

Korrespondenssianalyysi - kaavaliite (v. 1.04)

Jussi Hirvonen

20.5.2018

Paperin versiot		
Versio	muutokset	päivämäärä
0.1	harjoittelua - drawmatrix	14.7.2017
1.01	harjoittelua - matriisiyhtälöt ja ca-peruskaavat	5.8.2017
1.02	pientä korjailua, turhan poistoa	28.1.2018
1.03	lisiäillään yksinkertaisen ca:n kaavoja, taulukoita R-paketilla furniture	6.4.2018
1.04	ca:n loput matriisiyhtälöt, drawmatrix-kokeiluja	20.5.2018

Sisältö

1	Kaavat ja matemaattisen merkinnät - työkalut	1
2	Yksinkertaisia kaavoja leipätekstiin	2
3	Taulukoita	3
4	Matriisit ja niiden havainnollistaminen	4
5	Matriisiyhtälöt	4

1 Kaavat ja matemaattisen merkinnät - työkalut

Tähän voisi kerätä Latex:in suositukset ja tavat esittää kaavoja, tässä on pientä epäselvyyttä mulla vielä. Esimerkkejä löytyy varsinaisesta luonnosdokkarista. Myös kaavito sun muut, kuvien insertointi on varsinaisessa luonnosdokkarissa.

2 Yksinkertaisia kaavoja leipätekstiin

Yksinkertaisen korrespondenssianalyysin esittelyssä tarvitaan muutama kaava, vaikka kointa niitä liitteen ulkopuolella välttää (mallia on otettu MG:n PCAiP-kirjasta, ss 25-).

Taulukon homogeenisyyden tai riippumattomuushypoteestin tutkiminen, miten paljon rivit (tai sarakkeet) eroavat toisistaan.

Tuttu χ^2 - testisuure saadaan, kun lasketaan yhteen jokaisen solun havaittujen ja odotettujen (riippumattomuushypoteesi) frekvenssien erotukset muodossa

$$\chi^2 = \frac{(\text{havaittu} - \text{odotettu})^2}{\text{odotettu}}$$

Tämä voidaan esittää ca:han sopivammalla tavalla parilla muunnoksella, jolloin saamme riveittäin vastaavat termit rivisummalla painotettuna:

$$\text{rivisumma} \times \frac{(\text{havaittu riviprofiili} - \text{odotettu riviprofiili})^2}{\text{odotettu riviprofiili}}$$

Kun jaamme nämä tekijät havaintojen kokonaismäärällä n , rivisumma muuntuu rivin massaksi, ja niiden summa muotoon $\frac{\chi^2}{n}$.

$$\frac{\chi^2}{n} = \phi^2.$$

Tunnusluku ϕ^2 on korrespondenssianalyysissä kokonaisinertia (total inertia). Se kuvaa, kuinka paljon varianssia taulukossa on ja on riippumaton havaintojen lukumäärästä. Tilastotieteessä tunnusluvulla on useita vaihtoehtoisia nimiä (esim. mean square contingency coefficient), ja sen neliöjuurta kutsutaan ϕ - kertoimeksi.

Tässä siirrytään kahden luokittelumuuttujan taulukosta suhteellisten frekvenssien taulukkoon, ja pieni pohdinta taulukoista yleensä olisi paikallaan.

Frekvenssitaulukossa (jossa kaikki taulukon luvut on jaettu havaintojen lukumäärällä n) riviprofilien 1 ja 3 (euklidinen) etäisyys on

$$\sqrt{(p_{11} - p_{31})^2 + (p_{12} - p_{32})^2 + (p_{13} - p_{33})^2 + (p_{14} - p_{34})^2 + (p_{15} - p_{35})^2}$$

Rivien χ^2 - etäisyys on painotettu euklidinen etäisyys, jossa painoina ovat riviprofilin odotetut arvot. Ne ovat riippumattomuushypoteesin mukaisesti riviprofilien keskiarvoprofilin vastaavat alkioit r_i .

$$\sqrt{\frac{(p_{11} - p_{31})^2}{r_1} + \dots + \frac{(p_{15} - p_{35})^2}{r_5}}$$

Taulukko 1: Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä

	1	2	3	4	5	Test	P-V
	n = 295	n = 600	n = 593	n = 889	n = 732		
maa						Chi Square: 588.1	<.0
FI	47 (15.9%)	188 (31.3%)	149 (25.1%)	423 (47.6%)	303 (41.4%)		
HU	219 (74.2%)	288 (48%)	225 (37.9%)	190 (21.4%)	75 (10.2%)		
SE	29 (9.8%)	124 (20.7%)	219 (36.9%)	276 (31%)	354 (48.4%)		

Inertia voidaan esittää rivien ja “keskiarvorivin “ (sentroidin)

$$\chi^2$$

-etäisyyksien neliöiden painotettuna summana, jossa painoina ovat rivien massat m_i ja summa lasketaan yli rivien i .

$$\phi^2 = \sum_i (massa\ m_i) \times (profiilin\ i\ \chi^2 - etäisyys\ sentroidista)^2$$

3 Taulukoita

Kahden luokittelumuuttujan ristiintaulukointi (kontigenssitaulu), ja muitakin variantteja ja pohjia voi tehdä R-paketilla furniture (esim.), jossa output-formaatiksi voi valita latex tai latex2. Vaatii LateX-dokkarissa paketin booktabs.

Tämä meni oikealta yli (output = latex), vaan ei enään .

Toinen koe (output = latex2):

Taulukko 2: Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä

	V6				
	1	2	3	4	5
	n = 295	n = 600	n = 593	n = 889	n = 732
maa					
FI	47 (15.9%)	188 (31.3%)	149 (25.1%)	423 (47.6%)	303 (41.4%)
HU	219 (74.2%)	288 (48%)	225 (37.9%)	190 (21.4%)	75 (10.2%)
SE	29 (9.8%)	124 (20.7%)	219 (36.9%)	276 (31%)	354 (48.4%)

Yritetään vielä yksinkertaisempaa, taulukon luonti on vain mennyt pieleen.

Taulukko 3: Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä

	V6					
	1	2	3	4	5	Total
	n = 4	n = 4	n = 4	n = 4	n = 4	n = 4
maa						
FI	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)
HU	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)
SE	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)
Total	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)	1 (25%)

4 Matriisit ja niiden havainnollistaminen

Drawmatrix toimii, mutta vaatii säätöä. Voisi olla leipätekstissä hyvä matriisiyhtälöiden havainnollistamiseen.

Kummallinen kohdistus.

$$\begin{pmatrix} \boxed{A} & \boxed{B}^{-1} \end{pmatrix} \boxed{C}$$

Ainakin SVD - osuudessa voi hyödyntää tätä:

$$\begin{pmatrix} \boxed{S} & \boxed{s} & \boxed{s} & \boxed{D} \end{pmatrix} \boxed{U}$$

Yksinkertainen korrespondenssianalyysi on kahden luokittelumuuttujan määrittelmän frekvenssitaulukon analyysiä. Taulukon rivit ovat havaintoyksiköiden (individuals, havaintoyksikkö) aggregoituja summia, sarakkeet muuttujia.

Analyysissä osa riveistä tai sarakkeista voidaan jättää pois ratkaisun laskennasta ns. passiivisiksi, ja esittää kartalla täydentävinä pisteinä (supplementary points). Ne eivät vaikuta ratkaisuun, eli teknisesti niiden massa on nolla, mutta pisteiden esityksen (projektion) tarkkuus voidaan arvioida. Täydentävien profiilien on kuitenkin oltava yhteismitallisia taulukon datan kanssa. Mikä tahansa ei käy (kts. CAinP, vast.luku).

5 Matriisiyhtälöt

Tässä lähteenä Greenacren kirja (ca in practice) ja sen liite Theory of CA. Muistiinpanoja löytyy, joissa viitataan myös Biplots in practice - kirjaan. Kevään 2017 kurssin luentokalvoja on myös käytetty. Lisäilläään vielä käsitteitä LeRouxin ja Rouanetin kirjasta.

Korrespondenssianalyysin perusyhtälöt:

Datamatriisilla \mathbf{N} on I riviä ja J saraketta ($I \times J$). Alkiot ovat ei-negatiivisia (eli nollat sallittuja) ja samassa mitta-asteikossa. Jos mitta-asteikko on intervalli- tai suhdeasteikko, mittayksiköiden on oltava samoja (esim. euroja, metrejä). Taulukon alkioden summa on $\sum_i \sum_j n_{ij} = n$, missä $i = 1, \dots, I$ ja $j = 1, \dots, J$. GDA-kirjassa on tarkennettu tätä vaatimusta ei-negatiivisuudesta.

Korrespondenssimatriisi \mathbf{P} saadaan jakamalla matriisin \mathbf{N} alkiot niiden summalla n . Merkitään matriisin \mathbf{P} rivisummien vektoria $\mathbf{r} (= (r_1, \dots, r_I))$ ja sarakesummien vektoria $\mathbf{c} (= (c_1, \dots, c_J))$. Niitä vastaavat diagonaalimatriisit ovat $\mathbf{D_r}$ ja $\mathbf{D_c}$.

Korrespondenssianalyysin perusrakenne (algoritmi?) on tämä. Singulaariarvohajoitelma (singular value decomposition) tuottaa ratkaisun kun sitä sovelletaan standardoituun residuaalimatriisiin \mathbf{S} .

$$(1) \quad \mathbf{S} = \mathbf{D_r}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D_c}^{-1/2}$$

Residuaalimatriisi voidaan esittää myös ns. kontingenssi-suhdelukujen (contingency ratio) avulla.

$$\mathbf{D_r}^{-1}\mathbf{P}\mathbf{D_c}^{-1} = \left(\frac{p_{ij}}{r_i c_j} \right)$$

$$\mathbf{S} = \mathbf{D_r}^{1/2}(\mathbf{D_r}^{-1}\mathbf{P}\mathbf{D_c}^{-1} - \mathbf{1}\mathbf{1}^T)\mathbf{D_c}^{-1/2} \quad .$$

Toinen esitystapa on hyödyllinen, kun tarkastellaan CA:n yhteyksiä muihin läheisiin menetelmiin (log ratio analysis of compositional data, moniulotteinen skaalaus (?), lineaarinen diskriminanttianalyysi, kanoninen korrelaatioanalyysi, pääkomponenttianalyysi, kaksoiskuvat, yleensä SVD-perusteiset dimensioiden vähentämisen menetelmät).

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

ja toinen

$$s_{ij} = \sqrt{r_i} \left(\frac{p_{ij}}{r_i c_j} \right) \sqrt{c_j} \quad .$$

Mitäköhän tuosta pitäisi nähdä? Selitykset löytyvät em. teorialiitteestä.

Singulaariarvohajoitelma (singular value decomposition, SVD) matriisille \mathbf{S} on

$$\mathbf{S} = \mathbf{U}\mathbf{D_\alpha}\mathbf{V}^T$$

missä \mathbf{D}_α on diagonaalimatriisi, jonka alkiot ovat singulaariarvot suuruusjärjestyksessä $\alpha_1 \geq \alpha_2 \geq \dots$.

Matriisit \mathbf{U} ja \mathbf{V} ovat ortogonaalisia singulaarivektoreiden matriiseja. Singulaariarvo-hajoituksen merkitys dimensioiden vähentämiselle perustuu Eckart - Young - teoreemaan. Teoreema (30-luvulta?) kertoo, että saamme pienimmän neliösumman m - ulotteisen approksimaation matriisille \mathbf{S} (CAinP, ss. 244) matriisien \mathbf{U} ja \mathbf{V} ensimmäisten sarakkeiden ja ensimmäisten singulaariarvojen avulla.

$$\mathbf{S}_{(m)} = \mathbf{U}_{(m)} \mathbf{D}_{\alpha(m)} \mathbf{V}_{(m)}^T$$

Korrespondenssianalyysin ratkaisualgoritmissa tätä tulosta on muokattava niin, että rivien ja sarakkeiden massat huomioidaan pienimmän neliösumman approksimaatiossa painoina.

Näin saadaan standardikoordinaatit ja principal-koordinaatit riveille ja sarakkeille.

Rivien standardikoordinaatit

$$(2) \quad \Phi = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{U}$$

Sarakkeiden standardikoordinaatit

$$(3) \quad \Gamma = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{V}$$

Rivien principal-koordinaatit

$$(4) \quad \mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{U} \mathbf{D}_\alpha = \Phi \mathbf{D}_\alpha$$

Sarakkeiden principal-koordinaatit

$$(5) \quad \mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{V} \mathbf{D}_\alpha = \Gamma \mathbf{D}_\alpha$$

Pääakselien inertiat (principal inertias) λ_k

$$(6) \quad \lambda_k = \alpha_k^2, k = 1, \dots, K, K = \min\{I - 1, J - 1\}$$

Bilineaarinen korresepondenssimalli

Korrespondenssimatriisi \mathbf{P} voidaan esittää matriisi- ja alkiomuodossa ns. palautuskaavana (reconstitution formula).

$$(7) \quad \mathbf{P} = \mathbf{D}_r \left(\mathbf{1}\mathbf{1}^T + \Phi \mathbf{D}_\alpha^{\frac{1}{2}} \Gamma^T \right) \mathbf{D}_c$$

$$(8) \quad p_{ij} = r_i c_j \left(1 + \sum_{k=1}^K \sqrt{\lambda_k} \phi_{ik} \gamma_{jk} \right)$$

Tässä viitataan s. 101 (13.4), 109 (14.9), ja 109-110 (14.10 ja 14.11). Palautuskavoilla on monta esitystapaa bilineaarisessa mallissa.

Rivien ja sarakkeiden riippuvuus ja transitioyhtälöt. ss. 244, 108-109 skalaariversiot.

Pääkoordinaatit standardikoordinaattien funktiona (ns. barysentrinen ominaisuus - barycentric relationships)

$$(9) \quad \mathbf{F} = \mathbf{D}_r^{-1} \mathbf{P} \mathbf{\Gamma}$$

$$(10) \quad \mathbf{G} = \mathbf{D}_c^{-1} \mathbf{P}^T \mathbf{\Phi}$$

Pääkoordinaatit pääkoordinaattien funktioita:

$$(11) \quad \mathbf{F} = \mathbf{D}_r^{-1} \mathbf{P} \mathbf{G} \mathbf{D}_\lambda^{-\frac{1}{2}}$$

$$(12) \quad \mathbf{G} = \mathbf{D}_c^{-1} \mathbf{P}^T \mathbf{F} \mathbf{D}_\lambda^{-\frac{1}{2}}$$

Yhtälöt (9) ja (10) esittävät profilipisteet ideaalipisteiden (vertex points) painotettuina keskiarvoina, painoina profiilin elementit. Asymmetriset kartat (rivien tai sarakkeiden suhteen) perustuvat näihin yhtälöihin. Yhtälöiden (11) ja (12) kahdet pääkoordinaatit ovat perusta symmetrisille kartoille. Myös niitä yhdistää barisentrinen painotetun keskiarvon riippuvuus, mutta mukana ovat skaalaustekijät $\frac{1}{\sqrt{\lambda_i}}$. Ne ovat jokaisessa dimensiossa eri suuruisia.