \sin\vartheta - \delta)^2$ i $M = \max\limits_{i=1,\ldots,m} \|a^i\|$.

Na skup pravaca $\{\tilde{\ell}_1, \ldots, \tilde{\ell}_k, \tilde{\ell}_{k+1}\}$ nakon toga treba primijeniti KCL-algoritam.

Inkrementalni algoritam možemo zaustaviti korištenjem kriterija (8.13) pri čemu je $\mathfrak{D}_1$-udaljenost točke $a^i = (x_i, y_i)$ do pravca $\ell(\mathbf{p})$ definirana s $\mathfrak{D}_1(a^i, \ell(\mathbf{p})) = |\xi x_i + \eta y_i + \zeta|$ (Lema 8.1, str. 140), a ϵ-gustoća $\epsilon(\mathcal{A})$ skupa $\mathcal{A}$ određena je u skladu s Definicijom 8.2.

Algoritam 3 (Inkrementalni algoritam za prepoznavanje pravaca)

Input: $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, m\} \subset \Delta \subset \mathbb{R}^2$;

1: Stavi $\epsilon_B = .005$ i odredi $MinPts$ i $\epsilon(\mathcal{A})$ sukladno Definiciji 8.2;
2: Odredi TLS-pravac ℓ_1 skupa $\mathcal{A}$, sljedeći pravac ℓ_2 prema (8.46) i stavi $k = 2$;
3: Na centar-pravce ℓ_1, ℓ_2 primijeni KCL-algoritam i dobivenu particiju označi s $\hat{\Pi}^{(2)} = \{\hat{\pi}_1(\hat{\ell}_1), \hat{\pi}_2(\hat{\ell}_2)\}$;
4: Odredi QD$[\hat{\pi}_j]$, $j = 1, 2$ kao 90% kvantil skupa $V(\hat{\pi}_j)$ i QD$[\hat{\Pi}^{(2)}]$;
5: **while** QD$[\hat{\Pi}^{(k)}] > \epsilon$ **do**
6: Prema (8.46), odredi sljedeći centar-pravac ℓ_{k+1}
7: Na pravce $\hat{\ell}_1, \ldots, \hat{\ell}_k, \ell_{k+1}$ primijeni KCL-algoritam i dobivenu particiju označi s $\hat{\Pi}^{(k+1)} = \{\hat{\pi}_1(\hat{\ell}_1), \ldots, \hat{\pi}_{k+1}(\hat{\ell}_{k+1})\}$, a pravce s $\hat{L} = \{\hat{\ell}_1, \ldots, \hat{\ell}_{k+1}\}$;
8: Odredi QD$[\hat{\Pi}^{(k+1)}]$ i stavi $k := k + 1$;
9: **end while**

Output: $\{\hat{\ell}_1, \ldots, \hat{\ell}_k\}$.

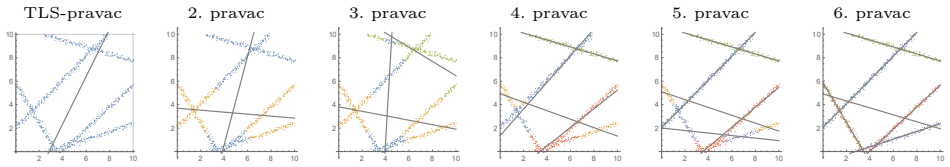
Primjer 8.5. *Inkrementalni algoritam primijenjen je na primjer s pet pravaca prikazanih na Slici 8.7. Rezultati izvođenja prvih* 6 *iteracija prikazani su na Slici 8.9.*

Budući da je ϵ-gustoća $\epsilon(\mathcal{A}) = 0.319$, algoritam se zaustavlja na particiji $\Pi^{(6)}$ za koju vrijedi $\texttt{QD}[\pi] < 0.319$ za svaki $\pi \in \Pi^{(6)}$.

Vektor gustoće za ovu particiju je:

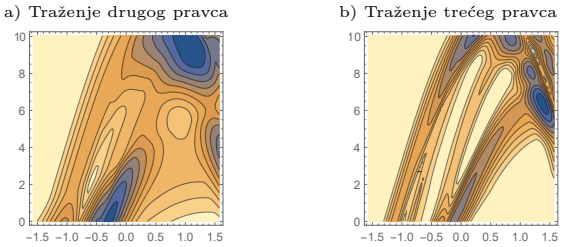
$$\rho(\Pi^{(6)}) = (20.71,\ 2.21,\ 21.04,\ 20.98,\ 20.22,\ 19.82)^T.$$

Kako je $\rho_2 = 2.21 < \frac{MinPts}{2\epsilon(\mathcal{A})} = 9.39$, ovaj pravac se ispušta i ostaje kao rješenje 5*-particija.*



Slika 8.9: Rezultati izvođenja prvih 6 iteracija inkrementalnog algoritma na primjeru s pet pravaca iz Slike 8.7

`ContourPlot` minimizirajuće funkcije iz (8.46) prilikom traženja drugog i trećeg pravca prikazan na Slici 8.10 pokazuje da problem (8.46) može imati više točaka lokalnog i globalnog minimuma.

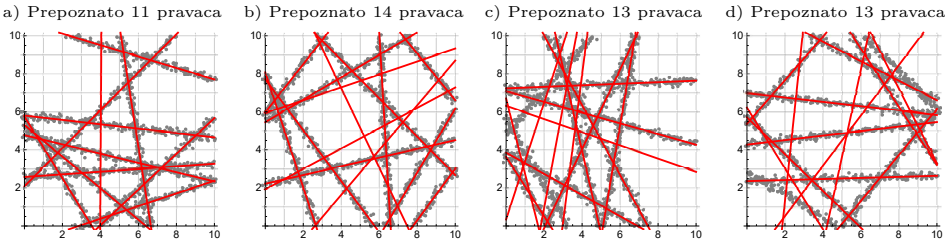


Slika 8.10: `ContourPlot` funkcije iz (8.46) za MLD problem iz Primjera 8.5

Particija	π_1	π_2	π_3	π_4	π_5	π_6	QD[Π]
$\Pi^{(2)}$	2.84	1.91	—	—	—	—	2.84
$\Pi^{(3)}$	1.32	1.61	1.50	—	—	—	1.61
$\Pi^{(4)}$	0.64	0.92	0.17	0.87	—	—	0.92
$\Pi^{(5)}$	0.23	0.56	0.18	0.43	0.72	—	0.56
$\Pi^{(6)}$	0.16	0.12	0.17	0.14	0.17	0.15	0.17

Tablica 8.1: Vrijednosti QD iz Primjera 8.5

Primjer 8.6. *Algoritam 3 ilustrirat ćemo na nekoliko primjera podataka koji potječu od po* 10 *pravaca. Na Slici 8.11 prikazani su rezultati izvođenja Inkrementalnog algoritam 3 (crvene linije).*



Slika 8.11: Rezultati implementacija Inkrementalnog algoritma 3

Output Inkrementalnog algoritma 3 u slučaju skupa podataka koji potječe od 10 pravaca vidljivih na Slici 8.11a sastoji se od 11 pravaca. Pri tome, 10 pravaca podudara se s originalnim pravcima od kojih su nastali podaci, a jedanaesti pravac nastao je slučajno.

Slično, output Inkrementalnog algoritma 3 u slučaju skupa podataka koji potječe od 10 pravaca vidljivih na Slici 8.11b sastoji se od 14 pravaca. Pri tome, 10 pravaca podudara se s originalnim pravcima od kojih su nastali podaci, a ostali su nastali slučajno.

Output Inkrementalnog algoritma 3 u slučaju skupa podataka koji potječe od 10 pravaca prikazanih na Slici 8.11c, odnosno na Slici 8.11d, sastoji se od po 13 pravaca među kojima se ne nalazi svih 10 pravaca od kojih su podaci nastali.

8.5.4 Traženje MAPart i prepoznavanje pravaca

Budući da je Inkrementalni algoritam 3 lokalnog karaktera, u pravilu će proizvesti više pravaca nego što se može očekivati s obzirom na konfiguraciju

skupa $\mathcal{A}$ (kao što se može vidjeti u Primjeru 8.6 pa se između dobivenih particija često ne nalazi i MAPart. To ukazuje na potrebu eliminacije nekih centar-pravaca iz ove particije.

Direktna primjena modificiranog CH ili DB-indeksa (vidi točku 8.3, str.135) pokazuje se neučinkovitom kao što pokazuje sljedeći primjer.

Primjer 8.7. *Na primjeru k-LOPart dobivenih KCL-algoritmom prikazanih na Slici 8.8 i k-LOPart dobivenih Inkrementalnim algoritmom 3 prikazanih na Slici 8.9 usporedit ćemo DBG-indeks s GO-indeksom. Pri tome, udaljenost $d(\ell_j, \ell_s)$ određujemo sukladno Definiciji 8.30, str. 142. U Tablici 8.2 može se primijetiti da je GO-indeks znatno pouzdaniji od DBG-indeksa.*

Indeks	Primjer iz Slike 8.8				Primjer iz Slike 8.9			
	$\Pi^{(2)}$	$\Pi^{(3)}$	$\Pi^{(4)}$	$\Pi^{(5)}$	$\Pi^{(2)}$	$\Pi^{(3)}$	$\Pi^{(4)}$	$\Pi^{(5)}$
DBG	2.28	2.49	2.40	$\approx 10^5$	2.28	2.37	2.18	1.97
GO	24.89	54.28	0.15	34.41	24.89	69.66	0.15	78.19

Tablica 8.2: Usporedba DBG i GO-indeksa na k-LOPart dobivenih KCL-algoritmom i Inkrementalnim algoritmom 3

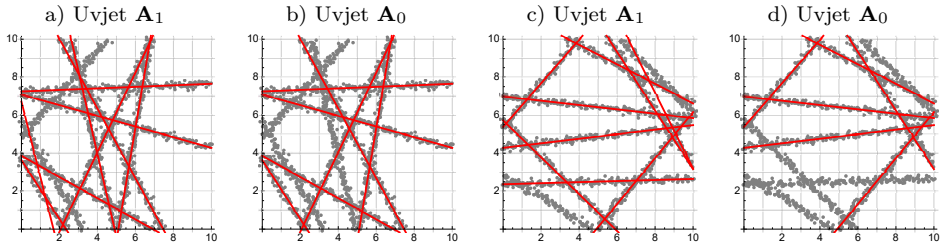
GO index za izbor MAPart testiran je u [91] na 100 različitih skupova podataka iz $\Delta = [0, 10] \times [0, 10]$ generiranih kao u točki 8.5.1, str. 146, a koji potječu od 2 do 5 pravaca. Za svih 100 skupova podataka stupanj prepoznavanja bio je 50%, a prosječno CPU-vrijeme 11.6 sec, od čega je 9.7 sec bilo je potrebno za DIRECT, a 1.5 sec za KCL-algoritam.

Primjena novog pristupa

U nastavku ćemo pokazati primjenu novog pristupa opisanog u točki 8.3.3, str. 137. U primjeru prikazanom na Slici 8.11a nakon provjere uvjeta $\mathbf{A}_1$ i ispuštanja jednog pravca koji ne ispunjavaju taj uvjet, dobivamo MAPart s 10 pravaca. Slično, i u primjeru prikazanom na Slici 8.11b nakon provjere uvjeta $\mathbf{A}_1$ i ispuštanja četiri pravca koji ne ispunjavaju taj uvjet, dobivamo MAPart s 10 pravaca.

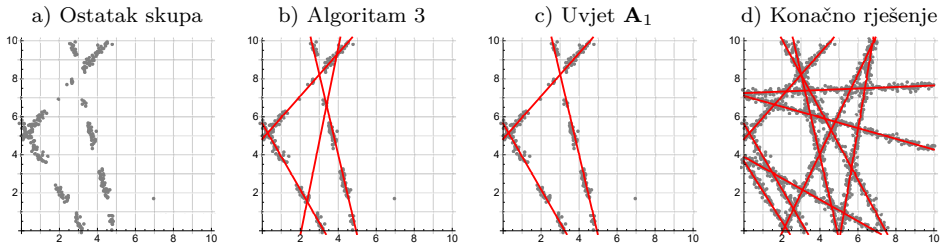
Output Algoritma 3 u slučaju skupa podataka koji potječe od 10 pravaca vidljivih na Slici 8.11c particija je s 13 pravaca. Nakon provjere uvjeta $\mathbf{A}_1$, ostaje 9 pravaca (vidi Sliku 8.12a), a dodatnom provjerom uvjeta $\mathbf{A}_0$ ostaje 7 pravaca (vidi Sliku 8.12b).

Slično, output Algoritma 3 u slučaju skupa podataka koji potječe od 10 pravaca vidljivih na Slici 8.11d je particija s 13 pravaca. Nakon provjere



Slika 8.12: Eliminacija neprihvatljivih centar-pravaca korištenjem uvjeta $\mathbf{A}_1$ i $\mathbf{A}_0$

uvjeta $\mathbf{A}_1$, ostaje 9 pravaca (vidi Sliku 8.12c), a dodatnom provjerom uvjeta $\mathbf{A}_0$ ostaje 6 pravaca (vidi Sliku 8.12d).



Slika 8.13: Ponovno pokretanje Algoritma 3 na ostatku skupa i konačno rješenje

Na ostatku skupa podataka koji se pojavio na Slici 8.12b i Slici 8.12d ponovo treba provesti Algoritam 3. Kako to izgleda na ostatku skupa iz Slike 8.12b prikazano je na Slici 8.13a-c. Na taj način i u ovom slučaju potpuno smo rekonstruirali svih deset pravaca (Slika 8.13d). Slično se može postupiti i u primjeru prikazanom na Slici 8.11d nakon eliminacija prikazanih na Slici 8.12c i Slici 8.12d.

8.6 Kružnica kao reprezentant skupa podataka

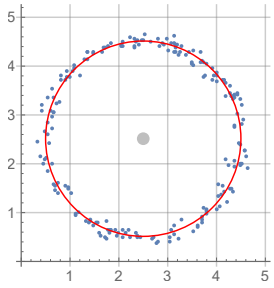
Kao što smo naveli na početku ovog poglavlja, kružnicu sa središtem u točki $S = (p, q)^T$ radijusa r možemo definirati na sljedeće načine

$$(i)\, K(S, r) = \{(x, y)^T \in \mathbb{R}^2 \colon (x - p)^2 + (y - q)^2 = r^2\},$$
$$(ii)\, K(S, r) = \{S + r(\cos t, \sin t)^T \colon t \in [0, 2\pi]\}.$$

Skup $\mathcal{A}$, smješten u pravokutnik $\Delta = [a, b] \times [c, d] \subset \mathbb{R}^2$, koji potječe od kružnice $K(S, r)$ sa središtem u točki $S = (p, q)^T$ radijusa $r > 0$ možemo konstruirati na sljedeći način. Najprije na segmentu $[0, 2\pi]$ odredimo

$m > 2$ uniformno distribuiranih brojeva $t_i \in [0, 2\pi]$. Nakon toga, svakoj točki $T_i = S + r(\cos t_i, \sin t_i)^T$ na kružnici dodajemo slučajnu pogrešku iz binormalne distribucije s očekivanjem $0 \in \mathbb{R}^2$ i kovarijacijskom matricom $\sigma^2 I$, $\sigma^2 = .01$, gdje je I jedinična matrica drugog reda. Na taj način osigurali smo da podaci koji dolaze od neke kružnice zadovoljavaju „svojstvo homogenosti", tj. podaci su pretežno homogeno rasuti oko te kružnice. Sukladno Definiciji 8.1, str. 130 skup $\mathcal{A}$ ima lokalnu gustoću $\rho(\mathcal{A}) = \frac{m}{2r\pi}$.

Primjer 8.8. *U kvadratu* $[0, 10]^2 \subset \mathbb{R}^2$ *izabrana je kružnica* $K(S, r)$ *sa središtem u točki* $S = (2.5, 2.5)^T$ *radijusa* $r = 2$. *U okolini kružnice* K, *generirat ćemo* $m = 200$ *slučajnih točaka na prethodno opisan način (vidi Sliku 8.14).*



Slika 8.14: Podaci koji potječu od kružnice

```
In[1]:= A ={}; S = {2.5,2.5}; r = 2; m = 200; sigma = .01;
        Do[
        t = RandomReal[{0, 2 Pi}];
        T = S + r {Cos[t], Sin[t]};
        A=Append[A,RandomReal[
             MultinormalDistribution[T, sigma IdentityMatrix[2]], {1}][[1]]
                               ];
          ,{ii,m}]
```

Može se postaviti inverzni problem, tj. problem prepoznavanja kružnice na osnovi poznavanja skupa podataka $\mathcal{A}$. U cilju rješavanja ovog problema možemo razmatrati problem traženje kružnice-centra kao najboljeg reprezentanta skupa podataka $\mathcal{A}$.

Najprije treba dobro definirati neku kvazimetričku funkciju koja će dati udaljenost točke $a \in \mathcal{A}$ do kružnice $K(S, r)$. Za to u literaturi postoje

sljedeći pristupi [22, 55, 65, 90]:

$$\mathfrak{D}_2(a,K) = |\sqrt{(x_i-p)^2+(y_i-q)^2}-r|^2 \quad \text{[Total Least Squares]} \tag{8.47}$$

$$\mathfrak{D}_1(a,K) = |\sqrt{(x_i-p)^2+(y_i-q)^2}-r| \quad \text{[Orthogonal Distances]} \tag{8.48}$$

$$\mathfrak{D}(a,K) = \left((x_i-p)^2+(y_i-q)^2-r^2\right)^2 \quad \text{[Algebraic Criterion]} \tag{8.49}$$

Upravo ova posljednja mogućnost (algebarska udaljenost) najviše je prisutna u literaturi i primjenama. Razlog za to treba tražiti u činjenici da je funkcija $\mathfrak{D}$ derivabilna po svim svojim varijablama.

Traženje najboljeg reprezentanta skupa $\mathcal{A}$ kao kružnice-centra može se provesti rješavanjem sljedećeg GOP:

$$\underset{(p,q)^T\in\Delta,\, r\in[0,R]}{\operatorname{argmin}} F(p,q,r), \quad F(p,q,r) = \sum_{i=1}^{m} \mathfrak{D}(a^i, K((p,q)^T, r)), \tag{8.50}$$

gdje za R možemo uzeti $R = \frac{1}{2}\max\{b-a, d-c\}$. Za rješavanje ovog problema može se primijeniti algoritam DIRECT za globalnu optimizaciju [46, 97].

U slučaju primjene algebarske $\mathfrak{D}$ udaljenosti može se primijeniti i Newtonova ili kvazi-Newtonova metoda ili metoda jednostavnih iteracija, ali tada moramo imati dobru početnu aproksimaciju $\hat{K}(\hat{S},\hat{r})$ tražene kružnice, što je u ovom slučaju lako postići. Naime, za $\hat{S}$ možemo izabrati centroid skupa podataka $\mathcal{A}$, a $\hat{r}$ tada možemo dobiti korištenjem svojstva aritmetičke sredine:

$$\sum_{i=1}^{m}\left(\|\hat{S}-a^i\|^2-r^2\right)^2 \geq \sum_{i=1}^{m}\left(\|\hat{S}-a^i\|^2-\tfrac{1}{m}\sum_{i=1}^{m}\|\hat{S}-a^i\|^2\right)^2. \tag{8.51}$$

Na taj način odredili smo početnu aproksimaciju $\hat{K}(\hat{S},\hat{r})$ tražene kružnice:

$$\hat{r}^2 = \tfrac{1}{m}\sum_{i=1}^{m}\|\hat{S}-a^i\|^2, \quad \hat{S} = \tfrac{1}{m}\sum_{i=1}^{m}a^i. \tag{8.52}$$

Primijetite da početnu aproksimaciju $\hat{K}(\hat{S},\hat{r})$ možemo zapisati kao:

$$\hat{r}^2 = \text{mean}[\{\|\hat{S}-a^i\|^2 \colon i=1,\ldots,m\}].$$

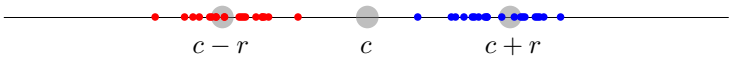
Zadatak 8.2. *Provjerite da je dobra aproksimacija radijusa ove kružnice zadana s* $\hat{r} \approx \sqrt{Tr(Cov(\mathcal{A}))}$.

Primjer 8.9. *Kružnica u* $\mathbb{R}$ *sa središtem u točki* $c \in \mathbb{R}$ *radijusa* $r > 0$ *definirana je s:*

$$K(c,r) = \{x \in \mathbb{R} \colon |c - x| = r\} = \{c - r, c + r\},$$

i kao što se vidi, sastoji se samo od dviju točaka.

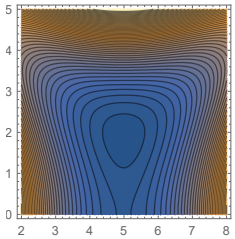
Skup $\mathcal{A}$ od $m = 40$ podataka koji potječe od ovakve kružnice $K(5,2) = \{3,7\}$ može se odrediti na sljedeći način (crvene i plave točkice na Slici 8.15):



Slika 8.15: Skup podataka $\mathcal{A}$ koji potječe od kružnice $K(c,r) \subset \mathbb{R}$

```
In[1]:= broj=20; a1=3; a2=7; sigma=.5; SeedRandom[13]
        pod1=RandomReal[NormalDistribution[a1, sigma], broj];
        pod2=RandomReal[NormalDistribution[a2, sigma], broj];
        A=Join[pod1,pod2]
```

U ovom je slučaju `GOP` (8.50) optimizacijski problem u $\mathbb{R}^2$ (dvije nezavisne varijable), a u slučaju primjene algebarske $\mathfrak{D}$ udaljenosti `ContourPlot` minimizacijske funkcije prikazan je na Slici 8.14



Slika 8.16: `ContourPlot` minimizacijske funkcije (8.50) iz Primjera 8.9

Primjenom *Mathematica*-naredbe `NMinimize` dobivamo $c^\star = 4.92314$, $r^\star = 2.17341$, pri čemu je odgovarajuća vrijednost minimizacijske funkcije (8.50), $F(c^\star, r^\star) = 211.303$.

Primjer 8.10. *Na osnovi kružnice* $K(S(p,q),r)$, $S = (2.5, 2.5)^T$, $r = 2$ *u Primjeru 8.8 konstruirali smo skup* $\mathcal{A}$ *s* $m = 200$ *podataka koji potječe od ove kružnice (vidi također Sliku 8.14, str. 154).*

Na osnovi skupa podataka $\mathcal{A}$ algoritmom `DIRECT` [97] možemo pronaći rješenje `GOP` (8.50). Dobivamo $(p^\star, q^\star, r^\star) = (2.510, 2.509, 1.999)$ i vrijednost funkcije cilja $F(p^\star, q^\star, r^\star) = 28.16$.

Primjenom neke lokalno optimizacijske metode (primjerice Newtonove ili Nelder-Meadove) također možemo dobiti rješenje jer raspolažemo s kvalitetnom početnom aproksimacijom. Uz početnu aproksimaciju:

$$\hat{S} = (2.603, 2.626)^T, \quad \hat{r} = 2.012, \qquad \hat{F} = 74.026$$

primjenom *Mathematica*-modula `FindMinimum[]` dobivamo:

$$(p^\star, q^\star, r^\star) = (2.496, 2.502, 2.012), \qquad F^\star = 28.16.$$

Zadatak 8.3. *Zadan je skup podataka iz Primjera 8.8. Korištenjem* Total Least Squares *kvazimetričke funkcije* (8.47) *točke $a \in \mathcal{A}$ do kružnice $K(S, r)$ rekonstruirajte kružnicu primjenom neke lokalno optimizacijske metode. Kako u ovom slučaju možete odabrati kvalitetnu početnu aproksimaciju za centar $\hat{S}$ i radijus $\hat{r}$ kružnice? Provjerite je li ova kvazimetrička funkcija osjetljiva na prisutnost "outliersa" među podacima.*

Uputa: U programskom sustavu Mathematica *možete koristiti* `FindMinimum`*, gdje možete izabrati Newtonovu metodu, Quasi-Newtonovu metodu ili neku drugu metodu.*

Zadatak 8.4. *Skup podataka iz Primjera 8.8 dopunite s* 10% *outliers. Korištenjem* Orthogonal Distances *udaljenosti* (8.48) *točke $a \in \mathcal{A}$ do kružnice $K(S, r)$ rekonstruirajte kružnicu primjenom neke lokalno optimizacijske metode. Kako u ovom slučaju možete odabrati kvalitetnu početnu aproksimaciju za centar $\hat{S}$ i radijus $\hat{r}$ kružnice?*

Uputa: U programskom sustavu Mathematica *možete koristiti* `FindMinimum` *s odgovarajućim izborom metode.*

Primjedba 8.3. *Za metode prepoznavanja jednog ili više kružnih lukova u ravnini bit će važno znati projekciju točke na kružnicu (vidi Sliku 8.17) i projekciju točaka kružnice na segment.*

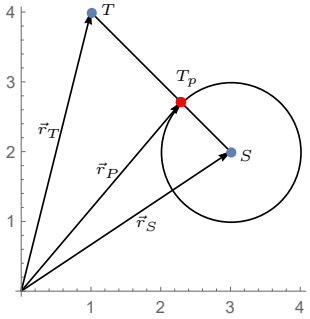
- *Radij-vektor projekcije P točke $T \in \mathbb{R}^2$ na kružnicu K zadan je s:*

$$\vec{r}_P = (1 - \lambda_T)\vec{r}_S + \lambda_T \vec{r}_T, \tag{8.53}$$

gdje je $\lambda_T = \frac{r}{\|\vec{r}_T - \vec{r}_S\|_2}$. Formula (8.53) *slijedi iz $\vec{r}_P - \vec{r}_S = \lambda_T(\vec{r}_T - \vec{r}_S)$;*

- *Neka su* $a_p^i = (u_i, v_i)^T$ *projekcije točaka* $a^i \in \mathcal{A}$ *na kružnicu-centar* $K(S,r)$. *Točke* a_0^i *preslikavamo na segment* $[0, 2\pi]$ *formulom:*

$$t_i = \begin{cases} \arctan \frac{v_i - q}{u_i - p}, & ako\ (u_i \geq p\,\&\, v_i \geq q)\ ili\ (u_i \leq p\,\&\, v_i \geq q) \\ \arctan \frac{v_i - q}{u_i - p} + \pi, & ako\ (u_i \leq p\,\&\, v_i \leq q)\ ili\ (u_i \geq p\,\&\, v_i \leq q) \end{cases}. \quad (8.54)$$



Slika 8.17: Projekcija točke T na kružnicu $K(S,r)$

8.7 Prepoznavanje više kružnica u ravnini

Neka je $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\} \subset \Delta$, $\Delta = [a,b] \times [c,d]$ skup podataka koji potječe od više unaprijed nepoznatih kružnica u ravnini, a koje treba prepoznati ili rekonstruirati. Podaci se mogu umjetno konstruirati na način opisan za podatke koji potječu od jedne kružnice. Pretpostavljamo da podaci koji dolaze od neke kružnice zadovoljavaju „svojstvo homogenosti", tj. pretpostavljamo da su podaci pretežno homogeno rasuti oko te kružnice (vidi Definiciju 8.1).

U literaturi se mogu pronaći metode za rješavanje ovog problema uz primjenu Houghovih transformacija [31]. U [90] ovaj problem promatra se kao problem grupiranja, pri čemu su centri klastera kružnice.

Traženje k-`GOPart` $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$, gdje je centar klastera $\pi_j^\star$ kružnica $K_j^\star(S_j^\star, r_j^\star)$ sa središtem u točki $S_j^\star$ i radijusom $r_j^\star$, svodi se na traženje optimalnih parametara $(p_j^\star, q_j^\star, r_j^\star)$, $j = 1, \ldots, k$, koji su rješenje sljedećeg `GOP` (usporedi s (3.40)):

$$\underset{(p,q)^T \in \Delta^{2k},\, r \in [0,R]^k}{\operatorname{argmin}} F(p,q,r),\ F(p,q,r) = \sum_{i=1}^{m} \min_{1 \leq j \leq k} \mathfrak{D}(a^i, K_j(p_j, q_j, r_j)), \quad (8.55)$$

gdje je $R = \frac{1}{2}\min\{b-a, d-c\}$, a $\mathfrak{D}(a^i, K_j(p_j, q_j, r_j))$ predstavlja udaljenost točke a^i do kružnice K_j. O načinu definiranja metričke funkcije $\mathfrak{D}$ pisali smo

u točki 8.6, str. 153. Kao što smo tamo spomenuli, koristit ćemo algebarsku udaljenost

$$\mathfrak{D}(a^i, K(S,r)) = (\|S - a^i\|^2 - r^2)^2. \tag{8.56}$$

8.7.1 Prilagođavanje *k*-means algoritma na slučaj kružnica-centara

Problem traženja k-`GOPart` s kružnicama kao klaster-centrima u slučaju poznatog broja k kružnica možemo rješavati primjenom `KCG`-algoritma (točka 8.1, str. 132) pri čemu su u ovom slučaju `G`-klaster-centri kružnice pa ćemo i algoritam zvati algoritam k-najbližih kružnica (`KCC`) (vidi [55, 90]).

Algoritam može započeti početnim centar-kružnicama ili početnom particijom. Posebnu pozornost treba obratiti izboru početne aproksimacije za `KCC`-algoritam.

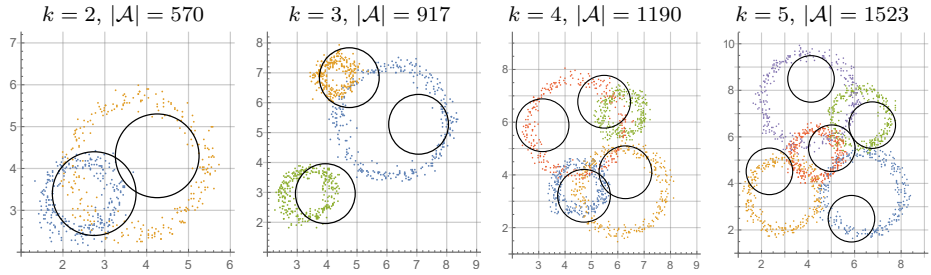
U slučaju poznatog broja k, algoritam ćemo započeti *Korakom A* tako da primjenom manjeg broja iteracija `DIRECT` algoritma potražimo aproksimaciju k-optimalne particije skupa $\mathcal{A}$

$$\underset{(p,q)^T \in \Delta^{2k}}{\operatorname{argmin}} \Phi(p,q), \quad \Phi(p,q) = \sum_{i=1}^{m} \min_{1 \le j \le k} \|a^i - (p_j, q_j)^T\|_2^2, \tag{8.57}$$

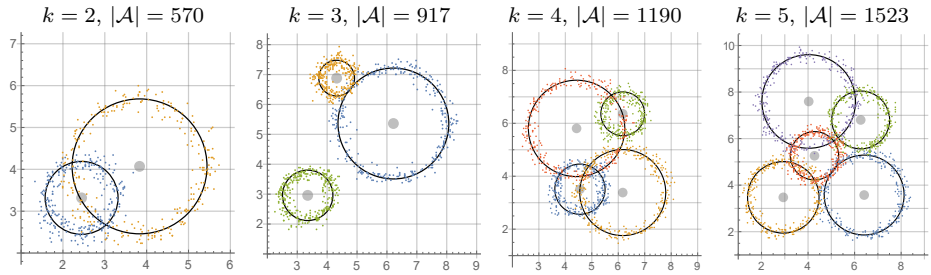
a nakon toga početne kružnice definiramo kao jedinične kružnice sa središtima u dobivenim centrima.

Najbolje kružnice-reprezentante u *Koraku B* možemo tražit nekom lokalno optimizacijskom metodom (Quasi-Newton, Nelder-Meade) jer raspolažemo dobrom početnom aproksimacijom (8.52).

Primjer 8.11. *Na Slici 8.18 prikazani su podaci koji potječu od dviju, triju, četiriju i pet kružnica. Za svaki skup podataka određene su odgovarajuće jedinične kružnice kao početne aproksimacije za* `KCC`*-algoritam. Rezultati izvođenja KCC algoritma prikazani su na Slici 8.19.*



Slika 8.18: Početne aproksimacije dobivene algoritmom `DIRECT`.



Slika 8.19: Optimalne particije dobivene `KCC`-algoritmom

8.7.2 Prilagođavanje inkrementalnog algoritma na slučaj kružnica-centara

Ako broj kružnica od kojih potječu podaci nije unaprijed poznat, potražit ćemo optimalne particije s $k = 2, 3, \ldots, \kappa$ klastera prilagođavanjem inkrementalnog algoritma i između njih pokušati odrediti `MAPart`.

Algoritam započinje izborom početne kružnice-centra $\hat{K}_1(\hat{S}_1, \hat{r}_1)$. Aproksimacija sljedeće kružnice-centra $\hat{K}_2(\hat{S}_2, \hat{r}_2)$ dobiva se rješavanjem `GOP`:

$$\underset{(p,q)^T \in [\alpha,\beta],\, r \in [0,R]}{\operatorname{argmin}} \sum_{i=1}^{m} \min\{\mathfrak{D}(a^i, \hat{K}_1(S_1, r_1)), \mathfrak{D}(a^i, K(S(p,q), r))\}. \quad (8.58)$$

Nakon toga na kružnice-centre $\hat{K}_1, \hat{K}_2$ primjenjuje se `KCC`-algoritam i time se dobiva lokalno optimalna 2-particija $\Pi^{(2)}$. Primijetite da je prilikom rješavanja problema (8.58) dovoljno izvesti manji broj koraka optimizacijskog algoritma `DIRECT` i tako dobiti dovoljno dobru početnu aproksimaciju za `KCC`-algoritam koji nakon toga daje lokalno optimalnu 2-particiju $\Pi^{(2)}$ s kružnicama-centrima $(K_1^\star, K_2^\star)$.

Općenito, poznavajući k kružnica-centara $\hat{K}_1, \ldots, \hat{K}_k$, sljedeću kružnicu-centar K_{k+1} potražit ćemo rješavanjem sljedećeg GOP

$$\underset{(p,q)^T \in [\alpha,\beta],\, r \in [0,R]}{\operatorname{argmin}} \sum_{i=1}^{m} \min\{\delta_k^i, \mathfrak{D}(a^i, K(S(p,q),r))\}, \tag{8.59}$$

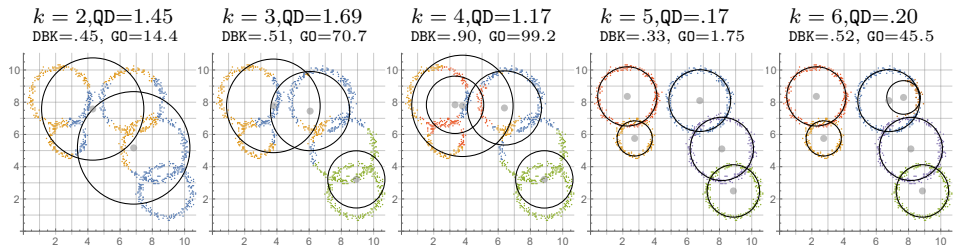
gdje je $\delta_k^i = \min\limits_{1 \le s \le k} \mathfrak{D}(a^i, K(S_s, r_s)$. Nakon toga, primjenom KCC-algoritma dobivamo k-LOPart $\Pi^\star = \{\pi_1^\star, \ldots, \pi_{k+1}^\star\}$ s kružnicama-centrima $K_1^\star, \ldots, K_{k+1}^\star$.

Inkrementalni algoritam zaustavlja se korištenjem kriterija (8.13), pri čemu je $\mathfrak{D}_1$-udaljenost točke $a \in \pi_j^\star$ do odgovarajuće kružnice-centra $K_j^\star(S_j^\star, r_j^\star)$ dana s:

$$\mathfrak{D}_1(a, K_j^\star(S_j^\star, r_j^\star)) = |\|S_j^\star - a\| - r_j^\star|. \tag{8.60}$$

Primjer 8.12. *Neka je $\mathcal{A}$ skup podataka zadan kao na Slici 8.20. Primjenom inkrementalnog algoritma pronađimo lokalno optimalne $2, 3, 4, 5$-particije.*

Inkrementalni algoritam pokrenut ćemo jediničnom kružnicom $\hat{K}_1(\hat{S}_1, 1)$, gdje je $\hat{S}_1 = Mean[\mathcal{A}]$. Tijek inkrementalnog algoritma prikazan je na slikama 8.20a-d. Sukladno nužnim uvjetima iz točke 8.3.3, str. 137 dobivena 5-LOPart je MAPart.



Slika 8.20: Particije skupa $\mathcal{A}$ dobivene inkrementalnim algoritmom

8.7.3 Traženje MAPart i prepoznavanje kružnica

Ako raspolažemo s više k-LOPart s kružnicama kao klaster-centrima, možemo pokušati između njih potražiti MAPart primjenom modificiranog CH ili DB-indeksa (točka 8.3, str. 135). U tu svrhu potrebno je dobro definirati udaljenost dviju kružnica, a tu možemo iskoristiti eksplicitnu formulu za Hausdorffovu udaljenost $d_H(K_1, K_2)$ dviju kružnica $K_1(S_1, r_1)$, $K_2(S_2, r_2)$ (vidi [90]):

$$d_H(K_1, K_2) = \|S_1 - S_2\| + |r_1 - r_2|. \tag{8.61}$$

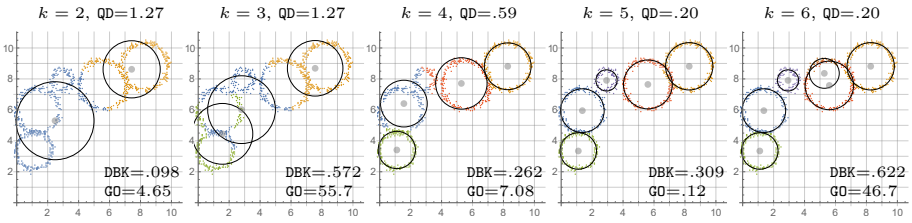
Zadatak 8.5. *Dokažite da je Hausdorffova udaljenost* $d_H(K_1, K_2)$ *dviju kružnica* $K_1(S_1, r_1)$, $K_2(S_2, r_2)$ *zadana s* (8.61).

Kao što smo već spomenuli, budući da su klasični indeksi konstruirani za sferične ili elipsoidne klastere, pokazalo se da je u ovom slučaju ipak bolje iskoristiti specijalnu strukturu `MCD` problema. Naime, na k-`LOPart` $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ možemo primijeniti `GO`-indeks, str. 136, ili iskoristiti nužne uvjete, str. 137, da k-particija $\Pi^\star$ bude ujedno i `MAPart` (vidi također točku 8.3):

$\mathbf{A}_0$: $\texttt{QD}[\Pi^\star] < \epsilon(\mathcal{A})$;

$\mathbf{A}_1$: $\rho(\pi^\star) \geq \frac{MinPts}{2\epsilon\mathcal{A})}$ za sve klastere $\pi^\star \in \Pi^\star$.

Primjer 8.13. *Na primjeru skupa* $\mathcal{A}$ *prikazanog na Slici 8.21 provedeno je prvih* 5 *iteracija inkrementalnog algoritma.*



Slika 8.21: Inkrementalni algoritam i izbor `MAPart`

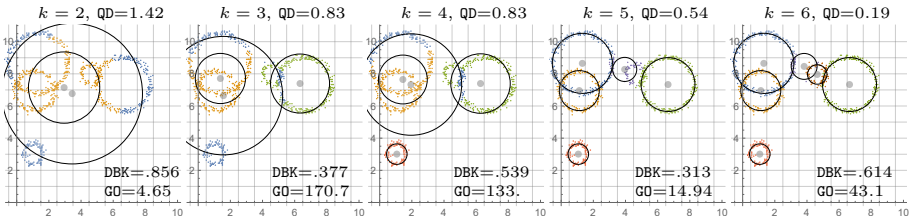
`DBG`-indeks pogrešno ukazuje na 4-particiju kao `MAPart`, dok `GO`-indeks ispravno ukazuje na 5-particiju kao `MAPart`.

Budući da je za $MinPts = \lfloor \log |\mathcal{A}| \rfloor = 6$, ϵ-gustoća skupa $\mathcal{A}$ je $\epsilon(\mathcal{A}) = 0.314$ pa inkrementalni algoritam treba zaustaviti kada $\texttt{QD} < \epsilon$. To se prvi put dogodilo za $k = 5$. Kako je $\frac{MinPts}{2\epsilon(\mathcal{A})} = 9.55$, a gustoće svih klastera 5-particije (23.23, 23.63, 24.16, 23.78, 23.47) veće su od $\frac{MinPts}{2\epsilon(\mathcal{A})}$, `QD` ispravno ukazuje na 5-particiju kao `MAPart`. Također, sukladno nužnim uvjetima iz točke 8.3.3, str. 137, kada bi iz 6-`LOPart` eliminirali kružnicu-centar $K((5.614, 7.607), 1.569)$ s klasterom lokalne gustoće $\rho = 6.59$, nakon provedbe `KCC`-algoritma dobili bismo `MAPart`.

Primjer 8.14. *Na primjeru skupa* $\mathcal{A}$ *prikazanog na Slici 8.22 provedeno je prvih* 5 *iteracija inkrementalnog algoritma.*

`DBG`-indeks ukazuje na 5-particiju kao `MAPart`, ali vidi se da nisu sve kružnice prepoznate. Dakle, to nije `MAPart`!

`GO`-indeks slično ukazuje na 5-particiju kao `MAPart`, iako jedna od kružnica nije prepoznata. Dakle, to nije `MAPart`!.



Slika 8.22: Inkrementalni algoritam i izbor `MAPart`

Budući da za $MinPts = \lfloor \log |\mathcal{A}| \rfloor = 6$, ϵ-gustoća skupa $\mathcal{A}$ iznosi $\epsilon(\mathcal{A}) = .314$, inkrementalni algoritam treba zaustaviti kada je $\texttt{QD} < \epsilon$. To se prvi put dogodilo za $k = 6$. Gustoće svih klastera 6-particije su

$$(20.76,\ 23.82,\ 22.29,\ 23.22,\ 7.71,\ 25.99).$$

Kako je $\frac{MinPts}{2\epsilon(\mathcal{A})} = 9.96$, uvjet $\mathbf{A}_1$ ukazuje da klasteru $\pi_5^\star$ tu nije mjesto. Kada se ispusti klaster $\pi_5^\star$ s njegovom centar-kružnicom i elemente klastera $\pi_5^\star$ razdijeli ostalim klasterima po principu minimalnih udaljenosti dobivamo `MAPart`, a sve kružnice su prepoznate!!!

8.8 Elipsa kao reprezentant skupa podataka

8.8.1 Elipsa kao Mahalanobis kružnica

Mahalanobis kvazimetrička funkcija $d_m\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ (vidi [62, 103]) definirana je s (6.6), str. 104:

$$d_m(u, v; \Sigma) = (u - v)^T \Sigma^{-1} (u - v) = \|u - v\|_\Sigma^2,$$

a za danu kovarijacijsku matricu $\Sigma > 0$, „jedinična" M-kružnica sa središtem u točki S:

$$E = \{x \in \mathbb{R}^2 \colon d_m(S, x; \Sigma) = 1\}, \tag{8.62}$$

predstavlja elipsu s poluosima ξ, η koje odgovaraju drugim korijenima svojstvenih vrijednosti matrice Σ.

Zadatak 8.6. *Pokažite da se elipsa* (8.3) *može zapisati kao „jedinična" M-kružnica, odnosno kao:*

$$\frac{[(x-p)\cos\vartheta + (y-q)\sin\vartheta]^2}{\xi^2} + \frac{[(x-p)\sin\vartheta - (y-q)\cos\vartheta]^2}{\eta^2} = 1.$$

Normalizirana Mahalanobis kvazimetrička funkcija $d_M\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ (vidi [62, 103]) definirana je s (6.9), str. 109:

$$d_M(u, v; \Sigma) := \sqrt{\det \Sigma}\,(u-v)^T \Sigma^{-1}(u-v) = \|u-v\|_\Sigma^2. \tag{8.63}$$

i za nju vrijedi sljedeća lema.

Lema 8.2. *Elipsa* $E(S, \xi, \eta, \vartheta)$ *zadana s* (8.3) *može se prikazati kao* M*-kružnica [58]:*

$$E(S, r, \Sigma) = \{u \in \mathbb{R}^2 \colon d_M(S, u; \Sigma) = r^2\}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \tag{8.64}$$

gdje je $r^2 = \sqrt{\det \Sigma} = \xi\eta$.

Obratno, M*-kružnica* $E(S, r, \Sigma)$ *odgovara ellipsi* $E(S, \xi, \eta, \vartheta)$*, gdje su poluosi* ξ, η *određene dekompozicijom na svojstvene vrijednosti:*

$$\operatorname{diag}(\xi^2, \eta^2) = U\big(\tfrac{r^2}{\sqrt{\det \Sigma}}\Sigma\big)U^T, \quad U = \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix}, \tag{8.65}$$

a kut ϑ *zadan je s:*

$$\vartheta = \tfrac{1}{2}\arctan\tfrac{2\sigma_{12}}{\sigma_{11}-\sigma_{22}} \in \begin{cases} [0, \pi/4), & \sigma_{12} \geq 0 \ \&\ \sigma_{11} > \sigma_{22} \\ [\pi/4, \pi/2), & \sigma_{12} > 0 \ \&\ \sigma_{11} \leq \sigma_{22} \\ [\pi/2, 3\pi/4), & \sigma_{12} \leq 0 \ \&\ \sigma_{11} < \sigma_{22} \\ [3\pi/4, \pi), & \sigma_{12} < 0 \ \&\ \sigma_{11} \geq \sigma_{22} \end{cases}. \tag{8.66}$$

Dokaz. Prva tvrdnja dokazuje se direktnom provjerom.

U cilju dokaza druge tvrdnje, ellipse $E(S, r, \Sigma)$ napišemo u obliku:

$$\begin{aligned} \tfrac{1}{r^2} d_M(S, u; \Sigma) = 1 \quad &\Rightarrow \quad \tfrac{\sqrt{\det \Sigma}}{r^2}(S-u)^T\Sigma^{-1}(S-u) = 1 \\ &\Rightarrow \quad (S-u)^T\big(\tfrac{r^2}{\sqrt{\det \Sigma}}\Sigma\big)^{-1}(S-u) = 1. \end{aligned}$$

Dekompozicijom na svojstvene vrijednosti matrice $\big(\frac{r^2}{\sqrt{\det \Sigma}}\Sigma\big)$ dobivamo traženu tvrdnju. □

Skup $\{u \in \mathbb{R}^2 \colon d_M(u, S; \Sigma) = r\}$ predstavlja normaliziranu elipsu s centrom u S i poluosima ξ', η' čiji je produkt $\xi'\eta' = 1$.

Primijetite također da je normalizirajući faktor $\sqrt{\det \Sigma}$ proporcionalan površini elipse ($\xi\eta\pi$), a veličina ovako definiranog „radijusa" r geometrijska je sredina poluosi ξ, η elipse (vidi Primjer 6.4, str. 110).

Problem traženja najboljeg reprezentanta skupa $\mathcal{A}$ u formi elipse (M-kružnice) može se definirati kao sljedeći GOP:

$$\underset{S\in\mathbb{R}^2,\, r\in\mathbb{R},\, \Sigma\in\mathbb{R}^{2\times 2}}{\operatorname{argmin}} \sum_{a\in\mathcal{A}} \mathfrak{D}(a, E(S, r, \Sigma)), \tag{8.67}$$

gdje je:

$$\mathfrak{D}(a, E) = \mathfrak{D}(a, E(S, r, \Sigma)) = (\|S - a\|_\Sigma^2 - r^2)^2 \tag{8.68}$$

takozvana *algebarska udaljenost* točke $a \in \mathcal{A}$ do elipse E. Slično kao kod obične kružnice, i u ovom slučaju može se definirati TLS-udaljenost ili OD-udaljenost (str. 155).

Rješenje GOP (8.67) može se potražiti korištenjem neke lokalno optimizacijske metode (Nelder-Meade, Quasi-Newton), budući da smo u mogućnosti pronaći vrlo dobru početnu aproksimaciju parametara S, r, Σ. Naime, kao aproksimaciju središta možemo izabrati centroid $\hat{S}$ skupa $\mathcal{A}$ (vidi Teorem 6.2), kao aproksimaciju kovarijacijske matrice možemo izabrati $\hat{\Sigma} = \frac{1}{m}\sum_{a\in\mathcal{A}} (\hat{S} - a)(\hat{S} - a)^T$, a dobra početna aproksimacija za radijus određena je s:

$$\hat{r}^2 = \tfrac{1}{m}\sum_{a\in\mathcal{A}} \|\hat{S} - a\|_{\hat{\Sigma}}^2, \tag{8.69}$$

jer, korištenjem svojstva aritmetičke sredine, vrijedi:

$$\sum_{a\in\mathcal{A}} \left(\|\hat{S} - a\|_{\hat{\Sigma}}^2 - r^2\right)^2 \geq \sum_{a\in\mathcal{A}} \left(\|\hat{S} - a\|_{\hat{\Sigma}}^2 - \tfrac{1}{m}\sum_{a\in\mathcal{A}} \|\hat{S} - a\|_{\hat{\Sigma}}^2\right)^2.$$

8.9 Prepoznavanje više elipsi u ravnini

Postoje brojne primjene problema prepoznavanja više elipsi u različitim područjima: računalni vid, medicina, biologija, poljoprivreda, astronomija itd. U literaturi se može pronaći više metoda za prepoznavanje elipsi uz primjenu Houghovih transformacija [31, 63]. U [2] navodi se metoda EDCircles koja prepoznaje više elipsi u realnom vremenu.

Mi promatramo problem prepoznavanja više elipsi na osnovi skupa podataka $\mathcal{A} \subset \mathbb{R}^2$ (*Multiple Ellipse Detection* (MED)). Neka je $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\} \subset \Delta$ skup podataka koji potječe od više unaprijed nepoznatih elipsi u ravnini, koje treba prepoznati ili rekonstruirati. Podaci se mogu umjetno konstruirati na sličan način kao za podatke koji potječu od više kružnica. U ovom slučaju procedura je znatno složenija i može se vidjeti u [37]. Pretpostavljamo da podaci koji dolaze od neke elipse zadovoljavaju „svojstvo homogenosti", tj. pretpostavljamo da su podaci pretežno homogeno rasuti oko te elipse (vidi Definiciju 8.1).

Za rješavanje ovog problema u [37] navedene su dvije metode: jedna, manje učinkovita, ovaj problem rješava kao center-based klastering problem pri čemu su centri klastera elipse, a druga, znatno učinkovitija, koristi `RANSAC` metodu (vidi primjerice [37]).

Traženje optimalne particije $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ gdje je centar klastera $\pi_j^\star$ elipsa, odnosno `M`-kružnica $E_j(S_j, r_j, \Sigma_j)$ s centrom $S_j = (p_j, q_j)^T$ i kovarijacijskom matricom $\Sigma_j = \begin{bmatrix} u_j & v_j \\ v_j & t_j \end{bmatrix}$ vodi na traženje optimalnih parametara $(p_j^\star, q_j^\star, r_j^\star, u_j^\star, v_j^\star, t_j^\star)$, $j = 1, \ldots, k$, koji su rješenje `GOP` (usporedi s (3.40)):

$$\underset{p_j, q_j \in \Delta,\, r_j \in [0,R],\, \Sigma_j \in M_2}{\operatorname{argmin}} \sum_{i=1}^{m} \min_{1 \le j \le k} \{\mathfrak{D}(a^i, E_j(S_j, r_j, \Sigma_j))\}, \tag{8.70}$$

gdje je $R = \frac{1}{2}\min\{b-a, d-c\}$, M_2 skup simetričnih pozitivno definitnih matrica drugog reda, a $\mathfrak{D}(a^i, E_j(S_j, r_j, \Sigma_j))$ predstavlja udaljenost točke $a^i \in \mathcal{A}$ do `M`-kružnice $E_j(S_j, r_j, \Sigma_j)$.

O načinu definiranja metričke funkcije $\mathfrak{D}$ pisali smo u točki 8.6, str. 153. Iz sličnih razloga, i u ovom slučaju koristit ćemo algebarsku udaljenost točke a^i do `M`-kružnice $E(S, r, \Sigma)$:

$$\mathfrak{D}(a^i, E(S, r, \Sigma)) = (\|S - a^i\|_\Sigma^2 - r^2)^2. \tag{8.71}$$

Primijetite da je optimizacijski problem (8.70) `GOP` s $6k$ nezavisnih varijabli. Ako na ovaj `GOP` primijenimo globalno optimizacijski algoritam `DIRECT`, vidjet ćemo da će potrebno `CPU`-vrijeme biti nerazumno veliko jer će algoritam pronalaziti svih $k!$ rješenja. Zato se rješenje ovog `GOP` može potražiti na sljedeći način:

1. Pronaći dobru početnu aproksimaciju za `GOP` (8.70);
2. Na ovu početnu aproksimaciju primijeniti modifikaciju k-means algoritma za `M`-kružnice-centre.

U slučaju poznatog broja elipsi od kojih su nastali podaci, početnu aproksimaciju za `GOP` (8.70) potražit ćemo tako da promatramo problem traženja optimalne particije s običnim kružnicama-centrima malenog fiksnog radijusa $r_0 > 0$. Na taj način umjesto rješavanja `GOP` (8.70) promatramo `GOP` s $2k$ nezavisnih varijabli:

$$\underset{p,q \in \Delta^k}{\operatorname{argmin}} F(p, q), \quad F(p, q) = \sum_{i=1}^{m} \min_{1 \le j \le k} \{\mathfrak{D}(a^i, S_j, r_0, I))\}, \tag{8.72}$$

gdje je $I \in \mathbb{R}^{2\times 2}$ jedinična matrica, a $S_j = (p_j, q_j)^T$.

Budući da je funkcija iz (8.72) Lipschitz-neprekidna [81], primjenom manjeg broja iteracija (recimo 10-20) algoritma `DIRECT` na `GOP` (8.72) dobit ćemo dobru početnu aproksimaciju za `GOP` (8.70). Na taj način dobro smo pozicionirali centre traženih elipsi.

8.9.1 Generiranje podataka koji potječu od više elipsi u ravnini

Skup podataka $\mathcal{A}$ koji potječe od k elipsi definirat ćemo na sljedeći način. Najprije izaberemo k centara $S_1, \ldots, S_k \in [1.5, 8.5]^2$ koji su međusobno udaljeni za barem 2.5. Poluosi ξ_j, η_j svake elipse uzet ćemo iz $\mathcal{U}(0.5, 2.5)$, a kut rotacije $\vartheta_j \in \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$. Tako dobivamo k elipsi $E_j(S_j, \xi_j, \eta_j, \vartheta_j)$, $j = 1, \ldots, k$ zadanih parametarski s (8.3).

Kako bismo osigurali homogenost podataka u okolini elipse E_j, najprije ćemo na njoj generirati $|\pi_j|$ uniformno distribuiranih točaka (vidi [37]) lokalne gustoće $\rho = \frac{|\pi_j|}{|E_j|} = 21$, gdje je $|E_j|$ duljina elipse E_j.

Primjedba 8.4. *Duljina (opseg) elipse E_j određena je s:*

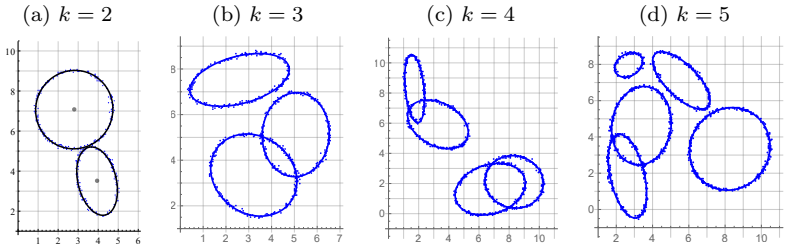
$$|E_j| = 4 \int_0^{\pi/2} \sqrt{\xi_j^2 \cos^2 t + \eta_j^2 \sin^2 t}\, dt. \tag{8.73}$$

Ovo je eliptički integral drugog reda i ne može se riješiti elementarno, ali za naše potrebe možemo iskoristiti dobro poznatu Ramanujan aproksimaciju

$$|E_j| \approx \pi(\xi_j + \eta_j)\left(1 + \frac{3h}{10+\sqrt{4-3h}}\right), \quad h = \frac{(\xi_j - \eta_j)^2}{(\xi_j + \eta_j)^2}. \tag{8.74}$$

Nakon toga, svakoj tako izabranoj točki na elipsi u smjeru normale dodamo slučajnu pogrešku iz binormalne distibucije s očekivanjem $0 \in \mathbb{R}^2$ i kovarijacijskom matricom $\sigma^2 I$, gdje je $\sigma^2 = 0.05$ (vidi [1, 37]). Tako smo definirali skup podataka $\mathcal{A}$ i njegovu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$, gdje je π_j skup podataka koji potječe od elipse E_j.

Primjer 8.15. *Na prethodno opisani način konstruirani su skupovi podataka s $k = 2, 3, 4, 5$ elipsi prikazanih na Slici 8.23.*

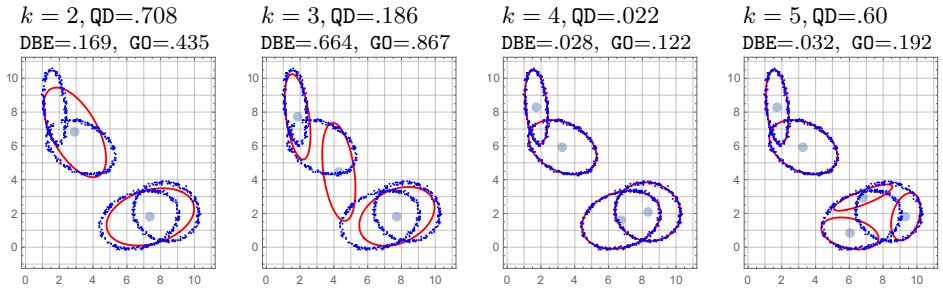


Slika 8.23: Podaci koji potječu od $k \in \{2, 3, 4, 5\}$ elipsi

8.9.2 Prilagođavanje *k*-means algoritma na slučaj M-kružnica centara

U slučaju poznatog broja k elipsi k-LOPart tražit ćemo prilagođenim KCG-algoritmom za slučaj elipsi (M-kružnica-centara) [37, 58] pa ćemo i algoritam zvati algoritam k-najbližih elipsi (KCE). Pri tome, važno je koristiti normaliziranu Mahalanobis kvazimetričku funkciju (8.63) jer će na taj način biti osigurano svojstvo monotonog pada funkcije cilja (vidi [62, 103]).

Algoritam može započeti početnim M-kružnicama-centrima ili početnom particijom. Posebnu pozornost treba posvetiti izboru početne aproksimacije za KCE-algoritam.



Slika 8.24: Optimalne k-particije dobivene KCE-algoritmom

Primjer 8.16. *Promatramo skup podataka $\mathcal{A}$ koji je nastao od* 4 *elipse (vidi Sliku 8.23c). Na skup $\mathcal{A}$ primijenit ćemo* KCE*-algoritam uz $k = 2, 3, 4, 5$ početnih* M*-kružnica centara dobivenih primjenom po* 10 *iteracija algoritma* DIRECT.

Svi korišteni kriteriji (DBE-indeks, GO-indeks, nužni uvjeti iz točke 8.3.3, str. 137) ukazuju da je 4-LOPart (vidi Sliku 8.24) ujedno i MAPart, dok ostale LOPart na slici to nisu. Zašto?

8.9.3 Prilagođavanje inkrementalnog algoritma na slučaj M-kružnica-centara

Ako broj elipsi od kojih potječu podaci nije unaprijed poznat, potražit ćemo optimalne particije s $k = 2, 3, \ldots, \kappa$ klastera prilagođavanjem inkrementalnog algoritma i među njima pokušati odrediti `MAPart`.

Algoritam započinje izborom početne M-kružnice-centra $\hat{E}_1(\hat{S}_1, \hat{r}_1, \hat{\Sigma}_1)$. Aproksimaciju sljedeće M-kružnice-centar potražit ćemo kao običnu kružnicu $\hat{E}_2(\hat{S}_2, \hat{r}_2, I_2)$ rješavajući sljedeći `GOP`:

$$\underset{(p,q)^T \in \Delta,\, r \in [0,R]}{\operatorname{argmin}} \sum_{i=1}^{m} \min\{\mathfrak{D}(a^i, \hat{E}_1(\hat{S}_1, \hat{r}_1, \hat{\Sigma}_1)), \mathfrak{D}(a^i, E(S(p,q), r, I_2)\}. \quad (8.75)$$

Nakon toga, na M-kružnice-centre $\hat{E}_1, \hat{E}_2$ primjenjuje se `KCE`-algoritam i time se dobiva lokalno optimalna 2-particija $\Pi^{(2)}$. Primijetite da je prilikom rješavanja problema (8.75) dovoljno izvesti manji broj koraka optimizacijskog algoritma `DIRECT` i tako dobiti dovoljno dobru početnu aproksimaciju za `KCE`-algoritam koji nakon toga daje lokalno optimalnu 2-particiju $\Pi^{(2)}$ s M-kružnicama-centrima $(E_1^\star, E_2^\star)$.

Općenito, poznavajući k M-kružnica-centara $\hat{E}_1, \ldots, \hat{E}_k$, aproksimaciju sljedeće M-kružnice-centra potražit ćemo u formi obične kružnice $E(S, r, I_2)$ rješavajući sljedeći `GOP`:

$$\underset{(p,q)^T \in \Delta,\, r \in [0,R]}{\operatorname{argmin}} \sum_{i=1}^{m} \min\{\delta_k^{(i)}, \mathfrak{D}(a^i, E(S(p,q), r, I_2)\}, \quad (8.76)$$

gdje je $\delta_k^{(i)} = \min\{\mathfrak{D}(a^i, \hat{E}_1(\hat{S}_1, \hat{r}_1, \hat{\Sigma}_1)), \ldots, \mathfrak{D}(a^i, \hat{E}_k(\hat{S}_k, \hat{r}_k, \hat{\Sigma}_k))\}$.

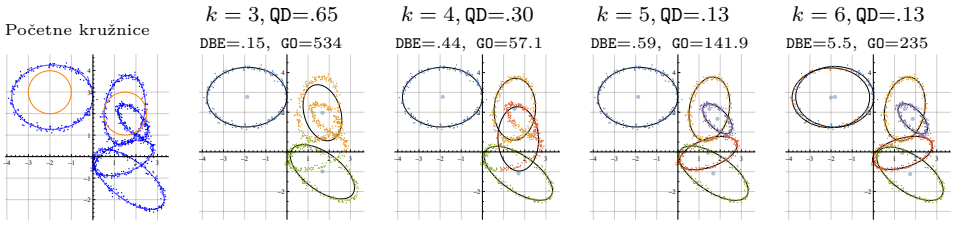
Nakon toga, primjenom `KCE`-algoritma, dobivamo $(k+1)$-`GOPart` $\Pi^\star = \{\pi_1^\star, \ldots, \pi_{k+1}^\star\}$ s M-kružnicama-centrima $E_1^\star, \ldots, E_{k+1}^\star$.

Inkrementalni algoritam zaustavlja se korištenjem kriterija (8.13) pri čemu je $\mathfrak{D}_1$-udaljenost točke $a \in \pi_j^\star$ do odgovarajuće M-kružnice-centar $E_j^\star$ dana s:

$$\mathfrak{D}_1(a^i, E_j^\star(S_j^\star, r_j^\star, \Sigma_j^\star)) = |\|S_j^\star - a^i\|_{\Sigma_j^\star} - r_j^\star|. \quad (8.77)$$

Primjer 8.17. *Neka je $\mathcal{A}$ skup podataka zadan kao na Slici 8.25. Primjenom inkrementalnog algoritma pronađimo lokalno optimalne $2, 3, 4, 5$-particije.*

Na osnovi skupa podataka $\mathcal{A}$ izračunat je broj $MinPts = \lfloor \log |\mathcal{A}| \rfloor = 6$ i ϵ-gustoća $\epsilon(\mathcal{A}) = .289$. Inkrementalni algoritam pokrenut ćemo s dvije crveno označene jedinične kružnice. Tijek inkrementalnog algoritma prikazan je na Slici 8.25.



Slika 8.25: Inkrementalni algoritam

Kao što se vidi u zaglavljima slika, kriterij (8.13) ispunio se u trenutku kada je algoritam pronašao 5 elipsi koje se podudaraju s originalnim, dakle, kada je detektirana `MAPart`.

Također, kada bi iz 6-`LOPart` eliminirali M-kružnicu-centar

$$E\big((-2.00, 2.73)^T, 1.66, \begin{bmatrix} 1.87 & 0.01 \\ 0.01 & 1.27 \end{bmatrix}\big)$$

s klasterom lokalne gustoće $\rho = 8.22$, nakon provedbe `KCE`-algoritma dobili bismo `MAPart`.

8.10 Rješavanje `MGD` problema primjenom `RANSAC` i `DBSCAN` metode uz korištenje `KCG`-algoritma

Za rješavanje općeg `MGD` problema (8.5) (točka 8.5, str. 130) može se primijeniti `RANSAC` (`RAN`dom `SA`mple `C`onsensus) metoda (vidi [35, 37]) uz već ranije spomenute elemente `DBSCAN` metode i odgovarajuće modifikacije k-means algoritma.

Kao i ranije, pretpostavljamo da je u pravokutniku $\Delta = [a, b] \times [c, d] \subset \mathbb{R}^2$, $a < b, c < d$ zadan skup podataka $\mathcal{A} = \{a^i = (x_i, y_i)^T : i = 1, \ldots, m\} \subset \Delta$, koji potječe od više unaprijed nepoznatih istovrsnih geometrijskih objekata u ravnini koje treba prepoznati ili rekonstruirati. Pri tome pretpostavljamo da podaci koji dolaze od nekog geometrijskog objekta γ zadovoljavaju svojstvo „homogenosti", tj. pretpostavljamo da je skup podataka π koji potječe od tog geometrijskog objekta pretežno homogeno rasut oko γ i da ima lokalnu gustoću $\rho = \frac{|\pi(\gamma)|}{|\gamma|}$ (vidi Definiciju 8.1, str. 130). Također, za zadani $MinPts = \lfloor \log |\mathcal{A}| \rfloor$ odredit ćemo ϵ-gustoću $\epsilon(\mathcal{A})$ skupa $\mathcal{A}$ kao 99.5% kvantil skupa $\{\epsilon_a : a \in \mathcal{A}\}$ (vidi Definiciju 8.2, str. 134).

Ako je γ određen s n parametara, slučajnim odabirom n točaka iz skupa $\mathcal{A}$ odredimo krivulju γ. Ako je $\gamma \subset \Delta$, smatrat ćemo da je to prihvatljivi kandidat za jedan od traženih geometrijskih objekata. Točke $a \in \mathcal{A}$ koje

se nalaze u području širine 2ϵ oko γ čine podskup $\pi(\gamma) \subset \mathcal{A}$. Postupak ponovimo N puta i zadržimo onaj par (γ_1, π_1) za koji je pripadni podskup točaka najveći. Podskup π_1 izdvojimo iz $\mathcal{A}$ i s ostatkom $\texttt{Complement}(\mathcal{A}, \pi_1)$ ponavljamo postupak tako dugo dok raspoloživi broj točaka ne padne ispod nekog unaprijed zadanog broja (primjerice, $12\,MinPts$).

Nakon toga principom minimalnih udaljenosti skup $\mathcal{A}$ razdijelimo na dobivene podskupove π_j tako da svaki $a \in \mathcal{A}$ pripadne najbližem γ_j i ispustimo sve parove (γ_s, π_s) za koje nije ispunjen uvjet:

$$\mathbf{A}_1: \quad \rho(\pi_s) \geq \frac{MinPts}{2\epsilon(\mathcal{A})}.$$

Na preostale parove (γ_r, π_r) primijenimo KCG-algoritam.

Sljedeća propozicija ukazuje na potreban broj pokušaja izbora po n točaka iz $\mathcal{A}$ kako bismo između prihvatljivih krivulja-kandidata mogli očekivati barem jednu blisku nekoj od originalnih krivulja.

Propozicija 8.3. *Neka je $\Pi = \{\pi_1, \ldots, \pi_k\}$ k-particija skupa $\mathcal{A}$. Vjerojatnost da će slučajnim odabirom n elemenata iz skupa $\mathcal{A}$ svi izabrani elementi biti iz jednog klastera particije dana je s:*

$$P(n) = \frac{\binom{|\pi_1|}{n} + \cdots + \binom{|\pi_k|}{n}}{\binom{|\mathcal{A}|}{n}}. \tag{8.78}$$

Osim toga, vrijedi $P(r+1) < P(r)$ za svaki $r = n, \ldots, \min\limits_{1 \leq j \leq k} |\pi_j|$ pri čemu niz $(P(r))$ brzo opada.

Propozicija pokazuje da možemo očekivati da ćemo u $N = \lceil \frac{1}{P(n)} \rceil$ pokušaja slučajnih izbora po n točaka iz skupa $\mathcal{A}$ dobiti barem jedan skup od n točaka iz jednog klastera.

Krivulja γ mogla bi se principom najmanjih kvadrata preciznije odrediti korištenjem više od n točaka, ali kao što pokazuje Propozicija 8.3 za to bi bilo potrebno znatno više slučajnih pokušaja izbora točaka.

Opisani postupak detaljnije ćemo obraditi u sljedećem odlomku na primjeru MCD problema. Drugi razmatrani problemi rješavali bi se na sličan način.

8.10.1 Rješavanje MCD problema primjenom RANSAC i DBSCAN metode uz korištenje KCC-algoritma

Pretpostavimo da je zadan skup $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\} \subset \Delta$, koji potječe od više unaprijed nepoznatih kružnica u ravnini koje treba prepoznati ili rekonstruirati. Nadalje, pretpostavimo da podaci koji dolaze

od neke kružnice $K(S,r)$ zadovoljavaju svojstvo „homogenosti" s lokalnom gustoćom ρ, a čitav skup $\mathcal{A}$ ima ϵ-gustoću $\epsilon(\mathcal{A})$ dobivenu uz $MinPts = \lfloor \log|\mathcal{A}| \rfloor$.

Kružnica $K(S(p,q),r)$ s parametrima p,q,r određena je s 3 različite točke $a^1 = (x_1,y_1)^T$, $a^2 = (x_2,y_2)^T$, $a^3 = (x_3,y_3)^T \in \mathcal{A}$ koje ne leže na nekom pravcu, tj. za koje vrijedi:

$$D = \det(a^1,a^2,a^3) = \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} \neq 0. \tag{8.79}$$

Uz ovaj uvjet parametri p,q,r dobivaju se eksplicitno kao što se može vidjeti u Algoritmu 4.

Ako pretpostavimo da skup podataka $\mathcal{A}$ dolazi od kružnica koje su potpuno sadržane u Δ i čiji radijusi nisu manji od 2ϵ, onda je *prihvatljiva kružnica-kandidat* $K_j(S_j,r_j)$ ona za koju vrijedi:

$$K_j(S_j,r_j) \subset \Delta \quad \& \quad r_j > 2\epsilon, \tag{8.80}$$

a dobiva se Algoritmom 4.

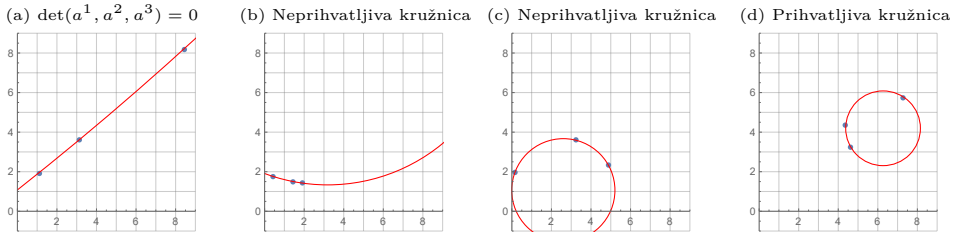
Algoritam 4 (Izbor prihvatljive kružnice)

Input: $\mathcal{A} \subset \Delta$, $\epsilon > 0$, $\mu > 0$, $n_p = 3$
1: Izaberi $a^1 = (x_1,y_1)^T$, $a^2 = (x_2,y_2)^T$, $a^3 = (x_3,y_3)^T \in \mathcal{A}$;
2: **if** $D = \det(a^1,a^2,a^3) \neq 0$, **then**
3: $p = \frac{1}{2D}\big(x_3^2(y_1-y_2) + (x_1^2 + (y_1-y_2)(y_1-y_3))(y_2-y_3) + x_2^2(-y_1+y_3)\big)$;
4: $q = \frac{1}{2D}\big(-x_2^2x_3 + x_1^2(-x_2+x_3) + x_3(y_1-y_2)(y_1+y_2) + x_1(x_2^2-x_3^2+y_2^2-y_3^2) + x_2(x_3^2-y_1^2+y_3^2)\big)$;
5: $r^2 = \text{mean}\{\|a^s - (p,q)\|^2, s = 1,2,3\}$;
6: $S = (p,q)^T$;
7: **end if**
8: **if** $|D| < \mu \vee r < 2\epsilon \vee K(S,r) \setminus \Delta \neq \emptyset$, **then**
9: GoTo Step 1;
10: **end if**
Output: $\{K(S(p,q),r)\}$

Kao što pokazuje sljedeći primjer, slučajan izbor 3 točke iz $\mathcal{A}$ ne dovodi nužno do prihvatljive kružnice.

Primjer 8.18. *Pretpostavimo da je* $\mathcal{A} \subset \Delta = [0,10]^2$. *Slučajnim izborom* 3 *točke iz* $\mathcal{A}$ *može se dogoditi da:*

- *točke leže na pravcu pa nije moguće odrediti kružnicu (Slika 8.26a),*
- *dobivena kružnica nije potpuno sadržana u* Δ *(Slika 8.26 b i c),*
- *su točke preblizu jedna drugoj pa dobivena kružnica ima premaleni radijus.*



Slika 8.26: Biranje prihvatljive kružnice-kandidata

Primjenom Algoritma 4 odredit ćemo N prihvatljivih kružnica-kandidata $K_j(S_j, r_j)$, $j = 1, \ldots, N$ i svakoj od njih pridružiti odgovarajući klaster

$$\pi_j = \{a \in \mathcal{A} \colon \mathfrak{D}(a, K_j) < \epsilon\},$$

gdje je $\mathfrak{D}$ algebarska udaljenost točke do kružnice zadana s (8.56). Kružnicu koja pripada najbrojnijem klasteru označit ćemo s $\tilde{K}$, a odgovarajući klaster s $\tilde{\pi}$. Par $(\tilde{K}, \tilde{\pi})$ može se korigirati rješavanjem optimizacijskog problema

$$\underset{p,q,r}{\operatorname{argmin}} \sum_{\tilde{a}^i \in \tilde{\pi}} \left(\|(p,q)^T - \tilde{a}^i\|^2 - r^2 \right)^2, \tag{8.81}$$

uz primjenu neke lokalno optimizacijske metode koristeći mean$[\tilde{\pi}]$ kao početnu aproksimaciju za središte kružnice i (8.52), str. 155 kao početnu aproksimaciju za radijus kružnice. Na taj način odredili smo prvu kružnicu i njen klaster $\{\hat{K}, \hat{\pi}\}$.

Algoritam 5 (Traženje jedne kružnice-centra)

Input: $\mathcal{A} \subset \Delta$, ϵ, N

1: Primjenom Algoritma 1 odrediti N prihvatljivih kružnica-kandidata $K_j(S_j, r_j)$;
2: Svakom K_j pridružiti skup $\pi_j \subset \mathcal{A}$ svih točaka koje su do na ϵ blizu K_j;
3: Odrediti $j_{max} \in \underset{1 \le j \le N}{\operatorname{argmax}} \rho_j$, gdje je $\rho_j = \frac{|\pi_j|}{2 r_j \pi}$; Odgovarajuću kružnicu označiti s $\tilde{K}$, a odgovarajući klaster s $\tilde{\pi}$;
4: Prema (8.81), odrediti korekciju $\hat{K}(\hat{S}, \hat{r})$ i definirati $\hat{\pi} := \{a \in \mathcal{A} \colon (a, \hat{K}) < \epsilon\}$

Output: $\{\hat{\pi}, \hat{S}, \hat{r}\}$

Nadalje, iz skupa $\mathcal{A}$ izvučemo klaster $\hat{\pi}$ i na preostalom skupu točaka $\mathcal{B} = \mathcal{A} \setminus \hat{\pi}$ ponovimo Algoritam 5 s nešto smanjenim (ali ne manjim od $\frac{1}{2}N$) brojem pokušaja kako predviđa Propozicija 8.3. Ovaj postupak ponavlja se tako dugo dok broj točaka skupa $\mathcal{B}$ ne padne ispod nekog unaprijed zadanog broja - tu su preostali tragovi podataka iz prethodnih iteracija.

Algoritam 6 (Rješavanje MCD problema korištenjem RANSAC metode)

Input: $\mathcal{A} \subset \Delta$, $MinPts$, ϵ, N

1: Stavi $\mathcal{B} = \mathcal{A}$, $N_k = N$, $trag = 12MinPts$, $Cir = \{\}$;
2: **while** $|\mathcal{B}| > trag$, **do**
3: Pozovi Algoritam 5 $[\mathcal{A}, \epsilon, N_k]$ i dobiveni rezultat označi s $\{\hat{\pi}, \hat{K}(\hat{S}, \hat{r})\}$;
4: Skupu Cir dodaj $\hat{K}(\hat{S}, \hat{r})$;
5: Stavi $\mathcal{B} = \mathcal{B} \setminus \hat{\pi}$; $N_k = \max\{\frac{4}{5}N_k, \frac{1}{2}N\}$;
6: **end while**
7: U skupu Cir zadrži one kružnice koje ispunjavaju uvjet $\mathbf{A}_1$;
8: Na skup $\mathcal{A}$ primijeni KCC-algoritam s početnim kružnicama iz prethodnog koraka i dobivene kružnice označi s $K_j^\star$, $j = 1, \ldots, k$;

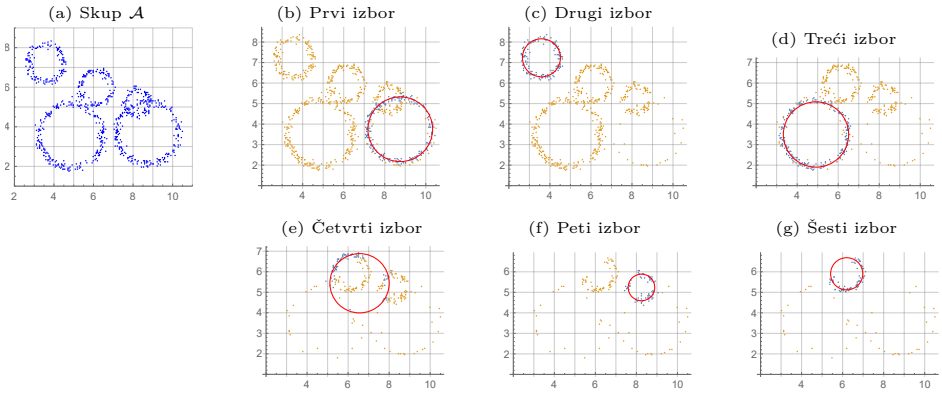
Output: $\{K_j^\star(S_j^\star, r_j^\star), j = 1, \ldots, k\}$

Na ovako dobiven skup parova klaster-kružnica $\{\{\hat{\pi}_j, \hat{K}_j\}, j = 1, \ldots, r\}$ provjeravamo uvjet $\mathbf{A}_1$ iz nužnih uvjeta iz točke 8.3.3, str. 137, i ispustimo sve parove $\{\hat{\pi}_j, \hat{K}_j\}$ za koje nije ispunjen ovaj uvjet.

Budući da preostali klasteri ne obuhvačaju čitav polazni skup podataka $\mathcal{A}$, a preostale kružnice ne zauzimaju optimalne pozicije, na skup $\mathcal{A}$ primijenit ćemo KCC-algoritam s preostalim kružnicama.

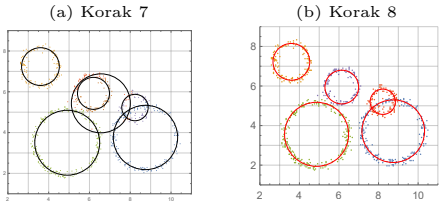
Tako dobivamo rekonstruirane kružnice $K_j^\star$, $j = 1, \ldots, k$. Cijeli postupak detaljnije je napisan u Algoritmu 6.

Primjer 8.19. *Opisani postupak ilustrirat ćemo na skupu podataka $\mathcal{A}$ prikazanom na Slici 8.27a. U ovom primjeru je $MinPts = 6$, $\epsilon(\mathcal{A}) = 0.335$, $\frac{MinPts}{2\epsilon(\mathcal{A})} = 11.937$, a $N = 18$ u skladu s Propozicijom 8.3. Djelovanje Algoritma 6 prikazano je na ostalim slikama.*



Slika 8.27: Algoritam 6 sukcesivno poziva Algoritam 5

Rezultat djelovanja Algoritma 6 prije Koraka 7 prikazan je na Slici 8.28a.



Slika 8.28: Završni koraci Algoritma 6

Dobiveni klasteri imaju lokalne gustoće:

$$\rho_i \in \{22.705,\ 23.888,\ 21.500,\ 8.918,\ 16.220,\ 13.392\},$$

što znači da će se sukladno uvjetu $\mathbf{A_1}$ iz nužnih uvjeta iz točke 8.3.3, str. 137 eliminirati samo kružnica čiji klaster ima lokalnu gustoću 8.918. Konačni rezultat dobiven KCC-algoritmom prikazan je na Slici 8.28b. Treba naglasiti visoku učinkovitost opisanog postupka: potrebno CPU-vrijeme za prepoznavanje 5 kružnica kreće se od $0.8 - 1.5$ sec, od čega je nešto manje od polovine vremena potrebno KCC-algoritmu.

Literatura

[1] S. J. Ahn, W. Rauh, H.-J. Warnecke, *Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola*, Pattern Recognition, **34**(2001) 2283–2303.

[2] C. Akinlar, C. Topal, *Edcircles: A real-time circle detector with a false detection control*, Pattern Recognition, **46**(2013) 725–740.

[3] F. Aurenhammer, R. Klein, *Voronoi diagrams*, In: J. Sack, G. Urrutia, editors, *Handbook of Computational Geometry, Chapter V*. Elsevier Science Publishing, 2000, 201–290.

[4] R. Babuška, P. J. van der Veen, U. Kaymak, *Improved covariance estimation for Gustafson-Kessel clustering*, In: *IEEE international conference on fuzzy systems*, 2002, 1081–1085.

[5] A. M. Bagirov, *Modified global k-means algorithm for minimum sum-of-squares clustering problems*, Pattern Recognition, **41**(2008) 3192–3199.

[6] A. M. Bagirov, *An incremental DC algorithm for the minimum sum-of-squares clustering*, Iranian Journal of Operations Research, **5**(2014) 1–14.

[7] A. M. Bagirov, J. Ugon, *An algorithm for minimizing clustering functions*, Optimization, **54**(2005) 351–368.

[8] A. M. Bagirov, J. Ugon, H. Mirzayeva, *Nonsmooth nonconvex optimization approach to clusterwise linear regression problems*, European Journal of Operational Research, **229**(2013) 132–142.

[9] A. M. Bagirov, J. Ugon, D. Webb, *An efficient algorithm for the incremental construction of a piecewise linear classifier*, Information Systems, **36**(2011) 782–790.

[10] A. M. Bagirov, J. Ugon, D. Webb, *Fast modified global k-means algorithm for incremental cluster construction*, Pattern Recognition, **44**(2011) 866–876.

[11] A. M. Bagirov, J. Yearwood, *A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems*, European Journal of Operational Research, **170**(2006) 578–596.

[12] M. Benšić, N. Šuvak, *Primijenjena statistika*, Odjel za matematiku, Sveučilište u Osijeku, 2012.

[13] J. C. Bezdek, R. Ehrlich, W. Full, *FCM: the fuzzy c-means clustering algorithm*, Computers & Geosciences, **10**(1984) 191–203.

[14] J. C. Bezdek, J. Keller, R. Krisnapuram, N. R. Pal, *Fuzzy models and algorithms for pattern recognition and image processing*, Springer, 2005.

[15] P. T. Boggs, R. H. Byrd, R. B. Schnabel, *A stable and efficient algorithm for nonlinear orthogonal distance regression*, SIAM J. Sci. Statist. Comput., **8**(1987) 1052–1078.

[16] R. J. Boscovich, *De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura eius ex exemplaria etiam sensorum impressa*, Bononienci Scientiarum et Artium Znstituto Atque Academia Commentarrii, **4**(1757) 353–396.

[17] B. Bozkaya, E. Erkut, G. Laporte, *A tabu search heuristic and adaptive memory procedure for political districting*, Europian Journal of Operational Reacearch, **144**(2003) 12–26.

[18] G. Bradski, *The opencv library, Dr. Dobb's Journal of Software Tools*, 2000.

[19] K. Brüntjen, H. Späth, *Incomplete total least squares*, Numerische Mathematik, **81**(1999) 521–538.

[20] T. Calinski, J. Harabasz, *A dendrite method for cluster analysis*, Communications in Statistics, **3**(1974) 1–27.

[21] R. Campello, *A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment*, Pattern Recognition Letters, **28**(2007) 833–841.

[22] N. Chernov, *Circular and linear regression: Fitting circles and lines by least squares*, volume 117 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, London, 2010.

[23] R. Cupec, R. Grbić, K. Nyarko, K. Sabo, R. Scitovski, *Detection of planar surfaces based on ransac and lad plane fitting*, In: *Proceedings of the 4th European Conference on Mobile Robots, ECMR'09*, 2009.

[24] R. Cupec, R. Grbić, K. Sabo, R. Scitovski, *Three points method for searching the best least absolute deviations plane*, Applied Mathematics and Computation, **215**(2009) 983–994.

[25] D. Davies, D. Bouldin, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **2**(1979) 224–227.

[26] P. G. de Cortona, C. Manzi, A. Pennisi, F. Ricca, B. Simeone, *Evaluation and optimization of electoral systems*, In: *SIAM Monographs on Discrete Mathematics*. SIAM, Philadelphia, 1999.

[27] I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, In: *Proceedings of the 10-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA*, 2004, 551–556.

[28] I. S. Dhillon, S. Mallela, R. Kumar, *A divisive information-theoretic feature clustering algorithm for text classification*, Journal of Machine Learning Research, **3**(2003) 1265–1287.

[29] Y. Dodge, editor, *Statistical data analysis based on the L1-norm and related methods, Proceedings of the Third International Conference on Statistical Data Analysis Based on the L1-norm and Related Methods*. Elsevier, 1997.

[30] Z. Drezner, H. W. Hamacher, *Facility Location: Applications and Theory*, Springer, 2004.

[31] R. O. Duda, P. E. Hart, *Use of the Hough Transformation to detect lines and curves in pictures*, Communications of the ACM, **15**(1972) 11–15.

[32] M. Ester, H. Kriegel, J. Sander, *A density-based algorithm for discovering clusters in large spatial databases with noise*, In: *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, 1996, 226–231.

[33] B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.

[34] L. A. Fernandes, M. M. Oliveira, *Real-time line detection through an improved Hough transform voting scheme*, Pattern Recognition, **41**(2008) 299–314.

[35] M. Fischler, R. Bolles, *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*, Communications of the ACM, **24**(1981) 381–395.

[36] H. Frigui, C. Hwang, F. C.-H. Rhee, *Clustering and aggregation of relational data with applications to image database categorization*, Pattern Recognitions, **40**(2007) 3053–3068.

[37] R. Grbić, D. Grahovac, R. Scitovski, *A method for solving the multiple ellipses detection problem*, Pattern Recognition, **60**(2016) 824–834.

[38] R. Grbić, E. K. Nyarko, R. Scitovski, *A modification of the DIRECT method for Lipschitz global optimization for a symmetric function*, Journal of Global Optimization, **57**(2013) 1193–1212.

[39] C. Gurwitz, *Weighted median algorithms for l_1 approximation*, BIT, **30**(1990) 301–310.

[40] D. E. Gustafson, W. C. Kessel, *Fuzzy clustering with a fuzzy covariance matrix*, In: *Proc. IEEE Conf. Decision Control*, San Diego, CA, 1979, 761–766.

[41] J. P. H. Chen, *0-1 semidefinite programming for graph-cut clustering: modelling and approximation*, In: P. M. Pardalos, P. Hansen, editors, *Data Mining and Mathematical Programming*, 2008, 15–39.

[42] F. Höppner, F. Klawonn, *A contribution to convergence theory of fuzzy c-means and derivatives*, IEEE Transactions on Fuzzy Systems, **11**(2003) 682–694.

[43] E. Hüllermeier, M. Rifqi, S. Henzgen, R. Senge., *Comparing fuzzy partitions: A generalization of the Rand index and related measures.*, IEEE Transactions on Fuzzy Systems, **20**(2012) 546–556.

[44] C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters, **38**(2010) 207–214.

[45] M. Jiang, *On the sum of distances along a circle*, Discrete Mathematics, **308**(2008) 2038–2045.

[46] D. R. Jones, C. D. Perttunen, B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, Journal of Optimization Theory and Applications, **79**(1993) 157–181.

[47] D. Jukić, R. Scitovski, *Matematika I*, Odjel za matematiku, Sveučilište u Osijeku, 2017.

[48] D. Jukić, R. Scitovski, H. Späth, *Partial linearization of one class of the nonlinear total least squares problem by using the inverse model function*, Computing, **62**(1999) 163–178.

[49] L. Kaufman, P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, Chichester, UK, 2005.

[50] J. Kogan, *Introduction to Clustering Large and High-dimensional Data*, Cambridge University Press, New York, 2007.

[51] V. Leemans, M.-F. Destain, *Line cluster detection using a variant of the Hough transform for culture row localisation*, Image and Vision Computing, **24**(2006) 541–550.

[52] F. Leisch, *A toolbox for k-centroids cluster analysis*, Computational Statistics & Data Analysis, **51**(2006) 526–544.

[53] S. Mardešić, *Matematička analiza u n-dimenzionalnom realnom prostoru I: brojevi, konvergencija, neprekinutost*, Školska knjiga, Zagreb, 1991.

[54] K. V. Mardia, P. E. Jupp, *Directional Statistics*, Wiley, 2000.

[55] T. Marošević, *Data clustering for circle detection*, Croatian Operational Research Review, **5**(2014) 15–24.

[56] T. Marošević, K. Sabo, P. Taler, *A mathematical model for uniform distribution voters per constituencies*, Croatian Operational Research Review, **4**(2013) 53–64.

[57] T. Marošević, R. Scitovski, *Multiple ellipse fitting by center-based clustering*, Croatian Operational Research Review, **6**(2015) 43–53.

[58] T. Marošević, R. Scitovski, *Multiple ellipse fitting by center-based clustering*, Croatian Operational Research Review, **6**(2015) 43–53.

[59] D. Matijević, N. Truhar, *Uvod u računarstvo*, Odjel za matematiku, Sveučilište u Osijeku, 2012.

[60] D. J. Maširević, S. Miodragović, *Geometric median in the plane*, Elemente der Mathematik, **70**(2015) 21–32.

[61] B. Mirkin, *Data clustering for Data Mining*, Chapman & Hall/CRC, 2005.

[62] A. Morales-Esteban, F. Martínez-Álvarez, S. Scitovski, R. Scitovski, *A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning*, Computers & Geosciences, **73**(2014) 132–141.

[63] P. Mukhopadhyay, B. B. Chaudhuri, *A survey of Hough transform*, Pattern Recognition, **48**(2015) 993–1010.

[64] Y. Nievergelt, *Total least squares: state-of-the-art regression in numerical analysis*, SIAM Review, **36**(1994) 258–264.

[65] Y. Nievergelt, *A finite algorithm to fit geometrically all midrange lines, circles, planes, spheres, hyperplanes, and hyperspheres*, Numerische Mathematik, **91**(2002) 257–303.

[66] A. Okabe, B. Boots, K. Sugihara, *Spatial Tessellations: Concepts and Applications of Voronoi diagrams*, John Wiley & Sons, Chichester, UK, 2000.

[67] J. M. Ortega, W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia, 2000.

[68] R. Paulavičius, Y. Sergeyev, D. Kvasov, J. Žilinskas, *Globally-biased DISIMPL algorithm for expensive global optimization*, Journal of Global Optimization, **59**(2014) 545–567.

[69] R. Paulavičius, J. Žilinskas, *Simplicial Global Optimization*, volume X of *Series: Springer Briefs in Optimization*, Springer-Verlag, Berlin, 2014.

[70] J. D. Pintér, *Global Optimization in Action (Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications)*, Kluwer Academic Publishers, Dordrecht, 1996.

[71] F. Ricca, A. Scozzari, B. Simeoni, *Weighted voronoi region algorithms for political districting*, Mathematical and Computer Modelling, **48**(2008) 1468–1477.

[72] F. Ricca, B. Simeoni, *Local search algorithms for political districting*, Europian Journal of Operational Reacearch, **189**(2008) 1409–1426.

[73] L. Rotaru, *Identifying the phenotypic resemblances of the vine breeds by means of cluster analysis*, Notulae Botanicae, **37**(2009) 249–252.

[74] P. J. Rousseeuw, M. Hubert, *Robust statistics for outlier detection*, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., **1**(2011) 73–79, DOI: 10.1002/widm.2.

[75] P. J. Rousseeuw, A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 2003.

[76] S. Rueda, S. Fathima, C. L. Knight, M. Yaqub, A. T. Papageorghiou, B. Rahmatullah, A. Foi, M. Maggioni, A. Pepe, J. Tohka, R. V. Stebbing, J. E. McManigle, A. Ciurte, X. Bresson, M. B. Cuadra, C. Sun, G. V. Ponomarev, M. S. Gelfand, M. D. Kazanov, C.-W. Wang, H.-C. Chen, C.-W. Peng, C.-M. Hung, J. A. Noble, *Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge*, IEEE Transactions on Medical Imaging, **10**(2013) 1–16.

[77] K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal, **50**(2008) 185–198.

[78] K. Sabo, R. Scitovski, *An approach to cluster separability in a partition*, Information Sciences, **305**(2015) 208–218.

[79] K. Sabo, R. Scitovski, P. Taler, *Uniform distribution of the number of voters per constituency on the basis of a mathematical model (in Croatian)*, Hrvatska i komparativna javna uprava, **14**(2012) 229–249.

[80] K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka - klasteri*, Osječki matematički list, **10**(2010) 149–178.

[81] K. Sabo, R. Scitovski, I. Vazler, *One-dimensional center-based l_1-clustering method*, Optimization Letters, **7**(2013) 5–22.

[82] K. Sabo, R. Scitovski, I. Vazler, M. Zekić-Sušac, *Mathematical models of natural gas consumption*, Energy Conversion and Management, **52**(2011) 1721–1727.

[83] V. Schwämmle, O. N. Jensen, *A simple and fast method to determine the parameters for fuzzy c-means cluster analysis*, Bioinformatics, **26**(2010) 2841–2848.

[84] R. Scitovski, *Numerička matematika*, Odjel za matematiku, Sveučilište u Osijeku, 3, izdanje, 2015, `https://www.mathos.unios.hr/images/homepages/scitowsk/Num.pdf`.

[85] R. Scitovski, *A new global optimization method for a symmetric lipschitz continuous function and application to searching for a globally optimal partition of a one-dimensional set*, Journal of Global Optimization, (2017), DOI: 10.1007/s10898-017-0510-4.

[86] R. Scitovski, M. B. Alić, *Grupiranje podataka*, Odjel za matematiku, Sveučilište u Osijeku, 2016.

[87] R. Scitovski, S. Kosanović, *Rate of change in economics research*, Economics analysis and workers management, **19**(1985) 65–75.

[88] R. Scitovski, S. Maričić, S. Scitovski, *Short-term and long-term water level prediction at one river measurement location*, Croatian Operational Research Review, **3**(2012) 80–90.

[89] R. Scitovski, D. Marković, D. Brajković, M. MiloložaPandur, *Linearna algebra I*, 2018, Recenzirani nastavni materijali, Odjel za matematiku, Sveučilište u Osijeku `https://www.mathos.unios.hr/images/homepages/scitowsk/LA1.pdf`.

[90] R. Scitovski, T. Marošević, *Multiple circle detection based on center-based clustering*, Pattern Recognition Letters, **52**(2014) 9–16, Accepted.

[91] R. Scitovski, U. Radojičić, K. Sabo, *A fast and efficient method for solving the multiple line detection problem*, Rad HAZU, Matematičke znanosti, (2019).

[92] R. Scitovski, K. Sabo, *Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters*, Knowledge-Based Systems, **57**(2014) 1–7.

[93] R. Scitovski, K. Sabo, *Application of the DIRECT algorithm to searching for an optimal k-partition of the set A and its application to the multiple circle*

detection problem, Journal of Global Optimization, **74(1),**(2019) 63–77, DOI: 10.1007/s10898-019-00743-8.

[94] R. Scitovski, K. Sabo, *Application of the DIRECT algorithm to solving the multiple ellipse detection problem*, Applications of Mathematics, (2019).

[95] R. Scitovski, K. Sabo, *DBSCAN-like clustering method for various data densities*, Pattern Analysis and Applications, (2019), DOI: 10.1007/s10044-019-00809-z.

[96] R. Scitovski, K. Sabo, *A combination of k-means and DBSCAN algorithm for solving the multiple generalized circle detection problem*, (2020), Submitted.

[97] R. Scitovski, K. Sabo, D. Grahovac, *Globalna optimizacija*, Odjel za matematiku, 2017, `https://www.mathos.unios.hr/images/homepages/scitowsk/GOP.pdf`.

[98] R. Scitovski, S. Scitovski, *A fast partitioning algorithm and its application to earthquake investigation*, Computers & Geosciences, **59**(2013) 124–131.

[99] R. Scitovski, I. Vidović, D. Bajer, *A new fast fuzzy partitioning algorithm*, Expert Systems with Applications, **51**(2016) 143–150.

[100] Y. D. Sergeyev, D. E. Kvasov, *Lipschitz global optimization*, In: J. Cochran, editor, *Wiley Encyclopedia of Operations Research and Management Science*, volume 4. Wiley, New York, 2011, 2812–2828.

[101] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal„ **27**(1948) 379–423, 623–656.

[102] H. Späth, *Algorithm 48: afast algorithm for clusterwise linear regression*, Computing, **29**(1981) 175–181.

[103] H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

[104] P. N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Wesley, 2006.

[105] M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research, **8**(2007) 65–102.

[106] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, Burlington, 2009, 4^{th} edition.

[107] J. C. R. Thomas, *A new clustering algorithm based on k-means using a line segment as prototype*, In: C. S. Martin, S.-W. Kim, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Springer Berlin Heidelberg, 2011, 638–645.

[108] N. Truhar, *Numerička linearna algebra*, Odjel za matematiku, Sveučilište u Osijeku, 2010.

[109] I. Vazler, K. Sabo, R. Scitovski, *Weighted median of the data in solving least absolute deviations problems*, Communications in Statistics - Theory and Methods, **41:8**(2012) 1455–1465.

[110] L. Vendramin, R. J. G. B. Campello, E. R. Hruschka, *On the comparison of relative clustering validity criteria*, In: *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 – May 2, 2009, Sparks, Nevada, USA*, SIAM, 2009, 733–744.

[111] I. Vidović, D. Bajer, R. Scitovski, *A new fusion algorithm for fuzzy clustering*, Croatian Operational Research Review, **5**(2014) 149–159.

[112] I. Vidović, R. Cupec, e. Hocenski, *Crop row detection by global energy minimization.*, Pattern recognition, **55**(2016) 68–86.

[113] I. Vidović, R. Scitovski, *Center-based clustering for line detection and application to crop rows detection*, Computers and Electronics in Agriculture, **109**(2014) 212–220.

[114] Ž. Turkalj, D. Markulak, S. Singer, R. Scitovski, *Research project grouping and ranking by using adaptive mahalanobis clustering*, Croatian Operational Research Review, **7**(2016) 81–96.

[115] J. Ward (Jr.), *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, **58**(1963) 236–244.

[116] I. Wolfram Research, *Mathematica*, Wolfram Research, Inc., Champaign, Illinois, 2016, version 11.0 edition.

[117] K.-L. Wu, M.-S. Yang, *A cluster validity index for fuzzy clustering*, Pattern Recognition Letters, **26**(2005) 1275–1291.

[118] K. S. Younis, *Weighted Mahalanobis distance for hyper-ellipsoidal clustering*, Ph.D. thesis, Air Force Institute of Technology, Ohio, 1999.

[119] M. Zekić-Sušac, M. Knežević, R. Scitovski, *Deep learning in modeling energy cost of buildings in the public sector*, In: F. M. Álvarez, A. T. Lora, J. A. S. Munoz, H. Quintián, E. Corchado, editors, *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, Springer Nature Switzerland AG 2020, 2020, 101–110, DOI 978-3-030-20055-8-10, 2020.

[120] M. Zekić-Sušac, R. Scitovski, A. Has, *Cluster analysis and artificial neural networks in predicting energy efficiency of public buildings as a cost-saving approach*, Croatian Review of Economic, Business and Social Statistics (CREBSS), **4**(2018) 57–66.

[121] C. Zhang, Y. Zhou, T. Martin, *A validity index for fuzzy and possibilistic c-means algorithm*, In: L. Magdalena, M. Ojeda-Aciegoand, J. L. Verdegay, editors, *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2008, 877–882.

Indeks