# 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 | lisäillään yksinkertaisen ca:n kaavoja, taulukoita R-paketilla furniture | 6.4.2018 |
| 1.04 | ca:n loput matriisiyhtälöt, drawmatrix-kokeiluja | 20.5.2018 |
| 1.05 | Siirretty Rmarkdown-dokkariksi | 13.6.2018 |
| 1.06 | utf8 inputenc | 5.8.2018 |

## Sisältö



# 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 koitan 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 $\chi^2$ - testisuure saadaa, kun lasketaan yhteen jokaisen solun havaittujen ja odotetettujen (riippumattomuushypoteesi) frekvenssien erotukset muodossa

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

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

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

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 $\phi^2$ 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 $\phi$ - 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) riviprofiilien 1 ja 3 (euklidinen) etäisyys on

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

Rivien $\chi^2$ - etäisyys on painotettu euklidinen etäisyys, jossa painoina ovat riviprofiilin odotetut arvot. Ne ovat riippumattomuushypoteesin mukaisesti riviprofiilien keskiarvoprofiilin vastaavat alkioit $r_i$ .

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

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 voidaa esittää rivien ja "keskiarvorivin " (sentroidin)

$$\chi^2$$

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

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

# 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<br>n = 4 | 2<br>n = 4 | 3<br>n = 4 | 4<br>n = 4 | 5<br>n = 4 | Total<br>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.

$$\left( A_{i} \; B^{-1} \right) C$$

Ainakin SVD - osuudessa voi hyödyntää tätä:

$$\left( S_{j} \; s_{k-1} \; s_{k} \; D \right) U$$

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. passiviisiksi, 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 kuintenkin oltava yhteismitallisia taulukon datan kanssa. Mikä tahansa ei käy (kts. CAinP, vast.luku).

Pinotut tai yhdistetyt matriisit ("stacked matrices"). Yksinkertainen korrespondenssianalyysi on kahden luokittelumuuttujan määrittämän taulukon (kontingenssitaulukko) analyysiä, mutta tutkimusasetelmaa voi melko helposti muuttaa useamman muuttujan analyysiksi. Menetelmän matemaattinen perusta ja ratkaisualgoritmi (SVD) toimivat, tulkinta vain muuttuu. Itse asiassa menetelmän yleisyys tekee sen vääränkin käytön mahdolliseksi.

Yksinkertaisin laajennus on lisätä alkuperäisen taulukon alle toinen taulukko. Rivit ovat esimerkissä maittan summattuja vastauksia, ja niiden alle voidaan lisätä joku toinen luokittelumuuttuja. Havaintojen määrä yhditetyssä ("pinotussa" ) taulussa kaksinkertaistuu. Miksi tämä ei ei vaikuta tuloksiin vääristävästi??

Merkitään edellisten analyysien kuuden maan ja viiden vastausvaihtoehdon taulukkoa matriisilla $\boldsymbol{A}_{IJ}$, missä $I$ on rivien ja $J$ sarakkeiden lukumäärä. Taulukoidaan ikäluokan (1 - 6) ja sukupuolen ($f$ = nainen, $m$ = mies) vuorovaikutusmuuttuja ($f1, \ldots, f6$ ja $m1, \ldots, m6$) samojen vastausvaihtoehtojen kanssa. Jos tätä taulukkoa merkitään matriisilla $\boldsymbol{B}_{I^{`}J}$, voimme muodostaa yhdistetyn matriisin

$$\begin{array}{c} \boxed{A}_{IJ} \\ \boxed{B}_{I^{`}J} \end{array}$$

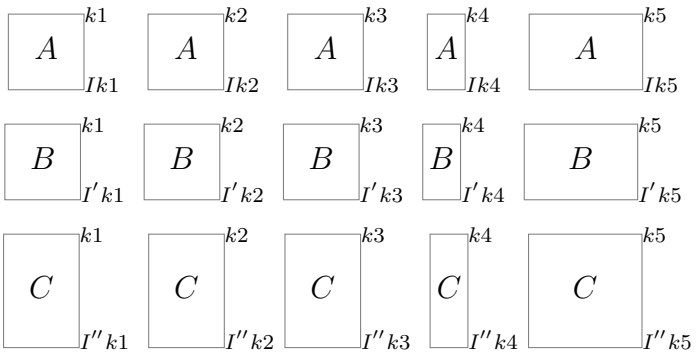
Miten päällekkäisten matriisien ympärille saisi sulut?

Rivien lukumäärä on molemmissa matriiseissa sama, koska luokkia sattuu olemaan kuusi sekä maa- että ikä- ja sukupuoli - luokittelumuuttujissa. Kun matriisit ovat dimensioiltaan ja myös muuttujien sisällön kannalta samankaltaiset, niitä kutsutaan yhteensopiviksi ("matched matrix"). Tällöin yksinkertaista korrespondenssianalyyisä voi soveltaa tutkimusongelmaan, jossa halutaan erotella jonkun ryhmän sisäinen vaihtelu ryhmien välisiestä vaihtelusta. (Greenacren ehdottama ABBA - analyysi).

$$\begin{array}{cc} \boxed{A}_{IJ} & \boxed{B}_{IJ} \\ \boxed{B}_{IJ} & \boxed{A}_{IJ} \end{array}$$

ABBA on erityistapaus yleisemmästä moniulotteisen taulukon (multiway table) analyysistä, jossa useita kahden muuttujan taulukoita "pinotaan" päällekkäin ja rinnakkain. Voimme ottaa yhden kysymyksen vastausten lisäksi analyysiin mukaan useamman kysymyksen vastaukset laajentamalla kahden päällekkäisen matriisin taulukkoa oikealle.

Teknisesti analyysi on yksinkertainen korrespondenssianalyysi, miten tämä tulkitaan?




# 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. Lisäillään vielä käsitteitä LeRouxin ja Rouanetin kirjasta.

**Korrespondenssianalyysin perusyhtälöt:**

Datamatriisilla $\boldsymbol{N}$ on $I$ riviä ja $J$ sarakketta ($IxJ$ ). 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 alkioiden summa on $\sum_i \sum_j n_{ij} = n$, missä $i = 1, \ldots, I$ ja $j = 1, \ldots, J$. GDA-kirjassa on tarkennettu tätä vaatimusta ei-negatiivisuudesta.

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

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

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

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

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

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

Toinen esitystapa on hyödyllinen, kun tarkastellaan CA:n yhteyksiä muihin läheisiin menetelmiin (log ratio analysis of compositional data, moniulotteinen skaalaus (?), lineaarinen diskriminanttianalyysi, kanoninen korrelaatioanalyysi, pääkomponettianalyysi, kaksoiskuvat, yleensä SVD-perusteiset dimensioden 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 $\boldsymbol{S}$ on

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

missä $\boldsymbol{D_\alpha}$ on diagonaalimatriisi, jonka alkiot ovat singulaariarvot suuruusjärjestyksessä $\alpha_1 \geq \alpha_1 \geq \ldots$

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

$$\boldsymbol{S}_{(m)} = \boldsymbol{U}_{(m)}\boldsymbol{D}_{\alpha(m)}\boldsymbol{V}^T_{(m)}$$

Korrrespondenssianalyysin 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) $$\boldsymbol{\Phi} = \boldsymbol{D_r}^{-\frac{1}{2}}\boldsymbol{U}$$

Sarakkeiden standardikoordinaatit

$$\boldsymbol{\Gamma} = \boldsymbol{D_c}^{-\frac{1}{2}}\boldsymbol{V} \tag{3}$$

Rivien principal-koordinaatit

$$\boldsymbol{F} = \boldsymbol{D_r}^{-\frac{1}{2}}\boldsymbol{U}\boldsymbol{D_\alpha} = \boldsymbol{\Phi}\boldsymbol{D_\alpha} \tag{4}$$

Sarakkeiden principal-koordinaatit

$$\boldsymbol{G} = \boldsymbol{D_c}^{-\frac{1}{2}}\boldsymbol{V}\boldsymbol{D_\alpha} = \boldsymbol{\Gamma}\boldsymbol{D_\alpha} \tag{5}$$

Pääakseleiden inertiat (principal inertias) $\lambda_k$

$$\lambda_k = \alpha_k^2, k = 1, \ldots, K, K = min\{I-1, J-1\} \tag{6}$$

Bilineaarinen korresepondenssimalli
Korrespondenssimatriisi $\boldsymbol{P}$ voidaan esittää matriisi- ja alkiomuodossa ns. palautuskaavana (reconstitution formula).

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

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

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)

$$\boldsymbol{F} = \boldsymbol{D}_r^{-1}\boldsymbol{P}\boldsymbol{\Gamma} \tag{9}$$

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

Pääkoordinaatit pääkoordinaattien funktiointa:

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

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

Yhtälöt (9) ja (10) esittävät profiilipisteet 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.