Vojko Antončič
Inštitut za sociologijo
Cveto Trampuž
Fakulteta za sociologijo,
politične vede in novinarstvo

DVE METODI ZA ANALIZO NOMINALNIH SPREMENLJIVK

Predstavljena je metoda, ki sloni na spektralni dekompoziciji matrike P in metoda, ki sloni na spektralni dekompoziciji matrike Q. Elementi matrike P so hkratne relativne frekvence. Elementi matrike Q so pogojne relativne frekvence.

binariziranje, hratne relativne frekvence, pogojne relativne frekvence, spektralna dekompozicija, metrične komponente

TWO METHODS FOR THE ANALYSIS OF NONNUMERICAL DATA

A method which is based on the spectral decomposition of matrix P and a method based on the spectral decomposition of matrix Q is discussed and evaluated; the elements of P are the intersection relative frequencies and the elements of Q are the conditional relative frequencies.

binary coding, intersection relative frequencies, conditional relative frequences, spectral decomposition, metric components

1. UVOD

V empiričnih socioloških raziskavah se moramo velikokrat ubadati z nominalnimi spremenljivkami, saj so začetni podatki, recimo podatki, ki jih dobimo s kakim anketnim instrumentom, pogosto nenumerični (ali kvečjemu kvazinumerični) in (najbrž) drugačni niti ne morejo biti. Ce jih hočemo uporabiti v kakšni taki analizi, kot je na primer kanonična korelacijska analiza, jih moramo najprej smiselno kvantificirati. Oglejmo si dve metodi, ki ju lahko uporabimo v ta namen. Imenujmo ju kar metoda A in metoda B.

2. METODA A

Naj bodo $U_1,U_2,\dots,U_k$ nominalne spremenljivke, ki jih upoštevamo v dani analizi. Spremenljivka U_i naj ima vrednosti $1,2,\dots,m_i$ $(i=1,2,\dots k)$, ki označujejo m_i nenumeričnih kategorij. Vseh kategorij je potemtakem

$$\sum_{i=1}^{k} m_i = m$$

Z njimi definiramo binarne spremenljivke $V_1, V_2, \ldots, V_m$. Naj bo

$$r_i = \sum_{h=1}^{i} m_h \qquad \text{in} \qquad s_i = r_i - m_1 \qquad (i = 1, 2, \ldots, k)$$

Binarno spremenljivko V_j z indeksom $j = s_i + h \leq r_i$ definiramo s h-to kategorijo spremenljivke U_i, in sicer takole:

$$V_j = \begin{cases} 1, & \text{če je } U_i = h \\ 0, & \text{sicer} \end{cases}$$

Vzemimo, da poznamo vrednosti spremenljivk U_i za n entitet. Za vsako entiteto določimo vrednosti binarnih spremenljivk in sestavimo matriko $\mathbf{A}_{n \times m}$: v j-tem stolpcu matrike $\mathbf{A}$ naj bodo vrednosti spremenljivke V_j $(j = 1, 2, \ldots, m)$.

Kategorije spremenljivk U_i kvantificiramo tako, da določimo neko smiselno transformacijo

$$\mathbf{K} = \mathbf{AT} \tag{1}$$

V ta namen izračunamo matriko

$$\mathbf{P} = [P_{ij}] = \frac{1}{n}\mathbf{A}^T\mathbf{A}$$

Elementi matrike $\mathbf{P}$ so *hkratne* relativne frekvence: relativna frekvenca P_{ij} pove, kolikšen je delež entitet, pri katerih ima spremenljivka V_i in *hkrati* tudi spremenljivka V_j vrednost 1. Če P_{ij} interpretiramo kot verjetnost, lahko zapišemo

$$P_{ij} = P(V_i = 1 \ \& \ V_j = 1) \qquad (i, j = 1, 2, \ldots, m)$$

Transformacijo $\mathbf{T}$ dobimo tako, da določimo vektorje $\mathbf{x}$, ki ustrezajo zahtevi

$$\max(\mathbf{Px}, \mathbf{x}) \qquad \text{pri pogoju} \qquad (\mathbf{x}, \mathbf{x}) = 1 \tag{2}$$

Pri tem je (,) oznaka za skalarni produkt dveh vektorjev. Kratek račun, ki ga poznamo iz komponentne analize, pokaže, da optimizacijski zahtevi (2) zadoščajo normirani lastni vektorji matrike $\mathbf{P}$. Naj bodo

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$$

lastne vrednosti matrike $\mathbf{P}$. Zanje velja:

1. Vse so nenegativne, saj je $\mathbf{P}$ Grammova matrika.

2. Po znanem izreku iz linearne algebre je vsota lastnih vrednosti dane matrike enaka vsoti njenih diagonalnih elementov, se pravi,

$$\sum_{j=1}^{m} \lambda_j = \mathrm{sled}(\mathbf{P})$$

Ker je za binarne spremenljivke pri vsakem i od 1 do k

$$\sum_{j=s_i+1}^{r_i} v_j = 1 \tag{3}$$

je tudi

$$\sum_{j=s_i+1}^{r_i} p_{jj} = 1$$

Zato je $\mathrm{sled}(\mathbf{P}) = k$, torej

$$\sum_{j=1}^{m} \lambda_j = k$$

3. Nadalje velja: zaradi relacije (3) je število linearno neodvisnih binarnih spremenljivk enako $m-k+1$, ali drugače povedano,

$$\mathrm{rang}(\mathbf{A}) = m-k+1$$

Ker je $\mathrm{rang}(\mathbf{P}) = \mathrm{rang}(\mathbf{A})$, sledi od tod, da ima matrika $\mathbf{P}$ samo $m-k+1$ pozitivno lastno vrednost.

Naj bodo $\mathbf{x}_j$, $j = 1,2,\ldots,m-k+1$, normirani lastni vektorji pri pozitivnih lastnih vrednostih matrike $\mathbf{P}$. V (1) vstavimo

$$\mathbf{T} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_{m-k+1}]$$

in izračunamo $\mathbf{K}$, se pravi, za vsako entiteto določimo vrednosti metričnih komponent. Ta metoda za kvantifikacijo nominalnih spremenljivk je opisana v knjigi »Uvod u analizu nominalnih variabli« (Momirović, 1988).

3. METODA B

Vpeljimo matriko

$$\mathbf{A}^* = \mathbf{E} - \mathbf{A}$$

Pri tem je $\mathbf{E}$ matrika, ki ima za elemente same enice. Matriki $\mathbf{A}$ in $\mathbf{A}^*$ konkatenirajmo v matriko

$$\mathbf{B} = [\mathbf{A} \mid \mathbf{A}^*]$$

Vpeljimo še diagonalno matriko

$$\mathbf{F} = \mathrm{diag}(\mathbf{B}^T\mathbf{B}) \tag{4}$$

Njen j-ti diagonalni element f_j $(j = 1,2,\ldots,2m)$ je število enic v j-tem stolpcu matrike $\mathbf{B}$. Za vsak j od 1 do m velja:

$$f_j + f_{j+m} = n \tag{5}$$

Namesto matrike P analizirajmo sedaj matriko

$$Q = [Q_{ij}] = F^{-1}B^TB$$

Njeni elementi so *pogojne* relativne frekvence. Ce jih interpretiramo kot verjetnosti, lahko za vse indekse $i,j=1,2,\ldots,m$ zapišemo

$$Q_{ij} = P(V_j=1 / V_i=1)$$

$$Q_{i,j+m} = P(V_j=0 / V_i=1)$$

$$Q_{i+m,j} = P(V_j=1 / V_i=0)$$

$$Q_{i+m,j+m} = P(V_j=0 / V_i=0)$$

Naj bodo

$$\eta_1 \geq \eta_2 \geq \ldots \geq \eta_{2m}$$

lastne vrednosti matrike Q. Kaj se dà povedati o teh lastnih vrednostih?

1. Vse so nenegativne, saj ima matrika Q enake lastne vrednosti kot Grammova matrika $F^{-1/2}B^TBF^{-1/2}$.

2. Iz (4) je razvidno, da so vsi diagonalni elementi matrike Q enaki 1. To pomeni, da je sled$(Q)=2m$ in zato tudi

$$\sum_{j=1}^{2m} \eta_j = 2m$$

3. Na dlani je, da je rang$(B)=$rang(A) in da ima matrika Q ravno toliko pozitivnih lastnih vrednosti kot matrika P.

4. Brez težav se prepričamo, da je

$$\eta_1 = m \tag{6}$$

Res: Naj bo e 2m-komponentni vektor enic. Ce v enačbo

$$Qy = \eta y \tag{7}$$

vstavimo $y = e$, s kratkim računom dobimo

$$me = \eta e \tag{8}$$

Od tod se vidi, da je število m lastna vrednost matrike Q. Ker pa so vse lastne vrednosti matrike Q nenegativne in je njihova vsota enaka 2m, je očitno, da je m največja lastna vrednost matrike Q. S tem je trditev (6) v celoti potrjena.

Iz enačbe (8) se vidi, da vsak lastni vektor matrike Q, ki pripada lastni vrednosti $\eta_1 = m$, lahko izrazimo v obliki ce; c je poljubno realno število. Metrična komponenta, ki jo dobimo s takim lastnim vektorjem, je trivialna, zato jo lahko izločimo. To napravimo tako, da od matrike $\mathbf{Q}$ odštejemo matriko

$$\mathbf{S} = c\,\eta_1\, y_1\, z_1^T$$

Pri tem je y_1 lastni vektor matrike $\mathbf{Q}$ pri lastni vrednosti η_1 in z_1 lastni vektor matrike $\mathbf{Q}^T$ ravno tako pri lastni vrednosti η_1; c pa je število, ki zadošča zahtevi

$$c(y_1, z_1) = 1 \tag{9}$$

Ce enačbo (7) premultipliciramo z $\mathbf{F}$, spoznamo, da je

$$z_1 = \mathbf{F}\, y_1 \tag{10}$$

Vzemimo $y_1 = e$, upoštevajmo (5) in (10), pa ugotovimo:

$$(y_1, z_1) = mn$$

Od tod sledi, da zahtevi (9) zadosti število

$$c = \frac{1}{mn}$$

Da izločimo prvo komponento, moramo potemtakem od matrike $\mathbf{Q}$ odšteti kvadratno matriko

$$\mathbf{S} = \frac{1}{n}\begin{bmatrix} F_1 & F_2 & \cdots & F_{2m} \\ F_1 & F_2 & \cdots & F_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ F_1 & F_2 & \cdots & F_{2m} \end{bmatrix}$$

Matrika $\mathbf{Q}-\mathbf{S}$ ima $m-k$ pozitivnih lastnih vrednosti: prva lastna vrednost matrike $\mathbf{Q}-\mathbf{S}$ je enaka drugi lastni vrednosti matrike $\mathbf{Q}$, druga lastna vrednost matrike $\mathbf{Q}-\mathbf{S}$ je enaka tretji lastni vrednosti matrike $\mathbf{Q}$, in tako dalje, zadnja pozitivna lastna vrednost matrike $\mathbf{Q}-\mathbf{S}$ je enaka zadnji pozitivni lastni vrednosti matrike $\mathbf{Q}$. Vsakič so pripadajoči lastni vektorji ene matrike enaki lastnim vektorjem druge matrike. Te trditve opiramo na izrek, ki ga Faddeev in Faddeeva (1963) uporabita pri potenčni metodi za določanje lastnih vrednosti.

Dà se pokazati (Trampuž in Antončič, 1981), da so metrične komponente, ki jih določimo s spektralno dekompozicijo matrike $\mathbf{Q}-\mathbf{S}$, ekvivalentne metričnim komponentam, ki jih določimo s spektralno dekompozicijo matrike

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}$$

Pri tem je:

$$Z = (A-\bar{A})\,D^*$$

$$\bar{A} = E\,D \qquad D^* = [\,D(I-D)\,]^{-1/2}$$

E je matrika dimenzije $n \times m$, vsi njeni elementi so enice; D pa je diagonalna matrika dimenzije $m \times m$, ki jo definiramo takole:

$$D = \frac{1}{n}\,\mathrm{diag}(A^TA)$$

Skratka: Z je matrika, v kateri so standardizirane vrednosti spremenljivk V_1, $V_2, \ldots, V_m$; R pa je korelacijska matrika. Matriki $Q-S$ in R imata enake lastne vrednosti. Naj bodo v

$$T_1 = [y_1 \; y_2 \; \cdots \; y_{m-k}]$$

normirani lastni vektorji matrike $Q-S$ pri pozitivnih lastnih vrednostih in v

$$T_2 = [y_1^* \; y_2^* \; \cdots \; y_{m-k}^*]$$

normirani lastni vektorji korelacijske matrike R pri istih lastnih vrednostih.

S T_1 določimo komponente

$$K_1 = A\,T_1$$

s T_2 določimo komponente

$$K_2 = Z\,T_2$$

Trditev, da so komponente K_1 ekvivalentne komponentam K_2, pomeni, da je

$$K_1 = K_2\,H$$

Pri tem je H neka diagonalna matrika.

4. NUMERIČEN PRIMER

Poglejmo na (izmišljenem) primeru, kako delujeta metodi A in B. Da ne bo preveč številk, smo sestavili primer, v katerem nastopata samo dve nominalni spremenljivki. Ena ima tri kategorije, ena pa dve kategoriji. Določili smo 20 vrednosti in jih binarizirali. Imamo torej primer, v katerem je $k=2$, $m=5$ in $n=20$.

Numerični preizkus metode A in metode B smo naredili na računalniku VAX. Uporabili smo programski jezik GENSTAT.

Vhodni podatki so tile:

Nominalni spremenljivki		Binarne spremenljivke				
U_1	U_2	V_1	V_2	V_3	V_4	V_5
1	1	1	0	0	1	0
3	1	0	0	1	1	0
1	1	1	0	0	1	0
3	2	0	0	1	0	1
1	2	1	0	0	0	1
2	1	0	1	0	1	0
2	2	0	1	0	0	1
1	1	1	0	0	1	0
3	2	0	0	1	0	1
2	2	0	1	0	0	1
2	1	0	1	0	1	0
3	1	0	0	1	1	0
1	1	1	0	0	1	0
2	2	0	1	0	0	1
2	2	0	1	0	0	1
1	1	1	0	0	1	0
3	1	0	0	1	1	0
1	2	1	0	0	0	1
2	2	0	1	0	0	1
2	1	0	1	0	1	0

Matrika P

0.3500				
0.0000	0.4000			
0.0000	0.0000	0.2500		
0.2500	0.1500	0.1500	0.5500	
0.1000	0.2500	0.1000	0.0000	0.4500

Lastne vrednosti matrike P

0.8598
0.5754
0.2927
0.2721
0.0000

Sled(P) = 2

Normirani lastni vektorji matrike P

1	2	3	4	5
-0.3995	-0.3028	0.7215	0.1678	-0.4472
-0.4400	0.4619	-0.4325	0.4539	-0.4472
-0.2280	-0.0469	-0.2295	-0.8326	-0.4472
-0.6459	-0.5298	-0.3152	0.0534	0.4472
-0.4216	0.6419	0.3747	-0.2642	0.4472

Komponente po metodi A

1	2	3	4
-1.0453	-0.8326	0.4064	0.2212
-0.8739	-0.5767	-0.5446	-0.7792
-1.0453	-0.8326	0.4064	0.2212
-0.6496	0.5950	0.1452	-1.0968
-0.8210	0.3391	1.0962	-0.0964
-1.0858	-0.0679	-0.7477	0.5073
-0.8615	1.1038	-0.0578	0.1897
-1.0453	-0.8326	0.4064	0.2212
-0.6496	0.5950	0.1452	-1.0968
-0.8615	1.1038	-0.0578	0.1897
-1.0858	-0.0679	-0.7477	0.5073
-0.8739	-0.5767	-0.5446	-0.7792
-1.0453	-0.8326	0.4064	0.2212
-0.8615	1.1038	-0.0578	0.1897
-0.8615	1.1038	-0.0578	0.1897
-1.0453	-0.8326	0.4064	0.2212
-0.8739	-0.5767	-0.5446	-0.7792
-0.8210	0.3391	1.0962	-0.0964
-0.8615	1.1038	-0.0578	0.1897
-1.0858	-0.0679	-0.7477	0.5073

Korelacijska matrika R

1.0000				
-0.5991	1.0000			
-0.4237	-0.4714	1.0000		
0.2423	-0.2872	0.0580	1.0000	
-0.2423	0.2872	-0.0580	-1.0000	1.0000

Lastne vrednosti matrike R

2.375
1.397
1.228
0.000
0.000

Sled(R) = 5

Normirani lastni vektorji matrike R

1	2	3	4	5
-0.3665	0.5129	-0.5053	0.5893	0.0000
0.4308	0.2311	0.6281	0.6053	0.0000
-0.0838	-0.8264	-0.1541	0.5350	0.0000
-0.5801	-0.0165	0.4040	0.0000	-0.7071
0.5801	0.0165	-0.4040	0.0000	-0.7071

Komponente po metodi B

1	2	3
-1.2020	0.8101	-0.3444
-0.8290	-1.7144	0.2904
-1.2020	0.8101	-0.3444
0.6844	-1.6582	-1.1750
0.3114	0.8663	-1.8098
-0.1327	0.2995	1.7685
1.3807	0.3556	0.3030
-1.2020	0.8101	-0.3444
0.6844	-1.6582	-1.1750
1.3807	0.3556	0.3030
-0.1327	0.2995	1.7685
-0.8290	-1.7144	0.2904
-1.2020	0.8101	-0.3444
1.3807	0.3556	0.3030
1.3807	0.3556	0.3030
-1.2020	0.8101	-0.3444
-0.8290	-1.7144	0.2904
0.3114	0.8663	-1.8098
1.3807	0.3556	0.3030
-0.1327	0.2995	1.7685

Korelacije med štirimi komponentami po metodi A in tremi komponentami po metodi B

A1	1.000						
A2	0.592	1.000					
A3	0.223	-0.003	1.000				
A4	-0.767	0.009	0.004	1.000			
B1	0.582	1.000	-0.028	0.015	1.000		
B2	-0.572	0.007	0.471	0.884	0.000	1.000	
B3	-0.578	-0.025	-0.882	0.467	0.000	0.000	1.000
	A1	A2	A3	A4	B1	B2	B3

5. PRIMERJALNA EVALVACIJA METOD A IN B

Po teoretični obravnavi in numeričnem preizkusu ene in druge metode lahko ugotovimo (vsaj) tole:

1. Komponente, ki jih dobimo po metodi A, so definirane v metriki, ki jo določajo hkratne relativne frekvence. Komponente, ki jih dobimo po metodi B, so definirane v metriki, ki jo določajo pogojne relativne frekvence.

2. Komponente, ki nam jih da metoda B, so nekorelirane; so »prave« glavne komponente. To ne velja za komponente, ki jih določimo po metodi A.

3. Metoda B je za eno komponento bolj parsimonična kot metoda A.

4. V prikazanem numeričnem primeru je za metodo A očitno naslednje: prva komponenta je presežek, ki samo moti, vendar nimamo analitičnega sodila za to, da bi jo izločili ali korigirali. Ker so spremenljivke AT med seboj korelirane, je morda še najbolje, da na njih naredimo komponentno analizo in na ta način izločimo, kar je v njih odveč.

5. Dà pa se pokazati, da pri poljubnem številu nominalnih spremenljivk velja tole: če so vrednosti vsake nominalne spremenljivke enakomerno porazdeljene, je mogoče analitično določiti največjo lastno vrednost matrike P in izločiti prvo komponento.

REFERENCE

Faddeev, D.K. in Faddeeva, V.N. (1963). Vičislitelnie metodi linejnoj algebri. Moskva: Gosudarstvennoe izdajateljstvo f-m literaturi.

Momirović, K. (1988). Uvod u analizu nominalnih variabli. Ljubljana: JUS, sekcija za metodologijo in statistiko ter RI Fakultete za sociologijo, politične vede in novinarstvo (Metodološki zvezki 2).

Trampuž, C. in Antončič, V. (1981). »Odnos izmedju skalogramske i komponentne analize.« Novi Sad: XIV godišnji sastanak Saveza statističkih društava Jugoslavije.