Jože ROVAN

KORESPONDENČNA ANALIZA

Zapiski predavanj

LJUBLJANA, 2006

1 Uvod

1.1 Oris metode

Korespondenčna analiza je multivariatna metoda namenjena prikazu povezav v večrazsežnih tabelah podatkov (npr. v kontingenčnih tabelah, Burtovi tabeli itd.). Tabele se preoblikuje v grafikone in skupaj s pripadajočimi numeričnimi kazalci pomenijo osnovo za vsebinsko analizo proučevanega pojava.

Pričnimo z najenostavnejšo možnostjo – z dvorazsežno kontingenčno tabelo z I vrsticami in J stolpci. V tem primeru dvorazsežni razsevni grafikon, značilen za korespondenčno analizo, vsebuje 2 množici točk: I točk, ki predstavljajo vrstice, in J točk, ki predstavljajo stolpce kontingenčne tabele.

Položaj točk v grafikonu odraža povezanost med proučevanima spremenljivkama. Tako točke vrstic, ki leže blizu skupaj, predstavljajo tiste vrstice kontingenčne tabele, ki imajo po stolpcih podobne pogojne porazdelitve (podobne vrstične profile). Na podoben način točke stolpcev, ki leže blizu skupaj, predstavljajo tiste stolpce, ki imajo po vrsticah podobne pogojne porazdelitve (podobne stolpične profile). In končno, tiste točke vrstice, ki leže blizu točk stolpcev, predstavljajo kombinacije skupin enot, ki se pojavljajo pogosteje, kot bi to pričakovali v primeru ko med proučevanima spremenljivkama ne bi bilo nobene povezanosti (v t.i. modelu neodvisnosti).

V korespondenčni analizi grafični prikaz običajno dopolnimo še s prikazom pripadajočih koordinat točk in s prikazom deležev inercij dimenzij optimalnega vektorskega prostora (deležev skupne informacije, ki je izražena s posamezno dimenzijo).

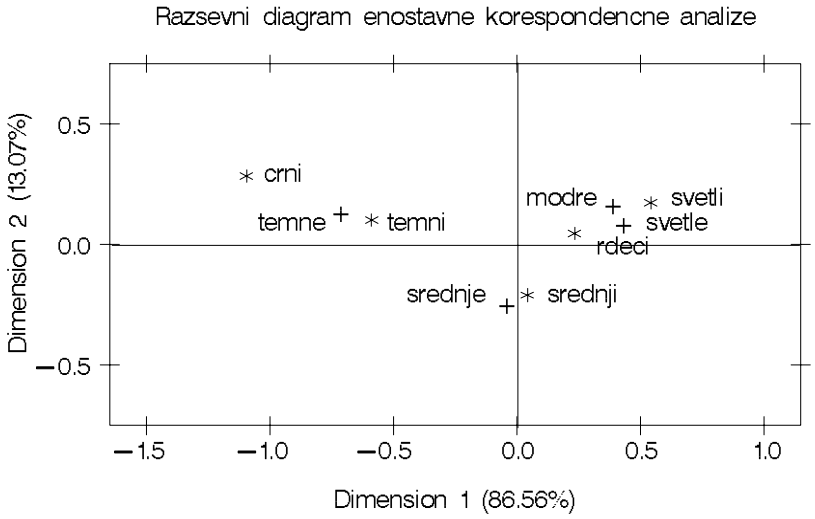
Primer 1

Analizirati želimo povezanost med barvo oči in barvo las 5387 učencev iz Caithnessa na Škotskem (Vir: R.A. Fisher, 1940). Dvorazsežna kontingenčna tabela (tabela 1.1) prikazuje vse možne kombinacije vrednosti spremenljivk barve oči in barve las in njihovo pogostnost v opazovani populaciji učencev.

Tabela 1.1: Prikaz števila učencev glede na barvo oči in barvo las (R.A. Fisher, 1940

		Barva las					
		svetli	rdeci	srednji	temni	crni	Total
Barva oci	svetle	688	116	584	188	4	1580
	modre	326	38	241	110	3	718
	srednje	343	84	909	412	26	1774
	temne	98	48	403	681	85	1315
Total		1455	286	2137	1391	118	5387

⇓ *preoblikovanje* ⇓



Slika 1.1 Prikaz profilov učencev glede na barvo oči in barvo las v optimalni ravnini

S korespondenčno analizo prikažemo odnose povezanosti med barvo oči in barvo las učencev, ki jih izraža kontingenčna tabela, kot odnose med točkami v razsevnem diagramu. Vsaka točka predstavlja skupino učencev z določeno barve oči ali las. Razlaga grafikona je podana v primeru 1, 7. nadaljevanje. ■

1.2 Izvori korespondenčne analize

Algebrske osnove korespondenčne analize so se pojavile neodvisno v delih številnih avtorjev in pod različnimi imeni in segajo nazaj v trideseta leta prejšnjega stoletja. Metode, teoretično sorodne korespondenčni analizi, najdemo pod imeni simultana linearna regresija (Hirshfeld, 1935), metoda recipročnih povprečij (Hill, 1974), dualno skaliranje (Gutmann, 1941, 1946, Nishisato, 1978) itd. Korespondenčno analizo je prav tako mogoče predstaviti kot poseben primer kanonične korelacijske analize [1].

Ime korespondenčna analiza je francoskega izvora – l'analyse des correspondances. Metodo je poimenovala skupina francoskih avtorjev, ki je na čelu z J.P. Benzécrijem v šestdesetih letih razvila predvsem geometrijsko plat korespondenčne analize. Prav možnost geometrijske predstavitve pojava pa je tista bistvena značilnost, ki loči korespondenčno analizo od že omenjenih sorodnih metod in ki predstavlja osnovo za interpretacijo proučevanih pojavov.

---

[1] Zelo popoln prikaz povezav med korespondenčno analizo in sorodnimi metodami najdemo v članku Tenenhousa in Younga (1985).

2 Enostavna korespondenčna analiza

2.1 Definicija nominalne spremenljivke

Nominalna spremenljivka je definirana kot množica V

$$V = \{v_k; k = 1,2,\ldots,K\} \tag{2.1}$$

pri čemer je v_k vrednost spremenljivke V in $K \geq 2$. Za množico V velja naslednje:

- $v_k \neq \varnothing \qquad k = 1, 2, \ldots, K$ (2.2)
- $v_k \cap v_{k'} = \varnothing \qquad k \neq k'; k, k' \in \{1, 2, \ldots, K\}$ (2.3)
- $\bigcup_{k=1}^{K} v_k = V$ (2.4)

kjer $\varnothing$ označuje prazno množico.

Nadalje, naj bo $\mathcal{O}$ množica enot

$$\mathcal{O} = \{o_h; h = 1, 2, \ldots, H\} \tag{2.5}$$

Z opisom množice $\mathcal{O}$ z množico V, ki je definiran s kartezičnim produktom $\mathcal{O} \otimes V$, prihaja do razvrstitve enot o_h $(h = 1, 2, \ldots, H)$ množice $\mathcal{O}$ v eno izmed podmnožic (skupin) $\mathcal{O}_k$ $(k = 1, 2, \ldots, K)$ množice $\mathcal{O}$. Elementi podmnožice $\mathcal{O}_k$ se ujemajo v vrednosti v_k. Glede na to, da za množico V ni definirana nobena relacija ureditve, pripadanje neke enote o_h neki od skupin $\mathcal{O}_k$ ne določa vrstnega reda enotam o_h množice $\mathcal{O}$.

Rezultati, dobljeni z opisom množice $\mathcal{O}$ z neko nominalno spremenjivko V, se običajno predstavijo z nizom

$$(o_1, v_k \mid o_1 \in \mathcal{O}_k), (o_2, v_{k'} \mid o_2 \in \mathcal{O}_{k'}), \ldots, (o_H, v_{k''} \mid o_H \in \mathcal{O}_{k''}),$$
$$k, k', k'' \in \{1, 2, \ldots, K\} \tag{2.6}$$

kjer je vsaki enoti o_h iz množice $\mathcal{O}$ pridružena njena vrednost v_k $(k = 1, 2, \ldots, K)$ nominalne spremenljivke in na podlagi le-te tudi opredeljena skupina $\mathcal{O}_k$, ki ji enota o_h pripada.

2.2 Kontingenčna tabela

Pretežni del metod, namenjenih analizi nominalnih spremenljivk, izhaja iz kontingenčnih tabel. Kontingenčna tabela posreduje hkratno frekvenčno porazdelitev proučevane množice enot glede na dve nominalni spremenljivki s po $I \geq 2$ in $J \geq 2$ vrednostmi. Ima obliko matrike reda $I \times J$, katere elementi so absolutne frekvence f_{ij} $(i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J)$ in predstavljajo število enot množice $\mathcal{O}$, za katere je značilna neka kombinacija vrednosti obeh nominalnih spremenljivk).

Nominalni spremenljivki naj bosta definirani kot $V = \{v_i; i = 1, 2, \ldots, I\}$ in $W = \{w_j;\ j = 1, 2, \ldots, J\}$. Kontingenčna tabela **F**, z elementi

$$f_{ij} = \text{Num}(\text{o}_h \in \mathcal{O}_i \cap \mathcal{O}_j)\ ;\ i = 1, 2, \ldots, I\ ;\ j = 1, 2, \ldots, J \tag{2.7}$$

ima naslednjo obliko:

$$\mathbf{F} = \left[f_{ij} \right] = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1j} & \cdots & f_{1J} \\ f_{21} & f_{22} & \cdots & f_{2j} & \cdots & f_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ f_{i1} & f_{i2} & \cdots & f_{ij} & \cdots & f_{iJ} \\ \vdots & \vdots & & \vdots & & \vdots \\ f_{I1} & f_{I2} & \cdots & f_{Ij} & \cdots & f_{IJ} \end{bmatrix} \tag{2.8}$$

Kontingenčni tabeli običajno pridružimo zbirni stolpec $\mathbf{f}_v$ z elementi $f_{i\bullet}$ in zbirno vrstico $\mathbf{f}_w$ z elementi $f_{\bullet j}$, kot tudi skupno število opazovanih enot $f_{\bullet\bullet}$, ki so definirani takole:

$$\begin{aligned} f_{i\bullet} &= \text{Num}\left(\text{o}_h \in \mathcal{O}_i\right) && i = 1, 2, \ldots, I \\ f_{\bullet j} &= \text{Num}\left(\text{o}_h \in \mathcal{O}_j\right) && j = 1, 2, \ldots, J \\ f_{\bullet\bullet} &= \text{Num}\left(\text{o}_h \in \mathcal{O}\right) && h = 1, 2, \ldots, H \end{aligned} \tag{2.9}$$

Na ta način dobimo razširjeno kontingenčno tabelo, predstavljeno kot bločno matriko naslednje oblike:

$$\left[\begin{array}{c|c} \mathbf{F} & \mathbf{f}_v \\ \hline \mathbf{f}_w' & f_{\bullet\bullet} \end{array}\right] \quad \text{oz.} \quad \left[\begin{array}{c|c} \left[f_{ij} \right] & \left[f_{i\bullet} \right] \\ \hline \left[f_{\bullet j} \right]' & f_{\bullet\bullet} \end{array}\right] \tag{2.10}$$

Elementi vektorja $\mathbf{f}_v$ (zbirnega stolpca) so absolutne frekvence $f_{i\bullet}$ ($i = 1, 2, \ldots, I$) (robne oz. marginalne frekvence) in predstavljajo število enot, ki pripada skupini $\mathcal{O}_i$ spremenljivke V. Določene so kot vsota absolutnih frekvenc po vrsticah kontingenčne tabele

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij} \qquad i = 1, 2, \ldots, I \tag{2.11}$$

Elementi vektorja $\mathbf{f}_w$ (zbirne vrstice) so absolutne frekvence $f_{\bullet j}$ ($j = 1, 2, \ldots, J$) (robne oz. marginalne frekvence) in predstavljajo število enot, ki pripada skupini $\mathcal{O}_j$ spremenljivke W. Določene so kot vsota absolutnih frekvenc po stolpcih kontingenčne tabele

$$f_{\bullet j} = \sum_{i=1}^{I} f_{ij} \qquad j = 1, 2, \ldots, J \tag{2.12}$$

Skupno število opazovanih enot $f_{\bullet\bullet}$ pa lahko določimo kot vsoto vseh absolutnih frekvenc kontingenčne tabele

$$f_{\bullet\bullet} = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \tag{2.13}$$

Primer 1 - 1. nadaljevanje

Elementi razširjene kontingenčne tabele (tabela 1.1) imajo naslednji pomen:

$f_{12} = 116$ - svetle oči in rdeče lase ima 116 učencev,

$f_{1\bullet} = 1580$ - svetle oči ima 1580 učencev,

$f_{\bullet 2} = 286$ - rdeče lase ima 286 učencev,

$f_{\bullet\bullet} = 5387$ - skupno število opazovanih učencev je 5387. ■

Marginalne frekvence (frekvence zbirne vrstice in zbirnega stolpca) predstavljajo omejitve v variiranju absolutnih frekvenc f_{ij}, saj v vsaki vrstici kontingenčne tabele (matrike **F**) lahko prosto variira samo $J - 1$ elementov, v vsakem stolpcu pa samo $I - 1$ elementov. Skupno število stopinj prostosti m kontingenčne tabele reda $I \times J$ je zato

$$m = (I - 1)(J - 1). \tag{2.14}$$

2.3 Relativne frekvence in ocene verjetnosti

Če elemente matrike **F**, torej absolutne frekvence f_{ij}, delimo s skupnim številom opazovanih enot $f_{..}$, dobimo matriko **P**, katere elementi so relativne frekvence p_{ij} :

$$\mathbf{P} = \frac{1}{f_{..}}\mathbf{F} = \left[\frac{f_{ij}}{f_{..}}\right] = \left[p_{ij}\right] \qquad \begin{matrix} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{matrix} \tag{2.15}$$

Predpostavimo, da je O slučajni vzorec iz neke znane končne populacije $\mathcal{O}$. V tem primeru imajo elementi matrike **F** polinomsko porazdelitev z verjetnostmi $\tilde{p}_{ij}$, pri čemer je $i = 1, 2, \ldots, I$ in $j = 1, 2, \ldots, J$. Za velik slučajni vzorec O elementi matrike relativnih frekvenc **P**, to je p_{ij}, predstavljajo nepristranske ocene verjetnosti $\tilde{p}_{ij}$, pri čemer se predpostavlja, da verjetnosti v vektorjih

$$\mathbf{p}_v = \frac{1}{f_{..}}\mathbf{f}_v \quad \text{in/ali} \quad \mathbf{p}_w = \frac{1}{f_{..}}\mathbf{f}_w \tag{2.16}$$

niso vnaprej fiksirane.

Razširjena matrika relativnih frekvenc (ocen verjetnosti) ima naslednjo obliko:

$$\left[\begin{array}{c|c} \mathbf{P} & \mathbf{p}_v \\ \hline \mathbf{p}'_w & 1 \end{array}\right] \quad \text{oz.} \quad \left[\begin{array}{c:c} \left[p_{ij}\right] & \left[p_{i\bullet}\right] \\ \hdashline \left[p_{\bullet j}\right]' & 1 \end{array}\right] \qquad \begin{matrix} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{matrix} \tag{2.17}$$

Elementi vektorja $\mathbf{p}_v$ (zbirnega stolpca) so marginalne relativne frekvence $p_{i\bullet}$ ($i = 1, 2, \ldots, I$) in predstavljajo delež števila enot *i*-te skupine po spremenljivki, ki je v čelu tabele, v skupnem številu opazovanih enot

$$p_{i\bullet} = \frac{f_{i\bullet}}{f_{..}} \qquad i = 1, 2, \ldots, I \tag{2.18}$$

Elementi vektorja $\mathbf{p}_w$ (zbirne vrstice) so marginalne relativne frekvence $p_{\bullet j}$ ($j = 1, 2, \ldots, J$) in predstavljajo delež števila enot *j*-te skupine po spremenljivki, ki je v glavi tabele, v skupnem številu opazovanih enot

$$p_{\bullet j} = \frac{f_{\bullet j}}{f_{..}} \qquad j = 1, 2, \ldots, J \tag{2.19}$$

Vsota vseh relativnih frekvenc (ocen verjetnosti) je razumljivo enaka 1

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = \frac{1}{f_{\bullet\bullet}}\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = 1 \qquad (2.20)$$

Prav tako sta enaki 1 tudi vsoti marginalnih relativnih frekvenc zbirne vrstice oz. zbirnega stolpca

$$\sum_{j=1}^{J} p_{\bullet j} = 1 \qquad (2.21)$$

$$\sum_{i=1}^{I} p_{i\bullet} = 1 \qquad (2.22)$$

Primer 1 - 2. nadaljevanje

Izračunajmo razširjeno matriko relativnih frekvenc za vse možne kombinacije vrednosti spremenljivk barve oči in barve las v opazovani populaciji učencev (tabela 2.1).

Tabela 2.1: Prikaz relativnih frekvenc učencev glede na barvo oči in barvo las (Vir: tabela 1.1)

		Barva las					
		svetli	rdeci	srednji	temni	crni	Total
Barva oci	svetle	,1277	,0215	,1084	,0349	,0007	,2933
	modre	,0605	,0071	,0447	,0204	,0006	,1333
	srednje	,0637	,0156	,1687	,0765	,0048	,3293
	temne	,0182	,0089	,0748	,1264	,0158	,2441
Total		,2701	,0531	,3967	,2582	,0219	1,0000

Elementi razširjene matrike relativnih frekvenc imajo naslednji pomen:

$p_{12} = \frac{f_{12}}{f_{\bullet\bullet}} = \frac{116}{5387} = 0,0215$ - delež učencev s svetlimi očmi in rdečimi lasmi je 0,0215,

$p_{1\bullet} = \frac{f_{1\bullet}}{f_{\bullet\bullet}} = \frac{1580}{5387} = 0,2933$ - delež učencev s svetlimi očmi je 0,2933,

$p_{\bullet 2} = \frac{f_{\bullet 2}}{f_{\bullet\bullet}} = \frac{286}{5387} = 0,0531$ - delež učencev z rdečimi lasmi je 0,0531. ■

2.4 Profili

Analizo lahko pričnemo ali s proučevanjem odnosov različnih skupin enot s stališča posameznih vrednosti prve spremenljivke glede na vrednosti druge spremenljivke ali pa obratno, s proučevanjem odnosov različnih skupin enot s stališča posameznih vrednosti druge spremenljivke glede na vrednosti prve spremenljivke. Tako lahko npr. pričnemo s proučevanjem odnosov različnih skupin učencev s stališča njihove barve oči glede na barvo las ali pa s stališča njihove barve las glede na barvo oči (glej tabelo 1.1). Ker lahko, kot bomo videli kasneje, korespondenčno analizo obravnavamo kot metodo najboljšega možnega hkratnega prikaza dveh množic podatkov, ki se nanašajo na vrstice oz. stolpce kontingenčne tabele, sta oba prikaza enakovredna.

Izberimo enega izmed obeh pristopov. Kontingenčna tabela predstavlja matriko podatkov z I vrsticami in J stolpci. Vrstice te matrike predstavljajo vektorje (oz. vektorje-točke) v vektorskem prostoru R^J, gre torej za I vektorjev v J-razsežnem vektorskem prostoru. Tako lahko npr. porazdelitev učencev določene barve oči glede na barvo las predstavimo kot točko v pet-razsežnem vektorskem prostoru. Če so komponente vektorjev v R^J absolutne frekvence f_{ij}, potem razlik med vrsticami matrike v splošnem ne moremo smiselno primerjati. Tako se npr. število učencev po barvi oči precej razlikuje (glej zbirni stolpec tabele 1.1) in so zato frekvence v vrsticah, ki prikazujejo porazdelitev manjših skupin učencev (s stališča barve oči) glede na barvo las nujno manjše, pripadajoči vektorji-točke pa bliže koordinatnemu izhodišču vektorskega prostora.

Zato bomo kot izhodišče analize namesto absolutnih frekvenc vzeli pogojne vrstične relativne frekvence $p_{ij,i\bullet}$, ki jih izračunamo tako, da:

- elemente kontingenčne tabele **F** (2.8), torej absolutne frekvence f_{ij}, delimo z odgovarjajočimi elementi zbirnega stolpca $\mathbf{f}_v$, torej z marginalnimi absolutnimi frekvencami $f_{i\bullet}$ ali
- elemente matrike **P** (2.15), torej relativne frekvence p_{ij}, delimo z odgovarjajočimi elementi zbirnega stolpca $\mathbf{p}_v$, torej z marginalnimi relativnimi frekvencami $p_{i\bullet}$.

$$p_{ij,i\bullet} = \frac{f_{ij}}{f_{i\bullet}} = \frac{p_{ij}}{p_{i\bullet}} \qquad \begin{matrix} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{matrix} \tag{2.23}$$

oz.

$$\mathbf{P}_{w.v} = \left[p_{ij,i\bullet} \right] = \mathbf{D}_{\mathbf{f}_v}^{-1}\mathbf{F} = \mathbf{D}_v^{-1}\mathbf{P}$$

pri čemer je $\mathbf{D}_{\mathbf{f}_v}$ diagonalna matrika marginalnih absolutnih frekvenc $f_{i\bullet}$.

$$\mathbf{D}_{\mathbf{f}_v} = \mathrm{diag}(f_{1\bullet}, f_{2\bullet}, \ldots, f_{I\bullet}) \tag{2.24}$$

in $\mathbf{D}_v$ diagonalna matrika marginalnih relativnih frekvenc $p_{i\bullet}$.

$$\mathbf{D}_v = \mathrm{diag}(p_{1\bullet}, p_{2\bullet}, \ldots, p_{I\bullet}) \tag{2.25}$$

Matriko pogojnih vrstičnih relativnih frekvenc $\mathbf{P}_{w.v}$

$$\mathbf{P}_{w.v} = \left[p_{ij,i\bullet} \right] = \begin{bmatrix} \mathbf{p}'_{w.1} \\ \mathbf{p}'_{w.2} \\ \vdots \\ \mathbf{p}'_{w.I} \end{bmatrix} \tag{2.26}$$

v korespondenčni analizi imenujemo matrika vrstičnih profilov, njene vrstice $\mathbf{p}_{w.i}$ ($i = 1, 2, \ldots, I$) pa vrstični profili.

Na podoben način kot v primeru razširjene matrike relativnih frekvenc (2.17) lahko sedaj opredelimo razširjeno matriko pogojnih vrstičnih relativnih frekvenc, ki ima naslednjo obliko:

$$\left[\begin{array}{c|c} \mathbf{P}_{w.v} & \mathbf{1} \\ \hline \mathbf{p}'_w & 1 \end{array}\right] \quad \text{oz.} \quad \left[\begin{array}{c|c} \left[p_{ij,i\bullet} \right] & [1] \\ \hline \left[p_{\bullet j} \right]' & 1 \end{array}\right] \qquad \begin{matrix} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{matrix} \tag{2.27}$$

V korespondenčni analizi imenujemo zgornjo matriko razširjena matrika vrstičnih profilov.

Če pa predpostavimo, da je iz populacije $\mathcal{O}$ izbran slučajni vzorec O in če so marginalne absolutne frekvence $f_{i\bullet}$ iz zbirnega stolpca $\mathbf{f}_v$ vnaprej fiksirane oz. marginalne relativne frekvence $p_{i\bullet}$ iz zbirnega stolpca $\mathbf{p}_v$ vnaprej poznane, potem bodo elementi kontingenčne tabele $\mathbf{F}$ porazdeljeni v polinomski porazdelitvi z verjetnostmi $\tilde{p}_{ij,i\bullet}$, ob omejitvah

$$\sum_{j=1}^{J} f_{ij} = f_{i\bullet} \qquad i = 1, 2, \ldots, I \tag{2.28}$$

in

$$\sum_{j=1}^{J} \tilde{p}_{ij,i\bullet} = 1 \qquad i = 1, 2, \ldots, I \tag{2.29}$$

Matrika ocen pogojnih verjetnosti $\mathbf{P}_{w.v}$ in razširjena matrika ocen pogojnih verjetnosti imata formalno isto obliko kot v izrazih (2.26) in (2.27), s tem da so sedaj $p_{ij,i\bullet}$ nepristranske ocene pogojnih verjetnosti in $p_{\bullet j}$ nepristranske ocene marginalnih verjetnosti.

Primer 1 - 3. nadaljevanje

Izračunajmo razširjeno matriko pogojnih vrstičnih relativnih frekvenc oz. razširjeno matriko vrstičnih profilov (tabela 2.2) – proučiti želimo odnose med skupinami učencev z različno barvo oči glede na barvo las.

Tabela 2.2 Prikaz profilov učencev glede na barvo las (Vir: tabela 1.1)

		Barva las					
		svetli	rdeci	srednji	temni	crni	Total
Barva oci	svetle	,4354	,0734	,3696	,1190	,0025	1,0000
	modre	,4540	,0529	,3357	,1532	,0042	1,0000
	srednje	,1933	,0474	,5124	,2322	,0147	1,0000
	temne	,0745	,0365	,3065	,5179	,0646	1,0000
Total		,2701	,0531	,3967	,2582	,0219	1,0000

Npr. $p_{12,1\bullet} = \dfrac{f_{12}}{f_{1\bullet}} = \dfrac{116}{1580} = 0,0734$.

Delež učencev z rdečimi lasmi med učenci s svetlimi očmi je 0,0734 . ■

Pojasnimo še oblikovanje stolpičnih profilov. Pogojne stolpične relativne frekvence $p_{ij,\bullet j}$ izračunamo tako, da:

- elemente kontingenčne tabele **F** (2.8), torej absolutne frekvence f_{ij}, delimo z odgovarjajočimi elementi zbirne vrstice $\mathbf{f}_w$, torej z marginalnimi absolutnimi frekvencami $f_{\bullet j}$ ali
- elemente matrike **P** (2.15), torej relativne frekvence p_{ij}, delimo z odgovarjajočimi elementi zbirne vrstice $\mathbf{p}_w$, torej z marginalnimi relativnimi frekvencami $p_{\bullet j}$

$$p_{ij,\bullet j} = \frac{f_{ij}}{f_{\bullet j}} = \frac{p_{ij}}{p_{\bullet j}} \qquad \begin{matrix} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{matrix} \tag{2.30}$$

oz.

$$\mathbf{P}_{v.w} = \left[\, p_{ij,\bullet j} \,\right] = \mathbf{F}\,\mathbf{D}_{\mathbf{f}_w}^{-1} = \mathbf{P}\,\mathbf{D}_w^{-1}$$

pri čemer je $\mathbf{D}_{\mathbf{f}_w}$ diagonalna matrika marginalnih absolutnih frekvenc $f_{\bullet j}$

$$\mathbf{D}_{\mathbf{f}_w} = \mathrm{diag}(f_{\bullet 1}, f_{\bullet 2}, \ldots, f_{\bullet J}) \tag{2.31}$$

in $\mathbf{D}_w$ diagonalna matrika marginalnih relativnih frekvenc $p_{\bullet j}$

$$\mathbf{D}_w = \mathrm{diag}(p_{\bullet 1}, p_{\bullet 2}, \ldots, p_{\bullet J}) \tag{2.32}$$

Matriko pogojnih stolpičnih relativnih frekvenc $\mathbf{P}_{v.w}$

$$\mathbf{P}_{v.w} = \left[\, p_{ij,\bullet j} \,\right] = \left[\mathbf{p}_{v.1} \quad \mathbf{p}_{v.2} \quad \cdots \quad \mathbf{p}_{v.J}\right] \tag{2.33}$$

v korespondenčni analizi imenujemo matrika stolpičnih profilov, njene stolpce $\mathbf{p}_{v.j}$ ($j = 1, 2, \ldots, J$) pa stolpični profili.

Na podoben način kot v primeru razširjene matrike relativnih frekvenc (2.17) lahko sedaj opredelimo razširjeno matriko pogojnih stolpičnih relativnih frekvenc , ki ima naslednjo obliko:

$$\left[\begin{array}{c|c} \mathbf{P}_{v.w} & \mathbf{p}_v \\ \hline \mathbf{1}' & 1 \end{array}\right] \quad \text{oz.} \quad \left[\begin{array}{c|c} \left[\, p_{ij,\bullet j} \,\right] & \left[\, p_{i\bullet} \,\right] \\ \hline \left[1\right]' & 1 \end{array}\right] \qquad \begin{array}{l} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{array} \tag{2.34}$$

V korespondenčni analizi imenujemo zgornjo matriko razširjena matrika stolpičnih profilov.

Če pa predpostavimo, da je iz populacije $\mathcal{O}$ izbran slučajni vzorec O in če so marginalne absolutne frekvence $f_{\bullet j}$ iz zbirne vrstice $\mathbf{f}_w'$ vnaprej fiksirane oz. marginalne relativne frekvence $p_{\bullet j}$ iz zbirne vrstice $\mathbf{p}_w'$ vnaprej poznane, potem bodo elementi kontingenčne tabele **F** porazdeljeni v polinomski porazdelitvi z verjetnostmi $\tilde{p}_{ij,\bullet j}$, ob omejitvah

$$\sum_{i=1}^{I} f_{ij} = f_{\bullet j} \qquad j = 1, 2, \ldots, J \tag{2.35}$$

in

$$\sum_{i=1}^{I} \tilde{p}_{ij,\bullet j} = 1 \qquad j = 1, 2, \ldots, J \tag{2.36}$$

Matrika ocen pogojnih verjetnosti $\mathbf{P}_{v.w}$ in razširjena matrika ocen pogojnih verjetnosti imata formalno isto obliko kot v izrazih (2.33) in (2.34), s tem da so

sedaj $p_{ij,\bullet j}$ nepristranske ocene pogojnih verjetnosti in $p_{i\bullet}$ nepristranske ocene marginalnih verjetnosti.

Primer 1 - 4. nadaljevanje

Izračunajmo razširjeno matriko pogojnih stolpičnih relativnih frekvenc (tabela 2.3) oz. razširjeno matriko stolpičnih profilov – proučiti želimo odnose med skupinami učencev z različno barvo las glede na barvo oči.

Tabela 2.3: Prikaz profilov učencev glede na barvo oči (Vir: tabela 1.1)

		Barva las					
		svetli	rdeci	srednji	temni	crni	Total
Barva oci	svetle	,4729	,4056	,2733	,1352	,0339	,2933
	modre	,2241	,1329	,1128	,0791	,0254	,1333
	srednje	,2357	,2937	,4254	,2962	,2203	,3293
	temne	,0674	,1678	,1886	,4896	,7203	,2441
Total		1,0000	1,0000	100,0000	1,0000	1,0000	1,0000

Npr. $p_{12,\bullet 2} = \dfrac{f_{12}}{f_{\bullet 2}} = \dfrac{116}{286} = 0,4056$

Delež učencev s svetlimi očmi med učenci z rdečimi lasmi je 0,4056 . ■

Posvetimo sedaj ponovno pozornost zbirnim vrsticam in stolpcem že omenjenih tabel. V razširjeni matriki pogojnih vrstičnih relativnih frekvenc (2.27) je zbirni stolpec vektor števil 1, ker je vsota pogojnih relativnih frekvenc (pogojnih verjetnosti) vsake vrstice enaka 1. Isti zaključek velja tudi za vsoto elementov zadnje, to je zbirne vrstice. Elementi te vrstice $p_{\bullet j}$ ($j = 1, 2, \ldots, J$) so namreč relativne frekvence (marginalne verjetnosti) skupin, določene na podlagi vrednosti spremenljivke, ki je v glavi tabele, glede na skupno število opazovanih enot. Elementi zbirne vrstice niso odvisni od vrednosti spremenljivke, ki je v čelu tabele. Zbirna vrstica $\mathbf{p}_w$ prestavlja torej specifičen profil, ki ima značaj povprečnega profila glede na vse opazovane enote, gledano s stališča spremenljivke v glavi tabele. Do simetričnih zaključkov bomo prišli, če opazujemo zbirni stolpec $\mathbf{p}_v$ v tabeli (2.34).

Zbirna vrstica $\mathbf{p}_w$ razširjene matrike vrstičnih profilov (2.27) oz. zbirni stolpec $\mathbf{p}_v$ razširjene matrike stolpičnih profilov (2.34) torej predstavljata specifična – povprečna profila, vselej glede na eno spremenljivko in neodvisno od druge

spremenljivke. Postavlja se vprašanje, kakšen je potem odnos med zbirno vrstico in vrstičnimi profili oz. zbirnim stolpcem in stolpčnimi profili.

Začnimo z zbirno vrstico $\mathbf{p}_w$. Izkaže se, da elemente zbirne vrstice $p_{\bullet j}$ ($j = 1, 2, \ldots, J$) ne moremo izračunati kot enostavne aritmetične sredine odgovarjajočih elementov vrstičnih profilov. Vzrok za to je v dejstvu, da smo pri določanju profilov kot vrstic pogojnih relativnih frekvenc (vrstic pogojnih verjetnosti) dodelili vsem profilom enak pomen. Zato moramo, če želimo izračunati povprečni profil (zbirno vrstico) na podlagi vrstičnih profilov, vsak profil tehtati v skladu z njegovim relativnim pomenom. Ta pomen pa je proporcionalen tisti marginalni frekvenci $f_{i\bullet}$ ($i = 1, 2, \ldots, I$), na podlagi katere je profil določen. Uporabiti moramo torej tehtano aritmetično sredino vrstičnih profilov, kjer so uteži elementi zbirnega stolpca $\mathbf{p}_v$, torej marginalne relativne frekvence (marginalne verjetnosti) $p_{i\bullet}$ ($i = 1, 2, \ldots, I$) glede na spremenljivko, ki je v čelu tabele (glej zbirni stolpec v matriki 2.17 oz. 2.34)

$$p_{\bullet j} = \sum_{i=1}^{I} p_{ij,i\bullet} p_{i\bullet} \qquad j = 1, 2, \ldots, J \tag{2.37}$$

Uteži $p_{i\bullet}$ se v korespondenčni analizi imenujejo mase.

Primer 1 - 5. nadaljevanje

Izračunajmo element $p_{\bullet 2}$ zbirne vrstice $\mathbf{p}_w$ razširjene matrike vrstičnih profilov (tabela 2.2)

$$p_{\bullet 2} = \sum_{i=1}^{I} p_{i2,i\bullet} p_{i\bullet} = 0,0734 \bullet 0,2933 + \ldots + 0,0365 \bullet 0,2441 = 0,0531 \qquad \blacksquare$$

Simetrični zaključki veljajo tudi za zbirni stolpec $\mathbf{p}_v$ oz. povprečni profil glede na spremenljivko, ki je v čelu tabele

$$p_{i\bullet} = \sum_{j=1}^{J} p_{ij,\bullet j} p_{\bullet j} \qquad i = 1, 2, \ldots, I \tag{2.38}$$

Povprečni profili, ki smo jih izračunali kot tehtana povprečja profilov (oz. vektorjev-točk) z utežmi, katerih vsota je enaka 1, dejansko predstavljajo

centroide oz. gravitacijske točke [2] .

2.5 χ^2- razdalja

Spomnimo se, da npr. profili vrstic odražajo položaj vektorjev točk v *J*-razsežnem vektorskem prostoru R^J. Zaželeno je, da oddaljenost med temi točkami odraža podobnost (relativni položaj) profilov na najboljši mogoči način.

Da bi to dosegli, bomo uporabili χ^2-razdaljo. Ker je izbor tega specifičnega tipa razdalje bistvena značilnost korespondenčne analize, bomo njeno določanje podrobno pojasnili.

Razdalja je vrednost, ki numerično izraža oddaljenost dveh vektorjev točk. V primeru, ko gre za evklidsko razdaljo, je izražena na naslednji način

$$d(\mathbf{a},\mathbf{b}) = \left[(\mathbf{a}-\mathbf{b})'(\mathbf{a}-\mathbf{b})\right]^{1/2} = \left[\sum_{t=1}^{T}(a_t - b_t)^2\right]^{1/2} \qquad (2.39)$$

kjer je:

$d(\mathbf{a},\mathbf{b})$ - evklidska razdalja in

$(\mathbf{a}-\mathbf{b})$ - vektor razlik vektorskih komponent vektorjev **a** in **b** .

Znano je, da ni zaželeno, da bi bila razdalja odvisna od merskih enot, v katerih so izražene vrednosti spremenljivk (komponente vektorjev-točk). Tej pomankljivosti se običajno izognemo tako, da že pred izračunom evklidske razdalje posamezne vrednosti spremenljivk delimo z njihovimi standardnimi odkloni. Tako določene standardizirane vrednosti vektorskih komponent so invariantne na izbor merskih enot.

Gornji pristop lahko v širšem smislu ekvivalentno definiramo kot pristop, pri

---

[2] Centroid $\bar{\mathbf{c}}$ je linearna kombinacija vektorjev-točk $\mathbf{c}_k$, za katere velja, da je vsota njihovih uteži α_k enaka 1

$$\bar{\mathbf{c}} = \sum_{k=1}^{K}\alpha_k\mathbf{c}_k \quad ; \quad \sum_{k=1}^{K}\alpha_k = 1$$

katerem prihaja do uporabe medsebojno različnih uteži glede na koordinatne osi (vrednosti spremenljivk se v tem primeru vnaprej ne standardizirajo). Tako določena razdalja se imenuje tehtana evklidska razdalja

$$d(\mathbf{a},\mathbf{b}) = \left[(\mathbf{a}-\mathbf{b})' \mathbf{D}_q (\mathbf{a}-\mathbf{b}) \right]^{1/2} = \left[\sum_{t=1}^{T} q_t (a_t - b_t)^2 \right]^{1/2} \qquad (2.40)$$

kjer je $\mathbf{D}_q$ diagonalna matrika, njeni diagonalni elementi $q_1, q_2, \ldots, q_T$ pa so pozitivna realna števila (npr. inverzne vrednosti varianc) in imajo vlogo uteži odgovarjajočih *T* razsežnosti.

Eden od najpogostejših primerov uporabe tehtane evklidske razdalje je statistika χ^2, namenjena testiranju hipoteze, da neka stvarna frekvenčna porazdelitev ustreza neki pričakovani frekvenčni porazdelitvi - $\mathrm{H}_0 : f_g = f_g^*$ ($g = 1, 2, \ldots, G$)

$$\chi^2 = \sum_{g=1}^{G} \frac{(f_g - f_g^*)^2}{f_g^*} \qquad (2.41)$$

kjer je:

f_g - opazovana absolutna frekvenca v *g*-ti skupini in

f_g^* - pričakovana absolutna frekvenca v *g*-ti skupini.

Ker gre za statistiko, ki temelji na primerjavi dveh vektorjev absolutnih frekvenc, lahko gornji izraz zapišemo v obliki

$$\chi^2 = (\mathbf{f} - \mathbf{f}^*)' \mathbf{D}_{\mathbf{f}^*}^{-1} (\mathbf{f} - \mathbf{f}^*) \qquad (2.42)$$

kjer je:

f - vektor opazovanih absolutnih frekvenc,

$\mathbf{f}^*$ - vektor pričakovanih absolutnih frekvenc in

$\mathbf{D}_{\mathbf{f}^*}$ - diagonalna matrika s pričakovanimi absolutnimi frekvencami.

Statistika χ^2 je torej kvadrat razdalje med vektorji opazovanih in pričakovanih absolutnih frekvenc v tehtanem evklidskem vektorskem prostoru, z inverznimi vrednostmi pričakovanih frekvenc kot utežmi.

Če delimo vektorje absolutnih frekvenc **f** in $\mathbf{f}^*$ s številom opazovanih enot $f_\bullet$, dobimo naslednje vektorje:

$$\mathbf{p} = \frac{1}{f_\bullet} \mathbf{f} \qquad (2.43)$$

$$\mathbf{p}^* = \frac{1}{f_\bullet}\mathbf{f}^* \tag{2.44}$$

kjer je:

$\mathbf{p}$ – vektor opazovanih relativnih frekvenc in

$\mathbf{p}^*$ - vektor pričakovanih relativnih frekvenc.

Statistika χ^2 dobi potem naslednjo obliko

$$\chi^2 = f_\bullet(\mathbf{p}-\mathbf{p}^*)'\mathbf{D}_{\mathbf{p}^*}^{-1}(\mathbf{p}-\mathbf{p}^*) = f_\bullet\sum_{g=1}^{G}\frac{(p_g - p_g^*)^2}{p_g^*} \tag{2.45}$$

kjer je:

$\mathbf{D}_{\mathbf{p}^*}$ – diagonalna matrika s pričakovanimi relativnimi frekvencami,

p_g - opazovana relativna frekvenca v *g*-ti skupini in

p_g^* - pričakovana relativna frekvenca v *g*-ti skupini.

Izraz (2.45) lahko obravnavamo kot produkt dveh faktorjev

$$\chi^2 = f_\bullet \times d^2(\mathbf{p},\mathbf{p}^*) \tag{2.46}$$

kjer je prvi faktor $f_\bullet$ skupno število opazovanih enot, drugi faktor pa

$$d^2(\mathbf{p},\mathbf{p}^*) = (\mathbf{p}-\mathbf{p}^*)'\mathbf{D}_{\mathbf{p}^*}^{-1}(\mathbf{p}-\mathbf{p}^*) \tag{2.47}$$

Izraz (2.47) predstavlja kvadrat tehtane evklidske razdalje, kjer so uteži inverzne vrednosti pričakovanih relativnih frekvenc. Ker je proporcionalen statistiki χ^2, se ta funkcija razdalje imenuje tudi χ^2-razdalja. Faktor proporcionalnosti v izrazu (2.46) je skupno število opazovanih enot $f_\bullet$, s čimer je velikost vzorca vključena v mero razlik med opazovanimi in pričakovanimi relativnimi frekvencami.

2.6 Določanje uteži vektorjem-točkam

Za korespondenčno analizo sta značilni dve vrsti tehtanja:

- tehtanje razsežnosti (osi oz. spremenljivk) in
- tehtanje vektorjev-točk.

Tehtanja izvornih razsežnosti vektorskega prostora smo spoznali v prejšnjem

razdelku, ko smo definirali χ^2- razdaljo . V nadaljevanju pa bomo prikazali še drugo vrsto tehtanja, to je tehtanje vektorjev-točk.

Tehtanje točk v skladu z njihovim pomenom s stališča uporabe neke konkretne metode se v statistiki pogosto uporablja. V našem primeru s tem, ko določamo različne uteži vektorjem-točkam, dejansko pripišemo različen pomen položaju posameznih točk v vektorskem prostoru.

Kot bomo videli kasneje, želimo v korespondenčni analizi določiti takšen vektorski prostor malih razsežnosti, ki se čim bolje prilega množici vektorjev-točk v vektorskem prostoru. Če so uteži točk medsebojno različne, lahko seveda upravičeno pričakujemo, da bo imel vektorski prostor malih razsežnosti takšen položaj, ki bo bliže položaju tistih točk, ki imajo večje uteži.

Kot smo že pojasnili, je položaj točk v korespondenčni analizi določen na podlagi profilov, ki predstavljajo koordinate vektorjev-točk v vektorskih prostorih R^I oz. R^J. Ker je vsota vrednosti vsakega profila enaka 1, imajo vse točke na ta način enak pomen (težo).

Da bi dosegli čim boljšo prilagoditev vektorskega prostora malih razsežnosti množici točk v prostorih R^I oz. R^J, je najbolje, da določimo vsakemu vektorju-točki takšno utež, ki bo proporcionalna tisti absolutni frekvenci, glede na katero je profil določen. Gre torej za marginalno frekvenco, katero smo uporabili kot imenovalec pri izračunu relativnih frekvenc oz. profilov, torej koordinat odgovarjajoče točke. Na ta način bomo ustrezno predstavili porazdelitev populacije.

V bivariatni korespondenčni analizi nastopata dva vektorska prostora R^I in R^J. V prostoru R^J so ustrezne uteži vektorjev-točk relativne frekvence (marginalne verjetnosti) $p_{i\bullet}$ $(i = 1, 2, \ldots, I)$ kategoriji [3] tiste spremenljivke, ki je v čelu kontingenčne tabele. Te uteži so proporcionalne absolutnim frekvencam omenjenih kategorij. Če se vrnemo k matriki vrstičnih profilov

---

[3] Kategorija je vrednost spremenljivke, dobljena tako, da je vsaki skupini, v katero so enote razvrščene po vrednosti ene spremenljivke ali več spremenljivk, prirejena nova vrednost, npr. moški, stari do 18 let (Košmelj et al.: Statistični terminološki slovar, 2001)

(tabela 2.2) iz primera 1, to je vektorskemu prostoru R^5, potem števila $p_{i\bullet}$ $(i = 1, 2, \ldots, 4)$ iz zbirnega stolpca tabele 2.1 predstavljajo uteži točk. Analogno so v prostoru R^I uteži vektorjev-točk relativne frekvence (marginalne verjetnosti) $p_{\bullet j}$ $(j = 1, 2, \ldots, J)$, določene glede na kategorije spremenljivke, ki je v glavi kontingenčne tabele. S stališča matrike stolpičnih profilov (tabela 2.3), v prostoru R^4, so to vrednosti zbirne vrstice tabele 2.1, torej vrednosti $p_{\bullet j}$ $(j = 1, 2, \ldots, 5)$.

Tako določene uteži, ki jih, kot smo že omenili, v korespondenčni analizi imenujemo mase, bomo uporabljali pri določanju optimalnih podprostorov malih razsežnosti kot tudi pri izračunavanju centroida, o čemer smo že govorili.

2.7 Inercija

2.7.1 Definiranje inercije

Spomnimo se izrazov (2.45) in (2.47), iz katerih sledi, da je

$$\chi^2 = f_{\bullet}(\mathbf{p} - \mathbf{p}^*)' \mathbf{D}_{\mathbf{p}^*}^{-1}(\mathbf{p} - \mathbf{p}^*) = f_{\bullet}\, d^2(\mathbf{p}, \mathbf{p}^*) \qquad (2.48)$$

Statistika χ^2 je torej v tem primeru produkt števila opazovanih enot $f_{\bullet}$ in kvadrata tehtane razdalje vektorja-točke $\mathbf{p}$, ki odraža opazovane relativne frekvence (opazovani profil), in vektorja-točke $\mathbf{p}^*$, ki odraža pričakovane relativne frekvence (pričakovani profil) oz. χ^2- razdalje.

Predpostavimo, da naša analiza namesto enega profila vključuje K profilov ($\mathbf{p}_k$; $k = 1, 2, \ldots, K$). To pomeni, da lahko za vsako skupino opazovanih enot, ki jim je skupno to, da se nanašajo na isto vrednost nominalne spremenljivke (isto kategorijo v_k), izračunamo statistiko χ_k^2:

$$\chi_k^2 = f_k(\mathbf{p}_k - \bar{\mathbf{p}})' \mathbf{D}_{\bar{\mathbf{p}}}^{-1}(\mathbf{p}_k - \bar{\mathbf{p}}) \qquad k = 1, 2, \ldots, K \qquad (2.49)$$

kjer je f_k število enot v k-ti skupini in $\bar{\mathbf{p}}$ vektor povprečnih relativnih frekvenc, torej povprečni profil oz. centroid

$$\bar{\mathbf{p}} = \sum_{k=1}^{K} \alpha_k \mathbf{p}_k \qquad (2.50)$$

Če seštejemo vrednosti statistik χ_k^2 $(k = 1, 2, \ldots, K)$, dobimo izraz

$$\chi^2 = \sum_{k=1}^{K} \chi_k^2 = \sum_{k=1}^{K} f_k (\mathbf{p}_k - \overline{\mathbf{p}})' \mathbf{D}_{\overline{\mathbf{p}}}^{-1} (\mathbf{p}_k - \overline{\mathbf{p}}) \tag{2.51}$$

Na ta način statistika χ^2 [4] odraža skupne razlike, ki nastajajo med profili (vektorji-točkami). Lahko pa ga opišemo tudi kot tehtano vsoto kvadratov razlik med vektorji $\mathbf{p}_k$ ($k = 1, 2, \ldots, K$) in vektorjem $\overline{\mathbf{p}}$ v $\mathbf{D}_{\overline{\mathbf{p}}}$ metriki.

Na podlagi statistike χ^2 lahko izračunamo t.i. skupno inercijo ali enostavno inercijo *In*(*K*) [5]

$$In(K) = \frac{\chi^2}{f_\bullet} \tag{2.52}$$

množice *K* profilov oz. vektorjev-točk, kjer je $f_\bullet$ skupno število opazovanih enot. Inercija je namreč mera oddaljenosti profilov od povprečnega profila oz. centroida.

Ker skupno inercijo lahko izrazimo tudi kot

$$In(K) = \sum_{k=1}^{K} \frac{f_k}{f_\bullet} (\mathbf{p}_k - \overline{\mathbf{p}})' \mathbf{D}_{\overline{\mathbf{p}}}^{-1} (\mathbf{p}_k - \overline{\mathbf{p}}) = \sum_{k=1}^{K} \frac{f_k}{f_\bullet} d_k^2 = \sum_{k=1}^{K} \alpha_k d_k^2 \tag{2.53}$$

pri čemer je

$$\sum_{k=1}^{K} \alpha_k = \sum_{k=1}^{K} \frac{f_k}{f_\bullet} = \frac{1}{f_\bullet} \sum_{k=1}^{K} f_k = 1 \tag{2.54}$$

sledi naslednji zaključek: skupna inercija je tehtana aritmetična sredina χ^2 – razdalj profilov $\mathbf{p}_k$ glede na povprečni profil $\overline{\mathbf{p}}$ (centroid).

Pri tem velja poudariti, da sta tako centroid $\overline{\mathbf{p}}$ kot tudi skupna inercija $In(K)$ neodvisna od neposrednih vrednosti absolutnih frekvenc, sta torej invariantna na množenje elementov kontingenčne tabele s poljubno konstanto. Absolutne frekvence pridejo do izraza le posredno, ko njihove relativne vrednosti določajo mase α_k ($k = 1, 2, \ldots, K$).

---

4 V statistični literaturi je statistika χ^2 običajno imenovana Pearsonov χ^2.

5 V statistični literaturi je skupna inercija običajno imenovana Pearsonov koeficient povprečne kvadratične kontingence (glej npr. Bishop et al., 1975, str. 385).

2.7.2 Določanje inercije v bivariatni korespondenčni analizi

Kot smo že omenili, skupno variacijo v vektorskem prostoru izraža statistika skupna inercija $In(K)$ (2.53). Gre za tehtano vsoto kvadratov oddaljenosti točk vektorskega prostora od njihovega centroida, ob uporabi odgovarjajočih uteži razsežnosti $\overline{p}_g$ ($g = 1, 2, \ldots, G$) in uteži točk α_k ($k = 1, 2, \ldots, K$).

V korespondenčni analizi je skupna inercija v *J*-razsežnem prostoru vrstičnih profilov $\mathbf{p}_{w.i}$ ($i = 1, 2, \ldots, I$)

$$In(I) = \sum_{i=1}^{I} p_{i\bullet}(\mathbf{p}_{w,i} - \mathbf{p}_w)' \mathbf{D}_w^{-1}(\mathbf{p}_{w,i} - \mathbf{p}_w) \tag{2.55}$$

oz.

$$In(I) = \text{sled}\left[\mathbf{D}_v(\mathbf{P}_{w.v} - \mathbf{1}\mathbf{p}_w')\mathbf{D}_w^{-1}(\mathbf{P}_{w.v} - \mathbf{1}\mathbf{p}_w')'\right] \tag{2.56}$$

Podobno je skupna inercija v *I*-razsežnem prostoru stolpičnih profilov $\mathbf{p}_{v.j}$ ($j = 1, 2, \ldots, J$)

$$In(J) = \sum_{i=1}^{I} p_{\bullet j}(\mathbf{p}_{v,j} - \mathbf{p}_v)' \mathbf{D}_v^{-1}(\mathbf{p}_{v,j} - \mathbf{p}_v) \tag{2.57}$$

oz.

$$In(J) = \text{sled}\left[\mathbf{D}_w(\mathbf{P}_{v.w} - \mathbf{p}_v\mathbf{1}')'\mathbf{D}_v^{-1}(\mathbf{P}_{v.w} - \mathbf{p}_v\mathbf{1}')\right] \tag{2.58}$$

Skupna inercija je enaka v obeh vektorskih prostorih (dokaz: glej npr. Rovan, 1991, stran 51)

$$In(I) = In(J) = \text{sled}\left[\mathbf{D}_v^{-1}(\mathbf{P} - \mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1}(\mathbf{P} - \mathbf{p}_v\mathbf{p}_w')'\right] \tag{2.59}$$

V skalarni obliki lahko skupno inercijo zapišemo na podlagi izraza 2.59 kot

$$In = \frac{\chi^2}{f_{\bullet\bullet}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - p_{i\bullet}p_{\bullet j})^2}{p_{i\bullet}p_{\bullet j}} \tag{2.60}$$

oz. statistiko χ^2 (2.51) kot

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*} \tag{2.61}$$

kjer je f_{ij}^* pričakovana absolutna frekvenca *ij*-te celice kontingenčne tabele v primeru ko med proučevanima spremenljivkama ne bi bilo nobene povezanosti

(v t.i. modelu neodvisnosti)

$$f_{ij}^{*} = \frac{f_{i\bullet} f_{\bullet j}}{f_{\bullet\bullet}} \tag{2.62}$$

Primer 1 - 6. nadaljevanje

Izračunajmo skupno inercijo *In* in statistiko χ^2 na podlagi relativnih frekvenc učencev glede na barvo oči in barvo las (Vir: tabela 2.1)

$$In = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - p_{i\bullet}p_{\bullet j})^2}{p_{i\bullet}p_{\bullet j}} =$$

$$= \frac{(0,1277 - 0,2933 \cdot 0,2701)^2}{0,2933 \cdot 0,2701} + \ldots + \frac{(0,0158 - 0,2441 \cdot 0,0219)^2}{0,2441 \cdot 0,0219} = 0,2302$$

$$\chi^2 = In \cdot f_{\bullet\bullet} = 0,2302 \cdot 5387 = 1240,04$$ ■

2.8 Določanje optimalnega podprostora

Korespondenčna analiza je, kot smo že omenili, namenjena prikazu povezav v večrazsežnih tabelah podatkov. V najenostavnejšem primeru bivariatne korespondenčne analize vrstice in stolpci kontingenčne tabele predstavljajo položaj različnih skupin enot glede na obe proučevani spremenljivki.

Kot izhodišče za izračun koordinat točk in pripadajočih algebrskih kazalcev korespondenčne analize običajno vzamemo matriko

$$\mathbf{D}_v^{-1/2}(\mathbf{P} - \mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1/2} \tag{2.63}$$

oz.

$$\left[\frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}}}\right] \qquad \begin{matrix} i = 1, 2, \ldots, I \\ j = 1, 2, \ldots, J \end{matrix} \tag{2.64}$$

pri čemer sta $\mathbf{D}_v$ (2.25) in $\mathbf{D}_w$ (2.32) diagonalni matriki marginalnih relativnih frekvenc, matrika $\mathbf{P} - \mathbf{p}_v\mathbf{p}_w'$ pa matrika hkrati centriranih vrstic in stolpcev

(izvedena iz matrike relativnih frekvenc **P**(2.15)) [6], z rangom

$$\text{rang}(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')\le L \tag{2.65}$$

pri čemer je L

$$L=\min(I-1,J-1) \tag{2.66}$$

Izračune izvedemo v štirih korakih:

1.) Najprej izračunamo singularno dekompozicijo matrike $\mathbf{D}_v^{-1/2}(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1/2}$:

$$\underset{(I\times J)}{\mathbf{D}_v^{-1/2}(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1/2}}=\underset{(I\times L)}{\breve{\mathbf{N}}}\,\underset{(L\times L)}{\mathbf{D}_\delta}\,\underset{(L\times J)}{\breve{\mathbf{M}}'} \tag{2.67}$$

z rangom

$$\text{rang}\left(\mathbf{D}_v^{-1/2}(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1/2}\right)=\text{rang}(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')\le L \tag{2.68}$$

pri čemer je

$$\breve{\mathbf{N}}'\breve{\mathbf{N}}=\breve{\mathbf{M}}'\breve{\mathbf{M}}=\mathbf{I} \tag{2.69}$$

in $\mathbf{D}_\delta$ diagonalna matrika singularnih vrednosti δ_l

$$\mathbf{D}_\delta=diag(\delta_1,\delta_2,\ldots,\delta_L) \tag{2.70}$$

urejenih od največje do najmanjše vzdolž glavne diagonale.

2.) Naj bo

$$\mathbf{N}=\mathbf{D}_v^{1/2}\breve{\mathbf{N}}\quad\text{in}\quad\mathbf{M}=\mathbf{D}_w^{1/2}\breve{\mathbf{M}} \tag{2.71}$$

Potem je, če upoštevamo izraz (2.67), singularna dekompozicija matrike centriranih vrstic in stolpcev $\mathbf{P}-\mathbf{p}_v\mathbf{p}_w'$ enaka

$$\mathbf{P}-\mathbf{p}_v\mathbf{p}_w'=\mathbf{N}\mathbf{D}_\delta\mathbf{M}'=\sum_{l=1}^{L}\delta_l\mathbf{n}_l\mathbf{m}_l' \tag{2.72}$$

pri čemer je $\mathbf{n}_l$ l-ti stolpični vektor matrike **N** in $\mathbf{m}_l$ l-ti stolpični vektor matrike **M**. V tem primeru so levi in desni singularni vektorji normalizirani na enotsko dolžino v $\mathbf{D}_v^{-1}$ in $\mathbf{D}_w^{-1}$ metriki

$$\underset{(L\times I)}{\mathbf{N}'}\,\underset{(I\times I)}{\mathbf{D}_v^{-1}}\,\underset{(I\times L)}{\mathbf{N}}=\underset{(L\times J)}{\mathbf{M}'}\,\underset{(J\times J)}{\mathbf{D}_w^{-1}}\,\underset{(J\times L)}{\mathbf{M}}=\underset{(L\times L)}{\mathbf{I}}\quad 7 \tag{2.73}$$

---

6 Pokažimo, da je matrika $\mathbf{P}-\mathbf{p}_v\mathbf{p}_w'$ matrika hkrati centriranih vrstic in stolpcev:

$$\mathbf{1}'(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')=\mathbf{1}'\mathbf{P}-\mathbf{1}'\mathbf{p}_v\mathbf{p}_w'=\mathbf{p}_w'-\mathbf{p}_w'=\mathbf{0}'$$
$$(\mathbf{P}-\mathbf{p}_v\mathbf{p}_w')\mathbf{1}=\mathbf{P}\mathbf{1}-\mathbf{p}_v\mathbf{p}_w'\mathbf{1}=\mathbf{p}_v-\mathbf{p}_v=\mathbf{0}$$

Stolpci matrike **N** določajo koordinatne osi točk, ki predstavljajo stolpične profile matrike **P**, stolpci matrike **M** pa določajo koordinatne osi točk, ki predstavljajo vrstične profile matrike relativnih frekvenc **P**.

3.) Izračunajmo koordinate vrstičnih profilov

$$\underset{(I\times K)}{\mathbf{C}} = \underset{(I\times I)}{\mathbf{D}_v^{-1}}\ \underset{(I\times L)}{\mathbf{N}}\ \underset{(L\times L)}{\mathbf{D}_\delta} \tag{2.74}$$

in koordinate stolpičnih profilov

$$\underset{(J\times K)}{\mathbf{G}} = \underset{(J\times J)}{\mathbf{D}_w^{-1}}\ \underset{(J\times L)}{\mathbf{M}}\ \underset{(L\times L)}{\mathbf{D}_\delta} \tag{2.75}$$

Vrstice matrik **C** in **G** predstavljajo koordinate centriranih vrstičnih in stolpičnih profilov, pri čemer je, če upoštevamo izraz (2.73),

$$\mathbf{C}'\mathbf{D}_v\mathbf{C} = \mathbf{G}'\mathbf{D}_w\mathbf{G} = \mathbf{D}_\delta^2 \tag{2.76}$$

Tehtana vsota kvadratov koordinat vrstičnih profilov vzdolž *l*-te osi in tehtana vsota kvadratov koordinat stolpičnih profilov vzdolž *l*-te osi sta obe enaki kvadratu *l*-te singularne vrednosti δ_l^2.

4.) Skupna inercija je vsota kvadratov vseh neničelnih singularnih vrednosti

$$In = \sum_{l=1}^{L} \delta_l^2 \tag{2.77}$$

kjer so $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_L > 0$ neničelni diagonalni elementi matrike $\mathbf{D}_\delta$, L pa rang matrike $\mathbf{P} - \mathbf{p}_v\mathbf{p}_w'$.

Teoretična rešitev minimizacije kriterijske funkcije za poljubno razsežnost optimalnega podprostora temelji na singularni dekompoziciji matrike in njeni optimalni aproksimaciji. Osnovna odlika singularne dekompozicije matrike je v tem, da lahko izpustimo nekaj zadnjih členov z najnižjimi singularnimi vrednostmi in tako dobimo aproksimacijo matrike v smislu metode najmanjših kvadratov.

V primeru singularne dekompozicije matrike hkrati centriranih vrstic in

---

7 Sigularno dekompozicijo matrike centriranih vrstic in stolpcev $\mathbf{P} - \mathbf{p}_v\mathbf{p}_w'$ (2.72), pri kateri so singularni vektorji normalizirani na enotsko dolžino v $\mathbf{D}_v^{-1}$ in $\mathbf{D}_w^{-1}$ metriki (2.73), imenujemo posplošena singularna dekompozicija matrike.

stolpcev

$$\mathbf{P} - \mathbf{p}_v\mathbf{p}_w' = \sum_{l=1}^{L} \delta_l \mathbf{n}_l \mathbf{m}_l' \tag{2.78}$$

dobimo za vrednost $L^* < L$ približek matrike **P**, t.j. matiko $\mathbf{P}^*$

$$\mathbf{P} \cong \mathbf{p}_v\mathbf{p}_w' + \sum_{l=1}^{L^*} \delta_l \mathbf{n}_l \mathbf{m}_l' = \mathbf{P}^* \tag{2.79}$$

Med vsemi matrikami ranga $L^* + 1$ (ali nižjega) matrika $\mathbf{P}^*$ minimizira kriterij posplošenih najmanjših kvadratov

$$\text{sled}\left[\mathbf{D}_v^{-1}(\mathbf{P} - \mathbf{P}^*)\mathbf{D}_w^{-1}(\mathbf{P} - \mathbf{P}^*)'\right] \tag{2.80}$$

Če se npr. osredotočimo na vrstice matrik **P** in $\mathbf{P}^*$, potem vrstice matrike $\mathbf{P}^*$ določajo tisti podprostor razsežnosti $L^* + 1$, ki je najbližji vrsticam matrike **P** s stališča tehtane vsote kvadratov razdalj. Seveda to velja tudi za stolpce matrik **P** in $\mathbf{P}^*$.

Naj bo $\mathbf{P}^* = \mathbf{p}_v\mathbf{p}_w'$. Ker je

$$\text{rang}(\mathbf{P}^*) = \text{rang}(\mathbf{p}_v\mathbf{p}_w') = \min \text{rang}(\mathbf{p}_v, \mathbf{p}_w) = 1 \tag{2.81}$$

je matrika $\mathbf{p}_v\mathbf{p}_w'$, ki ustreza modelu neodvisnosti, najboljši približek ranga 1 matrike **P**. Izraz (2.80) je v tem primeru enak skupni inerciji

$$\text{sled}\left[\mathbf{D}_v^{-1}(\mathbf{P} - \mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1}(\mathbf{P} - \mathbf{p}_v\mathbf{p}_w')'\right] = \frac{\chi^2}{f_{\cdot\cdot}} = In \tag{2.82}$$

ki izraža (s stališča posplošenih najmanjših kvadratov) odstopanje matrike **P** od matrike $\mathbf{p}_v\mathbf{p}_w'$. Prav to odstopanje pa želi prikazati korespondenčna analiza.

2.9 Določanje razsežnosti optimalnega podprostora in grafični prikaz točk v tem podprostoru

Določimo še razsežnost optimalnega podprostora. Kot smo to pojasnili v prejšnjem razdelku, je skupna inercija enaka vsoti kvadratov singularnih vrednosti (2.77). Na vsako glavno komponento otpade

$$\lambda_l = \delta_l^2 \quad ; \quad l = 1, 2, \ldots, L \tag{2.83}$$

skupne inercije, zato bomo ta del inercije imenovali inercija *l*-te glavne komponente λ_l. Za določitev optimalnega števila glavnih komponent oz.

razsežnosti optimalnega podprostora lahko uporabimo naslednji empirični kriterij:

$$L^* = \min\left\{ l \mid \lambda_1 + \lambda_2 + \ldots + \lambda_l \geq \frac{q}{100} \sum_{l=1}^{L} \lambda_l \right\} \tag{2.84}$$

kjer L^* izraža razsežnost optimalnega podprostora, q pa izbrani odstotni delež skupne inercije (npr. 90%), ki ga želimo aproksimirati s projekcijami točk v optimalnem podprostoru.

Kot smo že omenili v uvodu je ena temeljnih značilnosti korespondenčne analize možnost grafične predstavitve proučevanega pojava. V primerih, ko dosežemo zaželeni nivo aproksimacije že v dvorazsežnem podprostoru, za grafično predstavitev uporabljamo dvorazsežni razsevni grafikon. Če pa bi bila takšna predstavitev pomankljiva uporabimo zahtevnejše grafične metode, med katerimi sta po naših izkušnjah najprimernejša trirazsežni razsevni grafikon in Andrewsov grafikon (Rovan, 1991).

Primer 1 - 7. nadaljevanje

Kot smo že omenili, kot izhodišče za izračun koordinat točk in pripadajočih algebrskih kazalcev korespondenčne analize običajno vzamemo matriko $\mathbf{D}_v^{-1/2}(\mathbf{P} - \mathbf{p}_v\mathbf{p}_w')\mathbf{D}_w^{-1/2}$ (2.25).

Izračunajmo najprej singularne vrednosti δ_l, glavne inercije λ_l, vrednosti χ_l^2 ter odstotne deleže in kumulativne odstotne deleže glavnih inercij (oz. vrednosti χ_l^2) vseh treh glavnih komponent (tabela 2.4).

Tabela 2.4 Razčlenitev skupne inercije in statistike χ^2

Singularne vrednosti δ_l	Glavne inercije λ_l	Vrednosti χ_l^2	Odstotni deleži	Kumulativni odstotni deleži
0,4464	0,1992	1073,33	86,56	86,56
0,1735	0,0301	162,08	13,07	99,63
0,0293	0,0009	4,63	0,37	100,00
	0,2302	1240,04	100,00	

Če želimo določiti podprostor, katerega razsežnost omogoča dobro aproksimacijo skupne inercija, mora biti v našem primeru razsežnost optimalnega podprostora enaka $L^* = 2$ (2.84).

Izračunajmo koordinate vrstičnih profilov in koordinate stolpičnih profilov oz. vrednosti prvih dveh glavnih komponent iz matrik **C** in **G** (2.74 in 2.75) (tabela 2.5).

Tabela 2.5 Koordinate vrstičnih in stolpičnih profilov učencev glede na barvo oči in barvo las

Barva oči:	1. GK	2. GK
svetle	0,4407	0,0885
modre	0,4003	0,1654
srednje	-0,0336	-0,2450
temne	-0,7027	0,1339

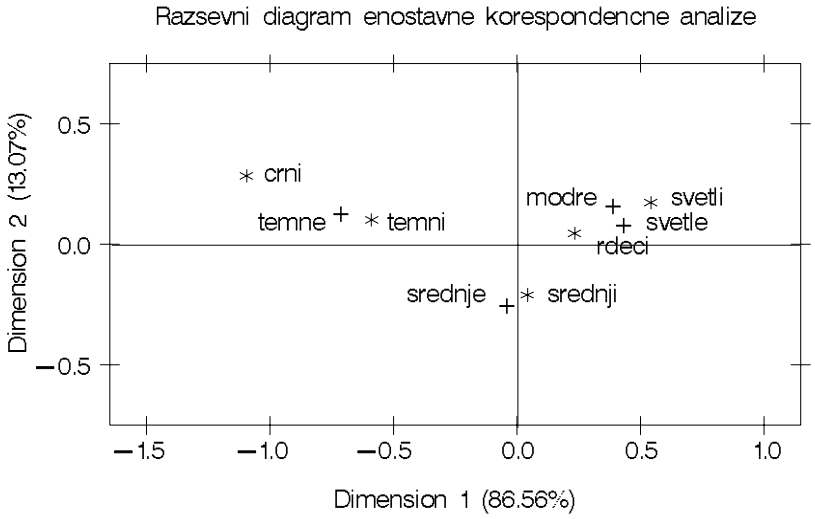
Barva las:	1. GK	2. GK
svetli	0,5440	0,1738
rdeči	0,2333	0,0483
srednji	0,0420	-0,2083
temni	-0,5887	0,1040
črni	-1,0944	0,2864

Prikažimo vrstice matrike **C** (kot že omenjeno se omejujemo na vrednosti prvih dveh glavnih komponent) kot točke v kartezičnem koordinatnem prostoru (znak +, slika 2.1). Na podlagi razporeditve točk, ki prikazujejo položaj skupin oseb, ki ustrezajo posameznim barvam oči, sklepamo naslednje: skupini oseb s svetlimi in modrimi očmi sta glede na porazdelitev njihove barve las dokaj podobni (imata podobne profile – glej tabelo 2.2). Gledano v smeri osi prve glavne komponente je položaj skupine oseb s temnimi očmi povsem nasproten omenjeni dvojici, položaj skupine oseb s srednjo barvo oči pa je približno na sredini med omenjeni skupinami. Gledano v smeri osi druge glavne komponente pa ima skupina oseb s srednjo barvo oči povsem nasproten položaj kot skupine oseb ostalih barv oči.

Prikažimo v istem grafikonu še vrstice matrike **G** (∗). Gledano v smeri osi prve glavne komponente opažamo precejšnjo razliko med položajem skupine oseb svetlih las in skupinama oseb rdečih in srednjih las, nato precejšnjo razliko do skupine oseb temnih las in potem znova do skupine oseb črnih las glede na barvo oči. Gledano v smeri osi druge glavne komponente ima skupina oseb s

srednjo barvo las povsem nasproten položaj kot skupine oseb vseh ostalih barv las glede na barvo oči.

Skupna kvaliteta prikaza položaja točk v sliki 2.1, izražena v obliki odstotnega deleža inercije (99,63%), ki je predstavljen z izbranim dvorazsežnim podprostorom, kaže na to, da v našem primeru omenjena slika skoraj idealno predstavlja položaj točk.



Slika 2.1 Prikaz profilov učencev glede na barvo oči in barvo las v optimalni ravnini (slika je enaka sliki 1.1)

V sliki 2.1 smo torej prikazali dve množici točk, ki odražata položaj profilov učencev glede na barvo oči in glede na barvo las. Analiza položaja posamezne množice točk kaže na stopnjo podobnosti med skupinami oseb v skladu z razdaljo med točkami iste množice, hkratna analiza položaja točk obeh množic pa kaže na povezanost ("correspondence") med točkami obeh množic. Potrebno pa je poudariti, da ni smiselno interpretirati razdalj med točkami različnih množic (v strogem pomenu besede), kajti takšne razdalje nismo definirali. Definirali smo samo χ^2-razdalje med točkami iste množice.

Povzemimo dosedanje vsebinske ugotovitve! Učenci s svetlimi in modrimi očmi imajo podobne profile glede na svojo barvo las. Položaj profila učencev z

rdečimi lasmi je približno na polovici med profili učencev s svetlo in srednjo barvo las glede na barvo oči. Največje razlike nastopajo med srednjo in temno barvo las oz. oči, kar je nedvomno genetsko pogojeno.

Za naš primer je zelo značilna tudi parabolična razporeditev točk, kar moramo pripisati t.im. "horseshoe" efektu (Kendall, 1971). Do tega efekta pogosto prihaja pri nominalnih spremenljivkah z močno medsebojno povezanostjo. Če preuredimo položaj kategorij obeh nominalnih spremenljivk kontingenčne tabele tako, da vrstni red kategorij ustreza rangu vrednosti prve glavne komponente (to smo storili tudi v našem primeru, glej tabeli 1.1 in 2.5), potem lahko pričakujemo visoke relativne frekvence v pasu "diagonalnih" elementov tabele. Čeprav so osi glavnih komponent medsebojno ortogonalne v algebrskem smislu, lahko med njimi obstajajo nelinearne stohastične zveze, posledica tega pa je "horseshoe" efekt. Ta efekt se pogosto pojavlja v grafičnih prikazih korespondenčne analize, vendar pa ne predstavlja nekega posebnega problema (M.J: Greenacre, 1984, str. 232). ■