# Statističke metode grupiranja u analizi podataka

**Uremović, Kristina**

**Master's thesis / Diplomski rad**

**2016**

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

*Permanent link / Trajna poveznica:* https://urn.nsk.hr/urn:nbn:hr:217:072997

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Kristina Uremović

# STATISTIČKE METODE GRUPIRANJA U ANALIZI PODATAKA

Diplomski rad

Voditelj rada:
prof.dr.sc.Siniša Slijepčević

Zagreb, 2016.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

*Zahvaljujem roditeljima na bezgraničnoj podršci i ljubavi iskazanoj tokom studranja.*
*Veliko hvala i svim ostalima koji su bili tu kad je trebalo.*

# Sadržaj

**Sadržaj** **iv**

**Uvod** **1**

**1 Teorija vjerojatnosti** **2**
1.1 Osnovne definicije . . . . . 2

**2 Pojam i grafičko detektiranje klastera** **7**
2.1 Pojam klaster analize . . . . . 7
2.2 Grafičko detektiranje klastera . . . . . 8

**3 Mjere bliskosti** **11**
3.1 Mjere bliskosti kategoričkih podataka . . . . . 11
3.2 Mjere različitosti i udaljenosti neprekidnih podatka . . . . . 13
3.3 Mjere sličnosti podataka sadržanih od neprekidnih i kategoričkih varijabli (mixed varijable) . . . . . 14
3.4 Mjere bliskosti strukturiranih podataka . . . . . 15
3.5 Mjere bliskosti između grupa . . . . . 17
3.6 Ponderiranje varijabli . . . . . 19
3.7 Standardizacija . . . . . 21
3.8 Izbor mjere bliskosti . . . . . 21

**4 Hijerarhijsko klasteriranje** **23**
4.1 Aglomerativne metode . . . . . 23
4.2 Metode dijeljenja . . . . . 28
4.3 Primjena hijerarhijskih metoda u klasteriranju . . . . . 30

**5 Metode optimizacije klastera** **38**
5.1 Kriteriji klasteriranja izvedeni iz matrice različitosti . . . . . 38
5.2 Kriteriji klasteriranja proizašli iz neprekidnih podataka . . . . . 40

5.3 Optimizacijski algoritmi . . . . . 44
5.4 Izbor broja klastera . . . . . 46

**6 Primjena klaster analize u medicini** **50**
6.1 Opisna statistika . . . . . 51
6.2 Primjena klaster analize . . . . . 51

**A Kodovi u R-u** **62**
A.1 Opisna statistika . . . . . 62
A.2 Klasterizacija . . . . . 64

**Bibliografija** **67**

# Uvod

Klaster analiza se bavi klasifikacijom, prepoznavanjem strukture i numeričkom taksonomijom. Traži se struktura podataka za grupiranje multivarijantnih opažanja u klastere na temelju dostupnih informacija koje opisuju podatke i njihove veze. Cilj je pronaći optimalan kriterij grupiranja kod kojeg su opažanja unutar svakog klastera slična, ali se različiti klasteri međusobno razlikuju.

U ovom ćemo radu dati definiciju klastera, teoretski i grafički. Kao uvod, a radi lakšeg daljnjeg čitanja, navodimo ključne definicije iz diskretne torije vjerojatnosti. U poglavlju 2 definiramo i pobliže objašnjavamo klaster analizu te kako nam prikazi podataka mogu pripomoći pri uočavanju klastera među danim podacima. Kroz poglavlje 3 objašnjavamo razne mjere bliskosti na kojima se temelji grupiranje podataka u klastere, a u poglavlju 4 metode klasterizacije temeljene na tim mjerama. Poglavlje 5 govori o tome na koji način se najbolje optimiziraju klasteri kako bi dobili što bolji uvid u danu bazu podataka.

Na kraju ćemo sve primjeniti na bazu podataka jedne bolnice.

Od našeg temeljnog interesa biti će klasteriranje podataka definiranih po redovima matrice podataka $\mathbf{X}$. Naime, nema određenog razloga zašto neke primjene klaster analize ne mogu biti provedene po stupcima matrice podataka $\mathbf{X}$.

Klaster analiza se esencijalno odnosi na otkrivanje grupa u podacima, te se metode klasteriranja ne treba miješati sa metodama diskriminacije i dodijeljivanja (assignment) gdje su grupe unaprijed poznate.

# Poglavlje 1

# Teorija vjerojatnosti

Radi lakšeg čitanja i razumijevanja daljenjeg teksta navodimo ključne definicije iz Teorije vjerojatnosti.

## 1.1 Osnovne definicije

Osnovni polazni objekt u teoriji vjerojatnosti jest neprazan skup $\Omega$ kojeg zovemo **prostor elementarnih događaja**, te on reprezentira skup svih ishoda slučajnog pokusa. Skup $\Omega$ i njegove elemente u daljnjem tekstu smatramo danima; oni su osnovni i nedefinirani pojmovi u teoriji vjerojatnosti. Točke $\omega$ skupa $\Omega$ zvat ćemo **elementarni događaji**.

**Definicija 1.1.1.** *Neka je $\Omega$ prostor elementarnih događaja. Familija $\boldsymbol{A}$ podskupova od $\Omega$ jest **algebra** skupova na $\Omega$ ako je*

1. $\emptyset \in \boldsymbol{A}$
2. $A \in \boldsymbol{A} \Rightarrow A^c \in \boldsymbol{A}$
3. $A_1, A_2, ..., A_n \in \boldsymbol{A} \Rightarrow \bigcup_{i=1}^{n} A_i \in \boldsymbol{A}$

**Definicija 1.1.2.** *Familija $\mathcal{F}$ podskupova od $\Omega$ jest $\sigma$**-algebra** skupova (na $\Omega$) ako je*

1. $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F},\ i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

**Definicija 1.1.3.** *Neka je $\mathcal{F}$ $\sigma$-algebra na skupu $\Omega$. Uređen par $(\Omega, \mathcal{F})$ zove se **izmjeriv prostor**.*

**Definicija 1.1.4.** *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor. Funkcija* $P : \mathcal{F} \longrightarrow \mathbb{R}$ *jest* ***vjerojatnost*** *ako vrijedi*

1. $P(A) \geq 0, A \in \mathcal{F}, P(\Omega) = 1$
2. $A_i \in \mathcal{F}, i \in \mathbb{N}\ i\ A_i \cap A_j = \emptyset\ za\ i \neq j \Rightarrow P(\bigcup_i^\infty A_i) = \sum_{i=1}^\infty P(A_i)$

**Definicija 1.1.5.** *Uređena trojka* $(\Omega, \mathcal{F}, P)$*, gdje je* $\mathcal{F}$ *σ-algebra na* $\Omega$ *i* $P$ *vjerojatnost na* $\mathcal{F}$*, zove se* ***vjerojatnosni prostor****.*

Ako je $\Omega$ prebrojiv, onda vjerojatnosni prostor $(\Omega, \mathcal{P}(\Omega), P)$ zovemo **prebrojiv diskretni vjerojatnosni prostor**. U slučaju konačnog ili prebrojivog skupa $\Omega$ točke iz skupa $\Omega$ zvat ćemo **elementarni događaji**.

**Definicija 1.1.6.** *Neka je* $(\Omega, \mathcal{P}(\Omega), P)$ *diskretan vjerojatnosni prostor.* ***Slučajna varijabla*** *proizvoljna je realna funkcija definirana na* $\Omega$ *.*

Slučajne se varijable označavaju velikim slovima latinice *X, Y, Z, ...* Dakle, $X$ je slučajna varijabla ako $X : \Omega \longrightarrow \mathbb{R}$. Intuitivno, slučajna varijabla je veličina koja se dobije mjerenjem u vezi s nekim slučajnim pokusom.

Za skup $A \subseteq \Omega$ sa $K_A$ označavamo **karakterističnu funkciju** skupa $A$. Funkcija $K_A : \Omega \longrightarrow \mathbb{R}$ definirana je sa

$$K_A = \begin{cases} 1, & \text{ako } \omega \in A \\ 0, & \text{inače} \end{cases} \tag{1.1}$$

Tada je očigledno

$$X = \sum_i a_i K_{A_i} \tag{1.2}$$

Lako se dokaže da prikaz slučajne varijable $X$ preko karakteristične funkcije nije jedinstven.

Slučajnu varijablu $X$ iz prethodne formule, najčešće ćemo označavati sa

$$X = \begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_1 & p_2 & p_3 & \dots & p_n \end{pmatrix} \tag{1.3}$$

U prvom retku matrice stoje sve moguće različite vrijednosti pokusa, dok su u drugom retku pripadne vjerojatnosti da $X$ poprimi te vrijednosti. Prema tome, svakoj slučajnoj varijabli na diskretnom vjerojatnosnom prostoru $(\Omega, \mathcal{P}(\Omega), P)$ na jednoznačan se način pridružuje gornja tablica, koju zovemo **distribucija slučajne varijable** $X$ ili **zakon razdiobe** od $X$.

**Definicija 1.1.7.** *Neka je* $(\Omega, \mathcal{P}(\Omega), P)$ *diskretan vjerojatnosni prostor i* $X_1, X_2, ..., X_k$ *slučajne varijable na* $\Omega$*. Kažemo da su* $X_1, X_2, ..., X_k$ ***nezavisne slučajne varijable*** *ako za proizvoljne* $B_i \subset \mathbb{R}$*,* $i = 1, ..., k$ *vrijedi*

$$\begin{aligned} P\{X_1 \in B_1, ..., X_k \in B_k\} &= P\{\omega \in \Omega; X_1(\omega) \in B_1, ..., X_k(\omega) \in B_k\} \\ &= P(\bigcap_{i=1}^{k} \{X_i \in B_i\}) \\ &= \prod_{i=1}^{k} P(\{X_i \in B_i\}). \end{aligned} \tag{1.4}$$

Iz definicije slijedi da su slučajne varijable $X_1, X_2, ..., X_k$ nezavisne ako i samo ako su za proizvoljne $B_i \subset \mathbb{R}$, $i = 1, ..., k$ događaji $\{X_1 \in B_1\}$, ..., $\{X_k \in B_k\}$ nezavisni.

**Definicija 1.1.8.** *Neka je* $\{X_t, t \in T\}$ *proizvoljna familija slučajnih varijabli. Kažemo da je to* ***familija nezavisnih slučajnih varijabli*** *ako su* $X_{t_1}, ..., X_{t_n}$ *nezavisne za svaki konačan podskup* $\{t_1, ..., t_n\}$ *različitih indeksa iz* $T$*.*

**Definicija 1.1.9.** *Neka je* $(\Omega, \mathcal{P}(\Omega), P)$ *diskretan vjerojatnosni prostor,* $\Omega = \{\omega_1, \omega_2, ...\}$ *i* $X$ *slučajna varijabla na* $\Omega$*, dakle* $X : \Omega \longrightarrow \mathbb{R}$*.*
*Ako red* $\sum_{\omega_k \in \Omega} X(\omega_k)P(\{\omega_k\})$ *apsolutno konvergira, onda njegovu sumu zovemo* ***matematičko očekivanje*** *ili kraće,* ***očekivanje slučajne varijable*** $X$ *i označavamo ga sa*

$$\mathbf{E}X = \sum_{\omega_k \in \Omega} X(\omega_k)P(\{\omega_k\}). \tag{1.5}$$

Primjetimo da u slučaju konačnog skupa $\Omega$ svaka slučajna varijabla ima očekivanje. Lako se vidi da, ako je funkcija $X$ konstantna, tj. za neko $c \in \mathbb{R}$ imamo da je $X(\omega_k) = c$ za sve $\omega_k \in \Omega$, tada je i $\mathbf{E}X = c$.

Neka je

$$X = \begin{pmatrix} a_1 & a_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix} \tag{1.6}$$

distribucija odnosno zakon razdiobe slučajne varijable $X$; prema tome $a_1, a_2, ...$ sve su različite realne vrijednosti koje $X$ poprima i $p_i = P(X = a_i)$. Tada vrijedi

**Teorem 1.1.10.** *Redovi* $\sum_{\omega_k \in \Omega} X(\omega_k)P(\{\omega_k\})$ *i* $\sum_{a_i} a_i p_i$ *istodnobno ili apsolutno konvergiraju ili apsolutno divergiraju. U slučaju apsolutne konvergencije suma im je ista, dakle vrijedi*

$$\mathbf{E}X = \sum_{a_i} a_i p_i. \tag{1.7}$$

Neka je X slučajna varijabla na diskretnom vjerojatnosnom prostoru $(\Omega, \mathcal{P}(\Omega), P)$ i neka je

$$X = \begin{pmatrix} a_1 & a_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix} \tag{1.8}$$

njezina **distribucija** odnosno **zakon razdiobe**.

**Definicija 1.1.11.** ***Funkcija gustoće*** *vjerojatnosti od X ili, kraće, gustoća od X jest funkcija* $f_X = f : X \longrightarrow \mathbb{R}$ *definirana s*

$$f(x) = P(\{X = x\}) = \begin{cases} 0 & ako\ x \neq a_i, x \in \mathbb{R} \\ p_i & ako\ x = a_i, x \in \mathbb{R} \end{cases} \tag{1.9}$$

*Neka je* $B \subset \mathbb{R}$*. Tada imamo*

$$P(X \in B) = P(X^{-1}(B^{-1} \cap \{a_1, a_2, ...\})) = \sum_{a_i \in B} P(X = a_i) = \sum_{a_i \in B} p_i = \sum_{x \in B} f(x). \tag{1.10}$$

**Definicija 1.1.12.** ***Funkcija distribucije*** *slučajne varijable X jest funkcija* $F_X = F : \mathbb{R} \longrightarrow [0, 1]$ *definirana sa*

$$F(x) = P(X \leq x) = P(\omega; X(\omega) \leq x), x \in \mathbb{R}. \tag{1.11}$$

**Definicija 1.1.13.** *Neka je X slučajna varijabla i neka* $\boldsymbol{E}X$ *postoji.* ***Varijanca*** *od X se definira sa*

$$VarX = \boldsymbol{E}[(X–EX)^2]. \tag{1.12}$$

***Standardna devijacija*** $\sigma_x$ *od X nenegativan je kvadratni korijen iz varijance, tj.* $\sigma_x = \sqrt{VarX}$.

**Definicija 1.1.14.** *Neka je* $(\Omega, \mathcal{P}(\Omega), P)$ *diskretan vjerojatnosni prostor.* ***Slučajni vektor*** *jest proizvoljna funkcija* $X : \Omega \longrightarrow \mathbb{R}^n$ *(u tom slučaju kažemo da je X n-dimenzionalan slučajni vektor).*

**Propozicija 1.1.15.** *Neka su X i Y slučajne varijable na* $\Omega$ *i neka postoji* $\boldsymbol{E}X^2$ *i* $\boldsymbol{E}Y^2$*. Tada postoji* $\boldsymbol{E}(XY)$ *i vrijedi*

$$|E(XY)| \leq (EX^2EY^2)^{1/2}. \tag{1.13}$$

Neka je $X = (X_1, X_2, ..., X_n)$ n-dimenzionalan slučajni vektor na $\Omega$ i neka postoji $\mathbf{E}X_i^2$. Stavimo $m_i = \mathbf{E}X_i$, $i = 1, ..., n$. Iz prethodne propozicije slijedi da postoje realni brojevi

$$\mu_{ij} = E(X_iX_j) – EX_iEX_j. \tag{1.14}$$

Simetričnu matricu ($\mu_{ij} = \mu_{ji}$)

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n1} & \mu_{n2} & \dots & \mu_{nn} \end{bmatrix} \tag{1.15}$$

zovemo **kovarijacijska matrica** slučajnog vektora $X = (X_1, ..., X_n)$ .

Također, iz prethodne propozicije slijedi da vrijedi nejednakost

$$\left|\mu_{ij}\right| \leq (\mu_{ii}\mu_{jj})^{1/2}. \tag{1.16}$$

Ako je $\mu_{ii} > 0$ i $\mu_{jj} > 0$, tada broj

$$\rho_{ij} = \frac{\mu_{ij}}{(\mu_{ii}\mu_{jj})^{1/2}}, i \neq j, i, j = 1, ..., n \tag{1.17}$$

zovemo **koeficijent korelacije** između slučajnih varijabli $X_i$ i $X_j$ . Prema istoj propoziciji vrijedi $\left|\rho_{ij}\right| \leq 1$.

**Definicija 1.1.16.** *Kažemo da su slučajne varijable X i Y (na* $\Omega$*)* ***nekorelirane*** *ako je* $cov(X, Y) = 0$.

# Poglavlje 2

# Pojam i grafičko detektiranje klastera

## 2.1 Pojam klaster analize

Klaster analiza se referira kao klasifikacija, prepoznavanje strukture i numerička taksonomija. Traži se struktura podataka za grupiranje multivarijantnih opažanja u klastere. Klasteri se formiraju na temelju dostupnih informacija koje opisuju podatke i njihove veze. Cilj je pronaći optimalan kriterij grupiranja kod kojeg su opažanja unutar svakog klastera slična, ali se različiti klasteri međusobno razlikuju. Pri tome se pretpostavlja da se može pronaći prirodan način grupiranja smislen za svakog istraživača. Međutim, u klaster analizi se unaprijed ne zna ni broj grupa, niti su grupe unaprijed poznate.

Da bi se opažanja grupirala u klastere, mnogi postupci počinju sa sličnostima između parova opažanja. Sličnosti su zasnovane na nekoj od mjera udaljenosti, osim u slučaju klasteriranja varijabli gdje je mjera sličnosti korelacija među varijablama.

Međutim, teško je dati formalnu definiciju klastera. Mnogi autori, na primjer Cormack (1971) i Gordon (1999), definiraju klaster pomoću unutarnje kohezije - homogenost i vanjske izolacije - separacija. Ali sama definicija neće pokriti sve moguće slučajeve. Nije u potpunosti jasno kako ćemo raspoznavati klastere na slikama, ali jedno je sigurno- udaljenosti između točaka će igrati veliku ulogu u detektiranju klastera.

U većini slučajeva provedbe klaster analize, polazi se od toga da svaki podatak čini svoj klaster, a u konačnici imamo set klastera koji sadrže sve podatke. Osnovni podaci za provedbu kalster analize su u većini primjera sadržani u $n \times p$ višedimenzionalnoj matrici, $\mathbf{X}$, koja sadrži sve vrijednosti varijabli svih podataka koje treba klasterirati; to je

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \tag{2.1}$$

Elementi $x_{ij}$ matrice $\mathbf{X}$ označavaju vrijednost $j$-te varijable $i$-tog podatka. Varijable matrice $\mathbf{X}$ često su mješavina neprekidnih, ordinalnih i/ili kategoričkih podataka, te se često događa da neki podaci nisu poznati pa imamo prazna mjesta u matrici. Miješane varijable, kao i one koje nedostaju, mogu zakomplicirati cijelu analizu. Također, često će se događati da redovi matrice $\mathbf{X}$ sadrže ponovljena mjerenja iste varijable, ali u drugačijim okolnostima.

Ponekad klaster analiza počinje konvertiranjem matrice $\mathbf{X}$ u $n \times n$ matricu sličnosti, različitosti ili udaljenosti (opći pojam za to je *neposredna blizina - proximity*), što će biti objašnjeno u jednom od sljedećih poglavlja.

## 2.2 Grafičko detektiranje klastera

Često se prisutnost klastera u cjelokupnoj bazi podataka može vidjeti iz jednostavnih grafova, poboljšanih izračunom funkcije gustoće. Jednodimenzionalni ili dvodimenzionalni klasteri mogu se identificirani različitim izračunima funkcije gustoće. Korisnost grafičkih prikaza je opravdana činjenicom da čovjek pomoću vizualne percepcije može otkriti uzorke koji bi tvorili klaster. Također, grafički prikazi mogu dati detaljniji uvid u podatke kojima raspolažemo. U ovom dijelu opisujemo relativno jednostavne statističke tehnike koje su često korisne pri davanju dokaza za ili protiv postojanja klaster strukture u danoj bazi podataka.

### Detektiranje klastera pomoću jedno- i dvodimenzionalnih prikaza podataka

Općenito pravilo ove metode je da je prisustvo nekog stupnja multimodalnosti u podacima dovoljno jak dokaz u korist nekog tipa klaster strukture. Baziramo se na jedno- i dvodimenzionalne grafove podataka.

### Histogram

Najbolji uvid u podatke nam daje histogaram, kao prvi korak u analizi podataka, naravno, ako su isti jednodimenzionalni.

### Scatterplots

Osnovni način prikazivanja dvodimenzionalnih podataka je *xy*-scatterplot. Scatterplotovi su dobri za prikaz manjeg broja podataka jer u suprotnom može doći do velikog broja preklapanja što može dodatno otežati uočavanje klastera.

**Definicija 2.2.1.** *Scatterplot matrica je kvadratna, simetrična mreža dvodimenzionalnih scatterplotova koja se satoji od p redaka i q stupaca, za svaku od p varijabli.*

Scatterplotove i histograme možemo poboljšati dodavanjem numeričke procjene dvodimenzionalne (jednodimenzionalne) funkcije gustoće podataka. Ako pretpostavimo određenu formu distribucije podataka, na primjer dvodimenzionalna (jednodimenzionalna) Gausova razdioba, procjena gustoće će tada biti reducirana na jednostavan slučaj izračunavanja vrijednosti parametara funkcije gustoće. Mi želimo da podaci govore sami za sebe te ćemo koristiti jednu od ponuđenih metoda neparametarske procjene. Zadovoljiti ćemo se kratkim prikazom procjene jezgre funkcije gustoće, počevši s procjenom u jednodimenzionalnom slučaju te prelazak na slučaj dvodimenzionalnih podataka, koji je važniji za klaster analizu.

U slučaju jednodimenzionalnih podataka imamo sljedeće:
Iz definicije vjerojatnosne funkcije gustoće, ako slučajna varijabla $X$ ima gustoću $f$, tada

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h). \tag{2.2}$$

Za svaki $h$, pravi (naive) procjenitelj funkcije $P(x - h < X < x + h)$ je omjer svih opažanja $X_1, X_2, ..., X_n$ iz intervala $(x - h, x + h)$, to jest

$$\hat{f}(x) = \frac{1}{2hn} \left[\text{broj varijabli } X_1, X_2, ..., X_n \text{u intervalu } (x - h, x + h)\right] \tag{2.3}$$

Ako uvedemo funkciju težina (weight function) $W$ danu s

$$W(x) = \begin{cases} \frac{1}{2} & \text{ako } |x| < 1 \\ 0 & \text{inače} \end{cases} \tag{2.4}$$

tada se pravi procjenitelj može napisati kao

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - X_i}{h}\right) \tag{2.5}$$

Nažalost, ovakav procjenitelj, odnosno funkcija ne zadovoljava svojstvo neprekidnosti. Međutim, takav procjenitelj implicira procjenitelja jezgre danog s:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.6}$$

gdje je $K$ funkcija jezgre, a $h$ propusnost ili parametar izglađivanja. Funkcija jezgre mora zadovoljavati

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{2.7}$$

Uobičajeno, ali ne uvijek, će funkcija jezgre biti simetrična funkcija gustoće (na primjer normalna). Procjenitelj jezgre je zbroj ispupčenja (skokova) opažanja. Funkcija jezgre određuje oblik skokova dok širina $h$ određuje njihovu širinu.
Tri funkcije jezgre koje se općenito koriste su:

1. pravokutna $K(x) = \frac{1}{2}$ za $\det x < 1$, 0 inače
2. trokutasta $K(x) = 1 - \det x$ za $\det x < 1$, 0 inače
3. Gaussova $K(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}$.

Procjenitelj gustoće jezgre funkcije (krenel density estimator), shvaćen kao zbroj skokova centriranih s obzirom na opažnja, lako se može proširiti na slučaj dvije dimenzije. Dvodimenzionalni procjenitelj podataka $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ definira se

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right) \tag{2.8}$$

gdje svaka koordinata ima svoj parametar glatkoće, $h_x, h_y$.

Za izračun dvodimenzionalnog procjenitelja gustoće uobičajeno se koristi funkcija jezgre standardne normalne dvodimenzionalne funkcije gustoće

$$K(x, y) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x^2 + y^2)\right] \tag{2.9}$$

Druga opcija je dvodimenzionalna Epanechikov jezgra dana s

$$K(x, y) = \frac{2}{\pi}(1 - x^2 - y^2), \tag{2.10}$$

U slučaju višedimenzionalnih podataka, sa tri li više varijabli, još uvijek koristimo scatterplotov svakog para varijabli kao početnu bazu ispitivanja podataka i otkrivanja klastra među njima. Odnosno, kada su scatterplotovi organizirani u *scatterplot matricu*. Svaka pozcija matrice predstavlja scatterplot dvije varijable. Bez obzira na svojstvo simetričnosti matrice s obzirom na glavnu dijagonalu, svaki puta crtamo i gornji i donji trokut matrice.

# Poglavlje 3

# Mjere bliskosti

Prilikom određivanja klastera u danim opažanjima od glavnog interesa je znati koliko 'blizu' su opažanja, to jest koliko su udaljena jedna od drugih. Mnogi primjeri počinju sa $n \times n$ matricom čiji elemnti predstavljaju udaljenosti, nazvane mjere udaljenosti ili sličnosti. Opći pojam za takve mjere je neposredna blizima-*proximity*. Dva opažanja su blizu kada im je različitost ili udaljenost mala ili sličnost velika. Neposredna blizina može biti određena direktno ili indirektno. Indirektno određivanje je izvedeno iz $n \times p$ matrice $\mathbf{X}$, objašnjene na početku.

Postoji širok krug mjera neposredne blizine. U ovom poglavlju ćemo se osvrnuti na neke od njih koje se najviše koriste. Krećemo sa mjerama primjenjivim na kategoričke varijable, zatim na neprekidne varijable i na kraju završavamo s mjerama primjenjivim na podatke koji sadrže i kategoričke i neprekidne varijable. Također, od interesa je objasniti i mjere neposredne blizine primjenjive na podatke koji sadrže opažanja mjerena više puta

## 3.1 Mjere bliskosti kategoričkih podataka

**Definicija 3.1.1.** ***Kategoričke varijable****, koje se često nazivaju i kvalitativne, sadrže podatke podijeljene u grupe ili poredane po veličini. Kategorički podaci razdvajaju ispitanike u jasno razgraničene grupe po određenoj karakteristici ili osobini. Na primjer spol (muški ili ženski), bračni status (neoženjen, oženjen, razveden, udovac).*

Ove mjere se primjenjuju na podatke koji sadrže kategoričke varijable. Mjere su sklairane na način da su sadržane u intervalu $[0, 1]$, iako su ponekad iskazane u postotcima od 0% do 100%. Dva opažanja $i$ i $j$ imaju koeficijent sličnosti $s_{ij}$ ako oba imaju jednake vrijednosti za neke varijable. Koeficijent sličnosti 0 implicira da se opažanja razlikuju u potpunosti.

## Mjere sličnosti za binarne podatke

Najčešći slučaj višedimenzionalnih kategoričkih podataka je onaj u komu su sve varijable binarne te imamo velik broj mjera sličnosti za takve tipove podataka. Sve mjere su definirane u terminima ulaska u kros-kvalifikaciju broja podudaranja i nepodudaranja u $p$ varijabli dvaju opažanja. Općenito, kros-kvalifikacija je prikaza u sljedećoj tablici:

| | | Individual i | | |
|---|---|---|---|---|
| | Outcome | 1 | 0 | Total |
| Individual j | 1 | $a$ | $b$ | $a+b$ |
| | 0 | $c$ | $d$ | $c+d$ |
| | Total | $a+c$ | $b+d$ | $p=a+b+c+d$ |

Slika 3.1: Broj binarnih ishoda dvaju podataka

Lista mjera sličnosti za binarne podatke je sljedeća:

1. Koeficijent podudaranja $s_{ij} = \frac{a+d}{a+b+c+d}$
2. Jaccard koeficijent $s_{ij} = \frac{a}{a+b+c}$
3. Rogers i Tanimoto $s_{ij} = \frac{a+d}{[a+2(b+c)+d]}$
4. Sneath i Sokal $s_{ij} = \frac{a}{[a+2(b+c)]}$
5. Gower i Legendre $s_{ij} = \frac{a+d}{\left[a+\frac{1}{2}(b+c)+d\right]}$
6. Gower i Legendre $s_{ij} = \frac{a}{\left[a+\frac{1}{2}(b+c)\right]}$

Razlog ovako dugog popisa je da se izbjegne neodumica u interpretaciji nula-nula podudaranja. U nekim slučajevima se nula-nula podudaranja poistovjećuju s jedan-jedan podudaranjem te i to treba biti sadržano u izračunatoj mjeri bliskosti. Međutim, treba voditi brigu i o tome da li odsutnost sadrži korisnu informaciju po pitanju sličnosti dvaju opažanja. Mjere koje ignoriraju odsutnost broja d su Jaccardov koeficijent ili koeficijent predložen od Sneath i Sokal. Kada se odsutnost smatra dobrom informacijom koristimo koeficijent podudaranja. Mjere 3 i 5 su dodatni primjeri simetričnih koeficijenata koji na isti način tretiraju pozitivna ($a$) i negativna ($b$) podudaranja. Koeficijenti se razlikuju samo u težinama koje pripisuju za podudaranja i nepodudaranja.

### Mjere sličnosti za kategoričke podatke sa više od 2 nivoa (levela)

Kategorički podaci u kojima varijable imaju više od 2 nivoa (npr. boja očiju) mogu se obraditi na sličan način kao i binarne varijable. Međutim, taj pristup nije naročito atraktivan zbog pojavljivanja negativnih podudaranja. Bolja metoda je dodijeliti ocjenu (score) $s_{ijk}$ nula ili jedan svakoj varijabli $k$ ovisno o tome da li su komponente $i$ i $j$ iste za tu varijablu ili ne. Nadalje, te ocjene su uprosiječene u odnosu na sve $p$ varijable te dobivamo traženi koeficijent sličnosti:

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} \tag{3.1}$$

Alternativna definicija sličnosti kategoričkih varijabli je da podijelimo sve moguće ishode *k-te* varijable u međusobno disjunktne podskupove kategorija, dodijeliti $s_{ijk}$ nula ili jedan ovisno o tome da li su dvije kategorije komponenata $i$ i $j$ u istom podskupu ili ne, te zatim odrediti omjer podskupova svih varijabli. Ovakav način izračuna mjere sličnosti se koristi u proučavanju kako su dva jezika povezana.

## 3.2 Mjere različitosti i udaljenosti neprekidnih podatka

Kada su sve varijable neprekidne, neposredna udaljenost između komponenti se zove *mjera različitosti (udaljenosti)*, gdje se mjera različitosti $\delta_{ij}$ naziva i mjera udaljenosti ako zadovoljava nejednakost trokuta:

$$\delta_{ij} + \delta_{im} \geq \delta_{jm} \tag{3.2}$$

za parove komponenti $ij$, $im$ i $jm$. Za $n \times n$ matricu različitosti, $\Delta$, čiji su elementi $\delta_{ij}$, gdje $\delta_{ii} = 0$ za sve $i$, kažemo da je metrička ako zadovoljava nejednakost (3.2) za sve trojke $(i, j, m)$. Iz nejednakosti trokuta slijedi da je različitost između komponenti $i$ i $j$ ista kao i između komponenti $j$ i $i$, te da ako su dvije točke blizu tada je i treća točka u sličnom odnosu s druge dvije. Metričke različitosti su prema definiciji nenegativne. Kada govorimo o mjerama udaljenosti tada $n \times n$ matricu udaljenosti označavamo sa $\mathbf{D}$, s elementima $d_{ij}$.

Za izvođenje matrice različitosti iz skupa neprekidnih višedimenzionalnih opažanja dane su mjere:

1. Euklidska udaljenost $d_{ij} = \left[\sum_{k=1}^{p} {w_k}^2 (x_{ik} - x_{jk})^2\right]^{1/2}$
2. Udaljenost blokova (City block distance) $d_{ij} = \sum_{k=1}^{p} w_k \left|x_{ik} - x_{jk}\right|$
3. Udaljenost Minkowskog $d_{ij} = \left(\sum_{k=1}^{p} {w_k}^r \left|x_{ik} - x_{jk}\right|^r\right)^{1/r}, r \geq 1$

4. Pearsonova korelacija $\delta_{ij} = \left(1 - \Phi_{ij}\right)/2$ sa

   $\Phi_{ij} = \sum_{k=1}^{p} w_k(x_{ik} - \bar{x_{i^*}})(x_{jk} - \bar{x_{j^*}}) / \left[\sum_{k=1}^{p} w_k(x_{ik} - \bar{x_{i^*}})^2 \sum_{k=1}^{p} w_k(x_{jk} - \bar{x_{j^*}})^2\right]^{1/2}$

   gdje je $\bar{x_{i^*}} = \sum_{k=1}^{p} w_k x_{ik} / \sum_{k=1}^{p} w_k$

5. Kutna separacija $\delta_{ij} = (1 - \Phi_i j)/2$

   sa $\phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} / \left(\sum_{k=1}^{p} w_k {x_{ik}}^2 \sum_{k=1}^{p} w_k {x_{jk}}^2\right)^{1/2}$

Dane mjere različitosti možemo podijeliti u dvije grupe: mjere udaljenosti i mjere korelacije. Najčešće korištena mjera udaljenosti je Euklidska udaljenost gdje su $x_{ik}$ i $x_{jk}$, redom, k-ta vrijednost varijable p-dimenzionalnog opažanja po komponentama $i$ i $j$. Formalno se zove i $l_2$ norma. Nadalje, City block distance ili $l_1$ norma opisuje pravocrtne udaljenosti. Obje mjere su specijalni slučajevi udaljenosti Minkowskog ili $l_r$ norme.

Mjera Pearsonove korelacije i kutne separacije su primjeri mjera različitosti izvedenih iz koeficijenata korelacije. Pearsonova mjera povlači Pearsonov koeficijent korelacije dok mjera kutne separacije povlači vektorsko množenje. S obzirom da za koeficijent korelacije vrijedi

$$-1 \leq \phi_{ij} \leq 1, \tag{3.3}$$

gdje vrijednost $'1'$ znači najjaču moguću pozitivnu povezanost, a vrijednost $' - 1'$ najjaču moguću negativnu povezanost, ti se koeficijenti mogu transformirati u različitosti, $\delta_{ij}$, sadržane u intervalu $[0, 1]$, što je vidljivo iz same definicije mjere.

U suštini, koeficijent korelacije ne može mjeriti razlike u veličini dvaju opažanja. Međutim, njegovo korištenje je opravdanao u situacijama kada su sva opažanja mjerena na istoj skali, te uzete precizne vrijednosti su važne samo u tome što predstavljaju.

## 3.3 Mjere sličnosti podataka sadržanih od neprekidnih i kategoričkih varijabli (mixed varijable)

Mnogo je pristupa izvođenju mjera bliskosti za baze podataka koje sadrže i neprekidne i kategoričke varijable. Jedna od mogućnosti je razdijeliti sve varijable i koristiti mjeru sličnosti za binarne podatke. Druga mogućnost je skalirati sve podatke, tako da svi budu na istoj skali, zamijenjujući vrijednosti varijabli njihovim pozicijama u promatranim objektima, te zatim koristiti mjere za neprekidne podatke. Treća mogućnosti je konstruirati mjeru različitosti za oba tipa varijabli i kombinirati ih sa ili bez ponderiranja u jedan koeficijent.

Mi ćemo se koncentrirati na mjeru sličnosti koju je predložio Grower (1971.). Opći oblik njegove mjere je

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} / \sum_{k=1}^{p} w_{ijk}, \tag{3.4}$$

gdje je $s_{ijk}$ sličnost između *i-te* i *j-te* komponente mjerene u *k-toj* varijabli, $w_{ijk}$ je obično jedan ili nula, ovisno da li je usporedba točna. Vrijednost $w_{ijk}$ je nula ako nedostaje ishod u *k-toj* varijabli na jednoj ili objema komponenatama *i* i *j*. U suštini, $w_{ijk}$ može biti postavljeno na nulu ako je *k-ta* varijabla binarna i može se izbaciti negativna podudaranja. Za binarne i kategoričke varijable sa više od dvije kategorije, koeficijent sličnosti, $s_{ijk}$ poprima vrijednost jedan kada dvije komponente imaju istu vrijednost, inače je nula. Za neprekidne varijable, Grower predlaže korištenje mjere sličnosti

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}, \tag{3.5}$$

gdje je $R_k$ domet opažanja *k-te* varijable.

## 3.4 Mjere bliskosti strukturiranih podataka

Za ponovljena mjerenja dobivena u različitim uvijetima ispitivanja, recimo A, B ili C, referentni vektor je oblika $(A, ..., A, B, ..., B, C, ..., C)'$. Ovdje promatramo mjere neposredne blizine (bliskosti) upravo za takve strukturirane podatke; to jest mjere koje uzimaju u obzir da sve vrijednosti varijable proizlaze iz istog prostora podataka te priznaju i koriste postojanje referentne varijable. Shvaćanje ponovljenih mjerenja kao matrice podataka sa pripadajućim referentim vektorom korisno je pri određivanju odgovarajućeg sažetka vrijednosti varijabli promatranog objekta i rezultirajućih mjerenja različistosti među promatrim objektima. Također, to nam pomaže i pri modeliranju očekivane vrijednosti (mean) i kovarijance ponovljenih mjerenja pomoću smanjenog(reduciranog) skupa parametara.

Najjednostavniji i najčešće upotrebljavan pristup iskorištavanja referentne varijable je konstrukcija smanjenog seta relevantnih sažetaka promatranog objekata koji se dalje upotrebljava kao baza za definiranje sličnosti među objektima. Što definira odgovarajuću zajedničku mjeru ovisiti će o kontekstu materije istraživanja. Podrobnije ćemo objasniti pristupe odabira sažete mjere i rezultirajuće mjere bliskosti za referentne vektore koje najčešće susrećemo - 'vrijeme', 'uvijeti eksperimenta' i 'ishodišne činjenice'.

Kada je referentna varijabla vrijeme i poznata nam je funkcionalna forma krivulje individualnog vremena, tada procjene parametara dobivenih primjenom linerarne ili nelinearne regresije na individualna vremena može predstavljati jedan skup takvih sažetaka.

Kada referentna varijabla raspodjeljuje ponovljena mjerenja u broj klasa, tada je uobičajen izbor sažete mjere objekta po klasi očekivanje. Sažeti pristup može biti proširen korištenjem

ne samo sažete mjere koja nas zanima, nego i preciznosti tih procjena u konstrukciji bliskosti.

Često strukturirani podaci proizlaze kada možemo pretpostaviti da varijable slijede poznati faktorski model. Ukratko, pod tako zvani potvrdni model faktorske analize svaka varijabla ili stavka može biti dodijeljena jednom od setova temeljnih činjenica ili ideja. Faktori ne mogu biti promatrani direktno, ali su nagoviješteni brojem stavki tako da je svaka mjerena na istoj skali.Varijabla kategoričkog referentnog faktora se, pri konstrukciji prikladnih sažetaka po razini faktora, može upotrijebiti kao i kategoričko svojstvo referentne varijable.

U konačnici, primjetimo da pristup sažetaka, uobičajeno korišten za neprekidne varijable, nije ograničen samo na varijable na ljestvici. Isti princip se može koristiti u radu s kategoričkim varijablama. Razlika leži u tome da sada sažete mjere moraju sadržavati relevantna stajališta glede distribucije kategoričkih varijabli ponovljenih mjerenja. Sažetci poput kvantila, omjera posebnih kategorija ili način distribucije će biti očigledni odabiri.

Retci matrice **X**, koji predstavljaju određene liste elemenata, odnosno sve varijable pružaju kategorički ishod te se ti ishodi mogu urediti u jednu dimenziju, općenito se odnose kao nizovi. Nizovi stvaraju skup kategoričkih ponovljenih mjerenja, kojima je struktura određena referentnim vektorom koji implicira poziciju varijable u dimenziji u kojoj poravnanje zauzima mjesto (npr. pozicija u riječi, vremenu). Neki pristupi ponovljenim mjerenjima u vremenskoj domeni mogu biti korišteni za konstrukciju bliskosti, ali poslijednja zanimanja za nizove, posebno u području genetike, potaknula su razvitak algoritama za određivanje mjera različitosti koje posebno koriste prirodu poravnavanja kategoričkih podataka.

Tako zvana analiza nizova je područje istraživanja korišteno u sociologiji i psihologiji, a u centar interesa stavlja probleme događaja i postupaka u njihovom vremenskom kontekstu te uključuje i mjerenja sličnosti između nizova. Najpoznatija mjera različitosti nizova je *Levenshtein udaljenost* (Levenshtein, 1966.) koja traži minimalan broj operacija potrebnih da se jedan niz kategorija transformira u drugi, gdje se pod operacijama misli na presjek, izbacivanje ili supstitucija jedne kategorije. Takve operacije su primjenjive samo na poravnane uređene skupove kategorija te prebrojavanje operacija vodi ka mjeri različitosti nizova. Svakoj operaciji se dodijeljuje njena težina (uobičajeno je dodijeliti duplu težinu za supstituciju u odnosu na presjek ili izbacivanje). Ponekad se takva mjera naziva i *'uređivanje udaljenosti'*.

Optimalan algoritam spajanja mora pronaći minimalan broj operacija potrebnih za spajanje jednog niza u drugi. Jedan od takvih algoritama je Needleman-Wunsch algoritam (Needleman i Wunsch, 1970) koji se koristi u bioinformatici.

*Jaro mjera sličnosti* (Jaro, 1995) je mjera sličnosti nizova kategorija često korištena za brisanje duplikata u području evidencije veza (record linkage, linkage=povezivanje). Daje korist informaciji poravnavanja brojeći broj, $m$, podudaranja značenja i broj, $t$, premještanja. Dvije kategorije se smatraju podudarajućim ako nisu udaljenije od $p/2 - 1$

pozicija od svake pozicije na poravnavajućoj skali. Prijelaz je zamjena dviju kategorija u nizu. Tada se Jaro sličnost definira:

$$s^{Jaro} = \frac{1}{3}\left(\frac{2m}{p} + \frac{m-t}{m}\right). \tag{3.6}$$

## 3.5 Mjere bliskosti između grupa

Do sada smo se bavili mjerenjem bliskosti među podacima. Međutim, za klaster analizu je bitno i kako mjeriti bliskosti između grupa podataka. Postoje dva pristupa definiranju takvih mjera bliskosti. Prvi kaže da se bliskost dviju grupa može definirati preko prikladnog sažetka bliskosti među podacima svake grupe. Drugi pristup kaže da se svaka grupa može opisati predstavnikom svih opažanja te grupe, birajući prikladan sažetak statistike svake varijable, te se bliskost između grupa definira kao bliskost između predstavnika opažanja.

### Bliskost između grupa izvedena iz matrice bliskosti

Brojne su mogućnosti izvođenja bliskosti između grupa iz matrice bliskosti podataka. Možemo uzeti najmanju različitost između dva podatka, po jedan iz svake grupe, što se u kontekstu udaljenosti definira kao najbliža udaljenost susjeda i temelj je metode klasterizacije zvane *jedna veza* (single linkage). Suprotno tomu bilo bi definirati udaljenost između grupa kao najveću udaljenost između dva podatka, po jedan iz svake grupe. To je poznato kao najdalja udaljenost susjeda i sadrži bazu klaster metode *potpune povezanosti* (complete linkage). Međutim, uvijek umjesto krajnosti možemo uzeti prosječnu udaljenost među podacima obiju grupa. Takve mjere se koriste u *group average clustering metodi*.

### Bliskost između grupa bazirana na grupnim sažetcima neprekidnih podataka

Jedno od očitih načina kako konstruirati međugrupna mjerenja različitosti neprekidnih podataka je da zamijenimo srednje vrijednosti grupa (znata kao centroid) za vrijednosti pojedinih varijabli u formuli Euklidske udaljenosti ili city block udaljenosti. Ako, na primjer, grupa A ima vektor očekivanih vrijednosti (mean vector) $\bar{\mathbf{x}}'_A = (\bar{x}_{A1}, ..., \bar{x}_{Ap})$, a grupa B $\bar{\mathbf{x}}'_B = (\bar{x}_{B1}, ..., \bar{x}_{Bp})$ tada se Euklidska udaljenost definira kao:

$$d_{AB} = \left[\sum_{k=1}^{p} (\bar{x}_{Ak} - \bar{x}_{Bk})^2\right]^{1/2}. \tag{3.7}$$

Prikladnije mjere mogu biti one koje podrazumijevaju znanje o varijaciji unutar grupe. Jedna mogućnost je koristiti Mahalanobisovu generaliziranu udaljenost $D^2$ (Mahalanobis, 1936) danu sa:

$$D^2 = (\bar{x}_A - \bar{x}_B)' \mathbf{W}^{-1} (\bar{x}_A - \bar{x}_B), \tag{3.8}$$

gdje $\mathbf{W}$ udružuje matrice kovarijanci dviju grupa. Mahalanobisova udaljenost raste s rastućom udaljenosti između centara dviju grupa i opada s unutar-grupnom varijacijom. Uzimajući u obzir i unutar-grupne korelacije, Mahalanobisova udaljenost uzima u obzir i oblik grupe.

Korištenje Mahalanobisove udaljenosti, $D^2$, implicira da je ispitivač spreman pretpostaviti da su matrice kovarijacije približno jednake u obje grupe. Kada to nije tako, $D^2$ je neprikladna međugrupna mjera, te postoji nekoliko alternativnih rješenja takvih situacija. Chaddha i Marcus (1968) su predložili tri takve mjere, te takvu mjeru udaljenosti između grupa definiraju s:

$$\delta_{AB} = \max_t \frac{2\mathbf{b}_t'\mathbf{d}}{(\mathbf{b}_t'\mathbf{W}_A\mathbf{b}_t)^{1/2} + (\mathbf{b}_t'\mathbf{W}_B\mathbf{b}_t)^{1/2}}, \tag{3.9}$$

gdje su $\mathbf{W}_A$ i $\mathbf{W}_B$ $p \times p$ matrice kovarijanci grupa A i B, redom, $\mathbf{d} = \bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B$ i $\mathbf{b}_t = (t\mathbf{W}_A + (1-t)\mathbf{W}_B)^{-1}\mathbf{d}$.

Druga alternativa je *normal information radius* (NIR) kojeg su predložili Jardine i Sibson(1971). Mjera je definirana sa:

$$NIR = \frac{1}{2}\log_2 \left\{ \frac{det\left[\frac{1}{2}(\mathbf{W}_A + \mathbf{W}_B)\right] + \frac{1}{4}(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)'(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)}{det(\mathbf{W}_A)^{1/2} det(\mathbf{W}_B)^{1/2}} \right\}. \tag{3.10}$$

Kada su $\mathbf{W}_A = \mathbf{W}_B = \mathbf{W}$, gornja formula se reducira na:

$$NIR = \frac{1}{2}\log_2 \left(1 + \frac{1}{4}D^2\right), \tag{3.11}$$

gdje je $D^2$ Mahalanobisova udaljenost.

## Međugrupne bliskosti bazirane na grupnim sažetcima kategoričkih podataka

Mnogi autori su se bavili istraživanjem mjera različistoti grupa kategoričkih podataka. Balakrishnan i Sanghvi (1968) su predložili formu indeksa različitosti:

$$G^2 = \sum_{k=1}^{p} \sum_{l=1}^{c_k+1} \frac{(p_{Akl} - p_{Bkl})^2}{p_{kl}}, \tag{3.12}$$

gdje su $p_{Akl}$ i $p_{Bkl}$ omjeri $l$-te kategorije $k$-te varijable u grupi A, odnosno B. Dok je $p_{kl} = \frac{1}{2}(p_{Akl} + p_{Bkl})$, a $c_k + 1$ je broj kategorija $k$-te varijable i $p$ je broj varijabli.

Kurczynski (1969) predlaže prilagođavanje Mahalanobisove udaljenosti, gdje kategoričke varijable zamjenjuju kvantitativne varijable. U njenoj najopćenitijoj verziji, mjera je definirana:

$$D_p{}^2 = (\mathbf{p}_A - \mathbf{p}_B)' \mathbf{W}_p{}^{-1} (\mathbf{p}_A - \mathbf{p}_B), \tag{3.13}$$

gdje $\mathbf{p}_A = (p_{A11}, p_{A12}, ..., p_{A1c_1}, p_{A21}, p_{A22}, ..., p_{A2c_2}, ..., p_{Ak1}, p_{Ak2}, ..., p_{Akc_k})$ sadrži proporcije uzoraka u grupi A, $\mathbf{p}_B$ se definira na sličan način, a $\mathbf{W}_p$ je $m \times m$ zajednička matrica kovarijanci uzoraka gdje je $m = \sum_{k=1}^{p} c_k$. S obzirom na to kako su elementi matrice $\mathbf{W}_p$ izračunati, možemo izvesti razne oblike ovakve mjere različitosti. Naime, ako sve varijable imaju multinomijalnu distribuciju i međusobno su nezavisne, tada se mjera različistosti definirana formulom (3.13) podudara s mjerom definiranom u (3.12).

## 3.6 Ponderiranje varijabli

Prilikom određivanja bliskosti između dva objekta, pondreriati varijablu znači odrediti njenu važnost među ostalim varijablama. Međutim, uvijek se postavlja pitanje 'Kako odabrati pondere (težine)?' Već pri samom odlučivanju koje varijable uključiti u studiju, a koje ne, radimo prvi korak ponderiranja s obzirom da se varijablama koje nisu odabrane dodijeljuje težina nula. Također, slično se vidi da je i standardizacija poseban slučaj ponderiranja varijabli.

Izabrane težine za varijable reflektiraju važnost koju im ispitivač dodjeljuje za klasifikaciju. Tako odabrane težine mogu biti rezultat osobnog stava ispitivača ili uzimajući u obzir neke aspekte matrice podataka, $\mathbf{X}$, zasebno. U prvom slučaju, kada ispitivač dodjeljuje težine, to može biti učinjeno specificiranjem težina direktno ili indirektno. Metode koje su predložili Sokal i Rohlf (1980) i Gordon (1990) su primjeri indirektnog dodijeljivanja težina. Oni objedinjuju opažene različistosti između odabranih objekata te također promatraju vrijednosti varijabli tih objekata. Nakon toga modeliraju različitosti koristeći temeljne varijable i težine koje ukazuju njihovu relativnu važnost. Težine koje najbolje odgovaraju uočenim različitostima bivaju izabrane.

Zajednički pristup određivanju težina iz matrice podataka, $\mathbf{X}$, je da se definira težina $w_k$ *k-te* varijable koja je obrnuto proporcionalna nekoj mjeri varijabilnosti u toj varijabli. Ovakav odabir težina implicira da se važnost varijable smanjuje kada njena varijabilnost raste. Pri definiranju težine koristi se nekoliko mjera varijabilnosti. Za neprekidne se varijable najčešće koristi ili recipročna vrijednost standardne devijacije ili recipročna vrijednost njenog opsega (range). Težine bazirane na rasponu uzorka, iliti opsegu, su najefektnije.

Uključivanje varijabilnosti težina je ekvivalentno nečemu što se naziva *standradizacija* varijabli. Na to ćemo se osvrnuti u sljedećem potpoglavlju.

Prethodni pristup pretpostavlja da je važnost varijable obrnuto proporcionalna potpunoj varijabilnosti te varijable. Potpuna varijabilnost varijable sadrži varijaciju oboje u grupi i između grupa koje postoje među setovima individualnih podataka. Cilj klaster analize je otkriti takve grupe. Stoga se može zaključiti da se važnost varijable ne treba smanjiti zbog razlike među grupama. Definiranje težine varijable obrnuto proporcionalno mjeri potpune varijabilnosti može imati ozbiljne nedostatke pri razrijeđivanju razlika između grupa varijabli koje najbolje diskriminiraju.

Nadalje, ako znamo grupe, koristeći unutar grupnu standardnu devijaciju *k-te* varijable za definiciju težina bi uvelike olakšalo ovaj problem. Ili, općenitije, za jednake kovarijance, Mahalanobis-ova generalizirana udaljenost se može koristiti za definiranje udaljenosti između dvaju objekata $i$ i $j$ sa vektorima mjerenja $\mathbf{x}_i$ i $\mathbf{x}_j$ danima kao

$$D_{ij}{}^2 = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{W}^{-1} (\mathbf{x}_i - \mathbf{x}_j), \tag{3.14}$$

gdje je $\mathbf{W}$ udružena među-grupna kovarijacijska matrica. Međutim, u klaster analizi, pripadanje pojedinoj grupi nije poznato prije same analize podataka. Ali, postoje alternativni načini kako iz baze podataka iščitati pripadanje pojedinim grupama. Jedan od alternativnih načina određivanja važnosti varijable iz baze podataka je predložio De Soete (1986). On predlaže pronalaženje težina, po jedna za svaku varijablu, koje pridonose ponderiranoj (vaganoj) Euklidskoj udaljenosti koja minimizira kriterij za odlazak iz ultrametrike. Ovakav pristup je motiviran dobropoznatom vezom između udaljenosti koja zadovoljava nejednakost ultrametrike i postojanje jedinstvenog hijerarhijskog stabla. Kasnije se dokazalo da ovakav pristup uvelike pomaže pri identificiranju varijabli koje su važne za klasteriranje objekata.

Druga metoda konstruiranja težina iz matrice podataka je *selekcija varijabli*. Ideja je, kao u multiploj regresiji, da se za identifikaciju podskupa inicijalnih varijabli uključenih u klaster analizu može koristiti empirijska selekcija. Procedura rezultira težinama vrijednosti jedan za selektiranu varijablu i nula za isključenu varijablu.

Davanje nedvosmislenog savjeta kako vagati (ponderirati) varijable pri konstrukciji mjera različitosti je teško, ali neke točke se mogu postaviti. Kao prvo, težine bazirane na subjektivnom stavu onog što je bitno mogu reflektirati postojanje klasifikacije podataka. Međutim, to nije ono što se općenito traži od klaster analize. Češće, metode klaster analize se primjenjuju na baze podataka u nadi da će se prethodno neuočene grupe pojaviti. Stoga je općenito poželjno da se smanji dojam subjektivnog opažanja i davanja važnosti na početni izbor varijabli koje treba zabilježiti. Tom selekcijom se odražava važnost ispitivačevog stava za svrhu klasifikacije podataka. Drugo, izvođenje klaster analize mjerama udaljenosti na težinama dobivenim empirijski ovisi o klaster stukturi. Međutim, težine dobivene mjerenjem ne-važnosti dobivene iz međugrupnih varijabilnosti (razlika)

imaju najviše potencijala da oporave grupe naknadnom klaster analizom. Dvije najčešće korištene strategije koje ubacuju većinu varijabli u standardni klaster algoritam baziran na udaljenostima, u nadi da niti jedna važna neće biti izostavljena, te uključivanje težina baziranih na standardnim devijacijama varijabli, čine se nedjelotvornim.

## 3.7 Standardizacija

U većini slučajeva klasteriranja, varijable koje opisuju objekte za klasteriranje nisu mjerene u istim jedinicama. Ustvari, često su i varijable drugačijeg tipa. Kada su sve varijable mjerene na neprekidnoj skali, najčešće sugeriran pristup problemu drugačijih mjernih jedinica, prije bilo kakve analize, je standardizacija svake varijable na jedinicu varijance. U takvim okolnostima se koriste razne mjere varijabilnosti. Kada se koristi standardna devijacija cijelog seta objekata koje treba klasterirati, metoda se najčešće naziva *autoskaliranje*, *standard scoring* ili *z-scoring*. Alternativne opcije su dijeljenje medijanom apsolutne devijacije ili opsegom, a potonji se pokazao da nadilazi autosklairanje u mnogim primjenama klasteriranja.

Standardizacija varijabli na jedinicu varijance se može shvatiti kao poseban slučaj ponderiranja (vaganja). Ovdje su težine recipročne vrijednosti mjera izabranih da mjere varijance varijabli - općenito standardna devijacija ili raspon uzorka neprekidnih varijabli. Prema tome, kada standardizacija nadilazi analizu, ispitivač pretpostavlja da važnost varijable opada sa rastućom varijabilnosti. Kao rezultat standardizacije, posebnog slučaja ponderiranja, neke preporuke s obzirom na izbor težina prenose se na standardizaciju: ako ispitivač ne može odrediti prikladnu mjernu jedinicu te standardizacija varijabli postaje neophodna, preferira se standardizacija varijabli korištenjem unutar-grupne mjere varijabilnosti rađe nego one potpune varijabilnosti. U kontekstu klasteriranja, najbolji način bavljenja s problemom prikladne mjerne jedinice bio bi korištenje klaster metode koja je nepromijenjiva pod utjecajem skaliranja, tako da se izbjegne pitanje standardizacije uopće.

## 3.8 Izbor mjere bliskosti

Postoji gotovo beskrajan broj koeficijenata sličnosti i različistosti. Koeficijent se podrazumijeva u kontekstu opisne statistike koje je dio, uključujući prirodu podatka, te željeni tip analize. Prvenstveno, priroda podatka treba snažno utjecati na izbor mjere bliskosti. U nekim okolnostima, na primjer, kvantitativni podaci mogu biti najbolje prikazani preko binarnih.

Nadalje, izbor mjere treba ovisiti i o mjernoj skali podataka. Koeficijenti sličnosti trebaju biti korišteni kada su podaci binarni. Tada se izbor mjere bliskosti centrira oko

'the treatment of co-absence'. Za neprekidne podatke, udaljenost ili korelacijski tip mjere različitosti treba biti korišten s obzirom na veličinu i tip (shape) objekata koji nas zanimaju.

Konačno, izbor klaster metode može imati utjecaj na izbor koeficijenata. Na primjer, izbor između nekoliko koeficijenata bliskosti sa sličnim svojstvima se može izbjeći koristeći klaster metodu koja ovisi samo o 'ranking' bliskosti, ne njihovim apsolutnim vrijednostima. Kao zaključak se povlači da za sve okolnosti nije moguće dati odgovor koju mjeru je najbolje koristiti.

# Poglavlje 4

# Hijerarhijsko klasteriranje

Podaci u hijerarhijskoj klasifikaciji nisu naprijed podijeljeni u određeni broj klasa ili grupa (klastera). Naime, klasifikacija se sastoji od niza particija koje mogu započeti sa jednim klasterom koji sadrži sve podatke, sve do $n$ klastera sastojanih od samo jednog podatka. Tehnike hijerarhijske klasifikacije se mogu podijeliti na *aglomerativne* metode, koje $n$ pojedinačnih podataka spajaju u grupe, i metode *dijeljenja*, koje dijele skup od $n$ podataka u manje, finije grupe. Oba pristupa se mogu protumačiti kao pokušaj pronalaženja točnog broja koraka u svakom stadiju spajanja ili dijeljenja podataka gdje obje metode rade na nekom tipu matrice udaljenosti (bliskosti).

Hijerarhijskim metodama, spajanjem ili dijeljenjem, jednom učinjeno se više ne može vratiti na početno stanje, niti na prethodni korak. Odnosno, kada aglomerativne metode spoje dva pojedinačna podatka u grupu, ista se više ne može razdvojiti, te kada metoda dijeljenja razdvoji podatke, isti se više ne mogu spojiti.

S obzirom da sve aglomerativne hijerarhijske metode u konačnici smanjuju bazu podataka na samo jednu grupu sastojanu od svih podataka, a metode dijeljenja u konačnici dijele čitav set podataka u $n$ grupa od kojih se svaka sastoji od samo jednog podatka, ispitivač, želeći imati riješenje sa optimalnim brojem klastera, mora odlučiti kada zaustaviti algoritam.

Hijerarhijske klasifikacije, proizašle bilo iz aglomerativnih metoda ili metoda dijeljenja, predstavljaju se dvodimenzionalnim dijagramom zvanim *dendogram*, koji ilustrira spajanja ili razdvajanja učinjena svakim korakom analize.

## 4.1 Aglomerativne metode

Aglomerativne metode su najčešće korištene hijerarhijeske metode. Iz njih proizlazi niz particija podataka: prvi se sastojii od $n$ pojedinačnih članova klastera, dok zadnji sadržava jednu grupu sačinjenu od svih $n$ podataka. Osnovne operacije svih takvih metoda su u

suštini slične, a možemo izdovijiti dva specijalna primjera, *jednostruke veze - single linkage* i *centroid linkage*. Ove metode u svakom koraku spajaju pojedinačne podatke ili grupe podataka koje su najbliže. Razlika između metoda proizlazi zbog različitih definicija udaljenosti (ili sličnosti) između pojedinačnog podatka i grupe sadržane od nekoliko pojedinačnih podataka, ili između dviju grupa pojedinačnih podataka.

## Jednostruke veze i centroid linkage

Metoda jednostrukih veza je najjednostavnija metoda hijerarhijskog klasteriranja, također zvana i metoda najbližeg susjeda. Bazirana je na korištenju matrice bliskosti, a udaljenost između grupa A i B definira se kao minimalna udaljenost dvaju podataka.

$$D(A, B) = min\left\{d(y_i, y_j)\right\}, \tag{4.1}$$

za $y_i \in A$ i $y_j \in B$. Pri tome se za mjeru udaljenosti $d(y_i, y_j)$ u uzima Euklidska udaljenost, ili neka druga udaljenost između vektora $y_i$ i $y_j$. U svakom koraku metode najbližeg susjeda traži se udaljenost između svaka dva klastera i klasteri s najmanjom udaljenosti se spajaju. Time se broj klastera smanjuje za jedan. Nakon što su dva klastera spojena, postupak se ponavalja u idućem koraku. Računa se udaljenost među svakim parom klastera i par s najmanjom udaljenosti se spaja u jedan klaster. Rezultat metode se prikazuje dendogramom koji prikazuje sve korake u hijerarhijskom postupku, uključujući i udaljenosti kod kojih su klasteri spojeni. Metoda jednostrukih veza, iliti najbližeg susjeda, služi kako bi ilustrirala opći postupak hijerarhijskih metoda.

Dok metoda jednostrukih veza radi direktno na matrici bliskosti, centroid metoda zahtjeva pristup originalnim podacima. Također, u većini slučajeva se za udaljenost među podacima uzima Euklidnska udaljenost, ali računajući je na originalnim podacima i svaki put uzmajući najmanju vrijednost te par s tom vrijednosti spajamo u klaster. U svakom idućem koraku ponavljamo postupak, spajamo par s najmanjom udaljenosti u novi klaster. Rezultat metode se i u ovom slučaju prikazuje dendogramom.

## Standardne aglomerativne metode

**Potpuna veza** (najdalji susjedi) je potpuna suprotnost od jednostruke veze, u smislu da se udaljenost između dva klastera A i B sada definira kao maksimalna udaljenost dvaju pojedinačnih podataka u A i u B.

$$D(A, B) = max\left\{d(y_i, y_j)\right\}, \tag{4.2}$$

za $y_i \in A$ i $y_j \in B$. Usvakom se koraku određuje udaljenost za svaki par podataka te se par s najmanjom udaljenošću spaja u klaster.

**Prosječna udaljenost**, također poznata kao UPGMA (unweighted pair-group method using the average approach), je metoda gdje se udaljenost dva klastera definira kao prosjek $n_A n_B$ udaljenosti između $n_A$ točaka u A i $n_B$ točaka u B:

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=!}^{n_A} \sum_{i=1}^{n_B} d(y_i, y_j), \tag{4.3}$$

U svakom se koraku udružuju dva klastera s najmanjom udaljenosti izračunatoj prema gornjoj fromuli.

**Centorid** metoda udaljenost između dvaju klastera A i B definira kao Euklidsku udaljenost između dvaju vektora sredina (koji se još naziva centroidom):

$$D(A, B) = d(\bar{y}_a, \bar{y}_B), \tag{4.4}$$

pri čemu su $\bar{y}_A$ i $\bar{y}_B$ vektori sredina za opažanja iz A, odnosno opažanja iz B, a $d(\bar{y}_A, \bar{y}_B) = \sqrt{(\bar{y}_A - \bar{y}_B)'(\bar{y}_A - \bar{y}_B)}$. U svakom se koraku spajaju dva centroida s najmanjom udaljenošću. Nakon što su dva klastera A i B združena, centroid novog klastera se računa kao vagana aritmetička sredina:

$$\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}. \tag{4.5}$$

Nadalje, potrebno je naglasiti da ova metoda radije koristi originalnu matricu podataka nego li matricu bliskosti.

**Medijan** metoda radi na principu da za medijan (polovište) pravca koji spaja dva podatka, A i B, a kako bi se izbjeglo ponderiranje centroida, određuje točku za računanje novih udaljenosti klastera $A, B$ u odnosu na druge klastere:

$$m_{AB} = \frac{1}{2}(\bar{y}_A + \bar{y}_B). \tag{4.6}$$

U svakom se koraku klasteri s najmanjom medijalnom udaljenosti spajaju u novi klaster. Treba naglastiti da medijan nije uobičajeni medijan u statističkom smislu. Terminologija potječe od medijana trokuta.

**Wardova metoda** je treći tip ovakve metode u kojoj je spajanje dvaju klastera bazirano na kriteriju greške sume kvadrata. Odnosno, Wardova metoda, poznata i kao inkrementalna suma kvadrata, upotrebljava (kvadrirane) udaljenosti unutar klastera i (kvadrirane) udaljenosti između klastera. Ako je AB klaster dobiven kombiniranjem klastera A i B, tada je zbroj udaljenosti unutar klastera (elemenata od centroida):

$$W_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A), \tag{4.7}$$

$$W_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)'(y_i - \bar{y}_B), \tag{4.8}$$

$$W_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB}), \tag{4.9}$$

pri čemu je $\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}$, a $n_A$, $n_B$ i $n_{AB} = n_A + n_B$ su brojevi točaka u A, B i AB, redom. S obzirom da su zbrojevi udaljenosti ekvivalentni sumama kvadrata odstupanja točaka klastera od njihovih centroida (within-cluster sum of squares), označene su s $W_A$, $W_B$, $W_{AB}$. Wardova metoda združuje dva klastera A i B koji minimiziraju prirast u W:

$$I_{AB} = W_{AB} - (W_A + W_B). \tag{4.10}$$

Može se pokazati da prirast $I_{AB}$ iz prethodne formule može poprimiti dva ekvivalentna oblika:

$$I_{AB} = n_A(\bar{y}_A - \bar{y}_{AB})'(\bar{y}_A - \bar{y}_{AB}) + n_B(\bar{y}_B - \bar{y}_B - \bar{y}_{AB})'(\bar{y}_B - \bar{y}_B - \bar{y}_{AB}), \tag{4.11}$$

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}(\bar{y}_A - \bar{y}_B)'(\bar{y}_A - \bar{y}_B). \tag{4.12}$$

Prema zadnjoj formuli slijedi da je minimiziranje prirasta u W ekvivalentno minimiziranju udaljenosti među klasterima. Kada bi A sadržavao samo $y_i$. a klaster B samo $y_j$, tada bi $W_A$ i $W_B$ bile jednake nuli, a prethodne dvije formule bi se reducirale na:

$$I_{ij} = W_{AB} = \frac{1}{2}(y_i - y_j)'(y_i - y_j) = \frac{1}{2}d^2(y_i, y_j). \tag{4.13}$$

Wardova je metoda povezana i s metodom centroida. Kvadrira li se udaljenost $d(\bar{y}_A, \bar{y}_B)$ iz (4.4) i usporedi s (4.12), jedina je razlika u koeficijentu $n_A n_B/(n_a + n_B)$ za Wardowu metodu. Veličina klastera utječe na Wardovu metodu, ali ne i na metodu centroida. Napiše li se $n_A n_B/(n_A + n_B)$ u obliku:

$$\frac{n_A n_B}{n_A + n_B} = \frac{1}{\frac{1}{n_A} + \frac{1}{n_B}} \tag{4.14}$$

može se uočiti da se povećanjem $n_A$ i $n_B$ povećava i $n_A n_B/(n_A + n_B)$. Napiše li se koeficijent u obliku:

$$\frac{n_A n_B}{n_A + n_B} = \frac{n_A}{1 + \frac{1}{n_A n_B}} \tag{4.15}$$

vidi se da se povećavanjem $n_B$, uz fiksni $n_A$, povećava i $n_A n_B/(n_A + n_B)$. Zbog toga će Wardova metoda spajati manje klastere ili klastere jednake veličine češće od metode centroida .

## Fleksibilna beta metoda

Pretpostavimo da su se klasteri A i B upravo spojili u novi klaster AB. Opći oblik formule udaljenosti između kalstera AB i klastera C dali su Lance i Williams (1967):

$$D(C, AB) = \alpha_A D(C, A) + \alpha_B D(C, B) + \beta D(A, B) + \gamma \left| D(C, A) - D(C, B) \right|. \tag{4.16}$$

gdje su udaljenosti $D(C, A)$, $D(C, B)$ i $D(A, B)$ elementi matrice udaljenosti prije udruživanja klastera A i B. Udaljenosti klastera AB u odnosu na druge klastere, kao što je predočeno gornjom formulom, upotrijebiti će se zajedno s međusobnim udaljenostima ostalih klastera pri formiranju nove matrice udaljenosti. Polazeći od te matrice odabrat će se par klastera s najmanjom udaljenosti, te će se taj par udružiti u sljedećem koraku. Da bi se gornji izraz pojednostavio, Lance i Willianms sugeriraju slijedeća ograničenja na vrijednosti parametara:

$$1.\ \alpha_A + \alpha_B + \beta = 1$$

$$2.\ \alpha_A = \alpha_B$$

$$3.\ \gamma = 0$$

$$4.\ \beta < 1.$$

S $\alpha_A = \alpha_B$ i $\gamma = 0$ dobiti će se $2\alpha_A = 1 - \beta$ ili $\alpha_A = \alpha_B = (1 - \beta)/2$, te se treba izabrati samo vrijednost koeficijenta $\beta$. Zbog fleksibilnosti koeficijenta $\beta$, rezultirajuća se hijerarhijska metoda naziva *fleksibilna beta metoda*.

Izbor vrijednosti od $\beta$ određuje karakteristike fleksibilne beta metode klasteriranja. Lance i Williams su predložili da se koristi mala negativna vrijednost za $\beta$, kao primjerice $\beta = -0.25$. Ukoliko postoje outliersi (netipične vrijednosti) u podacima, uporaba manjih vrijednosti za $\beta$, kao $\beta = -0.5$, može vjerojatnije izolirati netipične vrijednosti u klasterima.

Udaljenosti definirane u prethodnom poglavlju za aglomerativne metode klasteriranja mogu se izraziti kao specijalni slučajevi od (4.16). Tražene vrijednosti parametara navedene su u donjoj tablici, gdje su U, C, P, M dopustiva svojstva i odnose se na nema preokreta, konveksnost, proporcionalno s obzirom na točku i monotonost, redom. Dok

Tablica 4.1: Hijerarhijska aglomerativna klaster metoda: dopustiva svojstva i Lance-Williams parametri

| Metoda | U | C | P | M | $\alpha_i$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|---|---|---|
| Jednostruka veza | N | N | Y | Y | $\frac{1}{2}$ | 0 | $\frac{-1}{2}$ |
| Potpuna veza | N | N | Y | Y | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ |
| Prosječna udaljenost | N | N | N | N | $n_i/(n_i+n_j)$ | 0 | 0 |
| Centroid | Y | N | N | N | $n_i(n_i+n_j)$ | $-n_in_j(n_i+n_j)^2$ | 0 |
| Medijan | Y | N | Y | N | $\frac{1}{2}$ | $\frac{-1}{4}$ | 0 |
| Ward | N | Y | N | N | $(n_k+n_i)/(n_k+n_i+n_j)$ | $-n_k/(n_k+n_i+n_j)$ | 0 |

su $\alpha_i$, $\beta$ i $\gamma$ Lance-Williamsovi parametri. Za metodu centroida, medijana i Wardovu metodu udaljenosti u (4.16) moraju biti kvadrirane udaljenosti (ako pretpostavljamo Euklidsku udaljenost). Za ostale metode mogu i ne moraju biti kvadrirane.

Ilustrirajmo izbor vrijednosti parametara za metodu jednostruke veze (metodu najbližeg susjeda). Uzme li se $\alpha_A = \alpha_B = \frac{1}{2}, \beta = 0$ i $\gamma = -\frac{1}{2}$, formula 4.16 poprima oblik:

$$D(C,AB) = \frac{1}{2}D(C,A) + \frac{1}{2}D(C,B) - \frac{1}{2}\left|D(C,A) - D(C,B)\right|. \tag{4.17}$$

Ako je $D(C,A) > D(C,B)$, tada je

$|D(C,A) - D(C,B)| = D(C,A) - D(C,B)$ i (4.16) se reducira na:

$$D(C,AB) = D(C,B). \tag{4.18}$$

S druge strane, ako je $D(C,A) < D(C,B)$, tada je

$|D(C,A) - D(C,B)| = D(C,B) - D(C,A)$, pa se (4.16) reducira na:

$$D(C,AB) = D(C,A). \tag{4.19}$$

Dakle, 4.16 možemo pisati kao:

$$D(C,AB) = min\left\{D(C,A), D(C,B)\right\}, \tag{4.20}$$

što je ekvivalentno definiciji udaljenosti za metodu jednostruke veze.

## 4.2 Metode dijeljenja

Aglomerativne hijerarhijske metode počinju s $n$ klastera, a postupak završava s jednim klasterom koji sadržava svih $n$ podtaka. Suprotno toj metodi radi metoda dijeljenja koja počinje s jednim klasterom od $n$ članova, a završava s $n$ klastera (za svaku jedinicu po jedan). Rezultat se može predočiti dendogramom. Metode dijeljenja, kao i aglomerativne

metode, imaju nedostatak da nakon što je napravljena particija skupa u klastere, ne postoji mogućnost premještanja jedinica iz jednog klastera u drugi klaster (čiji nije bio član u vrijeme diobe). Međutim, ako su nam od interesa veći klasteri tada metode dijeljena imaju prednost pred aglomerativnim metodama, u kojima se veći klasteri dostižu samo nakon velikog broja koraka.

Općenito, postoje dvije skupine algoritama dijeljenja: monothetic i polythetic. U monothetic pristupu podjela grupe na dvije podgrupe je zasnovana na jednoj varijabli, dok se u polythetic pristupu koristi $p$ varijabli da bi se napravilo razdvajanje. Ako su varijable binarne (a kvantitativne varijable se mogu pretvoriti u binarne) monothetic pristup se može jednostavno primjeniti.

## Monothetic metode dijeljenja

Monothetic pristup se može jednostavno primjeniti ako su varijable binarne. Podjela na dvije grupe se zasniva na prisutnosti ili odsutnosti atributa. Ovakav način teži minimizaciji broja particija koje tek trebaju biti napravljenje. Jedan primjer kriterija homogenosti je informacijski pokazatelj, $C$ (koji u ovom slučaju označava nered ili kaos), definiran sa $p$ varijabli i $n$ objekata:

$$C = pn \log n - \sum_{k=1}^{p} \left[ f_k \log f_k - (n - f_k) \log(n - f_k) \right], \tag{4.21}$$

gdje je $f_k$ broj pojedinaca koji sadrže *k-ti* atribut. Ako grupu $X$ treba razdvojiti u dvije grupe A i B, tada je redukcija u $C$ jednaka $C_X - C_A - C_B$. Idealan set klastera trebao bi sadržavati članove sa identičnim atributima i $C$ jednakim nula (s obzirom da su klasteri u svakom koraku razdvojeni s obzirom na posjedovanje atributa koji vodi do najveće redukcije u $C$).

Umjesto homogenosti klastera, atribut korišten u svakom koraku se može izabrati s obzirom na njegovu cjelokupnu povezanost sa svim ostalim atributima u tom koraku: to se naziva *analiza asocijacija (association analysis)*, posebno korištena u ekologiji. Svakim se korakom particija skupa odvija s obzirom na prisutnost ili odsutnost atributa čija je povezanost sa ostalima maksimalna. Varijabla (atribut) je izabrana tako da maksimizira $\chi^2$-vrijednost ili neki informacijski pokazatelj.

Jedna od najvećih prednosti monothetic metode dijeljenja je da je očito koja varijabla u kojem koraku dovodi do razdvajanja (particije). Međutim, opći problem ovakvih metoda je da posjedovanje određenog atributa, koji je rijedak ili se teško pronalazi u kombinaciji s ostalima, može dovesti ispitivača na krivi put.

Nova metoda dijeljenja klastera koju su predložili Piccarreta i Billari (2007) može se koristiti za nizove podataka kao što su povijesni tokovi života. Metoda koristi logiku analize klasifikacijskog i regresijskog stabla (CART), te također omogućuje korištenje nekih od najkorisnijih značajki CART analize, kao što su obrezivanje ukrštenim potvrđivanjem

kako bi identificirali prikladan broj klastera. Piccarreta i Billari definiraju dva nova tipa podataka izvedena iz originalnih nizova: *auxiliary varibles* i *state permanence sequences*. To znači da su, umjesto da imamo potpuno različite zavisne i nezavisne varijable (kao u CART), varijable koje definiraju rascjepe, kriterij za procjenjivanje homogenosti klastera, te podaci koji karakteriziraju klastere izvedeni iz niza podataka. Podjele načinjene ovom metodom su s ciljem dobivanja 'čistih' klastera.

### Polythetic metode dijeljenja

Polythetic metode dijeljenja su sličnije aglomerativnim metodama jer koriste sve varijable istovremeno, te mogu raditi na matrici udaljenosti (bliskosti). Za polythetic pristup promatra se postupak kojeg je predložio MacNaughton-Smith (1964). Da bi se grupa razdijelila radi se sa odcijepljenom grupom i ostatkom. Traži se član u ostatku čija je prosječna udaljenost (različitost) od ostalih članova u ostatku, umanjena za njegovu udaljenost od odcijepljene grupe, najveća. Ako je najveća udaljenost negativna, postupak se zaustavlja i podjela je potpuna. Odcijepljenu grupu se može započeti s članom s najvećom prosječnom udaljenosti od ostalih članova u grupi.

## 4.3 Primjena hijerarhijskih metoda u klasteriranju

Kako bi što bolje primjenili hijerarhijske metode, bilo aglomerativne ili metode dijeljenja, ispitivač mora voditi računa o (uz početni odabir mjere udaljenosti-bliskosti):

1. Grafički prikaz klasteriranja
2. Usporedba dendograma
3. Matematička svojstva metode
4. Izbor praticije
5. Hijerarhijski algoritmi

### Dendogrami i drugi grafički prikazi

**Definicija 4.3.1.** ***Dendogram****, također nazivan i dijagram stabla, je matematički i slikovni prikaz kompletnog klaster procesa. Čvorovi dendograma predstavljaju klastere, a duljina stabljike (visina) predstavlja udaljenost na kojoj su klasteri spojeni.*

Stabljike ponekad ne proizlaze iz nulte linije dendograma kako bi prikazale redosljed kojim su klasteri prvi put spojeni. Dendogrami kojima stabljike nisu numerirane nazivaju

se *neponderirani* ili *rangirani*. Većina dendograma ima dva ruba koji proizlaze iz svakog čvora (binarna stabla). Način na koji su čvorovi i stabljike uređeni naziva se *topologija* stabla.

Imena objekata pripisanih završnim čvorovima nazivaju se *oznake*. Unutarnji čvorovi se u većini slučajeva ne označavaju. Reprezenatitivni članovi klastera se mogu povezati sa unutarnjim članovima, nazvanim *primjerci* ili *centralni tipovi*, te su definirani kao objekti sa maksimalnom sličnosti unutar klastera (ili minimalnim razlikama). Poseban oblik centralnog tipa je *medoid* (objekt sa minimalnom apsolutnom udaljenosti u odnosu na druge članove klastera). Dendogram sam za sebe prikazuje proces kojem je hijerarhija napravljena, gdje oznake primjerka i unutarnjih čvorova predstavljaju specifične particije. Pritom je potrebno uočiti da ista baza podataka i klaster procedura mogu dati $2^{n-1}$ različitih dendograma čiji izgled ovisi o rasporedu i prikazu čvorova.

Broj različitih oblika dendograma je kroz godine narastao. Jedan od njih je *espaliers* - generalizirani dendogram kod kojeg duljina horizontalne linije prenosi informaciju o relativnoj homogenosti i razdvajanju klastera. Nadalje, kao drugi tip se javlja *piramidalni* - unaprijeđeni tip dendograma za prikazivanje preklapajućih klastera. A kao treći tip dendograma javlja se *aditivno stablo* (ili stablo duljina puta - path length tree) koji je generalizacija dendograma, a duljina staze (puta) između čvorova predstavlja udaljenost (bliskost) između objekata, te gdje vrijedi *aditivna nejednakost* (ili nejednakost četiri točke). Ovakava generalizacija ultrametričke nejednakosti je nužno i dovoljno svojstvo kako bi skup udaljenosti mogao biti prikazan u formi aditivnog stabla. Aditivna nejednakost je dana s:

$$d_{xy} + d_{uv} \leq \max \left| d_{xu} + d_{yv} + d_{yu} \right|, \forall x, y, u, v. \tag{4.22}$$

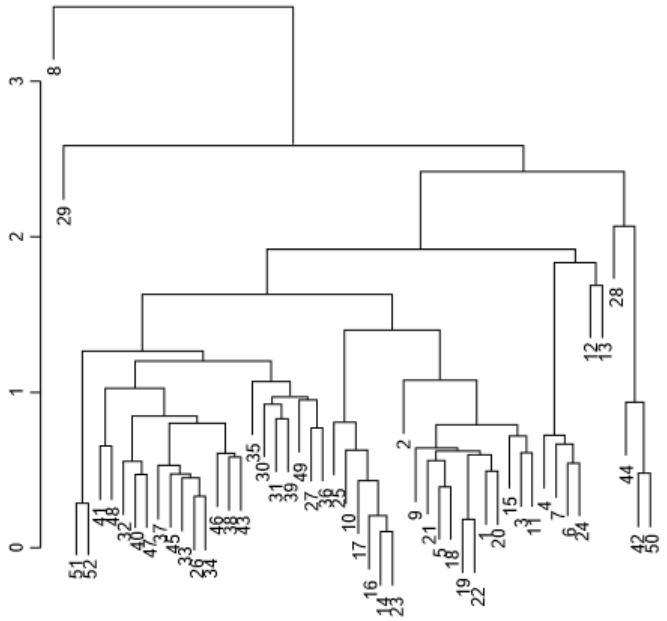
## Usporedba dendograma

Tehnike hijerarhijskog klasteriranja nameću hijerarhijsku strukturu podataka. Obično je potrebno uzeti u obzir je li takav prikaz zadovoljavajući ili on predstavlja neprihvatljivo narušavanje prvotnih veza među objektima s obzirom na njihove uočene udaljenosti. Dvije mjere korištene pri usporedbni dvaju dendograma sa matricom udaljenosti ili drugim dendogramom su *kofenetska korelacija (cophenetic correlation)* i *Goodmanova i Kruskalova* $\gamma$.

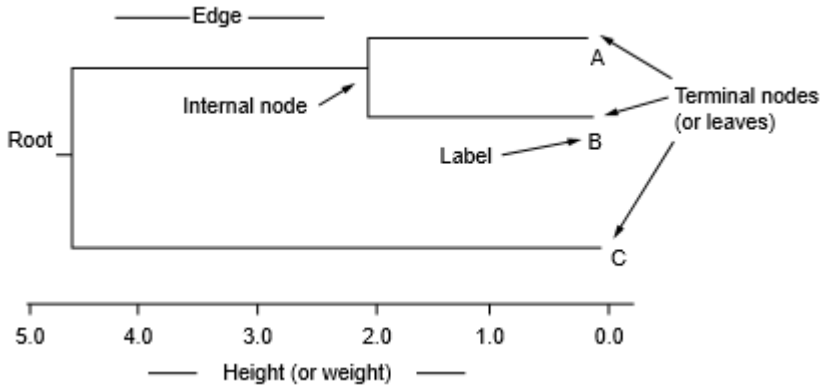
Početna točka obiju metoda je takozvana *kofenetska matrica*.

**Definicija 4.3.2.** *Kofenetska korelacija je mjera izobličenosti unesena u hijerarhijsku klaster analizu usporedbom očitih sličnosti u dendogramu s izvornim sličnostima između objekata.*

Odnosno, kofenetsku korelaciju dobivamo kao produkt korelacije između $n$ i $(n-1)/2$ elemenata $h_{ij}$ odgovarajuće kofenetske matrice (isključujući dijagonalne elemente). Sama




Slika 4.1: Dendogram




Slika 4.2: Piramidalni dendogram

matrica je uređena u vektorskoj formi. Elemetni matrice su visine, $h_{ij}$, na kojima su dva objekta združena u jedan klaster na dendogramu.

Druga, neparametarska mjera udruživanja je Goodmanova i Kruskalova $\gamma$ definirana kao $(S_+ - S_-)/(S_+ + S_-)$, gdje su $S_+$ i $S_-$ broj podudarnosti i razilaženja, redom. Usporedba podudarnosti i razilaženja u matrici se definira usporedbom svakog para od svih mogućih


Slika 4.3: Aditivno stablo

parova.

## Matematička svojstva metode

Iako hijerarhijske klaster metode same po sebi nisu definirane kao strogo matematičke, ipak sadrže neka matematička svojstva. Jedno od njih je i svojstvo *ultrametričnosti* kojeg su prvi definirali Hartigan, Jardine i Johanson (1967). Od tada se smatra da je svojstvo uvelike povezano sa raznim klaster metodama, uobičajeno s mogućnošću prezentiranja hijerarhije pomoću dendograma. Svojstvo ultrametričnosti kaže da je:

$$h_{ij} \leq \max(h_{ij}, h_{jk}), \forall i, j, k, \tag{4.23}$$

gdje je $h_{ij}$ udaljenost između kalstera $i$ i $j$. Alternativni način definiranja ovog svojstva je da između bilo koja tri objekta, dvije najveće udaljenosti među njima su jednake. Međutim, svojstvo ne mora nužno vrijediti za elemente matrice udaljenosti (bliskosti). Ali vrijedi u mnogim hijerarhijskim klaster metodama za visine $h_{ij}$ na kojima dva objekta bivaju spojena u jedan klaster.

Kao posljedica nuspjelog pokušaja zadovoljavanja svojstva ultrametrike, odnosno ako se jedan član ili klaster udruže s drugim klasterom na udaljenosti manjoj od one na kojoj su

se udružila prethodna dva klastera, događa se *inverzija* ili *obrat*. Inverzija je reprezentirana s križanjem (čvorom) u dendogramu.

Hijerarhijska klaster metoda u kojoj ne može doći do inverzije je monotona jer je udaljenost u svakom sljedećem koraku veća od udaljenosti u prethodnom koraku. Mjera udaljenosti ili metoda klasteriranja koja je monotona također se zove ultrametrična. Međutim, inverzija nije problem ako je glavni interes samo jedna određena particija, a ne cijela hijerarhijska struktura. Nadalje, Murtagh (1985) ističe da obrati mogu otežati interpretaciju hijerarhije u oba slučaja - bilo teoretskom proučavanju svojstava klastera kao i u primjenama gdje je hijerarhijska struktura unutrašnji dio modela. Razlog zbog kojeg se to događa je taj što se ugniježdena struktura ne održava. Obrati se mogu dogoditi u centroid i median klaster metodama.

Značajka koja veže klaster metode je njihova težnja da 'iskrive' (deformiraju) prostor. Za metodu klasteriranja koja ne mijenja svojstva prostornih udaljenosti kaže se da čuva prostor. Metoda koja ne čuva prostor može ga smanjiti ili povećati. Metoda smanjuje prostor (space-contracting method) ako su novooblikovani klasteri bliže pojedinačnim opažanjima, tako da pojedinačno opažanje više teži udruživanju s postojećim klasterom nego da s nekom drugom jedinicom formira novi klaster. Ta se tendencija naziva *ulančananje* (chaining). Metoda povećava prostor (space-dilating) ako su novonastali klasteri pomaknuti dakleko od ostalih opažanja pa pojedinačna opažanja više teže formiranju novih klastera s drugim jedinicama, nego udruživanju u postojeće klastere. U tom slučaju klasteri izgledaju različitiji nego jesu.

Dubien i Warde su opisali prostorna svojstva na sljedeći način. Pretpostavi li se da udaljenosti između tri klastera zadovoljavaju:

$$d(i, j) < d(i, k) < d(j, k). \tag{4.24}$$

Tada metoda čuva prostor ako je:

$$d(i, k) < d(ij, k) < d(j, k). \tag{4.25}$$

Metoda smanjuje prostor ako ne vrijedi prva nejednakost u 4.25, a proširuje prostor ako ne vrijedi druga nejednakost. Metoda jednostruke veze smanjuje prostor, s naznačenom tendencijom ulančavanja. Zbog toga je neki autori ne preporučuju. Metoda potpune veze jako povećava prostor, s tendencijom umjetnog udruživanja u klastere.

Druge hijerarhijske metode su između ekstrema reprezentiranih metodom jednostruke veze i metodom potpune veze. Metoda centroida i metoda prosječne veze čuvaju prostor, dok Wardova metoda sužava prostor. Felksibilna beta metoda sužava prostor za $\beta > 0$, čuva prostor za $\beta = 0$, a proširuje prostor za $\beta < 0$. Mali stupanj proširenja može pomoći pri definiranju granica klastera, no prevelika dilatacija može dovesti do formiranja prevelikog broja klastera. Preporučena vrijednost $\beta = -0.25$ predstavlja dobar kompromis.

Mnoga dopustiva svojstva predlažu Fisher i Van Ness (1971), kao što su kvaliteta, jednakost s drugim stvarima, te kao takva mogu pripomoći pri odabiru prikladne klaster metode. Jedno od svojstava (k-group) dobro-strukturirana dopustivost, koje se veže uz Lanceov i Williamsov parametar, predložio je Mirkin (1996) i naziva ga *clump dopustivost* (pramen dopustivosti). Mirkin ga definira: postoji grupiranje takvo da su sve udaljenosti unutar grupe manje od udaljenosti između grupa. Nadalje, on pokazuje da je clump dopustivost i čuvanje prostora ekvivalentno slijedećem svojstvu: za sve $x$ i $y$ takve da je $0 < x < 1$ i $y > 0$

$$\alpha(x, y) + \alpha(1 - x, y) = 1; \tag{4.26}$$

$$\beta(x, 1 - x, y) = 0; \tag{4.27}$$

$$|\gamma(y)| \leq \alpha(x, y), \tag{4.28}$$

gdje su $\alpha, \beta$ i $\gamma$ parametri iz Lance-Williams fleksibilne beta metode izraženi kao funkcije veličine klastera sa $x = n_k/n_+$, $y = n_i/n_+$ i $z = n_j/n_+$, gdje je $n_+ = n_i + n_j + n_k$.

Također, kasnije su predstavljena specijaliziranija, ali ipak korisna, svojstva dopustivosti:

1. Konveksna dopustivost: ako se objekti mogu prikazati u Euklidskom prostoru, tada se konveksne ljuske particija nikada ne sijeku

2. Dopustivost proporcionalnosti točke: replikacija točaka ne mijenja granice particija

3. Monotona dopustivost: monotona transformacija elemenata matrice udaljenosti ne mijenja grupiranja

## Izbor particije - problem broja grupa

Česta situacija je da ispitivača ne zanima cjelokupna hijerarhija, nego jedna ili dvije particije iz nje. Ovakav problem zahtjeva odlučivanje o broju grupa koje se prezentiraju na kraju, te u konačnici gdje 'prerezati' dendogram.

U standardnim aglomerativnim i polythetic metodama dijeljenja se particije postižu selektiranjem jednog od riješenja u ugniježdenom nizu klastera koji uključuje hijerarhiju, a to je ekvivalentno rezanju dendogarama na određenoj visini (ponekad se naziva i *najbolji rez*). To nam definira particiju takvu da su klasteri ispod te visine udaljeni najmanje za tu duljinu, a izgled dendograma nam sugerira broj klastera. Velike promjene u razinama fuzije se rade kako bi se prikazao najbolji rez. Puno fleksibilniji razvoj ove ideje je 'dinamičko rezanje stabla'. Pod tim se podrazumijeva da su različite grane stabla rezane na

različitim visinama. To je iterativni proces koji se zaustavlja kada je broj klastera ustaljen svim kombinacijama i rastavljanjima klastera, tvoreći uzastopne rezove pod-dendograma u klastere na temelju njihovih oblika.

Formalnije procedure 'izbora broja grupa' koje su naročito pogodne za hijerarhijske metode predložio je Mojena (1977). Prva je bazirana na relativnim veličinama različitih levela fuzija u dendogramu te je poznata pod nazivom *pravilo gornjeg repa*. Detaljnije, prijedlog je da izaberemo broj grupa koji odgovara prvom stadiju u dendogramu i zadovoljava:

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}, \tag{4.29}$$

gdje su $\alpha_0, \alpha_1, \alpha_2, ..., \alpha_{n-1}$ leveli fuzija koji odgovaraju $n, n-1, ..., 1$ klastera. Nadalje, $\bar{\alpha}$ i $s_{\alpha}$ su redom očekivanje i nepristrana standardna devijacija prijašnjih $j$ levela, a $k$ je konstanta. Mojena predlaže da vrijednosti od $k$ između $2.75 - 3.50$ daju najbolje rezultate, iako neki sugeriraju 1.25. Alternativno, može se koristiti $t$-distribucija (iako to podrazumijeva fundamentalnu normalnu distribuciju koja očito nije primjenjiva na fuzijske levele). Vizualni pristup bi bio da se identificiraju granice na grafu vrijednosti $(\alpha_{j+1} - \bar{\alpha})/s_{\alpha}$ naspram broja klastera $j$.

Druga metoda koju predlaže Mojena je bazirana na pristupu pokrenih sredina (moving average). Pravilo je da se koristi particija iz specifičnog klaster niza od $j = r$ do $j = n - 1$ klastera koja odgovara prvoj fazi $j$ i zadovoljava:

$$\alpha_{j+1} > \bar{\alpha} + L_j + b_j + ks_{\alpha}, \tag{4.30}$$

gdje su $\bar{\alpha}$ i $s_{\alpha}$ očekivanje i standardna devijacija vrijednosti fuzija baziranih na prethodnim $t$-vrijednostima; $L_j$ i $b_j$ su korekcije srednje vrijednosti ulaznog trenda u vrijednostima fuzija ($L_j$ je 'trend zaostajanja' u žargonu kontrole kvalitete, u nekim pretpostavkama jednak $(r-1)b_j/2$, gdje je $b_j$ kretajući nagib najmanjih kvadrata razina fuzija). Prednost ovakvog pravila je da podrazumijevana razina fuzija ne ulazi u osnove statistike, a nedostatak je da ispitivač sam izabire vrijednost za $r$. Uobičajeno je da se u oba slučaja odredi kriterijske vrijednosti i zatim odabere najniži broj klastera gdje je pravilo zadovoljeno.

## Hijerarhijski algoritmi

Potrebno je razlučiti hijerarhijske *metode* od hijerarhijskih *algoritama* za računanje grupiranja. Za bilo koju hijerarhijsku metodu može se koristiti nekoliko različitih računalnih algoritama da se postigne isti rezultat. Mnogi algoritmi daju ugniježdenu strukturu postepeno optimizirajući neki kriterij, dok neki rade globalno, kao na primjer minimizacija distorzije. Globalne metode su poznate i kao *direktni optimizirajući algoritmi*. Metodu za pronalaženje *škrtih stabala* (onih sa minimalnim brojem razina u hijerarhiji) predložili su

Sriram i Lewis (1993). Direktni optimizirajući algoritmi su korisni u situacijama kada neki od elemenata matrice udaljenosti nizu poznati.

Zahn (1971) daje grafičko-teoretske klaster algoritme bazirane na stablu s minimalnim rasponom (minimal spanning tree). Graf je skup čvorova i relacija između parova čvorova koji proizlaze iz rubova koji spaju čvorove. Skup opažanja i njihovih različitosti se predočava grafom čvorova i rubova, redom.

**Definicija 4.3.3.** *Rasponsko stablo (spannig tree) grafa je skup rubova koji pružaju jedinstven put između svakog para čvorova, a stablo s minimalnim rasponom je najkraće od svih takvih stabala.*

Stabla s minimalnim rasponom su povezana sa algoritmima jednostrukih veza.

# Poglavlje 5

# Metode optimizacije klastera

U ovom poglavlju podrazumijevamo klasu tehnika klasteriranja koje nam daju particiju pojedinaca u određeni broj grupa, bilo minimizacijom ili maksimizacijom određenog numeričkog kriterija. Takve metode optimizacije se razlikuju od onih opisanih u prethodnom poglavlju po tome što nužno ne formiraju hijerarhijsku klasifikacijsku strukturu podataka. Razlike između metoda u ovoj klasi se javljaju bilo zbog izbora kriterija klasteriranja, koji može biti optimiziran, te zbog izbora raznih optimizacijskih algoritama, koji mogu biti korišteni. U početnoj diskusiji o ovim metodama pretpostavlja se da je broj grupa prethodno fiksiran od strane ispitivača.

Osnovna ideja koja leži iza metoda ovog poglavlja je da se za svaku particiju od $\boldsymbol{n}$ pojedinaca u $\boldsymbol{g}$ grupa identificira indeks (vrijednost) $c(n, g)$, vrijednost koja mjeri neki aspekt 'kvalitete' te particije. Visoke vrijednosti nekih indeksa su povezane sa željenim riješenjem klasteriranja, dok je za ostale tražena vrijednost niska. Poistovjećujući indeks sa svakom particijom dopušta se usporedbu istih. Raznolikost takvih kriterija klasteriranja postoji u današnjici. Neki rade temeljeni na interindividualnim razlikama, dok drugi upotrebaljavaju originalnu matricu podataka.

## 5.1 Kriteriji klasteriranja izvedeni iz matrice različitosti

Pri konstrukciji indeksa određenog klastera, može se upotrijebiti koncept homogenosti i separacije. Informativna particija objekata trebala bi stvarati grupe takve da objekti unutar grupe imaju kohezivnu strukturu i sa grupama koje su dobro izolirane jedna od druge. Ovakav pristup je uobičajeno koristan pri definiranju kriterija klasteriranja koji radi na bazi *one-mode* matrice različitosti $\Delta$, sa elementima $\delta_{ij}$ koji mjere različitost između $i$-tog i $j$-tog objekta. Predložen je velik izbor kriterija klasteriranja, baziranih na $\delta_{ij}$, koji minimiziraju manjak homogenosti ili maksimiziraju separaciju grupa. Tablica predstavlja neke od mjera za indeks $r \in \{1, 2\}$.

1. Nedostatak homogenosti $h_1(m) = \sum_{l=1}^{n_m} \sum_{\substack{v=1 \\ v \neq l}}^{n_m} (\delta_{ml,mv})^r$
2. Manjak homogenosti $h_2(m) = \max\limits_{\substack{l,v=1,\ldots,n_m \\ v \neq l}} \left[(\delta_{ml,mv})^r\right]$
3. Manjak homogenosti $h_3(m) = \min_{v=1,\ldots,n_m} \left[\sum_{l=1}^{n_m} (\delta_{ml,mv})^r\right]$
4. Separacija $i_1(m) = \sum_{l=1}^{n_m} \sum_{k \neq m} \sum_{v=1}^{n_k} (\delta_{ml,kv})^r$
5. Separacija $i_2(m) = \min\limits_{\substack{l=1,..,n_m \\ k \neq m \\ v=1,\ldots,n_k}} \left[(\delta_{ml,kv})^r\right]$

Prve tri mjere, $h_1(m)$, $h_2(m)$ i $h_3(m)$, sve mjere manjak homogenosti, ili heterogenosti, $h(m)$, $m$-te grupe. Prvi indeks, $h_1(m)$, je suma svih (kvadriranih) različitosti između dva objekta grupe $m$; drugi, $h_2(m)$, je maksimum potonjeg. Kada je $r = 1$ i različitosti su metrike, tada se $h_2(m)$ može smatrati *dijametrom* klastera. Indeks $h_3(m)$ mjeri minimum sume svih (kvadriranih) različitosti između svih objekata grupe $m$ i jednog, referentnog člana grupe. Za $r = 1$ i metričke različitosti, indeks je poznat kao *zvijezdasti indeks* (star index), a ime je simbolično s obzirom na graf kojeg čine poveznice svih objekta sa referentnim objektom. Najmanja suma udaljenosti se postiže kada je referentni objekt lociran u centru 'zvijezde'. U ovom kontekstu, centar zvijezde se može interpretirati kao reprezentativni objekt ili *primjerak* grupe, a neki ga nazivaju i *medoid*. Mjere $i_1(m)$ i $i_2(m)$ mjere separaciju, $i(m)$, $m$-te grupe. Slično prvim dvama kriterijima heterogenosti, $i_1(m)$ mjeri sumu (kvadriranih) različitosti između objekata unutar grupe i jednog objekta izvan grupe, a $i_2(m)$ je minimum potonjeg.

Izabravši indeks koji mjeri grupni nedostatak homogenosti ili separacije, kriterij klasteriranja se može definirati preko pogodne agregacije nad grupama. Na primjer

$$c_1(n, g) = \sum_{m=1}^{g} h(m), \tag{5.1}$$

$$c_2(n, g) = \max_{m=1,..,g} [h(m)] \tag{5.2}$$

ili

$$c_3(n, g) = \min_{m=1,..,g} [h(m)]\,. \tag{5.3}$$

(Slično, iste zaključne funkcije mogu biti korištene za indekse separacije.) Prvi kriterij reflektira prosječan nedostatk homogenosti, dok zadnja dva mjere manjak homogenosti najgore i najbolje grupe, redom. Radeći s kriterijem manjka homogenosti, traženo rješenje klasteriranja je ono koje minimizira kriterij klasteriranja $c(n, g)$; dok za indekse separacije

vrijedi cilj maksimiziranja $c(n, g)$. Međutim, uočimo da kriterij klasteriranja $\sum_{m=1}^{g} h_1(m)$ ima ozbiljan nedostatak u smislu da broj različitosti koji tomu doprinosi ovisi o veličini grupe $n_m(\sum_{m=1}^{g} n_m = m)$, iako suma može postati jako velika samo jer određene grupacije u $m$ grupa proizvode mnoge različitosti koje se mogu promatrati). Ovo je dovelo do prijedloga da bi za indeks $h_1(m)$,

$$c_1{}^*(n, g) = \sum_{m=1}^{g} \frac{h_1(m)}{n_m} \tag{5.4}$$

bila prikladnija funkcija sažetka. Nadalje, kriterij klasteriranja može biti definiran i kao kombinacija mjera homogenosti i sepracije.

## 5.2 Kriteriji klasteriranja proizašli iz neprekidnih podataka

Najčešće korišten kriterij klasteriranja proizašao iz (two-mode) $n \times p$ matrice, $\mathbf{X}$, neprekidnih podataka koristi dekompoziciju $p \times p$ matrice disperzije, $\mathbf{T}$, dane s

$$\mathbf{T} = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}})(\mathbf{x}_{ml} - \bar{\mathbf{x}})', \tag{5.5}$$

gdje je $\mathbf{x}_{ml}$ $p$-dimenzionalni vektor opažanja $l$-tog objekta grupe $m$, a $\bar{\mathbf{x}}$ je $p$-dimenzionalni vektor ukupnih očekivanih vrijednosti (sample mean) svake varijable. Ta cjelokupna matrica disperzije se može particionirati u unutar-grupnu matricu disperzije

$$\mathbf{W} = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)', \tag{5.6}$$

gdje je $\bar{\mathbf{x}}_m$ $p$-dimenzionalni vektor srednjih vrijednosti unutar grupe $m$, i između-grupnu matrica disperzije

$$\mathbf{B} = \sum_{m=1}^{g} n_m (\bar{\mathbf{x}}_m - \bar{\mathbf{x}})(\bar{\mathbf{x}}_m - \bar{\mathbf{x}})', \tag{5.7}$$

tako da je

$$\mathbf{T} = \mathbf{W} + \mathbf{B}. \tag{5.8}$$

Za $p = 1$ (univariate data), zadnja jednadžba predstavlja podjelu cjelokupne sume kvadrata varijable u unutar- i između-grupnu sumu kvadrata, slično kao i jednodimenzionalna analiza varijance. U tom bi slučaju, prirodan kriterij za grupaciju bio izabrati particiju koja

odgovara minimalnoj vrijednosti unutar-grupne sume kvadrata ili, ekvivalentno, maksimalanu vrijednost između-grupne sume kvadrata.

## Minimizacija traga(W)

U višedimenzionalnom slučaju (za $p > 1$), izvod kriterija klasteriranja iz jednadžbe $\mathbf{T} = \mathbf{W}+\mathbf{B}$ nije tako jasan rez kao za $p = 1$, te je predloženo nekoliko alternativa. Očiti nastavak, za višedimenzionalan slučaj, minimizacije kriterija unutar-grupne sume kvadrata u jednodimenzionalnom slučaju je minimizacija suma između-grupne sume kvadrata nad svim varijablama. To znači minimizirati trag(**W**) (koji je ekvivalentan maksimizaciji trag(**B**)). Može se pokazati da je to ekvivalentno minimizaciji sume kvadriranih Euklidskih udaljenosti između pojedinaca i njihovih grupnih sredina, odnosno

$$E = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)'(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m) = \sum_{m=1}^{g} \sum_{l=1}^{n_m} {d^2}_{ml,m}, \tag{5.9}$$

gdje je $d_{ml,m}$ Euklidska udaljenost između $l$-tog objekta $m$-te grupe i očekivanja $m$-te grupe. Nadalje, kriterij se može izvesti i iz baze matrice udaljenosti

$$E = \sum_{m=1}^{g} \frac{1}{2n_m} \sum_{l=1}^{n_m} \sum_{v=1}^{n_m} {d^2}_{ml,mv}, \tag{5.10}$$

gdje je $d_{ml,mv}$ Euklidska udaljenost između $l$-tog i $v$-tog objekta $m$-te grupe. Prema tome minimizacija trag(**W**) je ekvivalentna minimizaciji kriterija manjka homogenosti ${c^*}_1(n, g)$ za Euklidsku udaljenost i $r = 2$ u definiciji $h_1(m)$.

## Minimizacija det(W)

U višedimenzionalnoj analizi varijance, jedan od testova za razlike vektora očekivanja grupa je baziran na omjeru determinanti cjelokupne i unutar-grupne matrice disperzije. Velike vrijednosti $det(\mathbf{T})/det(\mathbf{W})$ ukazuju da se grupni vektori očekivanja razlikuju. Takva saznanja navode Friedmana i Rubina (1967) da za kriterij klasteriranja stave maksimizaciju tog omjera. Pošto za sve particije od $n$ pojedinačnih podataka u $g$ grupa **T** ostaje ista, maksimizacija $det(\mathbf{T})/det(\mathbf{W})$ je ekvivalentna minimizaciji det(**W**).

## Maksimizacija trag($BW^{-1}$)

Daljnji kriterij kojeg su predložili Friedman i Rubin (1967) je maksimizacija traga matrice objedinjene iz produkta matrica između-grupne matrice disperzije i inverza unutar-grupne matrice disperzije. Ova funkcija je daljnji test-kriterij korišten u kontekstu višedimenzionalne

analize varijance, sa velikim vrijednostima trag($BW^{-1}$) indicirajući da se vektori očekivanja grupa razlikuju.

## Svojstva kriterija klasteriranja

Minimizacija trag(**W**) je najčešće korišteni kriterij klasteriranja od sva tri navedena u prethodnom poglavlju. Ali, opće je poznato da i to snosi određene posljedice. Prvenstveno, metoda ovisi o mjernoj skali. Različita riješenja mogu proizaći iz neobrađenih podataka i podataka standardiziranih u nekom pogledu. Da je to tako, može se vidjeti iz ekvivalentne definicije kriterija pomoću Euklidske udaljenosti u jednadžbi (5.9), te efekta vaganja na potonjem. Jasno je da je ovo od znatne praktične važnosti zbog potrebe standardizacije u mnogim primjenama. Daljnji problem korištenja ovog kriterija je taj da on može postaviti 'sferičnu' strukturu promatranih klastera čak i kada su 'prirodni' klasteri u originalnoj bazi podataka drugačijeg oblika.

Ovisnost o mjernoj skali metode trag(**W**) je bila Friedmanova i Rubinova (1967) motivacija potrage za alternativnim kriterijima koji nisu pod utjecajem skaliranja. Neovisnost kriterija o skali, baziranog na maksimizaciji $det(\mathbf{T})/det(\mathbf{W})$ ili trag($\mathbf{BW}^{-1}$), se može vidjeti formuliranjem tih funkcija u terminima svojstvenih vrijednosti $\lambda_1, ..., \lambda_p$ matrice $\mathbf{BW}^{-1}$, to jest

$$trag(\mathbf{BW}^{-1}) = \sum_{k=1}^{p} \lambda_k \tag{5.11}$$

i

$$\frac{det(\mathbf{T})}{det(\mathbf{W})} = \prod_{k=1}^{p} (1 + \lambda_k). \tag{5.12}$$

Jer su svojstvene vrijednosti matrice $\mathbf{BW}^{-1}$ iste, neovisno da je li matrica dobivena iz originalne matrice podataka **X** ili matrice težina $\mathbf{X}diag(w_1, ..., w_p)$, kriteriji optimizacije nisu pod utjecajem skakliranja. Naravno, implikacija ovoga je da takav kriterij nije prikladan za primjene klasteriranja kada ispitivač želi koristiti varijable u njezinim originalnim mjerama, ili ako želi uvesti težine bazirane na subjektivnoj prosudbi.

Kriterij minimiziranja det(**W**) je bio najčešće korišten jer ne ograničava klastere na sferične. Kao kontrast kriteriju trag(**W**), kriterij det(**W**) može identificirati eliptične klastere. Međutim, pokazalo se da oba kriterija, i trag(**W**) i det(**W**), daju grupe sadržane od približno istog broja objekata, a kriterij det(**W**), iako dopušta eliptičke klastere, pretpostavlja da klasteri imaju isti oblik (odnosno, istu orijentaciju i isti stupanj eliptičnosti). Naravno, to može prizvesti problem kada se podaci ne podudaraju s tim zahtjevima te je u tom slučaju potrebno koristiti drugačiji kriterij klasteriranja.

## Alternativni kriteriji za klastere drugačijih oblika i veličina

U pokušaju da se nadiđe problem 'sličnih oblika' kriterija det(**W**), Scott i Symons (1971) predlažu metodu klasteriranja baziranu na minimizaciji

$$\prod_{m=1}^{g} [det(\mathbf{W}_m)]^{n_m}, \tag{5.13}$$

gdje je $\mathbf{W}_m$ matrica disperzije unutar $m$-te grupe:

$$\mathbf{W}_m = \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)', \tag{5.14}$$

i $n_m$ je broj individua u $m$-toj grupi. (Metoda je ograničena samo na rješenja klastera gdje svaki klaster sazdrži najmanje $p + 1$ individuu. Restrikcija je potrebna da bi se izbjegle singularne matrice disperzije, čija bi determinanta bila nula.) Alternativni kriterij dan od Maronna i Jacovkis (1974) je minimizacija

$$\sum_{m=1}^{g} (n_m - 1)[det(\mathbf{W}_m)]^{1/p}. \tag{5.15}$$

Pri pokušaj da se nadiđe problem kriterija trag(**W**) i det(**W**), koji daju grupe jednakih veličina, Symons (1981) predlaže dva kriterija minimizacije:

$$\prod_{m=1}^{g} \left[det(W/n^2{}_m)\right]^{n_m} \tag{5.16}$$

(modifikacija kriterija determinante) i

$$\prod_{m=1}^{g} \left[det(\mathbf{W}_m/n^2{}_m)\right]^{n_m}, \tag{5.17}$$

(modifikacija kriterija danog s 5.13).

Većina kriterija klasteriranja danih gore su heuristički. To naravno ne znači da tu nisu uključene pretpostavka o klasnim strukturama. Zapravo, može se pokazati da su neki od kriterija ekvivalentni formalnijim statističkim kriterijima u kojima je optimizacija nekog kriterija ekvivalentna maksimizaciji ponašanja specifičnog baznog vjerojatnosnog modela. Takvi statistički modeli mogu pomoći boljem razumijevanju kako bi postojeći kriteriji klasteriranja bili uspješniji, te se mogu koristiti pri sugestijama daljnjijih kriterija za novonastale situacije. Na primjer, predloženi su kriteriji za klastere poznate po tome što su eliptički približno jednake veličine i oblika, ali orijentirani u različitim smjerovima.

Svi kriteriji spomenuti u ovom poglavlju su esencijalno prikladniji za podatke gdje su sve varijable mjerene na neprekidnoj skali. Kada varijable nisu neprekidne, prikladna matrica različitosti se može generirati koristeći mjere uvedene u poglavlju 3, te korištenjem kriterija klasteriranja koji radi na bazi matrice različistosti. Alternativno, matrica različitosti se može transformirati u matricu Euklidskih udaljenosti te klasteriranje bazirano na prikazu objekata u Euklidskom prostoru.

## 5.3 Optimizacijski algoritmi

Izabravši prikladan kriterij klasteriranja potrebno je odlučiti kako pronaći particiju u $g$ grupa koja optimizira dani kriterij. (Naravno da može postojati više od jedne particije koja optimizira dani kriterij klasteriranja.) Teoretski gledano, izračunali bi vrijednost kriterija za svaku moguću particiju te bi izabrali particiju koja daje optimalnu kriterijsku vrijednost. Međutim, u praksi to i nije tako očigledno. Broj različitih particija $n$ objekata u $g$ grupa dan je sa

$$N(n, g) = \frac{1}{g!} \sum_{m=1}^{g} (-1)^{g-m} \binom{g}{m} m^n. \tag{5.18}$$

Čak i za male $n$ i $g$ je velik broj mogućih particija. Prema tome, čak i s današnjim računalima, broj izračuna je prevelik tako da cjelokupno nabrajanje svake moguće particije nije moguće. Za neke je kriterije moguće identificirati optimalnu particiju bez nabrajanja svih mogućih. Za druge kriterije klasteriranja, eksplicitne tehnike optimizacije kao *dinamičko programiranje* ili *algoritmi grananja i određivanja* mogu biti korišteni da smanje nepotrebna nabrajanja, ali s takvim poboljšanjima su globalne potrage nepraktične.

Ovaj problem je doveo do razvitka algoritama dizajniranih da tragaju za optimalnom vrijednosti kriterija klasteriranja preko preraspodjele postojećih particija i čuvanja novih, samo ako se osigurava poboljšanje. Takvi algoritmi su poznati kao *penjajući algoritmi* (hill-climbing algoritmi), iako bi se u slučaju kriterija koji zahtjeva minimizaciju trebali zvati *hill descending*. Koraci pri implementaciji ovih algortama su:

1. Pronađi početnu particiju $n$ objekata u $g$ grupa
2. Izačunaj promjenu u kriteriju klasteriranja načinjenu pomicanjem svakog objekta iz njegove grupe u drugu
3. Napravi promjenu koja vodi najvećem poboljšanju u vrijednosti kriterija klasteriranja
4. Ponavljaj posljednja dva koraka sve dok nema pomaka objekta koji bi uzrokovao poboljšanje kriterija klasteriranja

Početna particija se može dobiti na mnogo načina. To bi na primjer moglo biti određeno na bazi prethodnog znanja ili bi moglo biti rezultat prethodnih aplikacija neke druge metode klasteriranja, možda jedne od hijerarhijskih metoda. Alternativno, početna particija se može izabrati slučajno (random) ili, kada se objekti mogu prikazati kao točke u Eklidskom prostoru, $g$ točaka može biti izabrano da izgledaju kao klaster. Rezultati optimizacijskih metoda mogu biti pod utjecajem izbora startne particije. Različite početne particije mogu dovesti do različitih lokalnih optimuma kriterija klasteriranja, iako je logično da sa dobro strukturiranim podacima očekujemo konvergenciju prema istom, nadajmo se globalnom, optimumu iz mnogih startnih konfiguracija. Marriott (1982) predlaže da spora konvergencija i vrlo različita grupiranja dobivena iz različitih početnih particija obično impliciraju da je $g$ krivo odabran, posebno ako nema posebnog dokaza klasteriranja. U takvim se situacijama savjetuje provođenje algoritma optimizacije više puta mijenjajući početnu particiju.

Jedan od ranijih hill-climbing algoritama predlaže iterativno ažuriranje particije istovremeno premještajući svaki objekt u grupu čijem je očekivanju najbliži, te potom preračunava očekivanje grupe. Iako to nije izričito navedeno, može se pokazati da, pod nekim uvijetima pravilnosti, je to ekvivalentno minimizaciji trag(**W**) kada Euklidske udaljenosti definiraju 'blizinu'. Takvi algoritmi, koji uključuju računanje srednje vrijednosti (centroida) svakog klastera, često se zovu *k-means* algoritmi. Algoritmi koji premještaju objekt u grupu čijem primjerku je najbliži u terminima neke mjere različitosti (udaljenosti), te nakon toga revaluiraju grupni primjerak, su od posebnog interesa u poslijednjih nekoliko godina. U kontrastu srednje vrijednosti grupe (centroid), grupni primjerak (medoid) odgovara stvarnom objektu u bazi podataka. Minimizacija različitosti primjeraka je ekvivalentno minimizaciji kriterija klasteriranja $c_1(n, g) = \sum_{m=1}^{g} h_3(m)$ za $r = 1$ i za danu matricu različitosti - ponekad poznato pod nazivom 'suma zvjezdastog kriterija'. Ovisno o njihovim začetnicima, takvi su algoritmi poznati kao *particioniranje oko medoida* ili PAM ili kao *k-median algoritmi*.

Iako prethodno navedena četiri koraka daju suštinu hill-climbing algoritma, postoje problemi pri njihovoj detaljnjijoj implementaciji. Implementacija k-means algoritma na minimizaciju trag(**W**) se razlikuje ako su objekti premješteni istovremeno ili pojedinačno. Objekti se mogu pojedinačno premještati na različite načine, na primjer slučajnim ili sistemskim redosljedom. Objekti mogu biti premješteni u najbližu grupu ili u onu koja rezultira najvećim poboljšanjem kriterija klasteriranja. Očekivanja grupa se mogu ažurirati nakon svakog pojedinačnog premještanja ili nakon premještanja određenog broja objekata. Konačno, varijacije k-means algoritma su razmjene parova dvaju objekata članova grupe.

## K-means metoda

Jedna od optimizacijskih, nehijerarhijskih metoda dijeljenja. U pristupu dijeljenjem, opažanja se razdvajaju u $g$ klastera bez uporabe hijerarhijskog pristupa zasnovanog na matrici udaljenosti ili sličnosti između svih parova točaka (podataka). *K-means* metoda dozvoljava

pomicanje članova iz jednog klastera u drugi, dozvoljava relokaciju, što nije dozvoljeno u hijerarhijskim metodama.

Prvo se odabere $g$ članova kao početne jedinice (seeds, sjeme). Oni se kasnije zamijenjuju s centroidima (vektorima sredina) klastera. Početne točke je moguće odabrati na više načina: na slučajan se način odabere $g$ članova (moguće udaljenih za specificiranu minimalnu udaljenost), izabere se prvih $g$ točaka (opet uz zahtjev minimalne udaljenosti), izabere se $g$ točaka međusobno najudaljenijih i slično.

Za metode izbora početnih točaka, broj klastera, $g$, mora biti zadan. Alternativno, može biti zadana minimalna udaljenost između početnih točaka te se tada svi članovi koji zadovoljavaju taj kriterij izabiru kao početni (seeds).

Nakon što su početne točke odabrane, svaka je preostala točka u skupu podataka pridružena klasteru s najbližom početnom točkom (zasnovanom na Euklidskoj udaljenosti). Čim klaster ima više od jednog člana, početna točka klastera se zamjenjuje njegovim centroidom. Nakon što su svi članovi pridruženi klasterima, za svaki se član provjerava je li bliži centroidu nekog drugog klastera nego centroidu vlastitiog klastera. Ako jest, premješta se u novi klaster, a centroid klastera se ponovno preračunava. Postupak se nastavlja sve dok nova poboljšanja više nisu moguća.

K-means metoda je osjetljiva na izbor polaznih nositelja (početnih točaka). Preporučljivo je da se postupak počne ponovno s drugačijim izborom početnih točaka. Ukoliko takav izbor rezultira potpuno drugačijim konačnim klasterima, ili ako je konvergencija ekstremno spora, može se zaključiti da nema prirodnih klastera podataka.

Metoda se može koristiti kao moguća potvrda hijerarhijskog postupka. Članovi se prvo klasteriraju hijerarhijskom metodom, a zatim se centroidi klastera koriste kao početne točke za k-means pristup koji dozvoljava relokaciju točaka iz jednog klastera u drugi.

Da bi se ilustrirala osjetljivost k-means metode na početni izbor nositelja koristit će se slijedeće četiri metode izbora:

1. Na slučajan se način bira $g$ opažanja čija je udaljenost barem $r$.
2. Odabere se prvih $g$ opažanja čija je udaljenost barem $r$.
3. Odabere se $g$ međusobno najudaljenijih opažanja.
4. Koristi se $g$ centroida iz rješenja s $g$ klastera dobivenih hijerarhijskom metodom prosječne veze.

## 5.4 Izbor broja klastera

U mnogim primjenama optimizacijskih metoda klasteriranja ispitivač je primoran procijeniti broj klastera u bazi podataka. Za većinu situacija je već predloženo nekoliko metoda

koje mogu pripomoći pri odluci. Većina ih je informativnog karaktera i uključuje crtanje vrijednosti kriterija klasteriranja naspram broja grupa. Velike promjene levela u grafu se obično uzimaju kao prijedlozi za određeni broj grupa (klastera). Baš kao i procedure prosuđivanja dendograma, ovakav pristup može biti vrlo subjektivan sa 'veliko' bivajući kao funkcija korisnikovih prethodnih očekivanja.

Međutim, predložen je velik broj formalnijih tehnika koje pokušavaju nadići problem subjektivnosti. Iako je predložen velik broj metoda, samo je ograničen broj istraživanja njihovih svojstava odrađen. Najdetaljnije usporedno istraživanje učinaka tehnika za određivanje broja grupa dali su Milligan i Cooper (1985), dok su zadnjih 15 indeksa za visoko-dimenzonalne binarne podatke dali Dimitriadou et al. (2002). Obje studije procjenjuju sposobnost formalnih (automatskih) metoda određivanja točnog broja klastera u nizu simuliranih podataka. Kao i sve simulacijske studije, njihovi zaključci se ne mogu generalizirani jer izvod metode može ovisi o (nepoznatoj) klaster strukturi kao i o klaster algoritmu korištenom pri određivanju članova grupe.

Dva vodeća izvođača u studiji Milligana i Coopera su bile tehnike koje su predložili Calinski i Harabasz (1974) i Duda i Hart (1973), a bile su namjenje za rad s neprekidnim podacima. Calinski i Harabasz (1974) predlažu uzimanje vrijednosti $g$, broj grupa, koja odgovra maksimalnoj vrijednosti $C(g)$, gdje je $C(g)$ dano s

$$C(g) = \frac{trag(\mathbf{B})}{(g-1)} / \frac{trag(\mathbf{W})}{(n-g)}. \tag{5.19}$$

Kao i sa svim tehnikama za određivanje broja grupa, evalulacija ovog kriterija za dani broj grupa $g$ zahtjeva znanje o članovima grupe pri određivanju matrica $\mathbf{B}$ i $\mathbf{W}$. Općenito, izabrani broj grupa ovisi o korištenoj klaster metodi (i njenoj implementaciji).

Duda i Hart (1973) nude kriterij za podjelu $m$-tog klastera u dva podklastera. Oni uspoređuju unutar-grupnu sumu kvadriranih udaljenosti između objekata i centroida, $J_1^2(m)$, sa sumom kvadriranih udaljenosti unutar-klastera, kada je klaster optimalno podijeljen u dva, $J_2^2$. Nul hipoteza je da odbacujemo tezu da je klaster homogen (i da je klaster podijeljen) ako

$$L(m) = \left(1 - \frac{J_2^2}{J_1^2} - \frac{2}{\pi p}\right)\left\{\frac{n_m p}{2\left[1 - 8/(\pi^2 p)\right]}\right\}^{1/2} \tag{5.20}$$

prelazi kritičnu vrijednost standardne normalne distribucije (ovdje $p$ označava broj varijabli, a $n_m$ je broj objekata u m-tom klasteru). Dakle, prijedlog Dudae i Harta prezentira lokalni kriterij. Ovo se može pretvoriti u globalni kriterij za određivanje da li je dodatna grupa prezentirana ili ne, imajući na umu skup testnih statistika, $\{L(m) : m = 1, ..., g\}$, za sve grupe. Nul hipoteza za homogene grupe se odbija u korist iduće grupe kada bar jedna testna statistika prelazi kritičnu vrijednost. Nadalje, treba imati na umu da se razine značajnosti ne interpretiraju kao i obično zbog višestrukih testiranja.

Daljnje pravilo koje radi na sumi kvadrata udaljenosti, također jedno od boljih preformansi iz studije Milligana i Coopera, je 'F-test' kojeg predlaže Beale (1969). Neka $S_g^2$ označava sumu kvadrata odstupanja od klasterovog centroida u uzorku. Tada je podjela $n$ objekata u $g_2$ klastera značajno bolja od podjele u $g_1$ klastera ($g_2 > g_1$) ako testna statistika

$$F(g_1, g_2) = \frac{(S_{g_1}^2 - S_{g_2}^2)/S_{g_2}^2}{[(n-g_1)/(n-g_2)]\,(g_2/g_1)^{2/p} - 1}. \tag{5.21}$$

prelazi kritičnu vrijednosti $F$-distribucije sa $p(g_2 - g_1)$ i $p(n - g_2)$ stupnjeva slobode.

Marriott (1971) predlaže moguću proceduru procijenjivanja broja grupa koristeći minimizaciju det(**W**) kao izabrani kriterij klasteriranja. On predlaže uzimanje vrijednosti $g$ za koju je $g^2 det(\mathbf{W})$ minimum. Za unimodalne distribucije, Marriott pokazuje da će to vjerovatno dovesti do prihvaćanja jednočlanske grupe (g=1), i za strogo grupirane podatke će dovesti do prikladne vrijednosti od $g$. Dodatno, za dano $g$, povezana statistika, $g^2 det(\mathbf{W})/det(\mathbf{T})$, čije vrijednosti opadaju s rastućim stupnjem klasteriranja, se može koristiti kao test za dokaz postojanja klaster strukture. Posebno, ako testna statistika ima vrijednost veću od 1, za sve moguće subdivizije, tada se može smatrati da objekti tvore jednu grupu. Svojstva uzorka testne statistike pod jedinstvenom hipotezom (specijalan slučaj ne postojanja klastera) se može ispitati Monte Carlo metodom. Pravilo su pronašli Milligan i Cooper (1985) te ima sklonost specificirati konstantan broj klastera.

Sve metode odabira broja grupa predložene do sada pretpostavljaju da su varijable mjerene na neprekidnoj skali, odnosno da su sve varijable neprekidne. Kao primjer metode koja se može koristiti za kategoričke podatke navodimo adaptaciju Goodman i Kruskal gama statistike koja se koristi u klasifikacijskim studijama. Ova procedura radi na matrici različitosti, gdje je svaka unutar-grupna udaljenost uspoređena sa svakom među-grupnom udaljenosti. U ovom se kontekstu par različitosti smatra skladnim (proturiječnim) ako je unutar-klasterska različitost strogo manja (strogo veća) od među-klasterske različitosti. Indeks skladnosti, $I(g)$, je definiran kao

$$I(g) = \frac{S_+ - S_-}{S_+ + S_-} \in [-1, 1]\,, \tag{5.22}$$

gdje su $S_+$ i $S_-$ brojevi skladnih i proturiječnih parova, redom. Broj grupa, $g$, se odabire tako da je $I(g)$ maksimum. Pravilo su pronašli Milligan i Cooper (1985).

Daljnja istraživanja koja su korisna pri određivanju broja grupa, te koja također rade bazirana na matrici različitosti, su *siluet plot* (silhouette plot) kojeg su predložili Kaufman i Rousseeuw (1990) i implementirali u R paker cluster. Za svaki objekt $i$ definiraju indeks $s(i) \in [-1, 1]$, koji uspoređuje razdvajanje objekta $i$ iz njegovog klastera sa heterogenosti klastera. Kada $s(i)$ ima vrijednosti blizu 1, heterogenost klastera objekta $i$ je puno manja nego njegovo razdvajanje i objekt $i$ se uzima kao 'dobro kalsificiran'. Slično, kada je $s(i)$ blizu -1 suprotna relacija povlači da je objekt $i$ uzet da bude 'pogrešno klasifici-

ran'. Kada je indeks blizu nule, nije jasno da li se objekt treba pripisati svom trenutnom klasteru ili susjednom klasteru. Siluete plot prikazuje vrijednosti $s(i)$ kao horizontalne granice, označene u opadajućem redosljedu za svaki klaster. Siluete plot je sredstvo za procjenu kvalitete riješenja klastera, omogućujući ispitivaču da identificira 'slabo' klasificirane objekte te tako razlikujući čist rez klastera od onih slabih. Siluete plotovi riješenja klastera, objedinjenih iz različitih izbora broja grupa, se mogu uspoređivati, a broj grupa izabran tako da kvaliteta riješenja klastera bude maksimizirana. U tom smislu *prosječna siluete širina* - prosjek $s(i)$ nad cijelom bazom podataka - se može maksimizirati tako da pruža formalniji kriterij za selekciju broja grupa.

Najnovija statistička literatura predlaže tako zvanu GAP-statistiku kao mjeru za određivanje broja klastera (Tibshirani et al., 2001). Tibshirani i kolege razvijaju pristup koji formalizira ideju pronalaska 'lakta' u grafu optimiziranog kriterija klasteriranja naspram broja klastera, $g$. Njihova ideja je standarizirati graf $\log[C(n,g)]$ naspram broja klastera, gdje je $C(n,g)$ kriterij klasteriranja koji je minimiziran iz $g$ klastera, uspoređujući ga sa njegovim očekivanjem pod null referentnom distribucijom. Za ovu priliku, dopuštajući da $E_n^*$ označava očekivanje pod uzorkom veličine $n$ iz referentne distribucije, oni predlažu da optimalna vrijednost za broj klastera bude vrijednost $g$ za koju je 'gap'

$$GAP_n(g) = E_n^*\{\log[c(n,g)]\} - \log[C(n,g)] \tag{5.23}$$

najveći. Njihova procedura transformira optimizirani kriterij klasteriranja u log-skalu, tako da maksimizacija apsolutnog nesrazmjera sa očekivanim vrijednostima iznosi maksimizaciju relativnog (faktor) nesrazmjera na originalnoj skali. Važnije, njihova procedura dopušta evaluaciju kvalitete jednočlanog riješenja klastera, te to omogućuje ispitivaču da postavi pitanje da li postoji ikakav dokaz o postojanju različitih klastera u bazi podataka ili ih je najbolje smatrati kao jednočlane homogene grupe.

Zaključno, savjetuje se neovisnost o jednom pravilu selektiranja broja grupa nego sinteza rezultata dobivenih primjenom više različitih tehnika. Također, kao i kod kriterija klasteriranja, neka pravila izbora broja klastera čine pretpostavke o strukturi klastera i daju dobre rezultate samo kada se te pretpostavke udovolje.

# Poglavlje 6

# Primjena klaster analize u medicini

U ovom poglavlju se bavimo primjenom klaster analize na bazu podataka jedne bolnice. Cilj bolnice je optimizirati korištenje dijagnostičkih uređaja. Podaci su stvarni. Cijela analiza je napravljena u **R** programu. Donja tablica prikazuje kakve podatke imamo.

Tablica 6.1: Primjer tablice s podacima

| Pregled | Spol | Starost(god) | Datum pregleda |
|---|---|---|---|
| UL | F | 35 | 01.01.2014. |
| MR | M | 56 | 02.01.2014 |
| RG | M | 23 | 01.01.2014 |
| CT | F | 15 | 01.02.2014. |
| Other | M | 66 | 02.02.2014. |
| ⋮ | ⋮ | ⋮ | ⋮ |

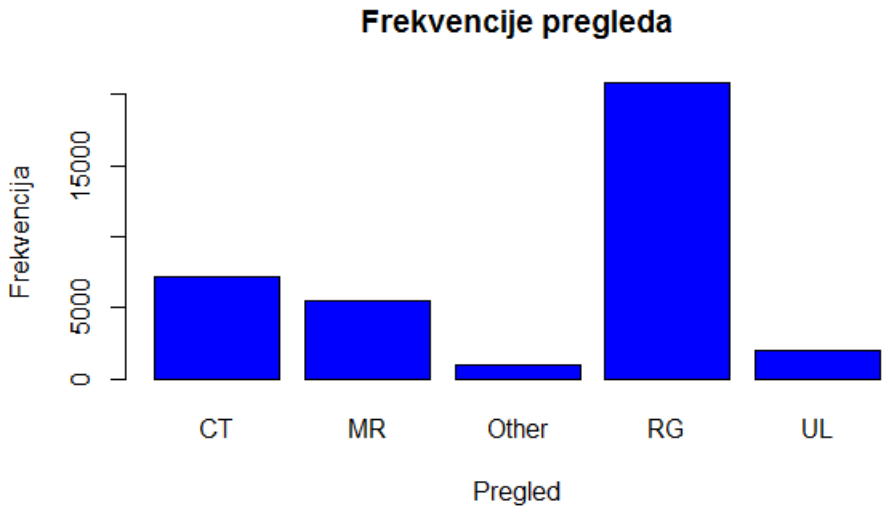
Pregledi su: UL (ultrazvuk), MR (magnetska rezonanca), RG (rendgen), CT (računalna tomografija), te Other (svi ostali pregledi koji ne koriste specifičan uređaj). Spolovi su označeni sa F (žensko) i M (muško). Datumi se kreću u rasponu od 01.01.2014. do 09.02.2014 (40 dana), a godine su u skupu $\{1, ..., 103\}$. Cjelokupna tablica sadrži 35 500 redaka, odnosno toliko pregleda.

Cilj analize je grupirati preglede po danima kako bi dobili algoritam i predikciju pomoću kojih se može uočiti 'navala' na određene preglede. Točnije, kako bi radnici bolnice bolje organizirali preglede i smanjili buduće liste čekanja.

Kao što vidimo, u podacima su prisutne kategorijske varijable: pregled, spol, te datum pregleda. Samim time ne možemo podatke obraditi na jednostavan način nego ćemo koristiti specifičnu funkciju u **R**-u za obradu kategoričkih varijabli. No, kako bi se bolje upoznali s podacima pogledajmo najprije opisnu statistiku.

## 6.1 Opisna statistika

U ovom dijelu ćemo se bolje upoznati s podacima kako bi daljnja klaster analiza istih bila lakša. Najprije pogledajmo frekvencije svih pregleda. Na donjoj slici 6.1 vidimo da su najčešće rađeni pregledi RG (rendgen), a svi svi ostali koji pripadaju u kategoriju Other su najslabije traženi te tako i odrađeni, a od četiri istaknuta .



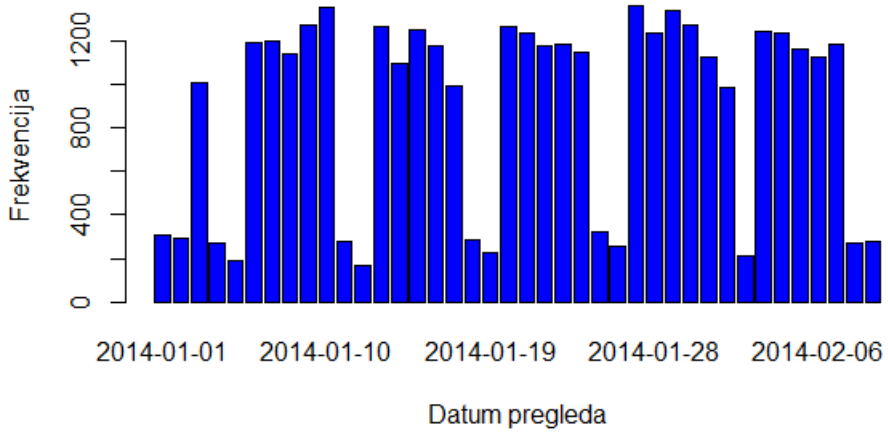

Slika 6.1: Frekvencije pregleda

Nadalje, pogledajmo slike 6.2 i 6.3 koje nam prikazuju frekvencije broja pregleda s obzirom na datuma pregleda i dob pacijenata. Prva nam kazuje kako se glavnina pregleda odvijala kroz tjedan, manje gužve su bile vikendima. Iz druge možemo isčitati da su pacijenti starosti oko 67 godina bili najzastupljeniji. Detaljni summary podataka nalazi se u prilogu s kodovima.

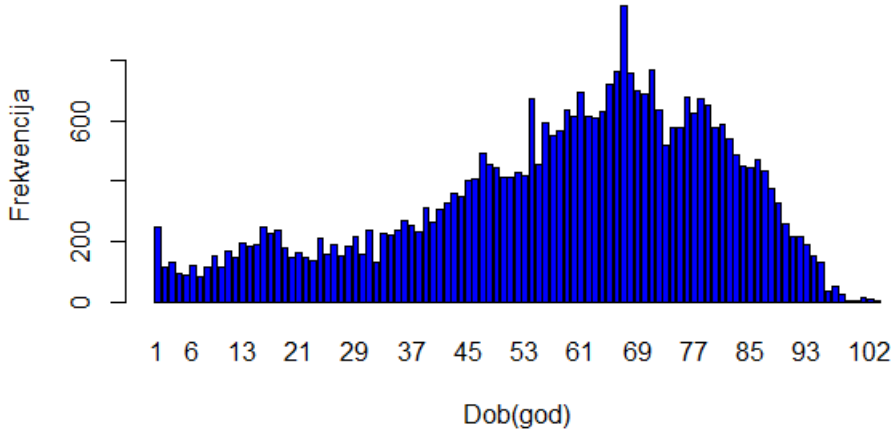
Daljnji graf Slika 6.4 nam pokazuje zastupljenost spolova, te vidimo da žene prednjače nad muškarcima. Udio žena je 0.541209. Dok nam graf Slika 6.5 pokazuje broj pregleda žena (crveni kružići) i muškaraca (plavi kružići) po godinama.

## 6.2 Primjena klaster analize

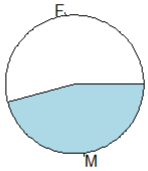
Kao što smo već napomenuli, podaci koje imamo sadrže mixed varijable, sadržani su od numeričkih i kategorijskih varijabli. Za obradu takvih podataka koristimo **R**-ovu funkciju



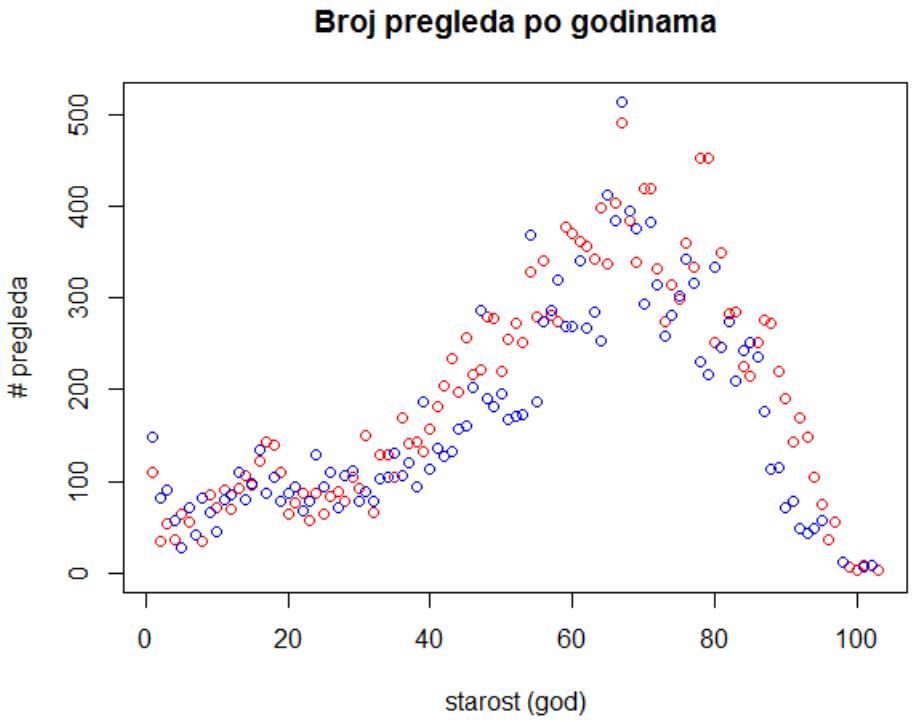

Slika 6.2: Frekvencije datuma pregleda




Slika 6.3: Frekvencije godina




Slika 6.4: Zastupljenost spolova




Slika 6.5: Dob pacijenata

*daisy*. Nadalje, bitno je napomenuti da nije bilo moguće napraviti analizu na svih 35 500 podataka zbog manjka memorije u **R**-u. Iz tog razloga je uzimano po 1000, 2000 i 3000 random pregleda te je klaster analiza rađena na takvom uzorku. Analiza kojom obrađujemo podatke je bazirana na k-means metodi u kojoj unaprijed određujemo broj klastera. Tako smo na ovom primjeru radili analizu s predviđenih dva, tri i četiri klastera. Svi korišteni kodovi su s objašnjenjima stavljeni u dodatak.

Prvenstveno krećemo sa dijagonalno simetričnom scatterplot matricom (Slika 6.6) frekvencija datuma pregleda, samih pregleda i broja pregleda po datumu. Pregledi su označeni brojevima 1 - 5 i to redom CT, MR, Other, RG, UL. Dana matrica potvrđuje zaključke dobivene opisnom statistikom. Vidimo da su pregledi RG najzastupljeniji i da se izvode svaki dan više puta. Također su pregledi svrstani u kateogriju Other ponovno najmanje izvođeni, te im je i dnevna frekvencija vrlo niska. Odnosno, nisu se izvodili svaki dan. Nadalje, ako govorimo o danima i broju pregleda po danu, vidimo da je većinom izvođeno oko 500 pregleda dnevno, a ponekad i više. Zbog toga dobivamo sliku na kojoj se jasno vide ti rezovi.

Klaster analizu provodimo k-means metodom na random generiranim uzorcima od 1000, 2000 i 3000 pregleda. Krećemo sa klasterima dobivenim na uzorku od 1000 pregleda. Na obje slike 6.7 i 6.8 vidimo da nam se klasteri uvelike poklapaju u oba slučaja, što znači da je sličnost među njima velika, a udaljenost mala. Klasterizacija uzorka u dva klastera na slici 6.7 nam daje dva jasno uočljiva klastera. Međutim, možemo uočiti da nam to i nije dobro jer oba klastera sadrže skoro sve podatke. Zbog toga je klasterizacija uzorka u tri klastera bolja jer jasnije razdvaja klastere na temelju sličnosti, odnosno različitosti. U oba slučaja vidimo da nam se na obje strane donja dva klastera izdvajaju. Kada bi pretpostavili da imamo četiri klastera, dobivamo sliku 6.9 na kojoj su klasteri točno grupirani po stupcima i sličnostima među podacima. Ujedno je to i najveći broj klastera do kojeg možemo pretpostavljati jer nam polazna tablica ima 4 stupca.

Ako uzmemo uzorak od 2000 pregleda i klasteriramo ga u dva klastera, tada dobivamo sliku 6.10. Uočavamo da dobivamo isti slučaj kao sa random uzorkom od 1000 pregleda. Također, slično dobivamo i za klasterizaciju 2000 pregleda u tri klastera, što možemo vidjeti na slici 6.11. Međutim, kada provodimo klasterizaciju u četiri klastera, Slika 6.12 nam se razlikuje od slučaja klasterizacije 1000 pregleda u 4 klastera. U ovom slučaju dobivamo da četvrti klaster sadrži grupacije s desne strane kao i jednu grupu s lijeve strane. Dakle, ipak na višim nivoima postoje podudaranja među klasterima.

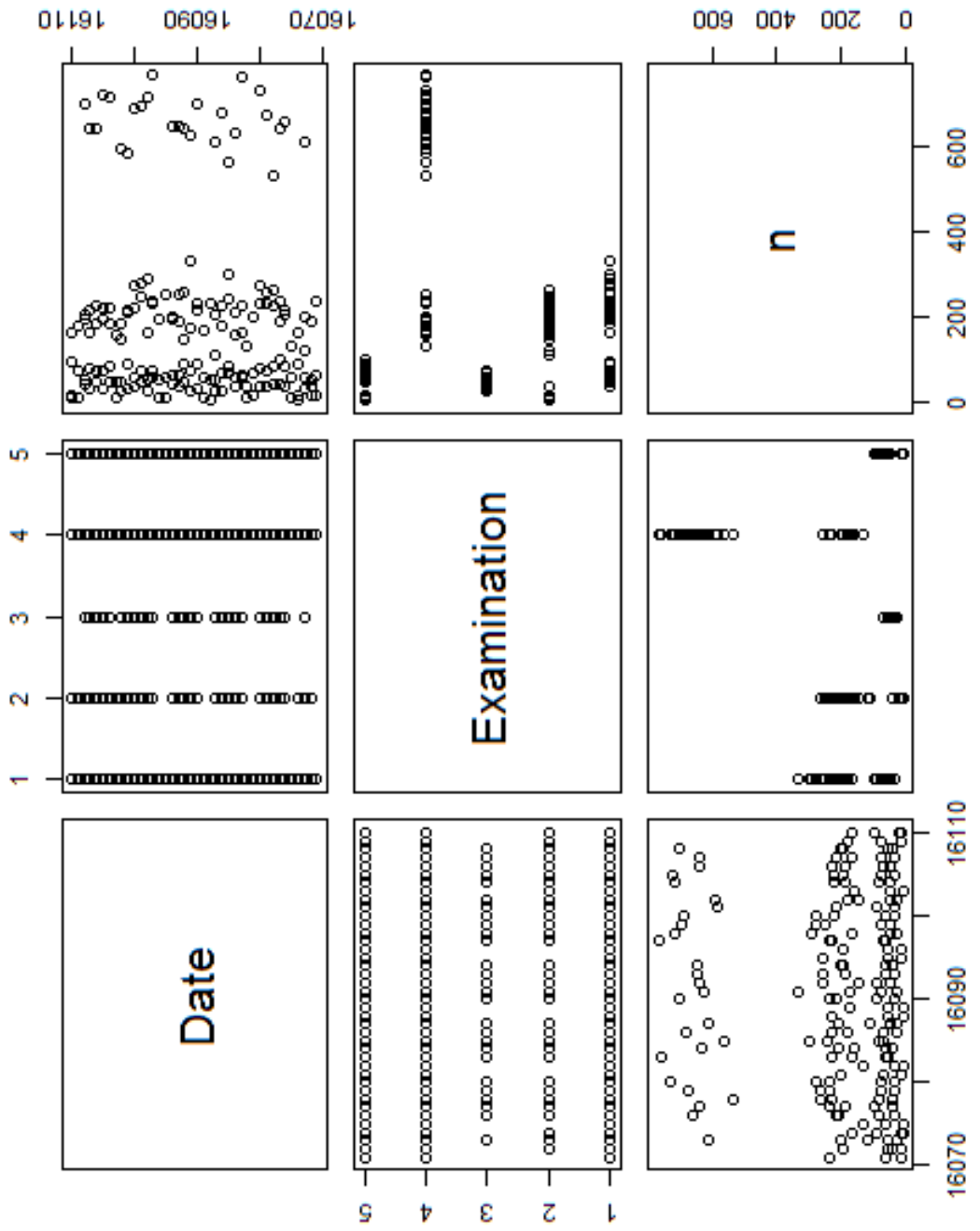
Ako pogledamo slučaj klasterizacije 3000 random izabranih pregleda u dva, tri i četiri klastera dobivamo slike: Slika 6.13, Slika 6.14 i Slika 6.15, redom. U slučaju s dva klasera dobivamo identične podklastere kao i u prethodna dva slučajna uzorka, ali dva velika klastera su drugačija. To se dogodilo jer biramo random uzorak te ne možemo utjecati na to koje podatke će program generirati. Klasterizacija uzorka u tri klastera se podudara sa prethodnim uzorcima, samo je simetrična. Razlika je uočljiva, međutim vidi se i određena

dosljednost među podacima. Svaki klaster sadrži drugačiju kombinaciju dva manja podklastera (lijevi i desni), međutim jasno uočavamo i dijelove gdje se podudaraju, to jest koliko su podaci zapravo slični. Slučaj klasterizacije 3000 pregleda u 4 velika klastera prikazan je na slici 6.15. Sličnosti se ponovno jasno uočavaju, a klasterizacija je dosljedna prethodim uzorcima.

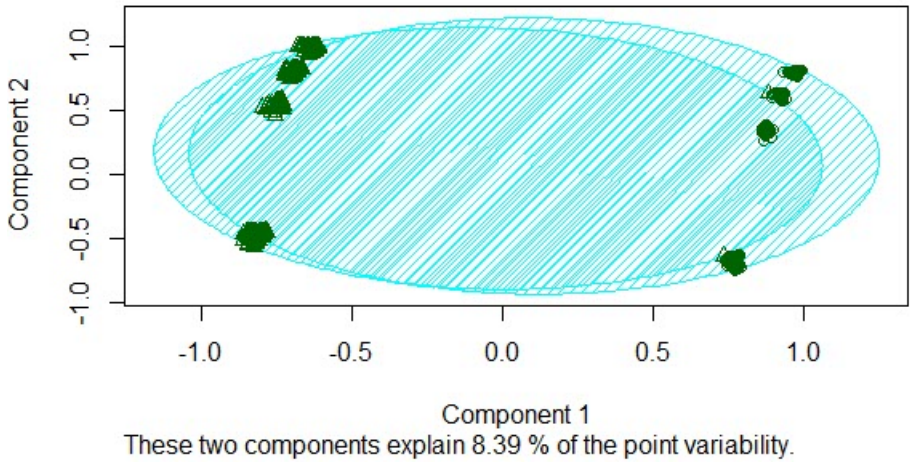
Na temelju tri obrađena uzorka možemo vidjeti određenu dosljednost među podacima. Svaki put dobivamo dva manja podklastera od koji su sastavljeni od jednog manjeg klaster i grupacije od čak tri, te su svi jaki klasteri s obrzirom da ne uočavamo odstupanja u sličnostima među podacima. Klasterizacijom uzoraka u dva, tri i četiri klastera dobivamo kombinacije većih klastera sastavljenih od manjih podklastera. Time vidimo i sličnosti među njima jer nam klasteri jasno pokazuju gdje se preklapaju.

Dobivena dva simetrična podklastera predstavljaju grupacije po spolu, a oba su sastavljena od četiri manja klastera koji prikazuju zastupljenost pregleda. Daljnom klaster analizom u dva, tri i četiri veća klastera dobivamo klastere grupirane na temelju sličnosti zastupljenosti pregleda po spolu.

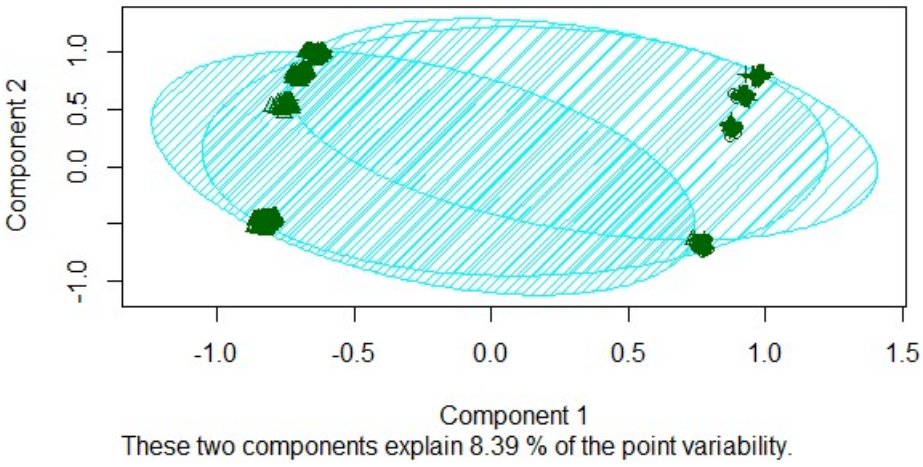
Najbolju aproksimaciju dobivamo na uzorku od 3000 pregleda, iako uzorak od samo 1000 pregleda ne odstupa značajno.



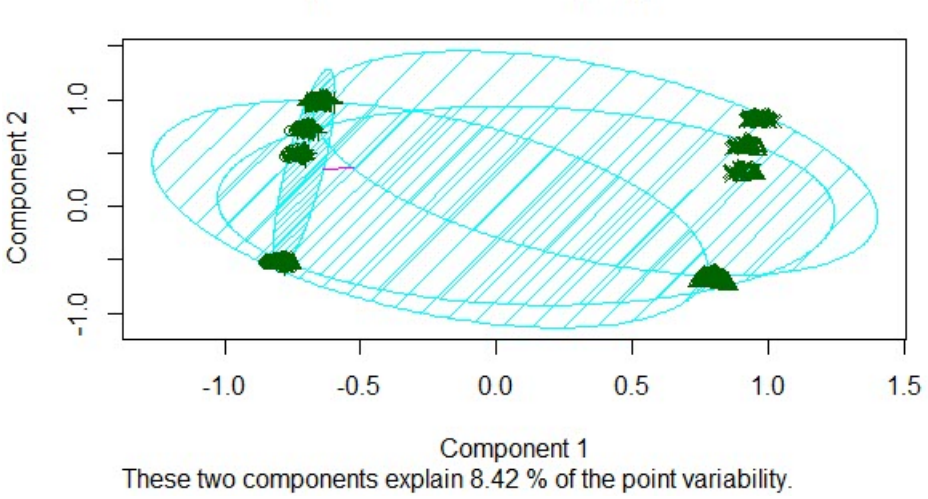

Slika 6.6: Scatterplot frekvencije datuma, pregleda i broja pregleda po datumu



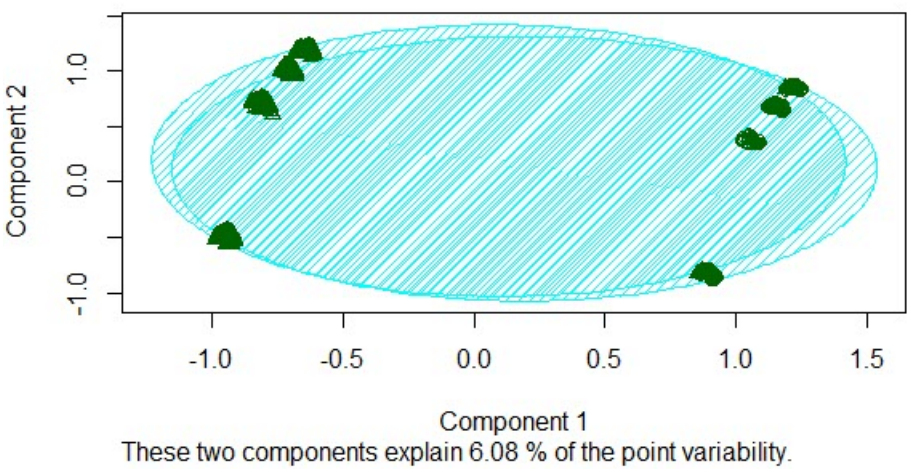

Slika 6.7: Klasterizacija uzorka od 1000 pregleda u 2 klastera



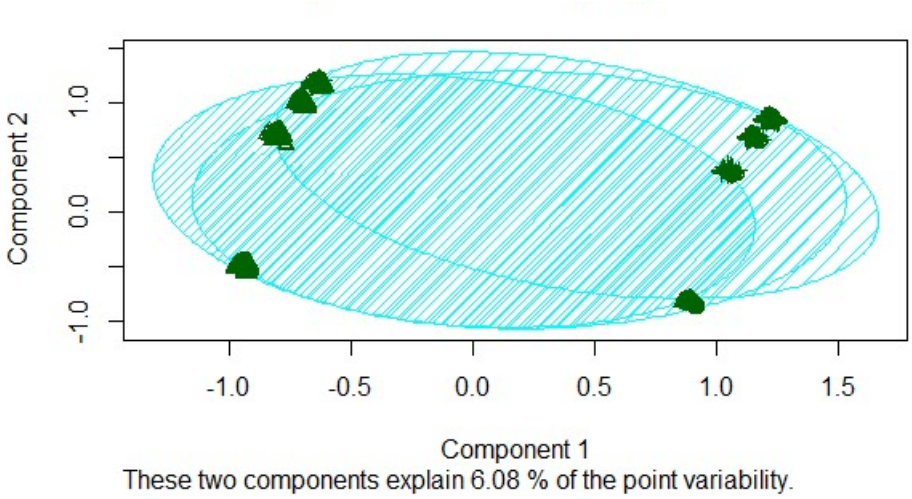

Slika 6.8: Klasterizacija uzorka od 1000 pregleda u 3 klastera



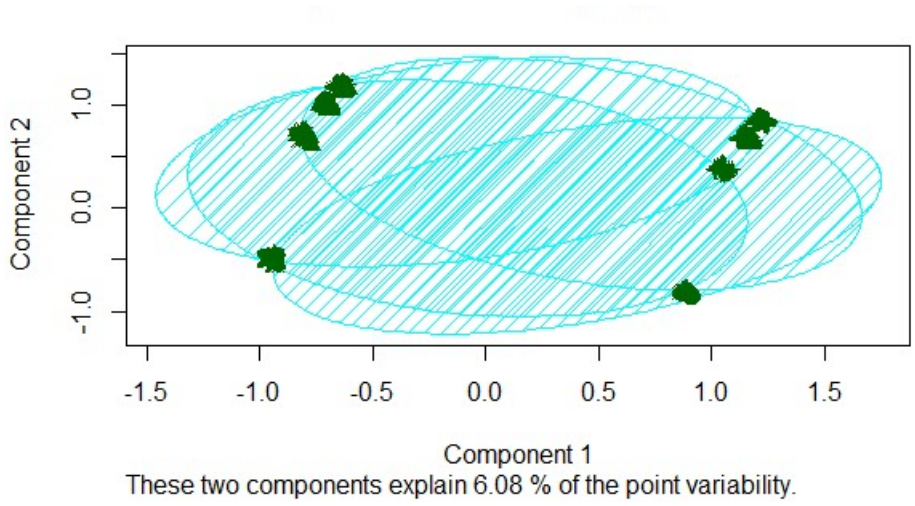

Slika 6.9: Klasterizacija uzorka od 1000 pregleda u 4 klastera



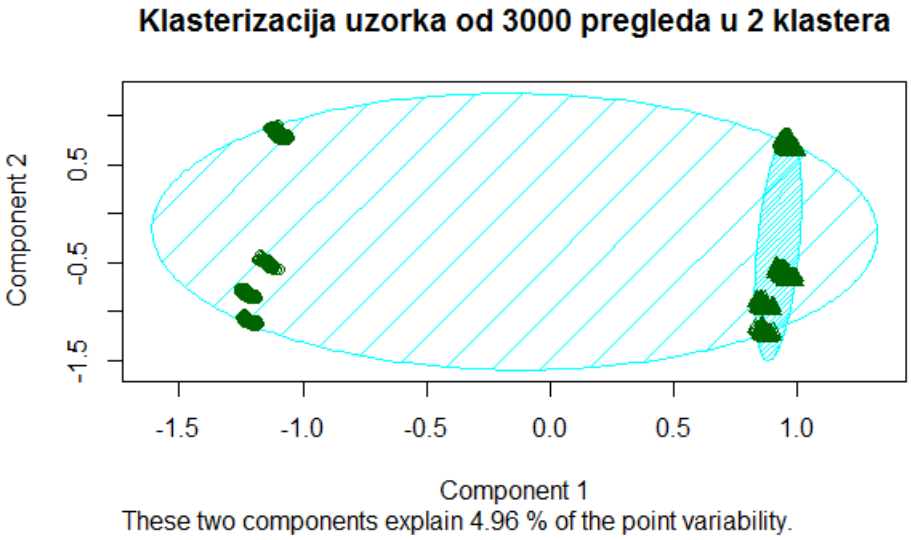

Slika 6.10: Klasterizacija uzorka od 2000 pregleda u 2 klastera



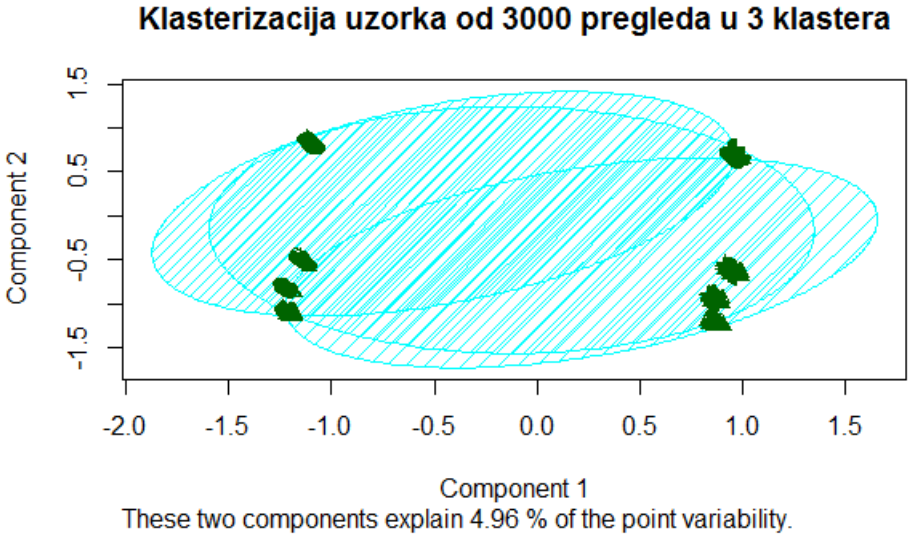

Slika 6.11: Klasterizacija uzorka od 2000 pregleda u 3 klastera



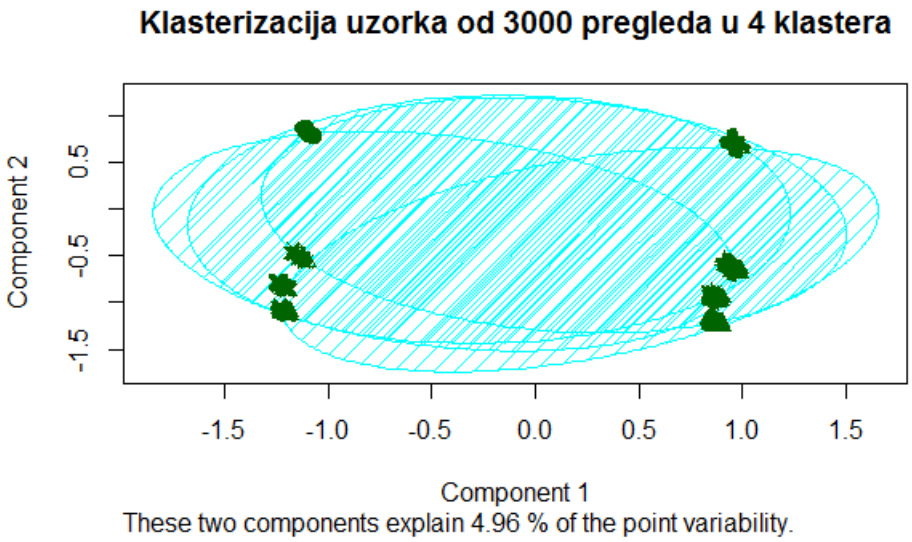

Slika 6.12: Klasterizacija uzorka od 2000 pregleda u 4 klastera




Slika 6.13: Klasterizacija uzorka od 3000 pregleda u 2 klastera




Slika 6.14: Klasterizacija uzorka od 3000 pregleda u 3 klastera




Slika 6.15: Klasterizacija uzorka od 3000 pregleda u 4 klastera

# Dodatak A

# Kodovi u R-u

Ovdje se nalazi dodatak s objašnjenim kodom korištenim za analizu, te summary svih podataka. Još jednom napominjem da zbog problema s memorijom u **R**-u nije bilo mogućnosti obraditi sve podatke odjednom. Kod je primjenjiv na veće baze podataka, a napravljen je u svrhu ovog diplomskog rada.

## A.1 Opisna statistika

```
rm(list=ls())
library(dplyr)
library(tidyr)
podaci = read.table("podaci1.csv", header = TRUE, sep=";")
data = tbl_df(podaci[, 2:5])
colnames(data)[4] = "Date"
data$Date = as.Date(data$Date, format="%m/%d/%Y")
summary(data)
```

Tablica A.1: Summary svih podataka

| Gender | Age | Examination | Date |
|---|---|---|---|
| Min. :0.0000 | Min. : 1.00 | CT : 7158 | Min. :2014-01-01 |
| 1st Qu.:0.0000 | 1st Qu.: 45.00 | MR : 5493 | 1st Qu.:2014-01-21 |
| Median :0.0000 | Median : 62.00 | Other: 1011 | Median :2014-01-30 |
| Mean :0.4586 | Mean : 58.08 | RG :20779 | Mean :2014-03-27 |
| 3rd Qu.:1.0000 | 3rd Qu.: 75.00 | UL : 2036 | 3rd Qu.:2014-06-02 |
| Max. :1.0000 | Max. :103.00 | | Max. :2014-12-01 |

```
    # # graf frekvencija pregleda muškaraca i žena s kružićima u boji
tmpF = data %>% filter(Gender==F)%>% group_by(Age) %>%
count() %>% arrange(desc(n))
tmpM = data %>% filter(Gender==M) %>% group_by(Age) %>%
count() %>% arrange(desc(n))
plot(tmpF$Age, tmpF$n, ylim=c(F,max(c(tmpF$n, tmpM$n))),main="Broj
pregleda po godinama", ylab="# pregleda", xlab="starost (god)", col="red")
points(tmpM$Age, tmpM$n, col="blue")

    podaci=read.csv('sve.csv', header = TRUE, sep=';')
Examine=podaci$Examination
FrekvencijeExam¡-data.frame(table(Examine))
nazivExam=FrekvencijeExam$Examine
barplot(FrekvencijeExam$Freq,names.arg=nazivExam, col = 'blue',
main = 'Frekvencije pregleda' ,xlab="Pregled",ylab="Frekvencija")
Genre=podaci$Gender
Age=podaci$Age
ED=podaci$Examination_date
ExamineDate=as.Date(ED,"%d.%m.%Y")
FrekvencijeED=data.frame(table(ExamineDate))
FrekvencijeGenre=data.frame(table(Genre))
FrekvencijeAge=data.frame(table(Age))
op =par(mfrow= c(1,2))
nazivED=FrekvencijeED$ExamineDate
barplot(FrekvencijeED$Freq,names.arg=nazivED, col = 'blue',
xlab="Datum pregleda",ylab="Frekvencija")
nazivAge=FrekvencijeAge$Age
barplot(FrekvencijeAge$Freq,names.arg=nazivAge,
col = 'blue',xlab="Dob(god)",ylab="Frekvencija")
par(op)
op1=par(mfrow=c(1,2))
nazivGenre=FrekvencijeGenre$Gender
barplot(FrekvencijeGenre$Freq,names.arg=nazivGenre,
col = 'blue4',xlab="Spol",ylab="Frekvencija")
pie(FrekvencijeGenre$Freq,labels=c("F","M"))
par(op1)
op =par(mfrow= c(1,2))
hist(ExamineDate, main= 'Datum pregleda',xlab = 'Datum',
ylab='Gustoća' ,breaks=20)
```

```
hist(Age, main='Dob', xlab = 'Dob(god)', ylab = 'Frekvencija')
par(op)
```

## A.2 Klasterizacija

```
rm(list=ls())
library(dplyr)
library(tidyr)
library(cluster)

    podaci=read.table('sve.csv', header = TRUE, sep = ';')

    ## funkcijom sample_n generiramo random uzorak iz početne tablice
uzorak2=sample_n(podaci, 1000, replace = F)

    ## koristimo fju daisy napravljenu za obradu kategorijskih podataka
## fja pam klasterira dane podatke k-mean metodom u prethodno
određen broj klastera k

    ## na 2 klastera
prvi4=daisy(uzorak2, metric = 'gower')
pamv =pam(prvi4, 2, diss = TRUE)
prvi4.clus= pamv$clustering
clusplot(prvi4, prvi4.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 1000 pregleda')

    ## na 3 klastera
prvi5=daisy(uzorak2, metric = 'gower')
pamv = pam(prvi5, 3, diss = TRUE)
prvi5.clus= pamv$clustering
clusplot(prvi5, prvi5.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 1000 pregleda u 3 klastera')

    ## na 4 klastera prvi6=daisy(uzorak2, metric = 'gower')
pamv = pam(prvi6, 4, diss = TRUE)
prvi6.clus= pamv$clustering
clusplot(prvi6, prvi6.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 1000 pregleda u 4 klastera')
```

```
    ## uzorak od 2000 pregleda
uzorak=-sample_n(podaci, 2000, replace = F)

    ## na 2 klastera
prvi=daisy(uzorak, metric = 'gower')
pamv =pam(prvi, 2, diss = TRUE)
prvi.clus= pamv$clustering
clusplot(prvi, prvi.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 2000 pregleda')

    ## na 3 klastera
prvi1=daisy(uzorak, metric = 'gower')
pamv = pam(prvi1, 3, diss = TRUE)
prvi1.clus= pamv$clustering
clusplot(prvi1, prvi1.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 2000 pregleda u 3 klastera')

    ## na 4 klastera
prvi7=daisy(uzorak, metric = 'gower')
pamv = pam(prvi7, 4, diss = TRUE)
prvi7.clus= pamv$clustering
clusplot(prvi7, prvi7.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 1000 pregleda u 4 klastera')

    ## uzorak od 3000 pregleda
uzorak1=sample_n(podaci, 3000, replace = F)
## na 2 klastera
prvi2=daisy(uzorak1, metric = 'gower')
pamv =pam(prvi2, 2, diss = TRUE)
prvi2.clus= pamv$clustering
clusplot(prvi2, prvi2.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 3000 pregleda u 2 klastera')

    ## na 3 klastera
prvi3=daisy(uzorak1, metric = 'gower')
pamv = pam(prvi3, 3, diss = TRUE)
prvi3.clus= pamv$clustering
clusplot(prvi3, prvi3.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 3000 pregleda u 3 klastera')
```

```
    ## na 4 klastera
prvi8=daisy(uzorak1, metric = 'gower')
pamv = pam(prvi8, 4, diss = TRUE)
prvi8.clus= pamv$clustering
clusplot(prvi8, prvi8.clus, diss = TRUE, shade = TRUE,
main = 'Klasterizacija uzorka od 3000 pregleda u 4 klastera')
```

# Bibliografija


[1] M. Aldenderfer, R. Blashfield, *Cluster analysis*, Sage, Los Angeles, 1985.

[2] J. M. Chambers, *Software for Data Analysis: Programming with R*, Springer, New York, 2009.

[3] B. Everitt et all, *Claster Analysis*, 5.izdanje, Wiley, SAD, 2011.

[4] D. Banks, et all, *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, Springer-Verlag, Berlin, 2002.

[5] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 1987.

[6] Clustering Mixed Data Types in R, dostupno na https://www.r-bloggers.com/clustering-mixed-data-types-in-r/ (kolovoz 2016.)

[7] Cluster Analysis, dostupno na http://www.statmethods.net/advstats/cluster.html (kolovoz 2016.)

# Sažetak

Kroz ovaj rad smo pobliže objasnili klaster analizu kao i njenu primjenu na podatke jedne bolnice. Upoznali smo se sa osnovnim pojmovima i tehnikama klasterizacije, te mjerama na kojima se podaci spajaju u klastere. Uz to smo objasnili metodu klasterizacije zasnovanu na unaprijed određenom broju klastera koju smo primjenili na dane podatke.

Klaster analiza podrazumijeva klasifikaciju i prepoznavanje strukture. Traži se optimalna struktura podataka za grupiranje opažanja u klastere koji se formiraju na temelju dostupnih informacija koje opisuju podatke i njihove veze. Cilj je pronaći optimalan kriterij grupiranja kod kojeg su opažanja unutar svakog klastera slična, ali se različiti klasteri međusobno razlikuju.

Analizirali smo podatke o pregledima u određenoj bolnici. Primjenom klaster analize, uočili smo različite klastere koji povezuju tipove pregleda s raznim karakteristikama pacijenata, poput spola.

# Summary

Through this dissertation, we have shown how one can use cluster analysis methods in medicine. We met the basic concepts of cluster analysis: definition, clustering methods and measures for forming clusters. In addition we explained the method by which we process the given data.

Cluster analysis includes the classification and identification of structures. It looks for optimal structure of the given data and then forms clusters based on the informations about the data. The goal is to find the optimal grouping criteria in which the observations within each cluster are similiar, a different clusters differ from each other.

We analyzed data on examinations in a certain hospital. By applying cluster analysis, we identified different clusters which relate different types of examination with characteristics of patients, such as sex.

# Životopis

Rođena sam 4. kolovoza 1992. godine u Požegi. Pohađala sam Osnovnu školu Fra Kaje Adžića u Pleternici u periodu od 1999.do 2007. Nakon osnovnoškolskog obrazovanja, 2007. godine upisujem Gimnaziju u Požegi, smjer Prirodoslovno-matematički, te isti završavam 2011. godine. Iste godine upisujem preddiplomski sveučilišni studij Matematika, nastavnički smjer, na Prirodoslovno-matematičkom fakultetu, Sveučilište u Zagrebu koji završavam 2014. godine. Nakon preddiplomskog nastavničkog smjera upisujem diplomski sveučilišni studij Financijska i poslovna matematika.