Korrespondenssianalyysi - kaavaliite (v. 1.03)

Jussi Hirvonen

4.4.2018

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	4.4.2018

Sisältö

1 Kaavat ja matemaattisen merkinnät - työkalut **1**

2 Yksinkertaisia kaavoja leipätekstiin **1**

3 Matriisit ja niiden havainnollistaminen **2**

4 Matriisiyhtälöt ilman kaavioita **3**

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.

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%)		

Kahden luokittelumuuttujan ristiintaulukointi (kontigenssitaulu)

En käytä kaavaliitteen notaatiota leipätekstissä, vaan yksinkertaisempaa tapaa. Aloitetaan taulukolla.

Koitetaan furniture-paketilla vääntää taulukko. Tämä menee oikealta yli (output = latex) .

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:

Bigskip

3 Matriisit ja niiden havainnollistaminen

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

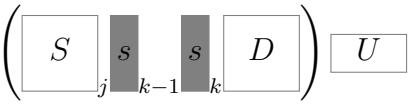
Kummallinen kohdistus.

$$\left(\boxed{A}_{i} \boxed{B}^{-1} \right) \boxed{C}$$

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%)

Ainakin SVD - osuudessa voi hyödyntää tätä:



4 Matriisiyhtälöt ilman kaavioita

Tässä lähteenä Greenacren kirja (ca in practice) ja sen liite Theory of CA. Muistiinpanoja löytyy, joissa viitataan myös Biplots in practice - kirjaan. Ei valmis, lähinnä kaavojen kirjoittelun harjoittelua.

Korrespondenssianalyysin perusyhtälöt:

Datamatriisin $\boldsymbol{N}$ alkiot ovat ei-negatiivisia (eli nollat sallittuja) ja samassa mitta-asteikossa (jos mitta-asteikko on intervalli- tai suhdeasteikko mittayksiköiden on oltava samoja), ja n on taulukon alkoiden summa. GDA-kirjassa on tarkennettu tätä vaatimusta ei-negatiivisuudesta.

Korrespondenssimatriisi $\boldsymbol{P}$ saadaan jakamalla matriisin $\boldsymbol{N}$ alkiot niiden summalla n (tai ehkä parempi merkintä N) . Merkitään matriisin $\boldsymbol{P}$ rivisummien vektoria $\boldsymbol{r}$ ja sarakesummien vektoria $\boldsymbol{c}$. Korrespondenssianalyysin termein nämä vektorit ovat rivi- ja sarakemassojen vektoreita, ja 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 $\boldsymbol{S}$.

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

tai

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

Toinen esitystapa on hyödyllinen, kun tarkastellaan CA:n yhteyksiä muihin läheisiin menetelmiin (erityisesti kai log ratio analysis of compositional data?). Ehkäpä siksi, että matriisin alkiolle elementtimuodossa saadaan vastaavasti kaksi esitystapaa. Ensimmäinen on

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

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

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

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

Näin saadaan standardikoordinaatit ja principal-koordinaatit riveille ja sarakkeille.

Rivien standardikoordinaatit

$$\boldsymbol{\Phi} = {\boldsymbol{D}_{\boldsymbol{r}}}^{-1/2} \boldsymbol{U} \tag{2}$$

Sarakkeiden standardikoordinaatit

$$\boldsymbol{\Gamma} = {\boldsymbol{D}_{\boldsymbol{c}}}^{-1/2} \boldsymbol{V} \tag{3}$$

Rivien principal-koordinaatit

$$\boldsymbol{F} = {\boldsymbol{D}_{\boldsymbol{r}}}^{-1/2} \boldsymbol{U} \boldsymbol{D}_{\boldsymbol{\alpha}} = \boldsymbol{\Phi} \boldsymbol{D}_{\boldsymbol{\alpha}} \tag{4}$$

Sarakkeiden principal-koordinaatit

$$\boldsymbol{G} = {\boldsymbol{D}_{\boldsymbol{c}}}^{-1/2} \boldsymbol{V} \boldsymbol{D}_{\boldsymbol{\alpha}} = \boldsymbol{\Gamma} \boldsymbol{D}_{\boldsymbol{\alpha}} \tag{5}$$