METODE REDANJA I VREDNOVANJA

LAVOSLAV ČAKLOVIĆ

SADRŽAJ

1. UVOD

U ovom tekstu dat ćemo kratki prikaz nekih multivarijatnih metoda iz statistike koje postaju sve zanimljivije u kontekstu rudarenja po podacima, modeliranju zavisnosti među varijablama (specijalno kategorijskih) i izvajanju informacijskih značajki potrebnih za prepoznavanje objekata. To su:

(1) Principal Component Analysis (PCA)
(2) Correspondence Analysis (jednostavna (CA) i kanonska (CCA))
(3) Metode vrednovanja (s naglaskom na metodu potencijala)

Date: Prosinac 2011.

Zajedničko tim metodama je da skup objekata nastoje poredati na nekoj skali. Ovisno o tipu i kvaliteti ulaznih informacija odlučujemo se za jednu (ili više) od metoda za analizu podataka. Skala dodatno može poslužiti za račun mjere sličnosti među objektima i njihovu kategorizaciju/klasterizaciju.

Metode su ilustrirane na dostupnim podacima iz literature i podacima dobivenim kliničkim istraživanjima. Neki dijelovi ovog teksta još su u fazi revizije i dio su još nedovršenog znanstvenog istraživanja.

2. Principal Component Analysis (PCA)

PCA je procedura (matematički alat) koji preslikava skup mjerenja ***koreliranih varijabli*** u skup vrijednosti ***nekoreliranih varijabli*** koje nazivamo *glavnim komponentama*. Osnovni pojmovi u priči su *kovarijanca* i *korelacija* dviju slučajnih varijabli X, Y definiranih na istom vjerojatnostnom prostoru (Ω, P). One se definiraju (redom) s

$$\mathrm{Cov(X, Y)} = E[X \cdot Y] - E[X] \cdot E[Y],$$
$$\mathrm{Corr(X, Y)} = \frac{\mathrm{Cov(X, Y)}}{\sigma_X \sigma_Y},$$

gdje je $E[X] := \sum_i p_i x_i \ (= \mu_X)$ očekivanje, a $\sigma_X := \sum_i p_i (x_i - \mu_X)^2$ standardna devijacija, a $p_i = P(\{i\})$ je elementarna vjerojatnost na Ω.

Za vektorsku slučajnu varijablu $X = (X_1, \ldots, X_n)$ matrica kovarijanci i matrica korelacije se definiraju s:

$$\mathrm{CovM}_{ij} := \mathrm{Cov(X_i, X_j)}, \quad \mathrm{CorrM}_{ij} := \mathrm{Corr(X_i, X_j)}, \quad i, j = 1, \ldots, n.$$

Geometrijski gledano nekoreliranost dviju (centriranih) varijabli znači okomitost vektora njihovih vrijednosti u skalarnom produktu $\langle\, x \mid y \,\rangle := \sum_i p_i x_i y_i$.

Glavne primjene PCA su:

(1) redukcija broja varijabli
(2) detekcija strukture zavisnosti među varijablama.

Ta struktura se najčešće izražava u terminima *component scores* (koordinate mjerenog objekta u bazi glavnih komponenata) i *loadings* (matrica svojstvenih vektora (glavnih komponenti)).

Glavne komponente (svojstveni vektori) mogu se računati ili iz matrice kovarijanci (ako varijable nisu ravnopravne i poznat je trade-off) ili iz matrice korelacije (ako analitičar tako želi). Prva glavna komponenta maksimizira funkciju varijance (inercije)

$$\sigma = \max_{\|\xi\|=1} \xi^\tau M \xi,$$

a geometrijski smisao tog izraza je moment inercije oblaka podataka u odnosu na ravninu koja je okomita na vektor ξ. Obično se svojstvene vrijednosti (inercije) poredaju po veličini i analitičar može originalne podatke aproksimirati njihovom projekcijom na potprostor određen s prvih k svojstvenih vektora. Zato se i govori o redukciji dimenzije. Ukupna varijanca je suma devijacija u odnosu na glavne komponente zbog simetričnosti matrice M. Vrijednost σ u gornjem izrazu se najčešće izražava u postocima od ukupne varijance i tada se naziva *postotkom objašnjene* varijance.

2.1. Analiza testiranja na Alzheimerovu bolest (AD).

Ovo je jednostavan primjer mjerenja 6 različitih karakteristika (varijabli) na skupu od 20 ispitanika. Rezultati mjerenja dani su u tablici 1. Studija je rađena sa svrhom nalaženja efikasnog testa za dijagnosticiranje Alzheimerove bolesti (AD).

n^o	A	B	C	D	E	F
$A1$	30	23	132	138	1	1
$A2$	30	25	136	116	1	1
$A3$	30	23	130	136	1	1
$A4$	30	26	138	138	1	1
$A5$	29	18	124	129	0	1
$A6$	29	25	116	110	0	1
$A7$	29	26	131	144	0	1
$A8$	30	22	107	112	0	1
$A9$	30	17	130	125	0	1
$A10$	30	22	138	135	0	1
$A11$	26	20	126	103	0	1
$A12$	26	7	103	93	0	0
$A13$	25	0	117	103	0	0
$A14$	26	12	99	93	0	0
$A15$	25	13	87	100	0	0
$A16$	22	2	55	81	0	0
$A17$	23	0	111	113	0	0
$A18$	24	10.5	92	110	0	0
$A19$	25	0	117	121	0	0
$A20$	26	1.5	103	100	0	0

$A =$ MSE, $B =$ REY, $C =$ IQVER
MSE = *Mental Status Exam*
REY = *Rey-Osterrieth test*
IQVER = *Verbal IQ test*

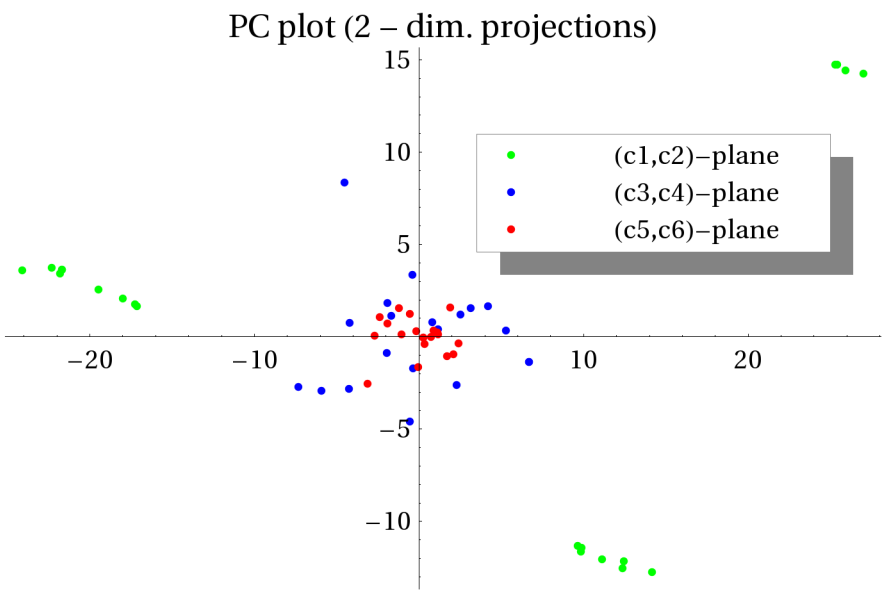
$D =$ IQPRF, $E =$ FRQ, $F =$ RARE
IQPRF = *Performance IQ test*
FRQ = *Indikator reakcije (frekventni ton)*
RARE = *Indikator reakcije (rijetki ton)*

TABLICA 1. Rezultati mjerenja 20 ispitanika (reci) pomoću šest testova na AD.

Pitanje koje se postavlja: "Je li za dijagnosticiranja AD potrebno izvršiti sva mjerenja ali samo neka?" Raspon mogućih vrijednosti za svaku varijablu je poznat i tablicu možemo reskalitati tako da posljednja 4 stupca s 1.2, 2.5, 30, 30 respektivno. Na novu tablicu je

Svojstvene vrijednosti i svojstveni vektori.					
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
369.77	96.70	15.71	7.66	2.52	1.00
0.124			−0.161		
					−0.974
0.462	−0.160	0.709	−0.497		
					0.105
0.159		−0.508	−0.512		
				−0.649	0.184
0.126		−0.415	−0.485		
				0.757	
0.421	0.901		0.106		
0.744	−0.402	−0.254	0.467		

TABLICA 2. Tablica svojstvenih vrijednosti (1. redak) i svojstvenih vektora (stupci) matrice kovarijanci podataka iz tablice 1. Nenapisane komponente svojstvenih vektora su zanemarive. Uočite dominaciju prvih dviju svojstvenih vrijednosti.



SLIKA 1. Glavne komponente za podatke iz tablice 1 računate preko matrice kovarijanci. Zelene točkice predstavljaju projekcije izmjerenih rezultata na prve dvije glavne komponente (scores). Prve dvije komponente (zeleno) objašnjavaju 95% ukupne varijance.

primjenjena PCA analiza. PCA daje glavne komponente (svojstveni vektori matrice kovarijanci) u tablici 2. Svojstvene vrijednosti nalaze se u prvom retku i poredane su po veličini.

Te su vrijednosti izračunate pomoću statističkog paketa `R` korištenjem standardne naredbe `princomp`.

```
> pcdat <- princomp(tdata, cor=FALSE)  # use covariance matrix
> summary(pcdat)
> pcdat$scores
```

Svojstveni vektori i drugi korijeni svojstvenih vrijednosti čuvaju se u varijablama `$loadings` i `$sdev`.

U tablici 3 dane su projekcije redaka (ispitanika) na prve dvije glavne komponente. Na slici 1 vide se tri zelena klastera koji odgovaraju ispitanicima A1–A4 (lijevo), A5–A11 (desno-dolje) i A12–A20 (desno-gore). Takvu istu klasterizaciju daju i varijable E i F što se jasno vidi iz tablice 1.

2.1.1. ***Zaključak.*** Zaključak se nameće sam po sebi. ***Varijable E i F su dovoljne za objasniti 95% ukupne varijance sadržane u podacima u tablici 1.***

3. SIMPLE CORRESPONDENCE ANALYSIS (CA)

Ovdje ćemo dati nekoliko jednostavnih primjena CA. Teorija i razne druge primjene dobro su opisane u Greenacre (1984, 2007) [5], [6] i Benzécri (1973) [1].

3.1. MOTIVACIJA. Klasični statistički problem vezan uz ***tablicu kontingencije*** je *redanje* (ordination). Pretpostavimo da su zadani objekti svrstani u n klasa $\xi_1, \ldots, \xi_n$ prema jednom obilježju i m klasa $\eta_1, \ldots, \eta_m$ prema drugom obilježju. U tablicu kontingencije[1]

[1]Još jedan (engleski) naziv je ***co-occurrence matrix***.

```
> pcdat$scores[,1:2]
        Comp.1     Comp.2
A1  -25.429839  14.719902      A11  -9.859274 -11.662757
A2  -25.929376  14.411289      A12  19.441125   2.535023
A3  -25.316144  14.726182      A13  21.654987   3.629045
A4  -27.007671  14.226345      A14  17.255772   1.743394
A5   -9.895583 -11.453230      A15  17.122426   1.633658
A6  -12.401870 -12.538562      A16  24.074184   3.574071
A7  -14.194074 -12.763400      A17  22.294244   3.710375
A8  -11.154078 -12.079827      A18  17.992792   2.043870
A9   -9.647114 -11.333008      A19  21.654987   3.629045
A10 -12.464422 -12.160462      A20  21.808924   3.409048
```

TABLICA 3. Projekcije redaka (ispitanika) na prve dvije glavne komponente.

Colour	Eye			
Hair	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

TABLICA 4. Tablica kontingencije: kosa-oči (boje).

za svaki par vrijednosti kategorijskih varijabli (ξ_i, η_j) zapisujemo broj $f_{ij} \geq 0$ koji može biti indikator (binarna vrijednost) ili učestalost događaja koje obje kategorije uzrokuju. To može biti broj ljudi s crnim očima i plavom kosom ili broj zaposlenih žena koje smatraju da su djeca sretnija ako je njihova mama s njima kod kuće. Retke tablice identificiramo s prvom kategorizacijom, a stupce s drugom.

Cilj analize takve tablice je poredati klase u nekom logičkom slijedu pomoću pridruženih im težina. Ako s x_i, y_j označimo tražene težine redaka i stupaca onda je (jedna) logična veza među njima

$$x_i \propto \frac{\sum_j f_{ij} y_j}{\sum_j f_{ij}}, \quad y_j \propto \frac{\sum_i x_i f_{ij}}{\sum_i f_{ij}} \tag{3.1}$$

ili matrično zapisano

$$x \propto D^{-1} F y, \quad y \propto E^{-1} F^\tau x \tag{3.2}$$

gdje je F tablica kontingencije, F^τ transponirana matrica, a

$$D = \text{diag}(f_{i\bullet}), \quad f_{i\bullet} = \sum_j f_{ij} \quad \text{(masa retka)},$$

$$E = \text{diag}(f_{\bullet j}), \quad f_{\bullet j} = \sum_i f_{ij} \quad \text{(masa stupca)}.$$

Vektori x i y su rješenja jednadžbi

$$x \propto D^{-1}FE^{-1}F^{\tau}x, \quad y \propto E^{-1}F^{\tau}D^{-1}Fy, \tag{3.3}$$

što znači da su to svojstveni vektori navedenih matrica.Te matrice su pozitivne matrice (jer je F pozitivna), pa Perronov teorem osigurava da je najveća (po modulu) svojstvena vrijednost (tih matrica) jednaka 1 i odgovarajući svojstveni vektor je $\mathbb{1} = [1, 1, \ldots, 1]^{\tau}$. Štoviše, svojstven par $(1, \mathbb{1})$ je jedinstven. Stoga je razumno tražiti netrivijalna rješenja za x i y koja se razlikuju od $\mathbb{1}$. Ova tehnika se naziva ***Correspondence Analysis***.

Jedan mogući načina računanja x i y je pomoću sljedećeg iterativnog postupka:

(1) Starta se s proizvoljnim vektorom $y = y_0 \neq \mathbb{1}$.
(2) Propisno ga standardiziramo[2] tako da vrijedi $\sum_j f_{\bullet j} y_j = 0, \quad \sum_j f_{\bullet j} y_j^2 = 1.$
(3) Neka je $x = D^{-1}Fy$ i x standardiziramo: $\sum_i f_{i\bullet} x_i = 0, \quad \sum_i f_{i\bullet} x_i^2 = 1.$
(4) Neka je $y = E^{-1}F^{\tau}x$.
(5) Ponavljaju se postupci (2) – (4) do zadovoljavajuće točnosti.

Mnogi numerički paketi i procedure koriste SVD (Singular Value Decomposition) što je opisano u poglavlju 3.2 (Modeliranje pomoću kovarijance).

3.2. MODELIRANJE POMOĆU KOVARIJANCE. Za razliku od gore opisanog iterativnog postupka promatrati ćemo problem maksimizacije kovarijance dane formulom (3.4) uz uvjete centriranja i normiranja (3.5) i (3.6) i naći kritične točke tog problema pomoću Lagrangeovih multiplikatora.

Ulazni podaci su u obliku dvodimenzionalne tablice $F = (f_{ij})$ s n ($i = 1, \ldots, n$) redaka i m stupaca ($j = 1, \ldots, m$). Nenegativan broj f_{ij} je učestalost (broj) izmjerenih entiteta koji posjeduju i-to i j-to obilježje.

Marginalne vrijednosti[3] $f_{i\bullet}, (i = 1, \ldots, n)$ (sume po recima) i $f_{\bullet j}, (j = 1, \ldots, m)$ (sume po stupcima) stavljamo kao dijagonale kvadratnih matrica $D \in \mathbb{R}^{n\times n}$ i $E \in \mathbb{R}^{m\times m}$ respektivno. Označimo li s u_n i u_m matrice stupce duljine n i m čiji su svi elementi jednaki 1 tada se suma svih elemenata matrice F može dobiti kao produkt matrica $N = u_n^{\tau} F u_m$.

Pretpostavimo da želimo naći težine redaka x i težine stupaca y tako da je kovarijanca

$$Cov(x, y) = \frac{1}{N} x^{\tau} F y \tag{3.4}$$

maksimalna moguća, uz uvjete

$$u_n^{\tau} D x = 0, \quad u_m^{\tau} E y = 0 \quad \text{(centriranje)} \tag{3.5}$$

$$x^{\tau} D x = N, \quad y^{\tau} E y = N \quad \text{(normiranje)}. \tag{3.6}$$

Interpretacija problema maksimizacije. Tablica F je tablica kontingencije dvodimenzionalne slučajne varijable, označimo je s (X, Y), definirane na vjerojatnostnom prostoru

$$\Omega = \{(i, j) \mid i = 1 \ldots, n; j = 1, \ldots, m, \}$$

s elementarnom vjerojatnošću

$$p_{ij} := \frac{1}{N} f_{ij}.$$

[2]Bez standardizacije postupak bi dao svojstveni vektor koji pripada maksimalnoj svojstvenoj vrijednosti 1, a to je $\mathbb{1}$.

[3]Nazivaju se još i *masama* redaka i stupaca

Tada je očekivanje

$$E[X] = \frac{1}{N}\sum_i f_{i\bullet}x_i = u_n^\tau Dx = 0,$$
$$E[Y] = \frac{1}{N}\sum_j f_{\bullet j}y_j = u_m^\tau Ex = 0,$$

odakle zaključujemo da je

$$\begin{aligned} Cov(X,Y) &= E[X\cdot Y] - E[X]\cdot E[Y] \\ &= \frac{1}{N}\sum_i\sum_j f_{ij}x_iy_j \\ &= \frac{1}{N}x^\tau Fy. \end{aligned}$$

Uvjet normiranja (3.6) daje

$$\begin{aligned} Var[X] &= \frac{1}{N}\sum_i f_{i\bullet}x_i^2 - E[X]^2 \\ &= \frac{1}{N}x^\tau Dx = 1 \end{aligned}$$

i na isti način

$$Var[Y] = 1,$$

što znači da su slučajne varijable X i Y koje zadovoljavaju uvjete (3.5) i (3.6) standardizirane.

3.3. SVD DEKOMPOZICIJA. Gornji problem maksimizacije je problem uvjetne maksimizacije bilinearne forme i pripadne stacionarne jednadžbe su:

$$Fy = \xi_x Dx + \mu_x Du_n \tag{3.7}$$
$$F^\tau x = \xi_y Ey + \mu_y Eu_m, \tag{3.8}$$

gdje su $\xi_x, \xi_y, \mu_x, \mu_y$ Lagrangeovi multiplikatori. Množenjem gornjih jednadžbi slijeva s u_n^τ i u_m^τ respektivno i uvažavanjem uvjeta (3.5) dobije se

$$u_n^\tau Fy = \xi_x u_n^\tau Dx + \mu_x u_n^\tau Du_n$$
$$u_m^\tau F^\tau x = \xi_y u_m^\tau Ey + \mu_y u_m^\tau Eu_m,$$

odnosno

$$u_m^\tau Ey = \mu_x \cdot N$$
$$u_n^\tau Dx = \mu_y \cdot N,$$

što daje $\mu_x = \mu_y = 0$. S druge strane je $\xi_x = \xi_y =: \sigma(x,y)$ što se dobije množenjem jednadžbi s x^τ i y^τ respektivno i uvažavanjem uvjeta (3.6). Time se stacionarne jednadžbe svode na jednostavniji sustav:

$$Fy = \sigma Dx \tag{3.9}$$
$$F^\tau x = \sigma Ey,$$

zajedno s uvjetima (3.5). Problem (3.9) je *singularni problem* ili problem nalaženja singularne dekompozicije što se vidi ako uvedemo supstitucije

$$x =: D^{-\frac{1}{2}}p, \quad y =: E^{-\frac{1}{2}}q, \quad Z := D^{-\frac{1}{2}}FE^{-\frac{1}{2}}. \tag{3.10}$$

Tada su p i q rješenja sustava

$$Zq = \sigma p \tag{3.11}$$
$$Z^\tau p = \sigma q,$$

što je ekvivalentno nalaženju SVD dekompozicije matrice Z

$$Z = P\Sigma Q^\tau. \tag{3.12}$$

Stupci matrica P, Q su oronormirani, oni su ujedno svojstveni vektori od $Z^\tau Z$ i ZZ^τ respektivno tj.

$$ZZ^\tau P = P\Sigma^2, \quad Z^\tau ZQ = Q\Sigma^2. \tag{3.13}$$

Nije teško vidjeti da je problem svojstvenih vrijednosti (3.3) ekvivalentan problemu (3.13).

Matricu P nazivamo *lijevom singularnom* matricom, a njene stupce *lijevim singularnim* vektorima, dok matricu Q nazivamo *desnom singularnom* matricom, a njene stupce *desnim singularnim* vektorima. Matrica Σ je dijagonalna i njene elemente nazivamo *singularnim vrijednostima* matrice Z. SVD dekompozicija matrice je standardna procedura svakog numeričkog paketa. U pravilu su singularne vrijednosti poredane u opadajućem nizu na dijagonali od Σ i one su korijeni svojstvenih vrijednosti od ZZ^τ (odnosno $Z^\tau Z$) što se vidi iz jednadžbe (3.13), a ima ih onoliko (različitih od nule) koliki je rang matrice Z.

Još jedan detalj je zaimljiv u vezi SVD. Ako jednadžbu (3.12) prepišemo na način

$$Z = \sum_{l=1}^{R} \sigma_l p_l q_l^\tau,$$

gdje je R rang matrice Z, a p_l, q_l lijevi i desni singularni vektori, onda su sumandi matrice ranga 1 i može se pokazati da je suma

$$Z^r = \sum_{l=1}^{r} \sigma_l p_l q_l^\tau,$$

najbolja aproksimacija matrice Z pomoću matrica ranga r u smislu

$$\|Z - Z^r\|^2 = \text{trace}(Z - Z^r)(Z - Z^r)^\tau = \min_X \|Z - X\|^2$$

gdje je X skup matrica ranga manjeg ili jednakog r. Dokaz se može naći u ???. Kvaliteta te aproksimacije je dana omjerom sume prvih r singularnih vrijednosti i sume svih svojstvenih vrijednosti. Taj omjer se interpretira kao *rekonstruiran omjer objašnjene varijance*.

3.4. RJEŠENJA STACIONARNIH JEDNADŽBI.

Teorem 3.1. *Sva rješenja (x, y, σ) stacionarnih jednadžbi su stupci matrica X i Y oblika*

$$X = \sqrt{N} D^{-\frac{1}{2}} P, \quad Y = \sqrt{N} D^{-\frac{1}{2}} Q$$

gdje su P i Q lijeva i desna singularna matrica u rastavu (3.12)*, a faktor $\sqrt{N}$ je prisutan zbog normiranja* (3.6)*.*

Matrica P (i X) je matrica tipa $n \times m$, a Q (i Y) je matrica tipa $m \times m$. Nadalje, $X^\tau DX = N \cdot P^\tau P = N \cdot I_n$ i $Y^\tau EY = N \cdot Q^\tau Q = N \cdot I_m$, tj. zadovoljeni su uvjeti centriranja i normalizacije.

Dokaz je direktna posljedica SVD dekompozicije i supstitucije (3.10).

U primjenama je najčešće $m \leq n$, tj. u tablici F je više redaka nego stupaca, što znači da Σ ima najviše m netrivijalnih singularnih vrijednosti. Ako je trojka (x, y, σ) rješenje sustava (3.9) onda ju nazivamo *singularnom trojkom*, a par (x, y) je *singularnim parom*.

Ukratko ćemo rezimirati rezultate dobivene SVD u terminima singularnih trojki. Dakle,

- postoji najviše m singularnih trojki $(x_s, y_s, \sigma_s),\ s = 1, \ldots, m$ gdje su x_s, y_s redom stupci matrica X i Y.
- Ako je $\sigma_s \neq \sigma_t$ onda su x_s i x_t D-ortogonalni, jednako kao što su y_s i y_t tj. $x_s^\tau D x_t = 0$ i $y_s^\tau E y_t = 0$.
- $(u_n, u_m, 1)$ je također singularna trojka (jer trivijalno zadovoljava (3.9)) i tu trojku nazivamo *trivijalnim* rješenjem jer ne ovisi o podacima danim u matrici F. Lako se vidi da je 1 najveća singularna vrijednost (v. [3]).
- Svaka druga singurarna trojka (x_s, y_s, σ_s) za koju je $\sigma_s < 1$ okomita je na trivijalnu, tj. $u_n^\tau D x_s = 0$ i $u_m^\tau E y_s = 0$. To znači da su sve singularne trojke (x_s, y_s, σ_s) stacionarne točke problema maksimizacije izuzev trivijalne jer ne zadovoljava uvjet centriranja.

3.5. CA ANALIZA PRIMJERA IZ TABLICE 4. U tablici je zabilježen je broj osoba s određenom bojom kose i bojom očiju od ukupno 592 osobe. Pitanje koje se postavlja je koje od navedenih boja kose i očiju se najčešće zajedno pojavljuju i može li se to izraziti na nekoj skali. CA analiza provedena je uz pomoć statističkog paketa R[4] korištenjem niže navedenih naredbi:

```
> library(ca)
> library(vcd)
> data(HairEyeColor)
> (HairEye <- margin.table(HairEyeColor, c(1,2)))
> res <- ca(HairEye)
> plot(res, main="Hair Color and Eye Color")
> title(xlab="Dim 1", ylab="Dim 2")
> summary(res)
```

s outputom:

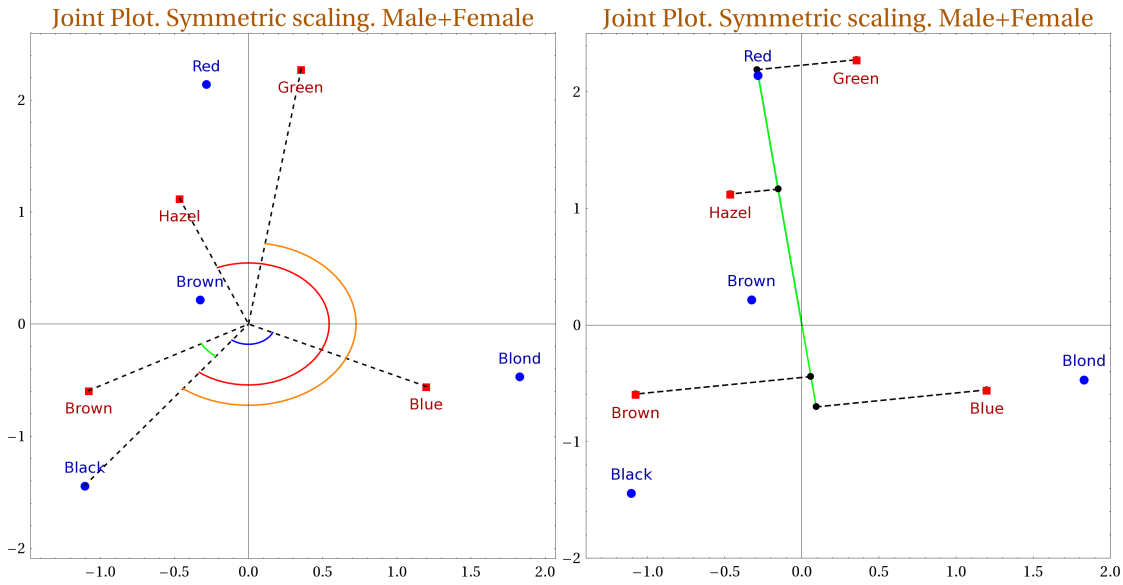
```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.208773  89.4  89.4  ************************
 2      0.022227   9.5  98.9  **
 3      0.002598   1.1 100.0
        -------- -----
 Total: 0.233598 100.0
```

Prva glavna komponenta objašnjava 89.4% ukupne varijance, a prve dvije komponente zajedno 98.9%. Glavne inercije (u stupcu `value`) su kvadrati singularnih vrijednosti i ujedno svojstvene vrijednosti od ZZ^τ.

3.5.1. ***Grafičko određivanje skale prioriteta.*** Kod grafičke interpretacije rezultata CA potreban je oprez. Na slici 2 dan je tzv. *biplot* gdje su x-os i y-os na slici prve dvije dimenzije i za retke i za stupce. To znači da su crvene i plave točke u različitim prostorima, reci u prostoru redaka, a stupci u prostoru stupaca. Međusobna daljenost crvenih (plavih) točkica ima smisla, dok odaljenost između crvene i plave točkice ***nema nikakvog smisla***.

[4] R je nekomercijalni statistički paket. Vidi http://www.r-project.org/.



SLIKA 2. Grafičko određivanje skale prioriteta. Kosa (reci)-plavo, oči (stupci)-crveno. Za crvenu bolju kose najčešće boje očiju su: `Green`, `Hazel`, `Brown`, `Blue`.

Međutim, pokazuje se da je smisleno za zadani redak (`Red`) promatrati kuteve koje tvore radijus vektori stupaca u odnosu na njegov radijus vektor. Što je kut manji to je korelacija[5] između tog retka i odabranog stupca veća. Nadalje, ako projiciramo stupce u biplotu na pravac određen radijus vektorom odabranog retka onda te projekcije definiraju skalu prioriteta koja se može normirati po potrebi.

Na slici 2 (desno) odabrana je boja kose `Red`. Boje očiju rangirane od više koreliranih prema manje koreliranim s tom bojom kose su: `Green`, `Hazel`, `Brown`, `Blue`. Na slici lijevo prikazano je to isto ali za boju kose `Black`, s tom razlikom što su naznačeni kutevi među vektorima, a ne projekcije. Koordinate redaka i stupaca dane su u tablici 5.

```
> res$rowcoord
           [,1]       [,2]
[1,] -1.1042772  1.4409170
[2,] -0.3244635 -0.2191109
[3,] -0.2834725 -2.1440145
[4,]  1.8282287  0.4667063
```

```
> res$colcoord
           [,1]       [,2]
[1,] -1.0771283  0.5924202
[2,]  1.1980612  0.5564193
[3,] -0.4652862 -1.1227826
[4,]  0.3540108 -2.2741218
```

TABLICA 5. Koordinate redaka (boja kose) i stupaca (boja očiju) u prve dvije dimenzije.

4. CANONICAL CORRESPONDENCE ANALYSIS (CCA)

CCA je prvi puta je prezentirana u Ter Braak (1986) [8] u ekološkom kontekstu. U njegovom primjeru bila su prisutne dva skupa podataka: podaci o pojavljivanju određenih

[5]To nije korelacija u klasičnom smislu riječi, ali se može opravdati naziv. Koristim tu riječ jer ovog trena ne znam bolju.

vrsta i njihovom obilju kao i podaci o vrijednostima određenog broja ekoloških varijabli čiji utjecaj na pojedine vrste se ispituje.

Prije daljnjeg čitanja ovog teksta možda je korisno pročitati primjer 4.3. Za razliku od CA ovdje se zadaju restrikcije na vrijednosti latentnih varijabli X i Y. One više nisu slobodne već su oblika $X = AU$ za retke i $Y = BV$ za stupce. Komponente od U predstavljaju učestalost kovarijabli koje sada imaju konkretnu interpretaciju. U primjeru 4.3 to su: *nemir, sumnja, poremećaj u razmišljanju* i *osjećaj krivnje*. Matrica A je tipa $n \times a$, a matrica B je tipa $m \times b$. Nadalje, pretpostavljamo da je A punog ranga po stupcima i da su vektori u_n i u_m u prostoru[6] stupaca od A odnosno B.

Kao i u prethodnom odjeljku traženje kritičnih točaka problema maksimizacije vodi na singularni problem

$$(A^\tau FB)V = (A^\tau DA)U\Sigma \tag{4.1}$$

$$(B^\tau F^\tau A)U = (B^\tau EB)V\Sigma. \tag{4.2}$$

Uvjeti standardizacije su

$$U^\tau A^\tau DAU = NI,$$
$$V^\tau B^\tau EBV = NI,$$

gdje je I jedinična matrica odgovarajuće dimenzije. Ako je $u_n = Ag$ i $u_m = Bh$, tada (g, h) određuje rješenje od (4.1) i (4.2) za $\sigma = 1$. Analogon matrice Z je sada

$$Z = (A^\tau DA)^{-\frac{1}{2}} A^\tau FB(B^\tau EB)^{-\frac{1}{2}}.$$

Ako je $Z = P\Sigma Q^\tau$ SVD dekompozicija od Z onda su

$$U = (A^\tau DA)^{-\frac{1}{2}} P, \quad V = (B^\tau EB)^{-\frac{1}{2}} Q$$

optimalna rješenja problema nalaženja maksimalne kovarijance, a odgovarajuće optimalne vrijednosti za bodove su

$$X = A(A^\tau DA)^{-\frac{1}{2}} PY, \quad Y = B(B^\tau EB)^{-\frac{1}{2}} Q.$$

X i Y su normirani i centrirani (osim za dominantnu singularnu vrijednosti $\sigma = 1$) i predstavljaju standardne koordinate koje još mogu biti reskalirane u svrhu bolje geometrijske prezentacije analize.

4.1. EKOLOGIJA. U ekološkim studijama F je matrica čiji elementi mjere pojavnost vrste j na lokaciji i. Prirodno je da na nekim lokacijama ima više (manje) jedinki pojedine vrste. Ako postoje razlozi za vjerovanje da je ta učestalost modelirana *okolinom* (environment-om) na tim lokacijama onda se svaka lokacija tipizira prisutnošću ili ne-prisutnošću nekih kovarijabli. U ekološkom kontekstu ta tipizacija se naziva *environmental gradient (scale)*. Vrijednosti 'bodova' lokacija na toj gradijentnoj skali za svaku lokaciju mogu se izračunati pomoću CA. Lokacije koje su bliske po bodovima bliske su i po biološkim resursima što se odražava na pojavnost određenih vrsti biljaka i/ili životinja.

CA daje nekoliko skala (dimenzija), za svaku singularnu vrijednost po jednu. Obično se singularne vrijednosti rangiraju po veličini, što odgovara postotku objašnjene varijance (raznolikosti) podataka. Najčešće se prve dvije dimenzije koriste za vizualizaciju rezultata analize (tzv. biplot). Jedan takav biplot prikazan je na sl. 4 desno.

[6]Procedura u R zapravo dodaje vektor u_n kao prvi vektor matrice. Time se poveća broj kovarijabli za jedan. U interpretaciji rješenja se to onda treba uvažiti.

4.2. ODREĐIVANJE ZNAČAJKI (FEATURE EXTRACTION). *Environmental gradijent* je informacijska značajka (u ekološkom problemu ordinacije) na kojoj je lako razlikovati objekte i grupirati ih po bliskosti. Ta informacijska značajka je apstraktna i ona je linearna kombinacija realnih značajki. U tom smjeru je objekte (vrste, lokacije) lakše razlikovati nego u bilo kojem drugom.

Bilo svjesno ili nesvjesno mi (ljudi) smo vrlo vješti u ekstrahiranju informacijskih značajki u svakodnevnom životu. Na primjer, nepogrešivo možemo identificirati spol osobe na daljinu i bez ispitivanja detaljnih podataka o toj osobi. To je zato jer prepoznajemo znakove koji sugeriraju spol: frizura, oblik tijela ili kombinacija jednog i drugog. Iako već dugo nismo vidjeli staru prijateljicu/prijatelja mi ćemo ju prepoznati uz neznatne poteškoće. Očito je da ne moramo biti u stanju procesirati svaku karakteristiku objekta kako bismo bili u stanju svrstati ga u određenu kategoriju.

Slobodno možemo zaključiti da:

- je potrebno mnogo varijabli za opisivanje složenog objekta;
- mi ne obrađujemo nužno sve te podatke kako bismo ga klasificirali;
- jedna varijabla nije dovoljna osim u slučaju kad se radi o ekstremno lakom zadatku;
- se čini kako naš zaključak baziramo na nekoliko *meta varijabli* koje su kombinacije osnovnih mjerenja (varijabli);
- je za složenije kategorizacije potrebno više meta varijabli.

Te meta varijable su upravo informacijske značajke. U statističkom učenju proces identificiranja tih meta varijabli poznat je kao *određivanje značajki*.

Postoje barem tri razloga zašto je određivanje značajki važan problem prediktivnog modeliranja i moderne analize podataka.

(1) ***Redukcija dimenzije***. Gotovo svi prediktivni modeli pate od predimenzioniranosti. Moćna kvaliteta određivanja značajki je redukcija dimenzije. Predimenzioniranost intuitivno shvaćamo kao višak irelevantnih podataka koji nisu bitni, što smanjuje efikasnost rasuđivanja. Isto tako, neke varijable mogu biti visoko korelirane u smislu da su jednako informativne. Redukcija dimenzije nas oslobađa viška takvih varijabli i olakšava daljnje numeričko procesiranje podataka.

(2) ***Automatsko rudarenje po podacima.*** U mnogim klasičnim primjerima informacijske značajke odabiru eksperti. Oni, iz vlastitog uvjerenja, biraju varijable za koje smatraju da su važne za model. Danas se sve više javlja potreba da to radi "crna kutija" koja je sama sposobna identificirati značajke. Dva su razloga za to: (1) jedan je potreba za obradom velikog broja podataka u kratkom vremenu i uz minimalnu manuelnu superviziju. (2) S druge strane, često nailazimo na podatke koje nismo u stanju razumjeti i modelirati. U tom slučaju jedini razumni pristup je dozvoliti podacima da govore sami za sebe.

(3) ***Vizualizacija.*** Ljudsko oko ima nevjerojatnu sposobnost prepoznavanja uzoraka dok s druge strane nismo u stanju 'osjetiti' podatke u prostoru većem od 3 dimenzije. Određivanjem vodećih dviju značajki moguće je podatke vizualizirati u ravnini i interpretirati njihov odnos uz pomoć elementarne geometrije.

Određivanje značajki nije samo po sebi cilj, ali je nužan korak u olakšavanju računa i izradni modela.

4.2.1. ***Human Brain Mapping Project (HBMP).*** Pojedini dijelovi mozga odgovorni su za obavljanje raznih zadataka. Isto tako neki zadatak može obavljati više dijelova mozga. Pomoću kanonske CA moguće je odrediti primarne (sekundarne, tercijarne,...) zadatke

za svaki dio mozga kao i primarne (sekundarne, tercijarne,...) dijelove mozga koji obavljaju isti zadatak. Analiziranjem i uspoređivanjem slika[7] aktivnosti mozga za vrijeme izvršavanja nekog zadatka (aktivno stanje) i za vrijeme odmora pokušavaju se identificirati najvažnije varijable (lokalizacija, pikselizacija) koje razlikuju aktivno stanje od kontrolnog stanja.

U ovom primjeru traži se model i jedan od načina za razlikovanje pojedinih aktivnosti mozga je određivanje meta varijabli i redukcija dimenzije. U najjednostavnijoj situaciji istraživač intuitivno doživljava kategorizaciju 'dijelova mozga' i 'zadatke' i sastavlja tablicu kontinencije koja povezuje te dvije kategorije. Nešto složeniji pristup bio bi da se dijelovi mozga i/ili zadaci kategoriziraju prema prisutnosti ili ne prisutnosti određenih kovarijabli kao što je to učinjeno u tablici 6 na primjer. Autoru ovog teksta nije poznato je li to netko učinio, a zanimljivo je i pitanje je li je regionalna klasifikacija mozga najprikladnija za mapiranje mentalnih funkcija. Osim ovog, također mi se čini zanimljivim projekt koji bi se nazivao *Human Consciousness Mapping Project* ako takav uopće postoji.

Možda je ovo pogodan trenutak za istaknuti heuristiku koja se uvukla u CA u obliku agregacije težina po formuli (3.3). Za neku drugu agregaciju analiza bi dala neke druge vrijednosti ali bi smisao dobivenih rezultata najvjerojatnije ostao isti (v. također odjeljak 5.8).

4.3. DIJAGNOSTIKA. PRIMJER.

4.3.1. ***Organizacija ulaznih podataka.*** Bolesnici koji boluju od *šizofrenije, manične depresije* i *poremećaja anksioznosti* podijeljeni su u 16 skupina ovisno o tome je li kod njih prisutan ili nije prisutan: *nemir, sumnja, poremećaj u razmišljanju* i *osjećaj krivnje*, v. tablicu 6.

Te četiri varijable su simptomi, nazivamo ih *kovarijable* (kofaktori, prediktori, dummy varijable). Te varijable igraju ulogu moderatora jer sugeriraju podjelu pacijenata po klasama. Kombinacije tih četiriju binarnih varijabli formiraju 16 različitih tipova simptoma i svaki tip odgovara jednom retku tablice. Ukupno je 620 pacijenata podijeljeno u 16 tipova simptoma i tri kriterijske grupe.

Pitanje: "Postoji li kakva ovisnost između tipova simptoma i pripadnosti pacijenta određenoj grupi i kako tu zavisnost prezentirati? "

4.3.2. ***Analiza.*** Desna strana u slici 3 predstavlja vektore bolesti (stupce) u prostoru generiranom s prve dvije principalne komponente (dimenzije). Svaki vektor "bježi" na svoju stranu što znači da stupci nisu u korelaciji. Slika na lijevoj strani pokazuje koordinate redaka u obje dimenzije. Crna linija prezentira promjene prve dimenzije ovisno o tipu simptoma, a crvena prezentira promjene druge dimenzije.

U prvoj dimenziji prepoznatljivo je cikličko ponašanje na prediktorima. y-vrijednosti za parove točaka 1-2, 3-4, 5-6, itd. se ne mijenjaju mnogo unutar tih parova. Među parovima su vidljive značajne razlike moderirane trećim prediktorom "thought disorders". Unutar parova razlike su slabo moderirane četvrtim prediktorom "delusions of guilt", a snažnije su moderirane prvim prediktorom "anxiety" jer druga polovica grafa (9–16) nastaje pomicanjem prve polovice grafa (1–8). Vidljiv je i utjecaj drugog prediktora "suspicion" jer se uzorak prve četvrtina grafa (1–4) ponavlja u cijelom grafu.

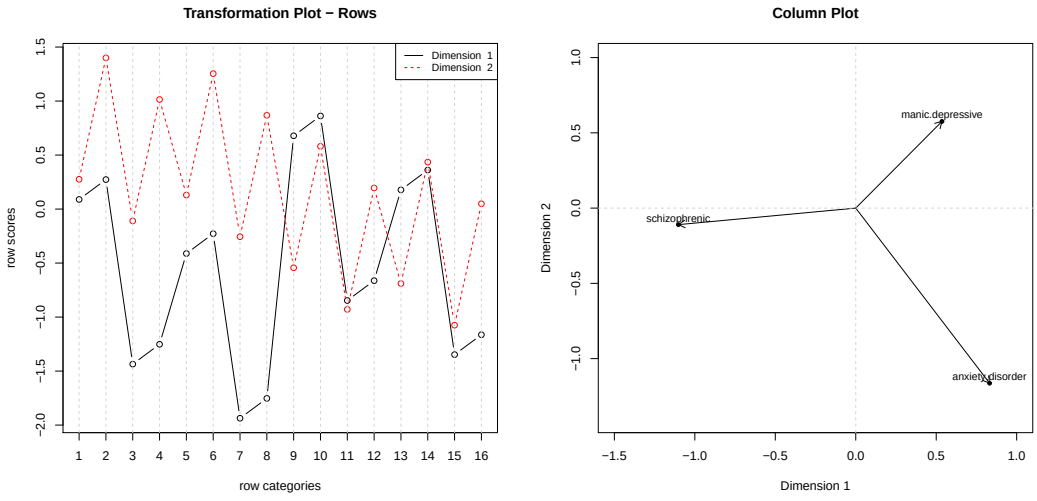
Druga dimenzija pokazuje alternirajuće ponašanje. Razlike unutar para moderirane su s "delusions of guilt", a lagani opadajući trend predstavlja utjecaj prvog prediktora "anxiety".

Gornji zaključak se još preciznije može vidjeti i iz singularnih vektora za vodeće dvije singularne vrijednosti (tablica 7). Primijetite da singularni vektori imaju 5 komponenti

[7]Dobivenih pomoću *Positron Emission Tomography* (PET) ili *Functional Magnetic Resonance Imaging* (fMRI) na primjer.

	Kovarijable					Bolesti		
Kat	anxiety	suspicion	thought disorders	delusions of guilt	Kat	schizophrenic	manic depressive	anxiety disorder
1	0	0	0	0	1	38	69	6
2	0	0	0	1	2	4	36	0
3	0	0	1	0	3	29	0	0
4	0	0	1	1	4	9	0	0
5	0	1	0	0	5	22	8	1
6	0	1	0	1	6	5	9	0
7	0	1	1	0	7	35	0	0
8	0	1	1	1	8	8	2	0
9	1	0	0	0	9	14	80	92
10	1	0	0	1	10	3	45	3
11	1	0	1	0	11	11	1	0
12	1	0	1	1	12	2	2	0
13	1	1	0	0	13	9	10	14
14	1	1	0	1	14	6	16	1
15	1	1	1	0	15	19	0	0
16	1	1	1	1	16	10	1	0

TABLICA 6. Kategorije tipova simptoma ovisno o vrijednostima kovarijabli (1–16) i broj bolesnika po parovima kategorija.



SLIKA 3. Grafička prezentacije analize podataka iz tablice 6. Desno – plot (stupci). Lijevo – Transformation plot za obje dimenzije (reci)

```
> res$left.singvec
            [,1]        [,2]
[1,]  0.01809171  0.12190282
[2,]  0.43853083 -0.58272541
[3,] -0.40369843 -0.07792225
[4,] -0.79800143 -0.19336085
[5,]  0.08709019  0.77595780
```

TABLICA 7. Lijevi singularni vektori za dvije vodeće singularne vrijednosti. Dominantna četvrta komponenta prve dimenzije odgovara trećoj kovarijabli "thought disorders". Prve dvije dimenzije objašnjavaju 65% i 87.2% ukupne varijance (kumulativno).

iako su u tablici 6 navedene 4 kovarijable. To je stoga jer je prije računanja SVD dekompozicije prostoru stupaca matrice kovarijabli dodan još i stupac u_n samih jedinica. Sada je vidljivo da su vodeće komponente prve dimenzije četvrta (-0.79800143), druga (0.43853083) i treća (-0.40369843) koje odgovaraju kovarijablama "thought disorders", "anxiety" i "suspicion", a vodeće komponente druge dimenzije su peta (0.77595780) i druga (-0.58272541) koje odgovaraju kovarijablama "delusions of guilt" i "anxiety".

Evo kôda pisanog u programskom jeziku R za analizu podataka iz tablice 6.

```
#! /usr/bin/Rscript
# maxwell.R
> library(anacor)
> data(maxwell)
> maxwell$table
> maxwell$row.covariates
> res <- anacor(maxwell$table,
  + row.covariates = maxwell$row.covariates,
  + scaling = c("Goodman","Goodman"));
> res
> res$row.scores
> res$col.scores
> res$singular.values
> res$right.singvec
> res$left.singvec
> plot(res, plot.type = "colplot", xlim = c(-1.5, 1),
  +arrows = TRUE, conf = NULL);
> plot(res, plot.type = "transplot", legpos = "topright")
> q();
```

4.4. CILJANI MARKETING. U tablici 8 dan je tipični scenario jedne marketinške studije. Lijeva tablica je tablica kontigencije. Desna tablica sadrži vrijednosti kovariabli za svakog kupca. Cilj analize je predvidjeti što će kupit novi kupac koji dolazi u dućan na temelju vrijednosti njegovih kovarijabli. U tu svrhu kupci se tipiziraju prema vrijednosti kovarijabli, a proizvodi se također mogu tipizirati prema nekim drugim kovarijablama koje nisu trenutačno prisutne u tablici.

Kupac	Igre	Vino	Cvijeće	God.	Dohodak	Spol
1	1	0	0	21	1T*	M
2	1	1	0	59	6.5T	F
3	0	1	1	31	4.5T	F

TABLICA 8. Ilustracija marketinške studije. Lijevo je matrica kontingencije. Desno su vrijednosti kovarijabli za svakog kupca koje mogu biti korištene u predviđanju buduće kupnje.

* T=tisuća kuna.

Primijetite da su `Igre` kupili i kupac 1 i kupac 2. Pažljiviji pogled na vrijednosti njihovih kovarijabli kaže da je tipični profil kupca 1 student sa skromnim prihodima, dok je tipični profil kupca 2 srednjovjekovna baka. Oni predstavljaju dvije različite sub-populacije zainteresirane za isti proizvod. Student je igricu vjerojatno kupio za sebe, a baka ju je najvjerojatnije kupila za svoju unučicu/unuka.

Drugi tip marketinške studije može biti da se potencijalni kupci zamole da rangiraju određen broj proizvoda prema nekim kriterijima (obično na skali 1–5). Nakon toga se sačini konsenzus tih individualnih preferencija i dobije se skala s bodovima proizvoda na toj skali. Udaljenosti na toj skali mogu se koristiti za grupiranje (klasterizaciju) proizvoda koja se grafički prezentira dendrogramom kao na slici 9. Takve su analize sofisticirane i praktički rađene u laboratorijskim uvjetima. Zamjerka takvim analizama je ta da kupac u trenutku kupovanja nekog proizvoda može mijenjati svoje preferencije koje se ne mogu laboratorijski kontrolirati. Studije prvog tipa zasnivaju se na realnim podacima sakupljenim u određenom vremenskom periodu na određenim lokacijama.

4.5. KLASIFIKACIJA MOŽDANOG UDARA. Jačina moždanog udara može biti jaka (`Sev`), umjerena (`Mod`) i slaba (`Mld`) ovisno o broju bodova na *NIHSS*[8] skali. S druge strane moždani udar može započeti glavoboljom (`Hdc`), epileptičkim napadajem (`Mld`), gubitkom svijesti (`Con`) ili nekim četvrtim načinom koji je klasificiran kao nepoznat (`Unk`). U tablici 9 dan je broj moždanih udara u svakoj od klasa moderiranih kovarijablama u desnoj tablici. Pitanje je ***može li se na temelju simptoma {Con, Hdc, Epi, Unk} klasificirati moždani udar kao {Mld, Mod, Sev} i može li se ta veza izraziti na nekoj skali***[9]?

Analiza koja slijedi napravljena je pomoću programskog paketa `R`. Funkcija `anacor` u tom programskom paketu vraća:

```
> res <- anacor(table, row.covariates = row.covariates,
  +  scaling = c("Benzecri","Benzecri"))

CA fit:
Sum of eigenvalues:  0.2096757
```

[8]The National Institutes of Health Stroke Scale (*NIHSS*) je metoda za utvrđivanje ozbiljnosti moždanog udara kod prijema bolesnika. Skala 0–42 (manje je bolje). Moždani udar smatra se blag (`Mld`) ako pacijent ima < 5 bodova na *NIHSS* ljestvici, umjeren (`Mod`) ako je broj bodova između 5 i 12 (uključivo i granice) i jak (`Sev`) ako ima > 13 bodova.

[9]Tu skalu ne treba miješati s *NIHSS* skalom koja mjeri (ne)autonomiju pacijenta na temelju subjektivne procjene liječnika

	Jačina MU				Row covariates			
	Mld	Mod	Sev		Con	Hdc	Epi	Unk
s2	50	78	57	s2	0	0	0	1
s3	1	4	4	s3	0	0	1	0
s5	29	26	4	s5	0	1	0	0
s7	1	0	0	s7	0	1	1	0
s9	0	5	19	s9	1	0	0	0
s11	0	0	4	s11	1	0	1	0

TABLICA 9. Klasifikacija moždanog udara moderirana načinom njegovog početka.

```
Total chi-square value: 59.129

Chi-Square decomposition:
             Chisq Proportion Cumulative Proportion
Component 1 56.038      0.932                 0.932
Component 2  3.090      0.051                 0.984

> res$col.scores[,1] # Scaled col scores
   Mld    Mod    Sev
-0.466 -0.160  0.635

> res$row.scores[,1] # Scaled row scores.
     s2      s3      s5      s7      s9     s11
 0.0046  0.3572 -0.5871 -0.4030  1.0800  1.2642
```

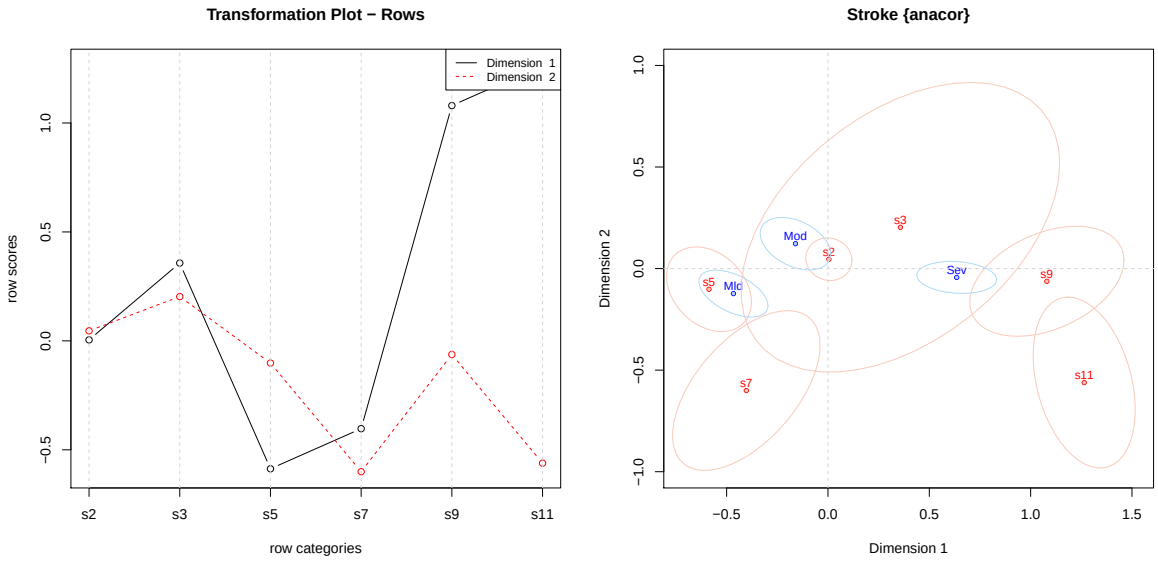
Vidljivo je da prva komponenta objašnjava 93.2% od ukupne varijance pa je redukcija dimenzije jako velika; reducirani prostor je jednodimenzionalan. Bodovna skala za jačinu

	`Mld`	`Mod`	`Sev`
x	-0.4664353	-0.1603499	0.6352364
$f(x)$	0	11.6692	42

TABLICA 10. Skala jačine moždanog udara i njena transformacija.

udara čuva se u varijabli `$col.scores` i ispisana je u retku x tablice 10. Transformiramo li tu skalu s $f(x) = 42 * (x + 0.4664353)/1.10167$, tada na transformiranoj skali umjereni moždani udar `Mod` ima vrijednost 11.6692, što grubo odgovara gornjoj granici za umjereni udar na *NIHSS* skali. To govori da cijela procedura nije besmislena jer je izvornu skalu (koja inače u praksi ne mora postojati) donekle sačuvala.

Iz slike 4 (lijevo) mogli bismo zaključiti sljedeće:



SLIKA 4. Transformation plot za tipove simptoma (retke) i jačine moždanog udara (stupce).

- trend rasta moderiran je varijablom `Con` (1) ;
- odnos susjednih parova moderiran je kovarijablom `Hdc` (2);
- odnos unutar parova 1-2, 3-4, 5-6 moderiran je kovarijablom `Epi` (3);

što je uglavnom u skladu s vrijednostima komponenti prvog lijevog singularnog vektora u tablici 11.

```
> res$left.singvec
            [,1]       [,2]
[1,]  0.07848346  0.2035717
[2,]  0.72025354 -0.6076077
[3,] -0.66417529 -0.7128864
[4,]  0.17452062 -0.2519706
[5,] -0.05906704  0.1329482
```

```
> res$right.singvec
            [,1]       [,2]
[1,] -0.5607791 -0.6311044
[2,] -0.2277015  0.7398938
[3,]  0.7960395 -0.2329473
```

TABLICA 11. Singularni vektori. Prva dimenzija objašnjava 93.2%, a druga 5.2% od ukupne varijance.

4.5.1. ***Zaključak. Skala jačine moždanog odara.*** Možda o svemu najbolje govori slika 4 (desno). Ako želimo saznati kako su rangirani tipovi simptoma (reci) u odnosu na `Sev` onda je odgovor: s11, s9, s3, s2, s7, s5. Želimo li to izraziti brojevnom skalom onda treba izračunati projekcije tih točaka na pravac koji prolazi ishodištem i točkom `Sev`, a to su koordinate tih točaka na prvoj dimenziji[10]:

4.6. CCA ANALIZA TESTIRANJA NA ALZHEIMEROVU BOLEST. Promatrajmo ponovno tablicu 1 u kojoj su dani rezultati testiranja 20 ispitanika pomoću 6 različitih tehnika. Po svojoj strukturi, tablica podsjeća na *table-covariable* zapis kod kanonske CA zbog binarnih

[10]U drugom retku tablice je skala u prvom retku afinom funkcijom transformirana u drugi redak.

Sev					
$s11$	$s9$	$s3$	$s2$	$s7$	$s5$
1.264	1.080	0.357	0.005	-0.403	-0.587
10	9	5	3	1	0

TABLICA 12. Skala jačine moždanog udara ovisno o njegovom početku.

vrijednosti u posljednja dva stupca. Neki ispitanici su zdravi, neki imaju MCI (Mild Cognitive Impairement) što je rizični faktor za Alzheimerovu bolest, a neki već boluju od te bolesti.

Raspon A-skale i B-skale je 0–30, obje IQ-skale imaju raspon 0–120, E i F su binarne varijable.

4.6.1. ***Tablica kontingencije.*** Kategorije stupaca prozvati ćemo: `No`, `Susp` i `Yes` (tablica 15) koje predstavljaju ispitanike za koje se vjeruje da nemaju AD, zatim one koji pokazuju izvjesne poremećaje i smatraju se suspektnima i ispitanike za koje se vjeruje da već boluju od AD.

Pogledajmo ponovno tablicu 13. Svaka varijabla (stupac) nudi vlastitu (trihotomnu) klasifikaciju kao što je navedeno u tablici 13. Na primjer ako ispitanik ima MSE vrijednost

	Kategorije stupaca		
Varijable	`No`	`Susp`	`Yes`
MSE	≥ 30	$26-29$	$0-25$
REY	≥ 23	$18-22$	$0-17$
IQPRF	≥ 135	$116-134$	$0-115$
IQVER	≥ 135	$116-134$	$0-115$

TABLICA 13. Kategorije stupaca za svaku varijablu.

iznad ili jednaku 30 pretpostavljamo da ne boluje od Alzheimerove bolesti i stavljamo ga u kategoriju `No` za tu varijablu. Ako je vrijednost MSE u intervalu $26 \leq$ MSE < 30 osobu kategoriziramo kao `Susp`, a ako je $0 \leq$ MSE < 26 tada pripada kategoriji `Yes`. Ove granice su ovdje dane samo da definiraju proceduru. Možda bi pažljiviji i argumentiran odabir granica dao još bolje rezultate.

Sljedeći korak je da se definiraju kategorije za sve varijable u razmatranju. Ako označimo s y (s, n) broj varijabli (za određenu osobu) koje ju karakteriziraju kao `Yes` (`Susp`,`No`), onda trojka (y, s, n) opisuje naklonost ispitanika ka svakoj od kategorija. Na primjer, ispitanik $A12$ ima MSE=26, REY=7, IQPRF=103, IQVER=93 što daje $(y, s, n) = (3, 0, 1)$. Sljedeći korak je odredit vrijednost v tripleta i na temelju te vrijednosti odrediti kategoriju. Izgleda razumno koristiti formulu $v = 2y + 1s + 0$. Ako je

if $v \geq 5 \rightarrow$ osoba spada u kategoriju `Yes`,
if $2 \leq v < 5 \rightarrow$ osoba spada u kategoriju `Susp`,

if $v < 2 \rightarrow$ osoba spada u kategoriju `No`.

Nije teško vidjeti da je ta klasifikacija ekvivalentna onoj u tablici 14. Primijenimo li na-

n^o y-ona	n^o s-sova	vrijednost v	kategorija
2 ili više	1 ili više	$v \geq 5$	`Yes`
2	0	$v = 4$	`Susp`
1	3	$v = 5$	`Yes`
1	2 ili manje	$3 \leq v < 5$	`Susp`
0	2 ili više	$2 \leq v \leq 4$	`Susp`
0	1 ili manje	$v \leq 1$	`No`
inače		$v < 2$	`No`

TABLICA 14. Klasifikacija stupaca.

vedena pravila klasifikacije na podatke iz tablice 1 dobije se tablica kontingencije 15 koju možemo analizirati pomoću kanonske CA.

F	`No`	`Susp`	`Yes`
`fr00`	0	0	9
`fr01`	1	5	1
`fr11`	4	0	0

kovarijable	1	2
`fr01`	0	1
`fr00`	0	0
`fr11`	1	1

TABLICA 15. Tablica kontingencije izvedena iz tablice 1 pomoću pravila definiranih u tablici 13 (ili tablici 14).

4.6.2. ***Analiza.*** Evo rezultata kanonske CA:

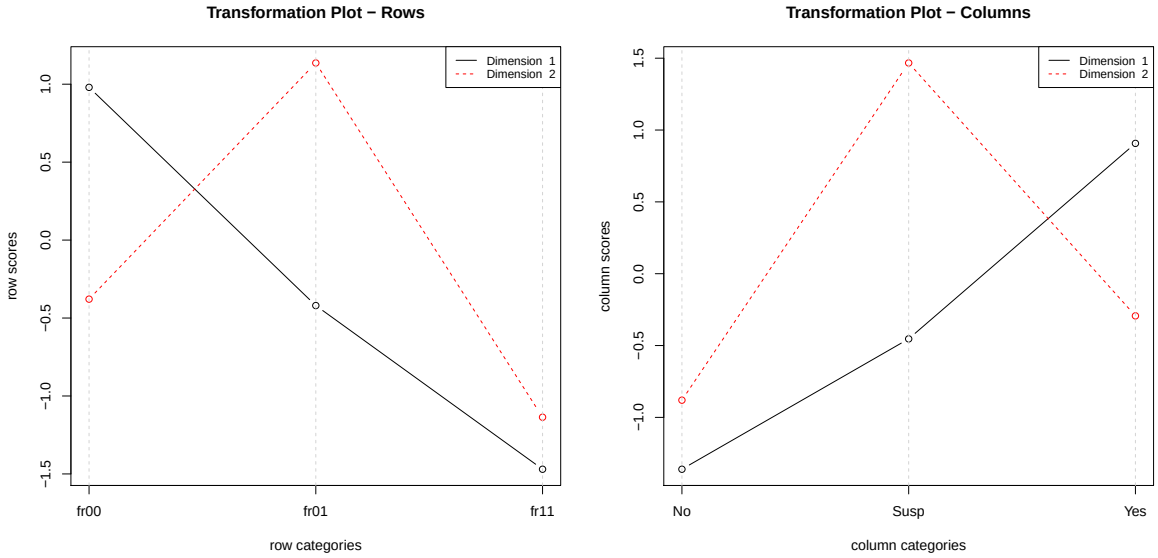
```
CA fit:
Sum of eigenvalues:  1.457143
Total chi-square value: 29.143

Chi-Square decomposition:
             Chisq Proportion Cumulative Proportion
Component 1 17.143      0.588                 0.588
Component 2 12.000      0.412                 1.000
```

Prve dvije dimenzije objašnjavaju 100% ukupne varijance, prva je dominantnija ali ne tako drastično kako to bude u sličnim analizama. Njihova relativna važnost je približno 6:4. Kako bi se to interpretiralo ne znam. Pogledamo li komponente lijevih singularnih vektora za obje dimenzije onda je vidljivo da je u prvoj dimenziji dominantnija komponenta `fr01`, a u drugoj `fr11` i približno su jednake. To bi to moglo navoditi na zaključak da je snaga prediktabilnosti tih varijabli u navedenom odnosu.

```
> res$left.singvec
           [,1]       [,2]
[1,] -0.4522351 -0.2673024
[2,]  0.7758004 -0.5885662
[3,]  0.4400196  0.7629805
```

TABLICA 16. Lijevi singularni vektori. U prvoj dimenziji dominantnija je komponenta `fr01`, a u drugoj `fr11`.



SLIKA 5. Transformation plot za retke i stupce. Uočite naglašenu simetriju u grafovima.

Na slici 5 nacrtani su grafovi bodova za retke (lijevo) i stupce (desno). Prisutna je naglašena simetrija što znači podudarnost klasa: `fr00` ≡ `Yes`, `fr01` ≡ `Susp`,`fr11` ≡ `No`. To je bilo i za očekivati zbog specifičnog izgleda tablice kontingencije 15.

4.6.3. ***Zaključak.*** Zaključak analize je da je klasifikacija ispitanika u kategorije {`No`, `Susp`, `Yes`} ***potpuno određena kombinacijama kovarijabli*** {`fr00`, `fr01`,`fr11`}. Drugim riječima ***dvije varijable*** FRQ ***i*** RARE ***su dovoljne*** za kategorizaciju ispitanika određenu s prve četiri varijable.

FRQ	RARE	Kategorija
0	0	`Yes`
0	1	`Susp`
1	1	`No`

TABLICA 17. Kategorizacija stupaca objašnjena pomoću varijabli FRQ i RARE.

4.7. KLASIFIKACIJA ZADATAKA ZA JAVNI ISPIT (DRŽAVNA MATURA). Cilj niže opisane procedure je na temelju povijesnih podatatka (zadataka sa već provedenih ispita) odrediti težinu zadaće koja se predlaže za ispit kao i broj bodova za svaki zadatak.

Ovo je samo ideja procedure koja je u fazi nastajanja, a bazirana je jednoj ideji od Ebela (1972). U njegovom problemu određuje se minimalni prag prolaznosti za svaki zadatak na način da se zadaci klasificiraju prema kategorijama *Težina* i *Relevantnost* u tablicu kontingencije (v. tablicu 18). Nakon toga, iskusan ispravljač zadataka određuje postotak učenika koji bi trebali riješiti zadatak za svaki par kategorija (ξ, η). Nakon toga se radi elementarna analiza tablice da bi se postigao traženi cilj.

Mi ćemo Ebelovu ideju nadograditi tako da ćemo definirati kovarijable za kategorizaciju *Težine* i primijeniti kanonsku CA. Kovarijable su bazični elementi koji svojom prisutnošću mijenjaju vrijednost težine svojom prisutnošću. To su u najopćenitijem slučaju viši i niži kognitivni procesi prisutni kod rješavanja zadatka.

> Napomena. *Ovdje provedena analiza služi samo za ilustraciju kako bi se mogao (trebao) zahvaćati problem u njegovoj složenosti. Podaci u tablici 18 nabacani su po osjećaju autora ovog teksta i ne predstavljaju rezultate nekog sustavnog istraživanja. Isto se odnosi i na kovarijable i njihovu vezu s kategorijama težine u tablici 19.*

4.7.1. ***Formulacija problema.*** Osnovni problem je kako odrediti težinu zadatka. Izgleda razumno to učiniti tako da se analizira struktura zadatka i pogleda što on sve zahtijeva od učenika. Te komponente su kovarijable koje mogu i ne moraju biti prisutne u svakom zadatku[11]. Te kovarijable mogu biti: *sposobnost računanja, uvježbanost, poznavanje teorije, sposobnost modeliranja...*. Ovisno o zastupljenosti kovarijabli zadatak može biti: *ekstra lagan, lagan, umjeren, težak, vrlo težak, ekstra težak* (v. tablicu 19).

	Relevantnost			
Težina zadatka	esencijalna	važna	prihvatljiva	upitna
ekstra lak	13	10	9	3
lak	10	7	5	2
umjer. težak	8	6	3	1
težak	7	4	2	1
vrlo težak	5	2	1	1
ekstra težak	4	1	1	1

TABLICA 18. Tablica kontingencije za par kategorija *(težina, relevantnost)*.

4.7.2. ***Analiza.*** Analiza je provedena pomoću naredbe `res<-anacor()`. Ispis (tablica 20) pokazuje da prve dvije glavne komponente objašnjavaju 92.4% od ukupne varijance.

Komponente singularnih vektora prve dvije dimenzije dane su na slici 6 (lijevo), a desno je projekcija podataka na prve dvije dimenzije. Prva dimenzija može poslužiti kao skala varijable *Težina*. Argumenti koji govore tome u prilog su relativno visok udio prve dimen-

[11] Tu bi se sad moglo diskutirati o postotku zastupljenosti svake od kovarijabli, no ostavimo to za razrađeniji model.

	Kovarijable (Vještine)			
Težina zadatka	Računanje	Uvježbanost	Teorija	Modeliranje
ekstra lak	0	0	0	1
lak	1	1	0	0
umjeren	1	0	1	0
težak	0	1	1	0
vrlo težak	0	1	1	1
ekstra težak	1	1	1	1

TABLICA 19. Kategorizacija težine zadataka pomoću kovarijabli.

```
CA fit:
Chi-Square decomposition:
            Chisq Proportion
Component 1 2.338      0.639
Component 2 1.046      0.286
Component 3 0.057      0.015
       Cumulative Proportion
Component 1            0.639
Component 2            0.924
Component 3            0.940
```

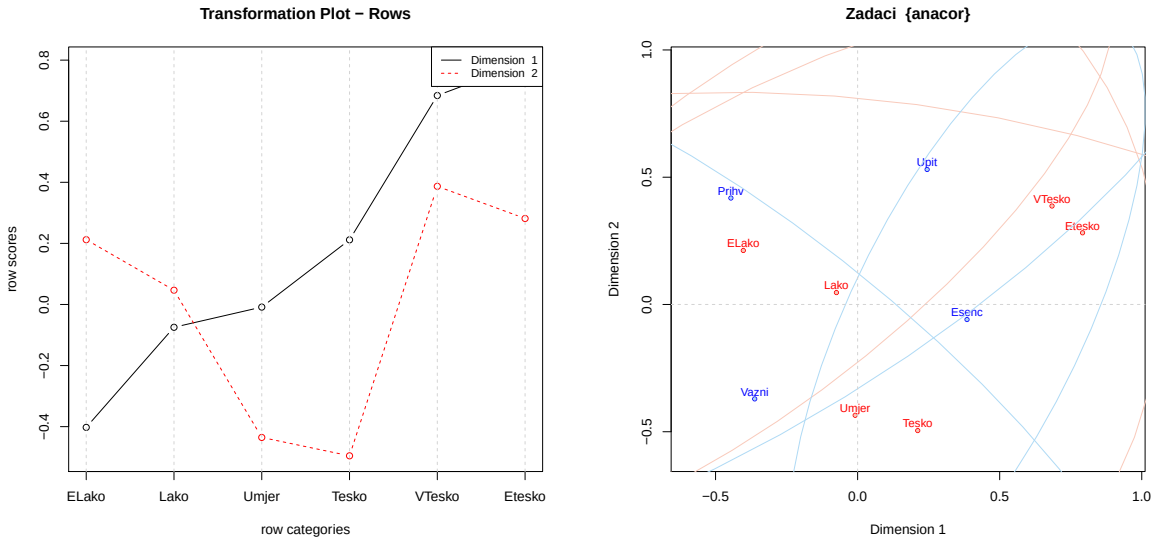
TABLICA 20. Zastupljenost komponenti u postocima.

```
> res$row.scores
                  D1          D2
ELako  -0.402383745  0.21213597
Lako   -0.074668484  0.04670467
Umjer  -0.008594139 -0.43529929
Tesko   0.211608298 -0.49547506
VTesko  0.684170180  0.38684198
Etesko  0.791683003  0.28158644
```

TABLICA 21. Vrijednosti bodova kategorijske varijable *Težina* (*row scores*) u prve dvije dimenzije. Prva dimenzija može poslužiti za kreiranje bodovne skale težine zadatka ovisno o utjecaju kovarijabli.

zije u ukupnoj varijanci (63.9%) i 'dobar' rast komponenata prve dimenzije što se vidi sa slike.

4.7.3. ***Određivanje skale težine.*** Skalu težine zadatka odredit ćemo iz stupca D1 u tablici 21. Pretpostavimo da želimo ekstra lakom zadatku dati 4 boda, a ekstra teškom 20



SLIKA 6.

```
> res$left.singvec
           [,1]        [,2]
[1,] -0.2529809  0.11505569
[2,] -0.2913914  0.08592649
[3,]  0.4688258 -0.03067557
[4,]  0.5471963 -0.65506198
[5,]  0.5760819  0.74116910
```

TABLICA 22. Komponente lijevog singularnog vektora ne pokazuju dominaciju neke od kovarijabli. Možemo samo reći da prisutnost teorije i modeliranja u zadatku povećavaju njegovu težinu više nego ostale kovarijable.

bodova. Funkcija koja transformira prvu dimenziju D1 na novu skalu je

$$F(x) = 4 + \frac{x - 4}{1.19407}$$

gdje nazivnik predstavlja razliku između maksimalne i minimalne vrijednosti komponenti u stupcu D1. Nova skala je dana u tablici 23.

Kategorijska vrijednost	ELako	Lako	Umjer	Tesko	VTesko	Etesko
težina	4	8.39125	9.27662	12.2272	18.5594	20

TABLICA 23. Bodovna skala za težinu zadatka.

4.7.4. ***Određivanje težine zadaće.*** U određivanju težine zadaće trebalo bi uvažiti utjecaj svake od kovarijabli na svaki zadatak. U principu bi kovarijable trebale biti jednako zastupljene ako se želi provjeravati cjelokupno gradivo s naglasakom na neke od kovarijabli ako se radi o specifičnim testovima.

..... nastavak slijedi

5. METODE VREDNOVANJA (RANGIRANJA)

5.1. MALO POVIJESTI. Začetak teorije vrednovanja seže u davnu 1785 godinu kad je Marquis de Condorcet kritizirao Bordinu metodu sume rangova za izbor u Francusku akademiju. Sredinom prošlog stoljeća problematika izbora i vrednovanja razvija se kao matematička disciplina pod nazivom *Decision Making*. U to vrijeme javlja se potreba za nalaženjem 'pravednih' zahtjeva (aksioma) koje bi svaka metoda izbora trebala zadovoljavati. Poznati američki ekonomist Kenneth Joseph Arrow dobio je i Nobelovu nagradu za svoj *Impossibility theorem* poznat još pod nazivom *Arrovljev paradoks*. Teorem je bio uzrok tadašnjeg razočarenja u zapadnu demokraciju jer je interpretiran kao nemogućnost postojanja 'demokratske' metode izbora jer svi razumni zahtjevi izbora 'vode u diktaturu'.

Jednim od najljepših pristupa u klasičnoj teoriji odlučivanja smatra se Savage-ovo poopćenje von Neumann-Morgensternovog teorema iz 50-tih godina. Njihova *teorija korisnosti* moderirana je ekonomskim pricipima i smatra se 'vrhuncem racionalog odlučivanja'. Rezultat ekonomskog presinga i na teoriju rezultirao je nesretnim homo economicusom koji trenutačno još uvijek pokušava 'spasiti' teoriju putem nove znanstvene discipline *Neuroekonomije*.

Već od samog začetka teorije korisnosti javljaju se i njeni protivnici. Jedni od najistaknutijih su bihejvioristi Daniel Kahnemann i Amos Tverski koji su za svoja istraživanja dobili i Nobelovu nagradu iz ekonomije 2002. godine. Dobar dio današnjih istraživanja još uvijek pokušava revitalizirati teoriju korisnosti jer je jednostavna i za razumijevanje i u primjeni.

Zadatak vrednovanja u najapstraktnijoj formi je da zadani skup objekata preslika u skup realnih brojeva (skaliranje, bodovanje, rangiranje). U praksi, svaki takav objekt ima određene kvalitete koje se mogu direktno ili indirektno mjeriti ili se objekti mogu uspoređivati u parovima (u odnosu na tu kvalitetu). Umjesto objekata mogu se promatrati i buduće akcije ili scenariji koji vode određenim ciljevima. Glavna poteškoća leži u agregaciji tih individualnih skala u zajedničku zbog suprotstavljenih interesa (ciljeva).

U nastavku iskazujemo dva teorema, jedan koji govori o *ekstenzivnom mjerenju* i drugi koji govori o mjerenju *razlike preferencija*. Iskazujemo ih sa svrhom da se vidi ozbiljnost i poteškoće problematike vezane uz mjerenje.

5.2. EKSTENZIVNO MJERENJE. Mjerenje u fizici i tehnici (tzv. *ekstenzivno mjerenje*) također funkcionira na uspoređivanju u parovima samo što to vrednovanje ima dodatne zahtjeve. Evo jednog teorema koji govori o tome:

Ekstenzivno ili opsežno mjerenje zahtijeva skup objekata koje mjerimo S, binarnu relaciju R i binarnu operaciju $\circ$ (konkatenacija) na S. Uređenu trojku $(S, R, \circ)$ nazivamo *relacijskim sustavom*. Osnovni problem je naći uvjete (nužne i dovoljne) za egzistenciju ordinalne funkcije vrijednosti V koja zadovoljava

$$aRb \iff V(a) \geq V(b) \tag{5.1}$$

$$V(a \circ b) = V(a) + V(b). \tag{5.2}$$

Takvu funkciju nazivamo *ekstenzivnom funkcijom vrijednosti* ili *ekstenzivnom mjerom*. Jedan od teorema koji govori o egzistenciji ekstenzivne mjere je:

Teorem 5.1 (Roberts & Luce (1968))**.** *Neka je $(S, R, \circ)$ relacijski sustav. Tada postoji funkcija vrijednosti V na S koja zadovoljava* (5.1) *i* (5.2) *ako i samo ako $(S, R, \circ)$ zadovoljava sljedeće aksiome*[12]*:*

[12]Sustav $(S, R, \circ)$ koji zadovoljava aksiome E1–E3 i E4 nazivamo ***ekstenzivnom strukturom***.

E1 (Slaba asocijativnost) *Za svaki* $a, b, c \in S$,

$$[a \circ (b \circ c)] \sim [(a \circ b) \circ c]$$

E2 (Slabi uređaj) (S, R) *je asimetrična i negativno tranzitivna.*

E3 (Monotonost) *Za svaki* $a, b, c \in S$,

$$aRb \iff (a \circ c)R(b \circ c) \iff (c \circ a)R(c \circ b).$$

E4 (Arhimedovost) *Za svaki* $a, b, c, d \in S$

$$aRb \implies \exists n \in \mathbb{N}\,((na \circ c)R(nb \circ d)).$$

Ekstenzivno mjerenje nije primjenjivo u psihologiji, subjektivnim procjenama ili kod analize podataka koji su jednim dijelom nedostupni ili ne postoje (missing data). Mjerenja u psihologiji su priča za sebe jer specijalni zahtjevi na skalu vode na zanimljive funkcionalne jednadžbe.

5.3. MJERENJE RAZLIKE PREFERENCIJA. Za razliku od ekstenzivnog mjerenja u kojem je dominirala algebarska operacija $\circ$ (konkatenacija) na skupu objekata, mjerenje razlike preferencija je 'svakodnevnije' u smislu da oponaša čovjekovo vrednovanje u svakodnevnom životu. Takav tip mjerenja je 'uzor' dobrom dijelu današnje programske podrške od koje izdvajamo *Expert Choice* (u pozadini je Saatyjeva metoda svojstvenog vektora) i *metodu potencijala* (Čaklović-Šego, `http://decision.math.hr`).

Osnačimo s $(S,\ \succcurlyeq\ )$ relaciju slabe preferencije na S. Na prvi pogled izgleda razumno reći da donositelj odluke preferira a u odnosu na b više nego nego što preferira c u odnosu na d ako i samo ako je više spreman odustati od b u zamjenu za a nego odustati od d u zamjenu za c. Tada govorimo o *zamjeni*, u oznaci $(a \leftarrow b)$. Relaciju slabe preferencije na zamjenama označimo s $\succcurlyeq_e$. Postavlja se pitanje egzistencije takve funkcije vrijednosti v na skupu objekata S koja zadovoljava:

$$a \succcurlyeq b \Leftrightarrow v(a) \geq v(b) \tag{5.3}$$

$$(a \leftarrow b) \succcurlyeq_e (c \leftarrow d) \Leftrightarrow v(a) - v(b) \geq v(c) - v(d). \tag{5.4}$$

Takvu funkciju, ako postoji, nazivamo ćemo *izmjerivom funkcijom vrijednosti*. Očito je v ordinalna funkcija vrijednosti, dok (5.4) zahtijeva da je razlika $v(a)-v(b)$ ordinalna funkcija vrijednosti na skupu zamjena usklađena s relacijom $\succcurlyeq_e$ na zamjenama.

Mnogi teoretičari smatraju da ne treba ljude prisiljavati na zamišljanje zamjena jer da oni imaju urođeni osjećaj za intenzitet preferencije. Zamišljamo li relaciju preferencije kao usmjeren graf onda je intenzitet preferencije broj koji je pridružen svakoj strelici u grafu. Takva prezentacija slabe preferencije u mnogome olakšava donositelju odluke da se koncentrira na parove i odmah svakom paru (a, b) pridruži intenzitet preferencije po svom osjećaju. Na kraju krajeva nije bitno kako ljudi zamišljaju zamjene, bitno je da to rade na konzistentan način.

Sljedeći teorem govori o egzistenciji izmjerive funkcije.

Teorem 5.2. *Sljedeći aksiomi A1–A6 (niže navedeni) su nužni za egzistenciju izmjerive funkcije vrijednosti* v*. Nadalje, aksiomi A1–A4 i A6 su i dovoljni.*

A1. (Slabi uređaj) Relacija $\succcurlyeq$ je relacija slabe preferencije na skupu objekata, a relacija $\succcurlyeq_e$ je relacija slabe preferencije na skupu zamjena.

A2. (Usklađenost $\succcurlyeq$ i $\succcurlyeq_e$) $\forall a, b \in S$

$$a \succcurlyeq b \Leftrightarrow (a \leftarrow b) \succcurlyeq_e (c \leftarrow c), \quad \forall c \in S.$$

A3. (Inverzija) $\forall a, b, c, d \in S$

$$(a \leftarrow b) \succcurlyeq_e (c \leftarrow d) \Leftrightarrow (d \leftarrow c) \succcurlyeq_e (b \leftarrow a).$$

A4. (Konkatenacija) $\forall a, b, c, d, e, f$

$$\left.\begin{array}{l}(a \leftarrow b) \succcurlyeq_e (d \leftarrow e)\\ (b \leftarrow c) \succcurlyeq_e (e \leftarrow f)\end{array}\right\} \implies (a \leftarrow c) \succcurlyeq_e (d \leftarrow f).$$

A5. (Rješivost) $(\forall b, c, d \in S)$ $(\exists a \in S)$ tako da je

(a) $$(a \leftarrow b) \sim_e (c \leftarrow d).$$

$(\forall b, c \in S)$ $(\exists a \in S)$ tako da je

(b) $$(b \leftarrow a) \sim_e (a \leftarrow c).$$

A6. (Arhimedovost) Svaki strogo omeđeni standardni niz je konačan.

Sljedeći teorem pokazuje da jedinstvenost izmjerive funkcije do na pozitivnu afinu transformaciju zahtijeva izvjesno bogatstvo skupa S.

Teorem 5.3. *Pretpostavimo da su ispunjeni zahtjevi rješivosti u formi aksioma 5.2.5(a) i 5.2.5(b). Tada je izmjeriva funkcija vrijednosti jedinstvena do na pozitivnu afinu transformaciju. Drugim riječima, ako su v i w dvije izmjerive funkcije vrijednosti onda postoje realni brojevi $\alpha > 0$ i β tako da je*

$$w(a) = \alpha v(a) + \beta, \quad \forall a \in S.$$

Izmjerivu funkciju vrijednosti v čija je pozitivna afina transformacija opet izmjeriva nazivamo *intervalnom skalom*.

5.4. LINEARNA FUNKCIJA VRIJEDNOSTI. Navesti ćemo jedan teorem koji govori o egzistenciji linearne funkcije vrijednosti na skupu predstavljenom uređenim n-torkama.

Teorem 5.4. *Neka je $\succcurlyeq$ slabi uređaj na $\mathbb{R}^n$. Tada postoji netrivijalna linearna funkcija vrijednosti na $\mathbb{R}^n$ koja je u skladu s $\succcurlyeq$ ako i samo ako vrijede sljedeća dva aksioma:*

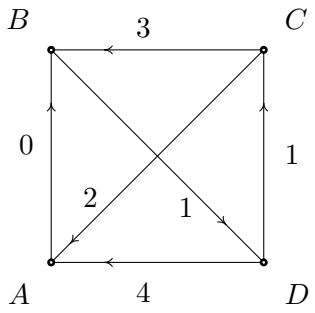
A1 (konstantan relativni trade-off) *postoji konstantan relativni trade-off među komponentama;*

A2 (monotonost) *postoji $a \in \mathbb{R}^n$ takav da je $a \succ 0$ i za svaki $b \in \mathbb{R}^n$ i svaki $\lambda > 0$ je*

$$b + \lambda a \succ b.$$

5.5. METODA POTENCIJALA. Metoda potencijala bazirana je na uspoređivanju u parovima. To znači da donositelj odluke (DO) mora biti u stanju za svaki par objekata koje uspoređuje jednom od njih dati prednost. Nadalje, toj prednosti mora dati i *intenzitet prednosti* mjeren na nekoj skali u njegovom mentalnom prostoru (subjektivno vrednovanje). Govoreno matematičkim rječnikom DO zadaje relaciju preferencije R ili usmjerni graf (*graf preferencija*) s težinama čiji su čvorovi zadani objekti, a lukovi su orijentirani od manje preferiranog čvora prema jače preferiranom (ako su uspoređivani). Intenzitet preferencije nazivamo *tokom*, u oznaci $\mathcal{F}$, po uzoru na strujni tok u električnoj mreži.

Za subjektivne preferencije uvriježeno je davati intenzitet u skupu $\{0, 1, 2, 3, 4\}$ što odgovara 'verbalnom intenzitetu': *jednaka*, *slaba*, *jaka*, *izrazita* i *apsolutna* preferencija. Jedan takav subjektivni graf preferencija dan je na slici 7. Cilj vrednovanja je iz zadanog toka preferencije rekonstruirati težine čvorova. Preciznije:



SLIKA 7. Graf preferencije. Brojevi uz lukove (strelice) predstavljaju intenzitet preferencije.

Problem. Za zadani graf preferencije i tok $\mathcal{F}$ na skupu lukova traži se funkcija X (*potencijal*) na skupu vrhova grafa tako da vrijedi

$$\mathcal{F}_\alpha \geq 0 \iff X(a) \geq X(b) \tag{5.5}$$

$$\mathcal{F}_\alpha \geq \mathcal{F}_\beta \iff X(a) - X(b) \geq X(c) - X(d) \tag{5.6}$$

gdje su $\alpha = (b, a)$ i $\beta = (c, d)$ lukovi.

Zahtjevi (5.5) i (5.6) nisu ništa drugo nego drugačije zapisani zahtjevi (5.3) i (5.4) tj. da X bude izmjeriva funkcija vrijednosti. Nije teško vidjeti da takva funkcija postoji ako je $\mathcal{F}$ konzistentan tok u smislu da je suma komponenata toka duž svakog ciklusa jednaka nuli, odnosno

$$y^\tau \mathcal{F} = 0$$

za svaki ciklus y. Ekvivalentno tome je da je $\mathcal{F}$ u prostoru razapetom stupcima matrice incidencije.

Efektivno računanje potencijala za nekonzistentne tokove svodi se na rješavanje jednadžbe

$$A^\tau A X = A^\tau \mathcal{F}, \tag{5.7}$$

gdje je A matrica incidencije grafa preferencije. U slučaju potpunog grafa rješenje je dano formulom

$$X_i = \frac{1}{n} \sum_{j=1}^{n} F_{ij}, \tag{5.8}$$

gdje je i indeks čvora, a F_{ij} vrijednost toka na luku (i, j)[13]. Suma s desne strane u (5.7) predstavlja razliku između ulaznog i izlaznog toka u i-tom čvoru. Za graf preferencije na slici 7 je:

$$X(A) = 6/4,\ X(B) = 2/4,\ X(C) = -4/4,\ X(D) = -4/4.$$

5.5.1. ***Condorcetov tok.*** Tipičan primjer toka preferencije je Condorcetov tok koji uvažava pravilo većine. Zamislimo da imamo n glasačkih listića i na svakom listiću je dana rang lista kandidata. Za dva kandidata a i b definira se $N(a, b)$ kao broj listića u kojima a stoji ispred b umanjen za broj listića u kojima b stoji ispred a. Ako je $N(a, b) > 0$ onda luk u grafu preferencije usmjerimo tako da izlazi iz b i ulazi u a. Ako je $N(a, b) = 0$ orijentacija

[13]Ovdje je i ulazni, a j izlazni čvor. Oznaka F_{ij} ako (i, j) nije luk shvaća se kao $-F_{ji}$.

luka je proizvoljna. Condorcet nije znao kako računati 'potencijal' pogotovo u slučaju ako je u grafu preferencije prisutan ciklus.

5.5.2. ***Mjera inkonzistentnosti.*** Tok $\mathcal{F}$ smatramo *konzistentnim* ako vrijedi

$$AX = \mathcal{F}$$

za neki $X \in \mathbb{R}^n$. U slučaju da $\mathcal{F}$ nije konzistentan tada kao mjeru inkonzistentnosti možemo uzeti sam kut $\deg(\mathcal{F})$ između toka $\mathcal{F}$ i njegove aproksimacije AX ili sinus tog kuta

$$\mu(\mathcal{F}) = \frac{\|AX - \mathcal{F}\|}{\|\mathcal{F}\|}. \tag{5.9}$$

gdje je X potencijal od $\mathcal{F}$ i $\|\cdot\|$ euklidska norma. Očito je

$$\mu(\mathcal{F}) = 0 \iff \mathcal{F} \text{ je konzistentan.}$$

Što je kut $\deg(\mathcal{F})$ veći to znači da donositelj odluke generira više netranzitivnih ciklusa.

Vrijedi i obrnuto, svaki ciklus koji se pojavljuje u grafu i duž kojeg suma tokova nije jednaka nuli pridonosi povećanju inkonzistentnosti jer daje jednu netrivijalnu komponentu od $\mathcal{F}$ koja je okomita na prostor $R(A)$ generiran stupcima matrice incidencije, a to povećava kut. U daljnjem ćemo za mjeru inkonzistentnosti uzimati stupanj tj. $\mu(\mathcal{F}) = \deg(\mathcal{F})$.

Korisno je definirati neku gornju ogradu prihvatljivosti toka μ^* za tu mjeru, tako da se tok za koji je $\mu(\mathcal{F}) < \mu^*$ smatra prihvatljivim. Na donositelju odluke je da sam procijeni želi li prihvatiti rangiranje određeno potencijalom od $\mathcal{F}$. Vrijednosti μ^* ovise o broju čvorova u grafu i dane su u tablici 24.

4	5	6	7	8	9	10	11	12	13	14	15
5.3	9.9	12.7	14.7	16.1	17.2	18.0	18.8	19.3	19.8	20.2	20.6

TABLICA 24. Vrijednosti gornje granice dopustive inkonzistentnosti μ^* u ovisnosti o broju čvorova n (Još neobjavljeni rezultati).

5.6. EFIKASNOST OPORAVKA OD MOŽDANOG UDARA. U tablici 25 dani su osnovni statistički parametri o četiri mjerene varijable za svakog pacijenta koji je bolnički liječen od moždanog udara: *BarthelPrijeCVI*[14], *NIHSS*, *BarthelOtpust*, *Rankin*[15]

5.6.1. ***Definicija efikasnosti oporavka.*** *Efikasnost oporavka* definira se kao razlika 'indeksa sposobnosti' u vrijeme izlaska iz bolnice i 'indeksa sposobnosti' kod dolaska u bolnicu. Što je ta razlika veća pacijent(+tretman) ima veću sposobnost oporavka. Od gornjih varijabli dvije su 'input' varijable (kod dolaska u bolnicu), a dvije su 'output varijable (kod izlaska iz bolnice). Agregacijom inputa i outputa dobit ćemo 'indeks sposobnosti'. Poteškoća je u tome što su sve varijable ordinalne, a agregacija ordinalnih varijabli je kontekstualno ovisna. Najčešći operatori ordinalne agregacije su: pravilo većine, median (Kemeny), Sugeno integral, ordinalna sredina s težinama i suma rangova (Borda). Ovdje smo testirali Bordinu metodu, Kemenyjev median i pravilo većine kombinirano s metodom potencijala. Ova treća metoda čini se najprikladnijom jer nabolje radzvaja preživjele od preminulih pacijenata.

[14]Barthel indeks je procjenska skala sposobnosti pacijenta nakon moždanog udara a dizajnirana je na procjeni stupnja neovisnosti bolesnika, različitosti u samopomoći te procjena uobičajene dnevne aktivnosti. Skala 0–100 (veće je bolje).

[15]Rankin skala je skala za mjerenje stupnja invalidnosti ili ovisnosti u svakodnevnim situacijama ljudi koji su pretpjeli moždani udar. Skala 0–6 (manje je bolje).

5.6.2. ***Tok efikasnosti.*** Za svakog pacijenta a i za svaku ordinalnu varijablu gleda se broj pacijenata koje on dominira, zatim broj pacijenata koji njega dominiraju i napravi se razlika ta dva broja (pravilo većine). To je razlika ulaznog i izlaznog toka za tu varijablu na pacijentu a. Nakon toga se zbroje razlike ulaznog i izlaznog toka za svaku varijablu. Dobiveni broj je desna strana u jednadžbi (5.7). Ako je graf potpun, tj. svi čvorovi su susjedni, onda koristimo formulu (5.8) za račun potencijala. Time smo izbjegli efektivno računanje toka. Ako smatramo da varijable nisu ravnopravne onda im možemo pridjeliti neke težine. Varijable u igri su:

$$(-\mathrm{BI}_0, \mathrm{NIHSS}, \mathrm{BI}_1, -\mathrm{R})$$

(BI_0 – Barthel indeks pri dolasku u bolnicu, BI_1 – pri odlasku), gdje smo stavili znak $-$ (minus) ispred nekih zbog suprotnih orijentacija skala kod BI i NIHSS,R. Isto tako treba uvažiti koja je input, a koja je output varijabla.

Kako odrediti težine varijabli nije apriori jasno. Ovdje je uzeta vrijednost $w = (1, 2, 3, 2)$ iz razloga kojeg ćemo kasnije objasniti. Metodom potencijala pacijenti su rangirani i dobivena je efikasnost (`Eff`) njegovog oporavka.

5.6.3. ***Informacijske ćelije.*** Jačina moždanog udara može biti jaka (`Sev`), umjerena (`Mod`) i slaba (`Mld`) ovisno o broju bodova na *NIHSS* skali. S druge strane moždani udar počinje glavoboljom (`Hdc`), epilepričkim napadajem (`Epi`), gubitkom svijesti (`Con`) ili nekim četvrtim načinom koji je klasificiran kao nepoznat (`Unk`).

Cilj ove studije je ustanoviti vezu između načina početka moždanog udara, njegove jačine i efikasnosti oporavka. Sekundarno pitanje je može li se efikasnost oporavka predvidjeti i pomoću vrijednosti varijabli kao što su: `Dob`, `LDL`, `HDL`,`L`, `GUK`. U tu svrhu promatraju se pacijenti grupirani u *informacijske ćelije* oblika: `Sev+Con`, `Sev+Con+D`, ... gdje `D` znači da je pacient preminuo. Za svaku takvu ćeliju izračunata je veličina ćelije te srednje vrijednosti za `Eff`, `Dob`, `LDL`, `HDL`,`L` i `GUK`. Izračunate vrijednosti dane su u tablici 26. Pri tome su ćelije s manje od 10 pacijenata izbačene s popisa.

5.6.4. ***Zaključak.*** Čini se da je efikasnost oporavka dobro definirana. Jedini kriterij za odabir težina u toku efikasnosti je taj da funkcija dobro razdvaja informacijske ćelije u kojima su preminuli pacijenti od ostalih. Također je testirana i Bordina metoda sume rangova za ordinalnu agragaciju i neke ad-hock aproksimativne metode za Kemenyjev medijan ali s daleko lošijim rezultatima.

Zahvaljujući vrijednostima svake informacijske ćelije na skali `Eff` moguće je računati njihovu udaljenost i klasterizaciju pomoću klasičnih metoda za klasterizaciju. Na slici 9 dan je dendrogram klasterizacije pomoću Wardove metode. Na slici se jasno vidi da je klaster preminulih pacijenata na pristojnoj udaljenosti od ostalih klastera. Druge vrijednosti za vektor težina w daje približno 'dobru' klasterizaciju kao što je ova na slici 9 dok

Varijabla	min	max	mean	median	varij.	# mjerenja
BarthelPrijeCVI	0.	100.	90.65	100.	454.75	392
NIHSS	0.	42.	10.58	8.	60.24	795
BarthelOtpust	0.	100.	52.84	50.	1489.56	374
Rankin	0.	6.	3.24	4.	3.28	380

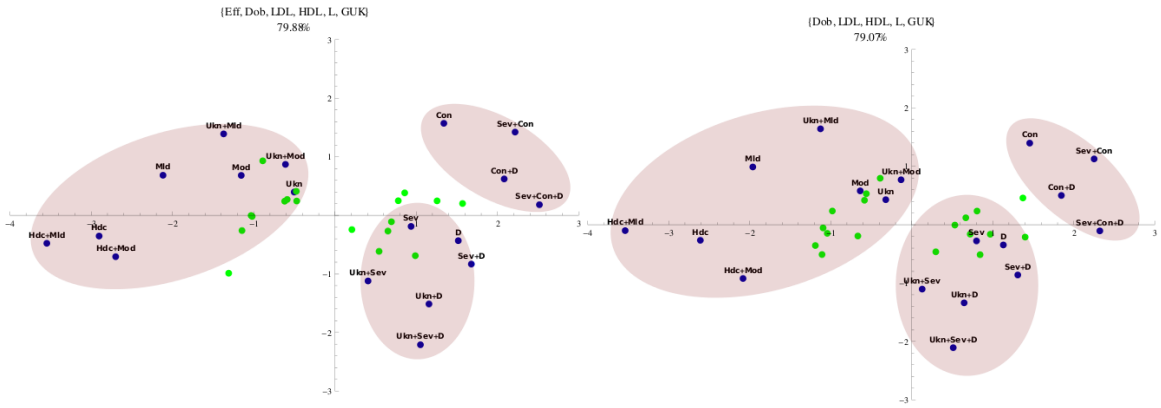
TABLICA 25. Osnovni statistički podaci o varijablama.

	Size	Eff	Dob	LDL	HDL	L	GUK
Hdc+Mod	26.	6.68	78.08	3.63	1.35	9.65	7.24
Hdc	60.	5.35	76.2	3.57	1.34	9.18	7.04
Mod	113.	5.17	79.16	3.37	1.27	9.21	7.67
Ukn+Mod	78.	4.68	79.58	3.35	1.26	9.21	7.95
Hdc+Mld	30.	4.55	74.87	3.57	1.36	8.55	6.72
Mld	81.	4.17	76.04	3.47	1.28	8.48	7.41
Ukn+Mld	50.	3.99	76.52	3.37	1.24	8.49	7.69
Ukn	185.	3.51	79.84	3.32	1.27	9.12	7.9
Con	28.	1.91	79.57	2.9	1.26	10.28	8.11
Sev	88.	1.53	81.85	3.12	1.31	9.94	8.21
Ukn+Sev	57.	1.47	83.12	3.26	1.33	9.55	8.03
Sev+Con	23.	1.11	79.87	2.87	1.27	10.94	8.55
Ukn+Sev+D	18.	0.36	84.5	3.58	1.31	9.43	9.04
Sev+D	32.	0.34	83.03	3.33	1.29	10.05	8.92
Sev+Con+D	13.	0.32	81.38	3.04	1.29	11.05	9.02
D	41.	0.29	83.32	3.23	1.27	9.52	8.61
Con+D	15.	0.29	80.33	3.	1.28	10.51	8.77
Ukn+D	25.	0.28	85.32	3.39	1.29	8.98	8.65

TABLICA 26. Srednje vrijednosti ćelija za svaku kontinuiranu varijablu. Ćelije su rangirane po stupcu Eff (efikasnost). Ćelije s manje od 10 pacijenata su izbačene s popisa. Separacija živih i preminulih pacijenata (koji se nalaze na dnu tablice) je vrlo dobra.

klasterizacija ćelija na temelju svih podataka iz tablice 26 nije tako dobra, što se vidi na slici 8, jer se u pojedinim klasterima nalaze i živi i preminuli pacijenti. To znači da se na temelju vrijednosti parametara kao što su Dob, LDL, HDL, L i GuK ***ne može odrediti*** efikasnost oporavka.

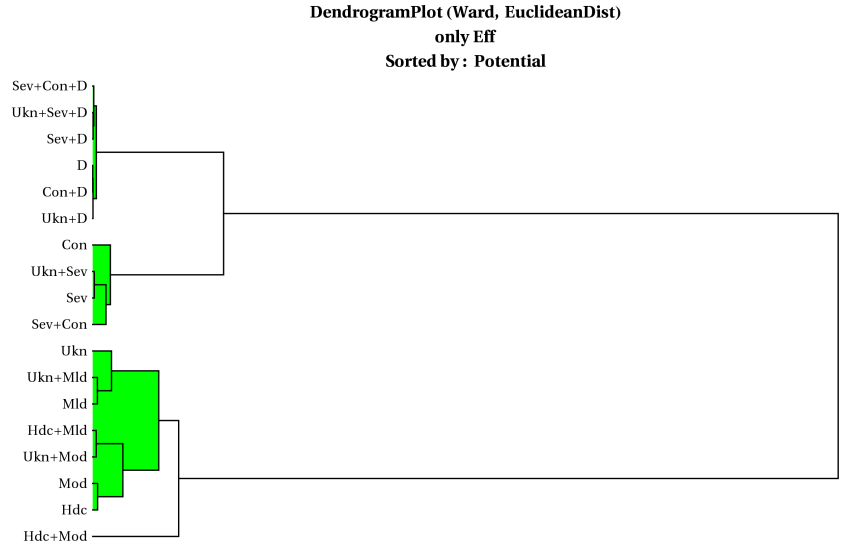
5.7. ALZHEIMER I METODA POTENCIJALA. Tablicu 1 možemo shvatiti na dva načina: (1) Kao rezultate mjerenja ispitanika pomoću raznih testova (stupci) ili (2) kao rezultate baždarenja mjernih instrumenata (testova) pomoću određenog broja ispitanika (reci). Ta je dualnost prisutna u svakom mjerenju. Ako su mjerni instrumenti pouzdani onda bi stupci koji mjere istu kvalitetu trebali biti ekvivalentni (ekvivalenciju određuje klasa dozvoljenih transformacija skale). Ako su ispitanici istih karakteristika onda bi reci trebali biti jednaki. Najčešći je slučaj da nije istina niti jedno niti drugo.



SLIKA 8. PC plot s varijablom `Eff` (lijevo) i bez nje (desno). Prisutna su 3 klastera u kojima se miješaju preživjeli i preminuli. Plave točke predstavljaju projekcije na Π_{12}, a zelene na ravninu Π_{34}. Prve dvije komponente objašnjavaju približno 80% ukupne varijance.

5.7.1. ***Rangiranje redaka.*** Svaka varijabla u svom stupcu već daje jedno rangiranje redaka. Potrebno je samo napraviti konsenzus ili agregaciju tih rangiranja po svim varijablama (stupcima). Ako težine stupaca nisu zadane težine (ovdje je to slučaj) onda je jedna od pretpostavki za agregaciju da su svi stupci ravnopravni (slučaj očajnog analitičara). Konsenzus po stupcima uz tu pretpostavku dano je u tablici 27 (zadnji stupac tablice).

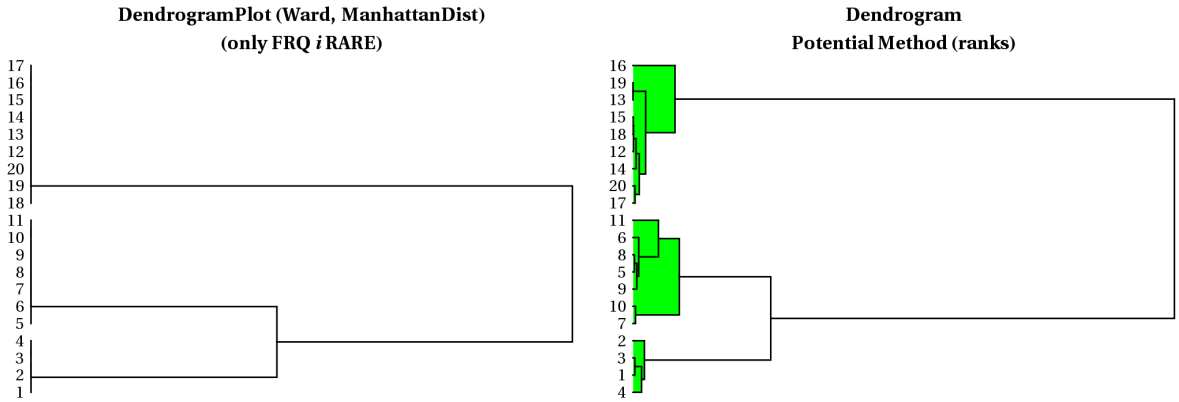
Klasterizacija na slici 10 desno je nastala tako da je matrica udaljenosti među ispitanicima računata na skali u posljednjem stupcu. Slika lijevo je posljedica klasterizacije sirovih podataka samo na temelju posljednje dvije varijable. Provedene su i klasterizacije u kojima je matrica sličnosti dana pomoću korelacije redaka i u kojima je matrica udaljenosti



SLIKA 9. Dendrogram klasterizacije ćelija (Ward) koristeći samo varijablu `Eff`. Udaljenost je izračunata na skali dobivenoj metodom potencijala. Može se primijetiti da su svi preminulu pacijenti u istom klasteru.

PM rang			
A4	0.0665125	A11	0.0502274
A1	0.0650859	A13	0.0411906
A3	0.0646674	A19	0.0411906
A2	0.0634226	A14	0.0408363
A7	0.0584799	A15	0.0402741
A10	0.0578927	A18	0.040264
A9	0.0549763	A12	0.0401621
A5	0.0543741	A20	0.0396988
A8	0.0541417	A17	0.0391073
A6	0.0535714	A16	0.0339243

TABLICA 27. Rezultat agregacije po svim varijablama. Rezultat ne ovisi o reskaliranju stupaca.



SLIKA 10. Klasterizacija nakon agregacije po stupcima (desno) i sirovih podataka samo na temelju posljednje dvije varijable (lijevo).

računata pomoću neke od standardnih metrika (Euklidska, Manhattan). Sve klasterizacije daju iste rezultate.

5.7.2. ***Zaključak.*** Posljednje dvije varijable su dovoljne za klasifikaciju ispitanika u tri klastera. Koji od njih predstavlja zdrave, suspektne i bolesne nije moguće odgovoriti bez dubljeg uvida u podatke. To mora reći čovjek.

5.8. SAMORANGIRANJE. Rezultati ovog odjeljka su ekspermentalnog karaktera i bazirani su na još neobjavljenom radu [9]. Problemi u realnom svijetu koji vode na samorangiranje su svi oni u kojima postoje povratne veze. To su na primjer: grupa donositelja odluke koji sami sebe rangiraju, naše individualne odluke koje uvažavaju emocije ili komunikacija između inteligentnih sustava koji izmjenjuju informacije i korigiraju sami sebe na bazi te razmjene.

Za ilustraciju samorangiranja mogu poslužiti podaci iz tablice 1. Ako tablicu shvatimo kao tablicu odlučivanja onda retke (ispitanike) možemo rangirati ako znademo težine pojedinih varijabli (stanja svijeta) u prvom retku. To je i bilo učinjeno u odjelju 5.7.1 na način da su sve varijable bile ravnopravne. U većini realnih situacija donositelji odluke daju ad-hock težine i opravdavaju ih na razne načine ili postupe na način kao što smo i mi to učinili pa svim stanjima svijeta daju jednake težine.

U mnogim metodama autori više uvažavaju kriterije koji jače razdvajaju alternative (na nekoj skali), a manje vrednuju one kriterije koji imaju 'slabu razlučivost'. Time priznaju da postoji povratna veza između izmjerenih rezultata i težina varijabli. Samorangiranje je samo formalizacija tog procesa u iterativnom postupku u kojem varijable rangiraju alternative na temelju unaprijed određenih početnih težina, zatim alternative rangiraju varijable na temelju istih podataka uvažavajući izračunate težine i taj se proces iterativno ponavlja.

Može se dokazati da taj proces konvergira uz vrlo slabe pretpostavke na podatke i konvergencija je to brža što je broj varijabli (ili alternativa) veći. Štoviše, rezultat samorangiranja ne ovisi o početno izabranim težinama. U tablici 28 dani su rezultati samorangiranja

Ispitanici			
A4	0.075771	A11	0.049385
A1	0.073283	A14	0.036314
A3	0.072608	A13	0.036272
A2	0.070711	A19	0.036272
A7	0.062489	A15	0.035488
A10	0.061642	A12	0.035317
A9	0.057015	A18	0.035261
A8	0.056010	A20	0.034586
A5	0.055941	A17	0.033396
A6	0.055061	A16	0.027178

Varijable	
C	0.239549
D	0.237776
A	0.143151
B	0.134254
E	0.122766
F	0.122504

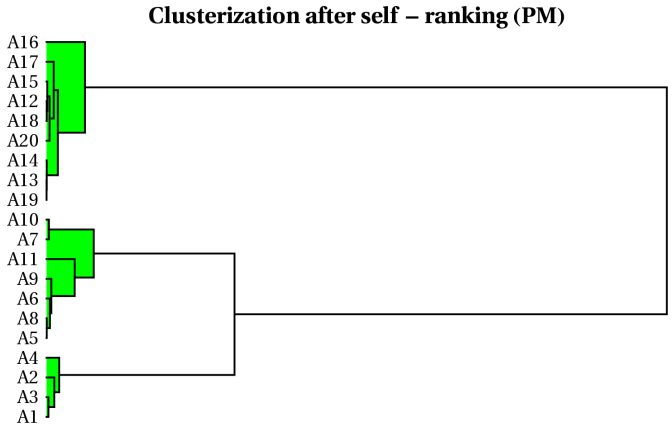
TABLICA 28. Rezultati samorangiranja ispitanika i varijabli na temelju podataka iz tablice 1. Usporedite s tablicom 27.

i ispitanika i varijabli. Rangiranja se donekle razlikuju pri dnu skale ali je dendrogram klasterizacije isti kao i dendrogram na slici 10. Sama procedura agregacije je ovdje nešto malo izmjenjena imajući u vidu da su testovi (varijable) *C* i *D* verbalnog karaktera i stoga neprecizni[16].

6. O STATISTIČKOM PAKETU R

`R` (`http://www.r-project.org/`) je programski jezik i okruženje za statističke procedure i grafiku. To je jedan od GNU projekata, sličan je programskom jeziku `S` koji je komercijalan i razvijen je u *Bell Laboratories*. `R` posjeduje razne statističke testove, analizu

[16]Radi se o tome da je parametar FN za varijable *A, B, E, F* postavljen na vrijednost 2, a za ostale varijable na 1. Detalji i obrazloženje parametra FN mogu se naći u Čaklović-Šego (2002) [10].



SLIKA 11. Dendrogram dobiven mjerenjem udaljenosti na skali dobivenoj samorangiranjem (tablica 28).

vremenskih serija, klasifikaciju, klasterizaciju i grafiku za predočavanje dobivenih rezultata. Za razliku od `S`, `R` je 'open source' projekt. Može se koristiti na većini UNIX platformi i sličnim sustavima uključujući FreeBSD i Linux, Windows i MacOS.

Jedna od boljih referenci za `R` [4]. Odlične reference za korištenje paketa `anacor` su [2] i [7].

LITERATURA

1. Jean-Paul Benzécri, *L'Analyse des donnees: I & II. La Taxonomie*, Dunod, Paris, 1973.
2. Jan de Leeuw and Patrick Mair, *Simple and Canonical Correspondence Analysis Using the R Package* `anacor`, **31** (2009), 1–18.
3. Zlatko Drmač, *On a curious property of the singular values of some F*, Personal communications (Drmač–Čaklović), 2011.
4. Brian S. Everitt and Torsten Hothorn, *A Handbook of Statistical Analyses Using R*.
5. M. Greenacre, *Theory and applications of correspondence analysis*, Academic Press, London, 1984.
6. _____, *Correspondence analysis in practice*, 2 ed., Chapman & Hall/CRC, Boca Raton, FL, 2007.
7. O. Nenadić and M. Greenacre, *Correspondence analysis in R, with two- and three- dimensional Graphics: The* `ca package`., Journal of Statistical Software **20** (2006), no. 3, 1–13.
8. Cajo J. F. Ter Braak, *Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis*, Ecology **67** (1986), no. 5, 1167–1179.
9. Lavoslav Čaklović, *Conflict Resolution. Risk-As-Feelings Hypothesis*, Labsi Working Papers (2011), no. 35, 1–16, (http://www.labsi.org/wp/labsi35.pdf).
10. Lavoslav Čaklović and Vedran Šego, *Potential Method applied on exact data*, Proceedings of KOI 2002 (Kristina Šorić, Tihomir Hunjak, and Rudolf Scitovski, eds.), Croational Operational Research Society, 2002, pp. 237–248.