Psihološka istraživanja, Vol. XVII (2) 2014.
179–190

UDK 303.71
159.9.072.43
TIP ČLANKA: Originalni naučni rad

Heuristički metod za detekciju grupa ispitanika s različitim faktorskim strukturama skupa merenih varijabli

Aleksandar Zorić[1]
Filozofski fakultet, Univerzitet u Beogradu

U radu je predložen heuristički metod za identifikaciju grupa u okviru ispitivanog uzorka, koje se međusobno razlikuju po faktorskoj strukturi varijabli na kojima je uzorak opisan. Dok klasične metode analize grupisanja, zasnivane na minimiziranju distance unutar grupa a maksimiziranju distance između grupa, tu distancu definišu u jedinstvenom prostoru varijabli na kojima je uzorak opisan, predložena metoda pokušava da uzorak podeli u grupe koje leže u različitim prostorima, ne nužno iste dimenzionalnosti. Dakle, osnovni cilj predloženog metoda grupisanja jeste pronalaženje kvalitativno različitih grupa ispitanika. Metod omogućuje nov uvid u analizirane podatke, ne proizvodeći nužno rešenje koje pojednostavljuje strukturu dobijenih podataka.

Ključne reči: analiza grupisanja, faktorska analiza, algoritam, kvalitativne razlike.

Uvod

Analiza grupisanja je statistička metoda koja se koristi za razvrstavanje uzorka entiteta opisanih na nekom skupu obeležja u određeni broj disjunktnih grupa, i to tako da entiteti u istoj grupi budu međusobno što sličniji, a istovremeno što različitiji od entiteta iz drugih grupa. To je dobro poznata postavka već sada klasičnog problema (videti u Legendre & Legendre, 1998), koji se u svojoj osnovi može svesti na definisanje mere distance (ili mere sličnosti), kako između dva entiteta, tako i između dve grupe entiteta. Ukratko, te iterativne procedure bi u prvom koraku izračunale udaljenost između svih parova entiteta, združivši ona dva između kojih je distanca najmanja. U sledećem koraku bi se ponovo izračunale distance, ovog puta između svih pojedinačnih entiteta, ali i svih pojedinačnih entiteta i (dvočlane) grupe koja je dobijena u prvom koraku, te bi se napravila nova grupa od entiteta s mi-

1 azoric@f.bg.ac.rs

nimalnom distancom. Taj algoritam bio bi nastavljen sve dok u poslednjem koraku svi entiteti ne bi bili združeni u jednu grupu. Takav pristup naziva se aglomerativnim jer se zasniva na ukrupnjavanju grupa. Na istraživaču je da prekine iteriranje u nekom trenutku i da dobijene grupe u tom koraku proglasi optimalnim rešenjem. Naravno, optimalnost u ovom slučaju nije ekstremum neke matematičke funkcije, već pre procena istraživača o tome koliko je dobijena podela korisna i interpretativna.

Opisani algoritam zahteva računanje i pamćenje distanci među svim entitetima, što u slučaju velikih uzoraka može predstavljati problem: broj distanci između n elemenata jeste kvadratna funkcija $n(n–1)/2$, tako da već između hiljadu elemenata broj distanci iznosi nešto manje od pola miliona. Da bi se razrešio ovaj problem, predložena je heuristička metoda koja se uobičajno zove k –aritmetičkih sredina (engl. k-means) (videti u Bock, 2007) (u programskom paketu SPSS, ovaj algoritam se naziva „Quick Cluster"– brzo grupisanje"). U prvom koraku se od istraživača zahteva da specifikuje željeni broj grupa (broj k, koji se spominje u nazivu metode), te zatim algoritam na slučaj uzima k tačaka iz prostora varijabli i svaki entitet pridružuje tački kojoj je najbliži i na taj način formira grupu. U sledećem koraku se tačke ponovo odrede, ali ovog puta ne na osnovu slučaja, već na osnovu određene mere centralne tendencije za grupu entiteta koji su pridruženi toj tački, te se entiteti ponovo razvrstaju na osnovu opisane procedure. Algoritam se ponavlja, iterira, sve dok nijedan entitet više ne menja tačku kojoj je najbliži, tj. grupu, ili dok se ne izvrši određeni broj iteracija.

Jasno je da je ovaj algoritam mnogo manje zahtevan, ne pamte se distance između svih entiteta, već se za svaki entitet izračuna udaljenost od izabranih k tačaka i on se pridruži njemu najbližoj tački u prostoru. Problem sa ovim algoritmom jeste taj što nije jednoznačno određen, već krajnji rezultat zavisi od koordinata početnih tačaka koje su na slučaj određene u prvom koraku.

U oba navedena klasična pristupa grupisanju entiteta distance se računaju uz pretpostavku da svi entiteti leže u jednom zajedničkom prostoru. Generalno se distanca između entiteta p i q, u oznaci d_{pq}, može predstaviti sledećom formulom:

$$d_{pq} = \left(\sum_{i=1}^{m} \left(x_{ip} - x_{iq} \right)^n \right)^{1/n}$$

gde x_{ip} označava vrednost ispitanika p na i-toj od ukupno m varijabli. U slučaju kada je $n = 2$, reč je o Euklidskoj distanci. Ono što je važno primetiti jeste to da se svi entiteti nalaze u jednom prostoru koji razapinje m varijabli na kojima je opisan analizirani uzorak.

S druge strane, priroda psiholoških problema akcenat stavlja na latentnu strukturu direktno izmerenih varijabli. Uobičajena je praksu da se faktorska

struktura najčešće analizira na celokupnim uzorcima, čak i u slučajevima kada se opravdano može posumnjati u homogenost uzorka u pogledu faktorske strukture analiziranih varijabli. Metoda predložena u ovom radu može se upotrebiti upravo za otkrivanje grupa entiteta koje imaju različite faktorske strukture. Samim tim, različite faktorske strukture označavaju i različite prostore u kojima se nalaze entiteti koji su predmet grupisanja. A to, samim tim, dovodi u pitanje i opravdanost računanja distanci, tj. kvantifikacije udaljenosti kada su dobijene razlike u tom slučaju prevashodno kvalitativne.

Najčešće korišćena metoda faktorisanja manifestnih varijabli jeste, zasigurno, analiza glavnih komponenti.

Neka matrica **Z** predstavlja rezultate merenja *n* ispitanika na *m* linearno nezavisnih varijabli i neka su rezultati transformisani u standardnu normalnu metriku, tako da je:

$$diag\left(\mathbf{Z}^t\mathbf{Z}\right)=\mathbf{I}$$

iz čega sledi da se matrica interkorelacija **R** dobija kao jednostavan produkt moment matrice podataka **Z**.

Metoda glavnih komponenata pronalazi vektor **x** koji sadrži koeficijente linearne kombinacije kolona iz **Z,** tako da novodobijeni vektor, glavna komponenta, ima maksimalnu varijansu. Ovaj metod tako dekomponuje matricu podataka u međusobno ortogonalne komponente, od kojih svaka ekstremizira varijansu koja je ostala nakon ekstrakcije prethodnih. To je, u stvari, ekvivalentno dekompoziciji matrice interkorelacija na svojstvene vrednosti i svojstvene vektore

$$\mathbf{X}^t\mathbf{R}\mathbf{X}=\mathbf{\Lambda}$$

pri čemu se u kolonama matrice **X** nalaze svojstveni vektori, a na dijagonali dijagonalne marice **Λ** svojstvene vrednosti. Vektori u matrici **X** predstavljaju koeficijente sklopa izvornih varijabli za odgovarajuću glavnu komponentu, dok svojstvene vrednosti predstavljaju varijanse glavnih komponenata.

Kako ova metoda algebarski daje onoliko komponenata koliko ima izvornih varijabli, u cilju smanjenja prostora potrebno je odrediti broj značajnih, važnih komponenata, tj. komponenata koje ne predstavljaju samo varijansu greške merenja.

Algoritam

Predloženi heuristički algoritam pokušava da razvrsta ispitanika u grupe tako da suma objašnjene varijanse zadržanih komponenti u svim grupama bude maksimum. Algoritam se svodi na šest koraka:

1. U prvoj iteraciji ispitancii se na slučaj raspodele u k grupa, tj. matrica $\mathbf{Z}$ se particioniše u k segmenata. Obeležimo sa $\mathbf{Z}_i$ i-ti segment, tj. podmatricu za i-tu grupu ispitanika.
2. U svakoj grupi se izdvoji r glavnih komponenata, tj. odrede se njihovi svojstveni vektori. Obeležimo sa $\mathbf{X}_p$, r svojstvenih vektora dobijenih na i-tom segmentu matrice $\mathbf{Z}$.
3. Za svakog ispitanika se izračuna njegova udaljenost od prostora koji razapinju glavne komponente dobijene na svakom od segmenata. Obeležimo grešku predviđanja j-tog segmenta na osnovu svojstvenih vektora dobijenih na i-tom segmentu sa

$$\mathbf{E}_{j.i} = \mathbf{Z}_j - \mathbf{P}_i\mathbf{Z}_j$$

gde je sa $\mathbf{P}_i$ označen projektor prostora koji razapinju komponente i-tog segmenta, a sa $\mathbf{R}_j$ matrica interkorelacija j-tog segmenta. Projektor $\mathbf{P}_i$ je definisan kao:

$$\begin{aligned}\mathbf{P}_i &= \mathbf{Z}_j\mathbf{X}_i\left(\mathbf{X}_i^t\mathbf{Z}_j^t\mathbf{Z}_j\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{Z}_j^t \\ &= \mathbf{Z}_j\mathbf{X}_i\left(\mathbf{X}_i^t\mathbf{R}_j\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{Z}_j^t\end{aligned}$$

U tom slučaju, matrica grešaka predviđanja podataka iz j-tog segmenta

$$\begin{aligned}\mathbf{E}_{j.i} &= \mathbf{Z}_j - \mathbf{Z}_j\mathbf{X}_i\left(\mathbf{X}_i^t\mathbf{R}_j\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{Z}_j^t\mathbf{Z}_j \\ &= \mathbf{Z}_j - \mathbf{Z}_j\mathbf{X}_i\left(\mathbf{X}_i^t\mathbf{R}_j\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{R}_j \\ &= \mathbf{Z}_j\left(\mathbf{I} - \mathbf{X}_i\left(\mathbf{X}_i^t\mathbf{R}_j\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{R}_j\right)\end{aligned}$$

na osnovu svojstvenih vrednosti dobijenih na i-tom segmentu je

$$e_{i.j}^2 = diag\left(\mathbf{E}_{j.i}\mathbf{E}_{j.i}^t\right)$$

dok je suma kvadriranih reziduala, tj. kvadrirana euklidska distanca svakog ispitanika od prostora koji razapinju komponente dobijene na i-toj grupi
4. Na osnovu ovih grešaka, svaki entitet se klasifikuje u grupu od čije je faktorske strukture najmanje udaljen.
5. Ukoliko se posle klasifikacije ispitanika veličina neke grupe nalazi ispod unapred zadate vrednosti (n_g), koja predstavlja minimalnu smislenu veličinu grupe, npr., jedna osmina veličine uzorka, ta grupa se ukida i ispitanici koji su bili u njoj reklasifikuju se u preostale grupe.
6. Algoritam u svakoj iteraciji računa ukupnu varijansu koju nosi r izdvojenih glavnih komponenti iz svakog od segmenata, što je, u stvari, suma svih prvih r svojstvenih vrednosti dobijenih na svakom od segmenata i pamti soluciju u kojoj se postiže maksimum – soluciju koja će na kraju algoritma biti proglašena rešenjem problema.
7. Algoritam se ponovo vraća na drugi korak i iterira sve dok nijedan entitet više ne menja grupnu pripadnost, što se retko dešava, ili dok se ne ponovi unapred zadat broj iteracija, npr., hiljadu.

Opisani algoritam napisan je u R programskom jeziku (R Development Core Team, 2014) i dostupan je na www.kal.rs/simulation. Pre pokretanje makroa *factax* treba podesiti vrednosti promenljivih *tot_step*, *clusters* i *k*, koje označavaju maksimalna broj iteracija, početni broj grupa i maksimalan broj glavnih komponenata, te u promenljivu *Dat* smestiti matricu podataka koja se analizira.

Finalna solucija algoritma zavisi, uglavnom, od izbora parametra *k,* koji označava broj glavnih komponenata koje treba zadržati u faktorskoj soluciji, broja inicijalnih grupa i prve slučajne klasifikacije ispitanika.

Broj glavnih komponenata *k* koje treba izdvojiti treba odrediti na osnovu faktorske solucije celog uzorka i, u tom slučaju, njegovu veličinu treba odrediti kao broj dobijenih faktora na celom uzorku, ili kao taj broj uvećan za jedan. Naravno, broj važnih komponenti posebno se određuje za svaku grupu i u finalnoj soluciji broj komponenti u nekoj grupi može neretko biti manji, ali veoma retko veći od zadate vrednosti *k*. Iz tog razloga ovaj broj predstavlja gornju granicu broja faktora koje je moguće dobiti u svakoj od grupa.

Početni broj grupa treba odrediti u skladu s veličinom uzorka, tj u skladu s pragom za disoluciju grupe. Naravno, treba imati na umu i interpretabilnost dobijenog rešenja.

Inicijalna slučajnost razvrstavanja entiteta u prvom koraku može se ublažiti dodavanjem i slučajnosti prilikom razvrstavanja entiteta u svakom koraku. Jedno za sada predloženo pravilo jeste to da se u dva od tri koraka razvrstavanje ne vrši deterministički u grupu čije je prostor glavnih komponenti najmanje udaljen od ispitanika, već da i to razvrstavanje bude probabilističko, npr., s verovatnoćama inverzno srazmernim tim udaljenostima. Na taj način manja udaljenost imaće veću verovatnoću pridruživanja, ali, naravno, manju od jedan.

Primeri

Rad predloženog algoritma pokazan je na tri primera. Prvi primer koristi simulirane podatke i ujedno pokazuje da algoritam zaista uspeva da detektuje grupe s različitom strukturom. U preostala dva primera prikazano je funkcionisanje predloženog postupka na realnim podacima iz istraživanja.

Primer 1

U ovom primeru korišćeni su simulirani podaci. Matrica podataka je generisana tako da se sastoji od dve grupe od po 150 ispitanika i 9 varijabli. I u jednoj i u drugoj grupi zadata je faktorska struktura sa dva faktora, s tom razlikom što je zadata faktorska struktura različita. Konstrukcija matrice podataka (**Z**) u jednoj grupi započinje od zadate matrice strukture **F**, koja se mno-

ži s normalno distribuiranim slučajnim brojevima koji predstavljaju faktorske skorove **V**. Proizvod te dve matrice predstavlja prave skorove (**T**) na koje se dodaje i slučajno distribuirana greška (**G**), koja predstavlja grešku merenja.

$$\mathbf{Z} = \mathbf{VF}' + \mathbf{G} = \mathbf{T} + \mathbf{G}$$

Matrice **T** i **G** se skaliraju tako da varijansa svake od kolona matrice **T** iznosi 0,7, a kolona matrice **G**, 0,3. Na taj način simulirano je da greška merenja iznosi oko 30%, tj. da je pouzdanost varijabli oko 0,70.

Matrice podataka za prvu i drugu grupu su posle toga spojene i rezultujuća matrica predstavlja ulazne podatke za testiranje algoritma. U Tabeli 1 date su svojstvene vrednosti i faktorske strukture posle varimax rotacije u svakoj od dveju grupa posebno, ali i za podatke u celini, nastale spajanjem te dve grupe.

I jedna i druga grupa imaju po dve svojstvene vrednosti veće od jedan, tako da i po Gutman-Kajzerovom (Guttman, 1953) kriterijumu, a i po Katelovom (Cattell, 1966) „scree" kriterijumu proizilazi da se radi o dva faktora, koliko ih je i bilo zadato. Prva dva faktora i u prvoj i u drugoj grupi objašnjavaju po 82% varijanse. Njihove strukture su različite: dok prva grupa ima visoka zasićenja na prvih pet varijabli, druga grupa ima visoka zasićenja na prvoj, trećoj, petoj, sedmoj i devetoj varijabli.

Kada se faktoriše zajednička matrica podataka, u kojoj su podaci grupa sastavljeni, dobijaju se pod istim kriterijumima tri faktora. Drugi „zajednički" faktor odgovara drugom faktoru prve grupe, dok su prvi i treći „zajednički" faktor, najsličniji faktorima druge grupe. Lako se može uvideti da dobijena zajednička struktura nije jasno povezana sa strukturama pojedinačnih grupa.

Tabela 1. *Svojstvene vrednosti (L) i faktorska varimax struktura (F) simuliranih podataka, svake od dve grupe posebno i uzetih zajedno.*

Prva grupa			Druga grupa			**Obe grupe uzete zajedno**			
L	**F1**	**F2**	**L**	**F1**	**F2**	**L**	**F1**	**F2**	**F3**
3,85	**0,88**	0,05	4,26	**0,85**	0,16	4,00	**0,83**	0,09	0,30
3,10	**0,86**	–0,01	2,71	0,11	**0,90**	1,75	0,28	0,11	**0,85**
0,40	**0,87**	0,05	0,42	**0,90**	–0,05	1,47	**0,89**	0,08	0,12
0,35	**0,88**	–0,01	0,38	0,11	**0,88**	0,37	0,33	0,05	**0,84**
0,32	**0,85**	0,11	0,31	**0,84**	0,19	0,36	**0,81**	0,16	0,27
0,28	0,08	**0,88**	0,29	–0,03	**0,87**	0,29	–0,15	**0,74**	0,49
0,25	0,07	**0,88**	0,26	**0,86**	0,04	0,27	0,44	**0,76**	–0,17
0,24	–0,05	**0,90**	0,19	0,20	**0,87**	0,26	–0,07	**0,81**	0,38
0,20	–0,01	**0,91**	0,18	**0,88**	0,09	0,24	0,44	**0,78**	–0,17

Predloženi algoritam, sa zadatim brojem faktora jednakim tri, izdvojio je dve grupe, sa 129 i 171 ispitanikom. U Tabeli 2 prikazane su svojstvene vrednosti i faktorske strukture nakon varimax rotacije u svakoj od detektovanih

grupa. Strukture sa po dva faktora uspešno su detektovane, ali Tabela 3 nam pokazuje da ispitanici nisu uspešno razvrstani u polazne grupe.

Tabela 2. *Svojstvene vrednosti (L) i faktorska varimax struktura (F) simuliranih podataka u dve grupe koje je detektovao predloženi algoritam.*

I grupa			II grupa		
L	**F1**	**F2**	**L**	**F1**	**F2**
4,01	**0,89**	–0,04	4,67	**0,76**	0,28
2,56	**0,80**	–0,10	1,88	0,12	**0,87**
0,83	**0,87**	–0,08	0,62	**0,86**	0,07
0,37	**0,85**	–0,07	0,42	0,12	**0,86**
0,35	**0,89**	–0,03	0,36	**0,80**	0,26
0,26	–0,21	**0,80**	0,33	0,18	**0,83**
0,23	0,05	**0,87**	0,29	**0,83**	0,13
0,22	–0,17	**0,86**	0,24	0,30	**0,77**
0,18	0,00	**0,80**	0,19	**0,85**	0,21

Tabela 3. *Matrica slaganja zadatih i detektovanih grupa na simuliranim podacima.*

Zadate grupe	Detektovane grupe		
	I	**II**	**Total**
I	72	78	**150**
II	57	93	**150**
Total	**129**	**171**	**300**

Prva grupa je detektovana sa 48% uspešnosti, što je na nivou slučajnog pogađanja, dok je uspešnost detekcije druge nešto veća (60%), ali još uvek daleko od zadovoljavajuće.

Prvi primer jasno pokazuje da predloženi algoritam uspeva da detektuje različite faktorske strukture među ispitanicima. Ono što je isto tako važno uvideti jeste to da dobijena faktorska struktura može predstavljati superponiranje faktorskih struktura određenih segmenata uzorka i da se detekcijom tih segmenata dobijaju jednostavnija rešenja.

Primer 2

Algoritam je testiran i na velikom uzorku od 5.105 ispitanika, reprezentativnom za punoletnu populaciju Rebuplike Srbije. Uzorak je prikupljen u intervalu od marta do septembra 2014. godine, kao deo standardnog mesečnog anketiranja terenskom anketom koje sprovodi agencija „Ipsos Strategic Marketing". Od ispitanika je traženo da ocenama na skali od 1 do 5 ocene aktuelne političke prvake. Zadržani su samo ispitanici koji su znali sve navedene političare, tj. oni koji su dali validnu ocenu za sve političke prvake. Ispitivanje faktorske strukture tako dobijenih podataka već je uveliko deo metodologije

politikoloških istraživanja (Armstrong et al., 2014; Box-Steffensmeier, Brady, & Collier, 2009; Craig, Martinez, & Kane, 1999). Za razliku od standardne prakse analize ovih podataka, začete još 70-ih godina XX veka (Weisberg & Rusk, 1970), koja za analizu prostora u kojem leže stranački prvaci koristi metodu multidimenzionalnog skaliranja, predloženi metod, kako je i opisano, koristi faktorsku analizu, ali je njen cilj jako sličan: otkrivanje dimenzija u kojima se nalaze te ocene.

Kako je uzorak jako veliki, a radi se o malom broju varijabli, s pet različitih odgovora, algoritam će pronalaziti ispitanike s linearno zavisnim odgovorima (npr., istim ocenama za sve političare), čime se kolabira dimenzionalnost prostora i veštački povećava varijansa prve komponente. Kako bi se to izbeglo, podaci su transformisani u imaž prostor (Guttman, 1953). To je transformacija koja skor na svakoj varijabli zamenjuje predviđenim skorom te varijable na osnovu svih ostalih varijabli u skupu (više videti u Momirović, Wolf, & Popović, 1999) . Tom transformacijom obezbeđuje se i normalnost distribucija varijabli transformisanih u imaž oblik, budući da normalnost distribucija netransformisanih varijabli u slučaju mere stavova može biti jako narušena.

Na celom uzorku su izdvojene dve glavne komponente (Tabela 4) koje zajedno obuhvataju 91% varijanse početnih varijabli. Faktori bi se najlakše, za potrebe ovog rada, mogli opisati kao „vlast" i „opozicija". Rešenje koje se sastoji od dve dimenzije (Bornschier, 2010), dobija se i u drugim evropskim državama. Iako ovo rešenje već deluje jako parsimonično, predloženi algoritam je izdvojio tri grupe ispitanika:

- Prva grupa u koju je svrstano 12% ispitanika smeštena je u trodimenzionalnom prostoru. Prve dve dimenzije se mogu na isti način opisati kao vlast i opozicija, dok treći faktor predstavljaju političari iz vlasti ali iz „drugog" plana (Stefanović, Mihailović, Mali)
- Druga grupa od 54% ispitanika ima dvofaktorsko rešenje koje je ekvivalentno rešenju iz celog uzorka, s tim da Pajtić i Koštunica nisu tako jasno saturirani „opozicionim" faktorom.
- Treća grupa, koja čini 34% uzorka, zapravo ima samo jedan faktor; u ovom slučaju Katelov test osulina („scree" test) mnogo je primenljiviji jer prva komponenta nosi gotovo celu varijansu podataka. Ovaj faktor zapravo predstavlja generalno podržavanje/nepodržavanje svih političkih prvaka.

Kako je jasno da su ocene stranačkih prvaka povezane sa ideologijama koje oni predstavljaju (Green, 1988), može se zaključiti da jedna trećina populacije ili nema jasan stav ili uopšte nema razvijen politički stav: unidimenzionalnost njihovog prostora upravo govori tome u prilog.

Predloženi algoritam pokazuje da sve tri grupe imaju lako interpretabilne i razumljive faktorske strukture, tj. da predloženi algoritam izdvaja, u naoči-

gled jasnom dvodimenzionalnom prostoru, skupine koje imaju prostore sa različitom dimenzionalnošću.

Tabela 4. *Svojstvene vrednosti (L) i faktorska varimax struktura (F) ocena političkih lidera celog uzorka sa izdvojena dva faktora, i faktorske strukture tri detektovane grupe ispitanika, sa po tri, odnosno dva i jednim faktorom.*

	Ceo uzorak (n= 5105)			I grupa (n = 589)				II grupa (n = 2768)			III grupa (n = 1748)	
	L	F1	F2	L	F1	F2	F3	L	F1	F2	L	F1
Koštunica, V.	9,81	0,53	**0,83**	5,22	**0,86**	0,46	–0,06	8,77	0,69	O,66	11,15	0,99
Tadić, B.	2,02	0,29	**0,94**	3,65	**0,96**	0,13	–0,13	2,72	0,19	**0,92**	1,11	0,97
Dinkić, M.	0,69	0,33	**0,92**	2,52	**0,92**	0,15	–0,5	0,68	0,30	**0,90**	0,32	0,97
Jovanović, Č.	0,13	0,23	**0,95**	0,42	**0,95**	0,05	–0,14	0,25	0,06	**0,94**	0,15	0,96
Đjilas, D.	0,10	0,26	**0,95**	0,34	**0,93**	–0,10	0,04	0,16	0,15	**0,92**	0,08	0,96
Pajtić, B.	0,06	0,43	**0,81**	0,26	0,55	–0,54	0,52	0,13	0,59	0,61	0,07	0,93
Nikolić, T.	0,05	**0,91**	0,22	0,20	0,03	**0,94**	0,02	0,10	**0,93**	0,13	0,05	0,81
Marković–Palma, D.	0,04	**0,84**	0,48	0,14	0,33	**0,88**	0,12	0,08	**0,92**	0,27	0,02	0,95
Vučić, A.	0,03	**0,93**	0,09	0,10	–0,07	**0,93**	0,11	0,04	**0,95**	0,02	0,02	0,71
Dačić, I.	0,03	**0,86**	0,43	0,09	0,21	**0,90**	0,06	0,03	**0,93**	0,23	0,02	0,95
Stefanović, N.	0,02	**0,82**	0,44	0,05	–0,08	0,06	**0,89**	0,02	**0,91**	0,0,8	0,02	0,95
Mihajlović, Z.	0,01	**0,87**	0,38	0,02	–0,14	0,26	**0,84**	0,02	**0,94**	0,24	0,00	0,96
Mali, S.	0,00	**0,80**	0,41	0,01	–0,0	–0,04	**0,89**	0,01	**0,89**	0,27	0,00	0,92

Primer 3

Podaci u ovom primeru prikupljeni su baterijom za procenu intelektualnih sposobnosti KOG9 (Wolf, Momirović, & Džamonja, 1992). Studenti psihologije, njih 159, ispitani su sa devet testova te baterije koja po konstruktorima meri tri dimenzije intelektualnog funkcionisanja. Te dimenzije, od kojih je svaka operacionalizovana sa po tri testa, mere efikasnost funkcionisanja perceptivnog, serijalnog i paralelnog procesora, koji su delovi predloženog kibernetskog modela intelektualnih sposobnosti. Naknadne provere strukture te baterije (Lazarević & Knezević, 2008) pokazuju da postulirani trofaktorski model, zbog slabije operacionalizacije perceptivnog procesora, lako kolabira na stabilniji i replikabilniji dvofaktorski model, koji se nekada opisuje i kao verbalno–neverbalni prostor intelektualnog funkcionisanja (Lalović, 2000).

Analiza celokupnog uzorka, primenom Katelovog „scree" testa, izoluje dva faktora (Tabela 5) u kojima se prepoznaju paralelni i serijalni procesor, iako je jedan test (**it1**) perceptivnog funkcionisanja visoko saturiran serijalnim faktorom, dok su preostala dva testa kojima je operacionalizovan ovaj procesor, kao i jedan test paralenog procesora (**d48**), ostali sa saturacijama koje ih jednoznačno ne opredeljuju nijednom od ova dva faktora.

Predloženi algoritam, kojem je u ovom slučaju zbog veličine uzorka zadat parametar bio razvrstavanje u maksimalno dve grupe sa strukturom od maksimalno tri faktora, detektovao je grupe u kojima se izdvajaju tri, odnosno jedan faktor.

U Tabeli 5 prikazane su faktorske strukture ovih grupa. Prvu grupu sa tri faktora čini 59 ispitanika i u njoj se prvi faktor jasno identifikuje kao faktor efikasnosti paralelnog procesora. Serijalni i perceptivni faktor su razmenili po jedan test; kao i na celokupnom uzorku test **it1** je postao visoko zasićen serijalnim procesorom, dok test **al4** ima jedinu visoku, i to negativnu, korelaciju s perceptivnim procesorom.

U drugoj izdvojenoj grupi od 90 ispitanika izdvojen je samo jedan faktor generalne inteligencije, s kojim su svi testovi visoko zasićeni.

U prvom primeru pokazano je da svrstavanja ispitanika u grupe može dati vrlo nestabilne rezultate, tj. da se ista ili slična faktorska struktura može dobiti i s veoma različitom klasifikacijom ispitanika. Imajući to u vidu, postavlja se pitanje kvantitativnog poređenja tih grupa, ali čisto deskriptivnim poređenjem njihovih distribucija na skoru koji predstavlja prosek svih devet testova dobija se nalaz da su to dve preklopljene distribucije, jednakih proseka, od kojih grupa s tri faktora ima skoro duplo manju standardnu devijaciju od grupe s jednofaktorskom strukturom.

Ovaj nalaz bi se možda mogao interpretirati na sledeći način: baterija KOG9 dobro operacionalizuje pretpostavljena tri faktora kognitivnog funkcionisanja u nivoima prosečnog intelektualnog funkcionisanja, dok se u slučaju merenja ekstremnih zona uspešnosti intelektualnog funkcionisanja svodi na faktor generalne inteligencije.

Tabela 5. *Svojstvene vrednosti (L) i faktorska varimax struktura (F) baterije KOG9 celog uzorka sa izdvojena dva faktora, i faktorske strukture dve detektovane grupe ispitanika, sa po dva odnosno jednim faktorom.*

	Ceo uzorak (n= 149)			I grupa (n = 59)				II grupa (n = 90)	
	L	F1	F2	L	F1	F2	F3	L	F1
s1	3,47	0,03	**0,87**	2,39	**0,79**	–0,22	0,17	4,18	0,65
it2	1,37	0,13	**0,86**	1,99	**0,84**	–0,06	0,20	1,17	0,69
d48	0,91	0,41	0,50	1,33	**0,64**	0,25	–0,18	0,97	0,65
al4	0,81	**0,53**	0,18	0,84	0,05	0,01	**–0,75**	0,66	0,70
alf7	0,71	**0,83**	0,08	0,75	–0,09	**0,75**	0,02	0,63	0,76
gsn	0,60	**0,86**	0,06	0,58	–0,33	**0,60**	0,37	0,49	0,78
it1	0,45	**0,62**	0,20	0,53	0,20	**0,86**	0,05	0,38	0,60
cf2	0,37	0,39	0,54	0,39	0,29	0,36	**0,60**	0,36	0,64
gt7	0,31	0,46	0,31	0,22	0,11	0,05	**0,81**	0,17	0,65

Zaključak

Predloženi algoritam podele jedinica posmatranja u grupe s različitom faktorskom strukturom, kako je pokazano u tri primera, konvergira ka zadatom cilju, pronalasku različitih struktura koje imaju različiti segmenti uzorka.

Najveća mana predloženog algoritma jeste njegova heurističnost, zasnovana na probabilističkom modelu traženja rešenja, što onemogućava njegovu unapred garantovanu replikabilnost. Ipak, to se može shvatiti i kao dobra osobina algoritma: istraživač je u toj situaciji svesniji direktne zavisnosti od slučajnosti, te to može uzeti u obzir pri interpretiranju dobijenih rezultata. Naime, u krajnjoj liniji i ceo uzorak je samo jedna od mogućih solucija uzorkovanja. Konvergiranje algoritama ka jednom rešenju posle više primena algoritma, pa čak i na istim podacima, može govoriti u korist stabilnosti, reproducibilnosti, ali i relevantnosti dobijene solucije.

Ideja klasifikacije implicitno podrazumeva kvalitativne razlike između grupa, a predloženi algoritam, kao što je pokazano, takve razlike i detektuje, ako one realno postoje. U situaciji kada je analizirani uzorak homogen, postoji mogućnost da algoritam ne konvergira ka jednoj grupi. Algoritam bi se tako mogao unaprediti i dodavanjem testiranja značajnosti dobijenih razlika između grupa, kao i neke od mera veličine efekta, što bi pomoglo da se odgovori na pitanje o tome da li su dobijene razlike uopšte toliko velike da uzorak zaslužuje razvrstavanje u više grupa.

Mogućnost primene algoritama je potencijalno velika, praktično kad god nas interesuje latentna struktura skupa merenih varijabli u datoj populaciji (na primer, struktura ličnosti ili stavova), možemo se zapitati i da li ta struktura važi za celu populaciju ili je zapravo neka rezultanta različitih struktura pojedinih segmenata populacije.

Pitanje interpretabilnosti nalaza, kao i u većini multivarijatnih statističkih metoda, ostaje jedno od presudnih pitanja kada se procenjuje važnost dobijenih rezultata. Nalazi koji se ne uklapaju u teorijske okvire nameću potrebu za novim i metodološki savršenijim istraživanjima; međutim, ukoliko se konzistentno repliciraju, mogu poslužiti i za preispitivanje početnih teorijskih okvira.

Reference

Armstrong, D. A., Bakker, R., Carroll, R., Hare, C., Poole, K. T., & Rosenthal, H. (2014). *Analyzing spatial models of choice and judgment with R*. Boca Raton, FL: Chapman & Hall/CRC.

Bock, H.-H. (2007). Origins and extensions of the k –means algorithm in cluster analysis. In P. Brito, P. Bertrand, G. Cucumel, & F. De Carvalho (Eds.), *Selected contributions in data analysis and classification* (pp. 161–172). Heidelberg: Springer.

Bornschier, S. (2010). The new cultural divide and the two-dimensional political space in Western Europe. *West European Politics*, *33*(3), 419–444. doi:10.1080/01402381003654387

Box-Steffensmeier, J. M., Brady, H. E., & Collier, D. (Eds.). (2009). *The Oxford handbook of political methodology*. New York: Oxford University Press. doi:10.1093/oxfordhb/9780199286546.001.0001

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Craig, S. C., Martinez, M. D., & Kane, J. G. (1999). The structure of political competition: dimensions of candidate and group evaluation revisited. *Political Behaviour*, *21*(4), 283–304.

Green, D. P. (1988). On the dimensionality of public sentiment toward partisan and ideological groups. *American Journal of Political Science*, *32*(3), 758. doi:10.2307/2111245

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrica*, *18*(4), 277–296.

Lalović, D. (2000). Povezanost inteligencije i brzine kognitivne obrade reči srpskog jezika. *Psihologija*, *33*(1–2), 75–104.

Lazarević, L. & Knezević, G. (2008). Provera faktorske strukture baterije za procenu intelektualnih sposobnosti KOG9. *Psihologija*, *41*(4), 489–505. doi:10.2298/PSI0804489L

Legendre, P. & Legendre, L. (1998). *Numerical ecology (2nd English ed.)*. Amsterdam: Elsevier Science BV.

Momirović, K., Wolf, B., & Popović, B. (1999). *Uvod u teoriju merenja, I. Metrijske karakteristike kompozitnih mernih instrumenata*. Priština: Univerzitet u Prištini.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Weisberg, H. F. & Rusk, J. G. . (1970). Dimensions of candidate evaluation. *The American Political Science Review*, *64*(4), 1167–1185.

Wolf, B., Momirović, K., & Džamonja, Z. (1992). *KOG 3 – Baterija testova inteligencije*. Beograd: Centar za primenjenu psihologiju.

DATUM PRIJEMA RADA: 28.11.2014. DATUM PRIHVATANJA RADA: 13.12.2014.

The heuristic method for detecting clusters of respondents with different factor structure of measured variables

Aleksandar Zorić
Faculty of philosophy, University of Belgrade

The new heuristic method is proposed for the identification of clusters of cases with different factor structure of measured variables. The classical cluster analysis methods try to minimize within and maximize between-groups distance, implicitly holding the assumption that all entities lie in one common space spanned with measured variables. The proposed algorithm, on the other hand, tries to divide the sample into clusters that are located in different spaces that can differ even in their dimensionality. The basic goal of the proposed method is to detect qualitative differences between clusters, and by that to obtain the new insights into the data.

Key words: cluster analysis, factor analysis, algorithm, qualitative differences.