Korrespondenssianalyysi - geometrinen ja graafinen data-analyysin menetelmä (1.0)

Jussi Hirvonen

Paperin versiot		
Versio	muutokset	päivämäärä
0.1	runko ja harjoittelua	5.10.2017
	runko ja sisältöäkin, oikoluku (W)	5.10.2017
0.2	uusiksi	31.1.2018
0.3	työversio; kaikki teksti ja jäsennys	5.2.2018
0.4	työversio; lähteitä, tutkielman rakenne-osan muokkausta	6.2.2018
0.5	työversio; teksti uusiksi - turhat pätkät pois (v2)	6.2.2018
1.0	(v2) 6.2.2018 - siistitään tätä vielä 6.2.2018	2.3.2018 utf-8

Sisältö

1 Johdanto

Korrespondenssianalyysi (correspondence analysis, CA) soveltuu erityisesti (mutta ei ainoastaan) luokitteluasteikon muuttujien riippuvuuksien analyysiin. Menetelmän matemaattiset perusteet muotoili Jean-Paul Benzécri Ranskassa 60-luvulla, ja sosiologi Pierre Bourdieun tutkimukset 1980-luvun alussa tekivät siitä kansainvälisesti tunnetun (historiasta kts. [1], [12] ja Suomen osalta [9]). Yksinkertainen tai "klassinen" korrespondenssianalyysi tarkastelee kahta luokitteluasteikon muuttujaa, useamman muuttujan korrespondenssianalyysissä muuttujia on enemmän (multiple correspondence analysis, MCA).

Menetelmä on ei-parametrinen, ja teknisesti sen voi määritellä luokitteluasteikon muuttujien pääkomponenttianalyysiksi. Tärkein väline riippuvuuksien tutkimisessa on yleensä kaksiulotteinen kuva, jossa kahden muuttujan tapauksessa esitetään molemmat muuttujat.

Tutkielman nimessä termi "geometrinen data-analyysi" on uudehkosta ranskalaisten tutkijoiden oppikirjasta [12]. He korostavat alkuperäisiä ideoita, joissa menetelmän perusta – abstrakti algebrallinen teoria – on keskeinen asia. Tämä varsin vaikea esitystapa on luultavasti hidastanut menetelmä yleistymistä, mutta on toki perusteltu. On hieman vaikeampi ymmärtää, miksi menetelmän esittäminen matriisiyhtälöinä tuomitaan. Sen perusta on tunnettu singulaariarvohajoitelma (singular value dekomposition), ja tavoitteena on muista monimuuttujamenetelmistä tuttu dimensioiden vähentäminen.

Tässä työssä pyrin esittämään korrespondenssianalyysin graafisena data-analyysin menetelmänä, ja esitän matemaattiset perusteet omana jaksona. Michael Greenacren oppikirjat ovat tehneet menetelmää tunnetuksi juuri tässä hengessä ([8]). Tilastollisesti perusteltujen graafisten analyysimenetelmien esittely on perusteltua, sillä ranskalainen "matriisilaskennan kritiikki" lienee oikeassa ainakin siinä, että matriisiyhtälöiden kautta menetelmä soveltaminen ei monelle tule ymmärrettäväksi. Kun menetelmän ongelmalliset kohdat juuri graafisessa esitystavassa ovat melko hyvin tiedossa (kts. esim. [3]), voi esityksen rakentaa oikean datan kautta eteneväksi. Oikea aineisto on oleellinen, sillä eräs menetelmän alkuperäinen tavoite oli juuri suurten aineistojen analyysissä ([12] s. 15, "on erittäin vaikeaa osoittaa kotiakvaarioissa, että verkko on tehokas").

Graafinen data-analyysi ei korrespondenssianalyysisäkään rajoitu vain yhden lopullisen kuvan esittämiseen, vaan se on vaiheittain etenevä tutkimusprosessi. Se on myös taitolaji, ja graafinen esitys elää analyysin mukana alun eksploratiivisista ja kuvailevista versioista lopullisiin johtopäätökset kiteyttäviin visualisointeihin. Tässä matriisiyhtälöiden ymmärtämisellä on roolinsa, sillä analyysin tarkentaminen vaati myös ymmärrystä siitä, miten kuvaa voi muokata tarkoituksenmukaiseksi.

Tutkielma kirjoitetaan lukijalle, joka tuntee jonkun verran tilastollisia menetelmiä ja haluaa tehdä data-analyysiä.

2 Tutkimusongelma

1. Tutkimusongelma

   Tutkielman ydin on esitellä kahden luokittelumuuttujan korrespondenssianalyysi ja sen eräitä laajennuksia. Painopiste on graafisten menetelmien esittely esimerkkiaineistolla, joka on riittävän laaja ja monipuolinen. Yksinkertainen kahden luokittelumuuttujan korrespondenssianalyysi antaa graafisen analyysin "...perussäännöt tulkinnalle. Kaikki muut korrespondenssianalyysin muodot ovat saman algoritmin soveltamista toisen tyyppiisiin datamatriiseihin, ja tulkintaa sovelletaan vastaavasti (with the consequent adaptation of the interpretation"([5], s. 437).

   Tulkinnat eivät kuitenkaan ole aivan yksinkertaisia. Greenacre ehdottaa artikkelissaan ([3], s. 20-) viittä sääntöä yksinkertaisen korrespondenssianalyysin ongelmakohtien välttämiseen. Ne liittyvät oleellisesti graafisen analyysin tulkintapulmiin ja ns. skaalausongelmaan (symmetriset ja epäsymmetrisen kuvat, standardi- ja pääkooridinaatit, standard and principal coordinates) .Näistä perusasioista käydään edelleen keskustelua (kts. esim. [2]).

   Graafiset menetelmät ovat myös taitolaji, kuvissa on askel kerralaan päästävä kokeiluista kohti selkeää esitystä jossa toivon mukaan on vain oleelllinen informaatio ([13].

   Korrespondenssianalyysi sopii kaikkien sellaisten aineistojen analyysiin, joissa data voidaan esittää jonkinlaisina tutkimusongelman kannalta järkevinä lukumäärinä. Teknisesti taulukon lukujen on oltava ei-negatiivisia (tosin LeRoux et. al. sallivat myös negatiiviset luvut tietyin ehdoin [12] , s. 60) ja mitta-asteikon on oltava sama (lukumääriä, metrejä, euroja jne). Luokitteluasteikko on tavallaan hyvin lievä oletus, ja menetelmä sovelutuu mainiosti esimerkiksi järjestysasteikon muuttujan analyysiin. Sen avulla nähdään selvemmin kyselytutkimusten Likert-asteikon "oikea" mitta-asteikko, useinhan se oletetaan tasaväliseksi. Yksinkertaisen korrespondenssianalyysin laajennukset mahdollistavat myös monipuolisempia tutkimusasetelmia ja lisäinformaation käytön.

2. Tavoitteet

   (a) Historiaa tiiviisti Sivuutan laajemman menetelmän historian käsittelyn, mutta sanon siitä toki jotain. Historiaa esitellääään lyhyesti ja yleisesti [12]. Kahman väitöskirjan [9] johdantoluku kertoo menetelmän käytöstä suomalaisessa sosiologiassa. Kahden luokittelumuuttujan korrespondenssianalyysistä löytyy

(simple ca) bibliografinen katsaus vuodelta 2004 [1]. CA ja sen kaltaiset menetelmät – lähisukulaiset – ovat olleet käytössä monilla tieteenaloilla. On tavallaan keksitty muutaman kerran uudestaan (Japanissa, "homogeenisyysanalyysinä (Gini)" , "spektirianalyysinä") ja sillä on yhteyksiä myös ns. proportional data (koostumus-data?)menetelmiin ([8]). Taustalla on melko yleinen data-analyysin tutkimusongelma.

(b) Yksinkertaisen korrespondenssianalyysin esittely

Tavoite on esitellä vaiheittain korrespondenssianalyysin peruskäsitteet ja graafisen esitykset niin, että ne voi ymmärtää perehtymättä vektoriavaruuksien isomorfimismeihin tai matriisialgebran perusteisiin. Samalla menetelmän matemaattinen rakenne kuvataan riittävän täsmällisesti, jotta siihen voidaan viitata kun graafisen menetelmien pulmakohtia selvennetään. Matemaattisen osan tavoite on myös auttaa näkemään korrespondenssianalyysi eräänä monimuuttujamentelmänä, ja yhtenä tapana analysoida moniulotteisia aineistoja. Graafinen data-analyysi vaatii myös kuvien muokkaamista, ja silloin on kyettävä hahmottamaan korrespondenssianalyysin tulostietojen sisältö.

3. Rajaus - mitä ei tutkita)

(a) ekologisessa tutkimuksessa suosittu "detrended ca"

(b) oppihistoria, erityisesti CA:n toisistaan riippumattomat "keksimiset"

(c) Bourdieun tutkimuset ja niiden inspiroima kansainvälinen tutkimus, joka jatkuu edelleen

(d) laskennalliset (algoritmi) ongelmat, esim. harvojen (sparse) data-matriisien käsittely

(e) sovellusalueet ohitan vain maininnalla (biologia, arkeologia, lääketiede, kirjallisuus/kielitiede)

(f) Otantatutkimus ja korrespondenssianalyysi

Perinteisesti korrespondenssianalyysissä on korostettu jyrkkää eroa "todennäköisyysteoreettiseen" tilastolliseen päättelyyn, ja tähän vastakkainasettelua pitää hieman kuvata. Monissa tapauksissa todennäköisyysteoreettiset käsitteet tulevat mukaan analyysiin, mutta sivuutan ainakin otantatutkimuksen (esim. analyysipainojen käyttö).

4. Puuttuva osa - data-analyysin lähestymistavat

Jätän viimeiseksi aiheen, joka on kiinnostava mutta vaikeasti hahmotettava. Korrespondenssianalyysi kuuluu tilastollisten menetelmien joukossa eksploratiivisiin (vs. konfirmatorinen), ei-parametrisiin mentelmiin. On katsottava dataa, datan ehdoilla. Tämä korostus on ollut vahva menetelmän ranskalaisilla kehittäjillä, ja ehkä muillakin. Jotain tästä pitää sanoa, mutta mitä?

3 Tutkielman rakenne ja esitystapa

Tutkielman sisältö ja rakenne

1. Yksinkertainen korrespondennsianalyysi

   (a) Johdatteleva esimerkki - kahden luokittelumuuttujan frekvenssitaulukko
   - peruskäsitteet: lähteinä MG:n oppikirjat ja luentomateriaalit, sovelutuvin osin myös [12].

   - CA:n oletuskartta (symmetrinen kartta)

   - tulkinnan perussäännöt; yksinkertainen "rautalanka-metodologia" tiiviisti ([12]).
   - "CA jargon": profiilit, massat, inertia, modaliteetit

   (b) Varianssianalyysi - yksiulotteinen korrespondenssianalyysi - ehkä hyvä didaktinen keino

   (c) Kaksoiskuvat (biplots) ja korrespondenssianalyysi
   Korrespondenssianalyysin pääväline – kartta – ei ole aivan yksinkertainen, ja tulkinnoissa voi ajautua harhateille. Korrelaatiodiagrammaa muistuttavaan kaksiulotteiseen kuvaan on projisoitu kaksi pistepilveä, ja kartan "lukeminen" on pisteryhmien välisten ja sisäisten etäisyyksien oikeaa ymmärrystä.

   Kaksoiskuva (biplot) on yleinen geometrinen menetelmä havaintojen ja muuttujien graafiseen esittämiseen samassa kuvassa. Korrespondenssianalyysin kartat ja niiden rajoitukset on hyvä ymmärtää tässä yleisemmässä kehikossa. Lähteinä ([7], [3].
   Tämä jakso voisi olla omana kokonaisuutena?
   Tämän osuuden jälkeen seuraavassa jaksossa sovelletaan näitä käsitteitä data-analyysissä.

2. Yksinkertaisen korrespondenssianalyysin laajennuksia

   Tässä jaksossa täydennetään alustavaa esimerkkianalyysiä, ja samalla esitellään korrespondenssianalyysin graafisia menetelmiä laajemmin. Konribuutio-kuvat ja inertian dekomponointi ennen muuta, mutta kuvien "säätämisessä" on myös paljon vaihtoehtoja.

   Tärkeimmät lähteet lähteet [6] ja [8].

(a) Täydentävät muuttujat (supplementary points)

(b) ”Stacked and concatenated tables/matrices”

(c) “ABBA”

(d) Osajoukon korrespondenssianalyysi (subset CA)

-

3. Usean muuttujan korrespondenssianalyysi

   (a) Indikaattorimatriisi ja Burtin matriisi - pulmia tulkinnassa ja tuloksissa

   (b) JCA - Joint Correspondence Analysis

4. Korrespondenssianalyysi regressiomallin kaltaisessa tutkimusasetelmassa (kanoninen CA) Suosittu ympäristö- ja biotieteissä, esimerkiksi lajiston esiintymisdataa jota halutaa analysoida “regressiotyyliin” ehdollisesti selittavien (mahdollisesti jatkuvien) muuttujien suhteen ([4] ja [8]).

5. Korrespondenssianalyysi ja koostumusdata (compositional data)

   Nämä kaksi jälkimmäistä eivät nyt ehkä ole ihan loppuun asti harkittuja. MG:n kirjasta löytyy muutakin.

6. Matched matrices

7. Square tables

8. Korrespondenssianalyysin ja muiden monimuuttujamenetelmien yhteyksistä

   (a) Korrespondenssianalyysi ja pääkomponenttianalyysi

   (b) Korrespondenssianalyysi ja moniulotteinen skaalaus

   (c) Luokittelu- ja erottelumenetelmät

   - lähteinä Mustosen ja Vehkalahden kirjat ([11], [14]). - vain lyhyesti, koska joskus CA:n graafisista esityksistä virheellisesti “tunnistetaan” ryhmiä tms.

4 Lähdeaineisto ja data

Koitan pitää lähdeluettelon kohtuullisen kokoisena.

4.1 Kirjallisuus ja artikkelit

Tutkielman tärkein lähdeaineisto ovat Greenacren oppikirja [7] ja muut artikkelit. Julkaisu kaksoiskuvista (biplots) [8] esittää korrespondenssianalyysille oleelliset kaaviot osana yleisempää graafista analyysitapaa.

Käytän keväällä 2017 HY:ssa järjestetyn kurssin luentokalvoja, ainakin muistilistan tapaan. Niissä ei liene mitään uutta, joten viittaukset ovat kirjallisiin lähteisiin. LeRouxin ja Rouanetin [12] kirja esittelee menetelmä ranskalaisen perinteen mukaisesti, uudella otsikolla (geometrinen data-analyysi). Uppsalan luentokalvot (LeRoux) ovat hyvä tiivistys. En tiedä kuinka laajasti tätä selostan, mutta metodologiset ideat ja metodin käytännön ohjeet kannattaa ainakin poimia. Koitan ainakin mainita muutaman suomalaisen tutkimuksen, ja viitata muutamaan löytämääni luetteloon. Kulttuurisia eroja Suomessa on tutkittu ranskalaisten esikuvien (Bourdieu) tyyliin kartan käsittein (cultural map) [10], ja Nina Kahman väitöskirjassa "Yhteiskuntaluokka ja maku" ([9] sovelletaan myös useamman muuttujan korrespondenssianalyysiä (MCA). Jari Oksanen (http://cc.oulu.fi/ jarioksa/) on julkaissut biotieteiden R-paketteja, ja osallistunut myös keskusteluun ns. "detrented CA" - menetelmän hyvistä ja huonoista puolista (kts. esim. [1] ja siinä mainitut lähteet). Olen silmäilly myös joitakin kotimaisia pro gradu-tutkielmia, mutta jätän ne tässä vaiheessa sivuun.

4.2 Data ja R-koodi

Aineisto on ISSP:n viimeisin (2012) "Family and Changing Gender Roles" - data. Se antaa hyvän mahdollisuuden laajentaa analyysiä yksinkertaisesta yhä isompaan ja mutkikkaampaan. Sama data vuodelta 2002 oli käytössä kevään 2017 MCA-kurssilla. Aion käyttää myös kurssin laskuharjoitusten r-koodia mallina.

Aineisto on vapaasti ladattavissa, mutta lähdeviitteistä se vielä puuttuu.

Ohjelmisto on R, en esittele muita laajasti käytettyjä tilasto-ohjelmia (esim. SAS, SPSS). Pieni viite kuitenkin, kun Survo tässäkin oli aikanaan eturivissä.

Kirjoitan tutkielman LateX:lla, ja data-analyysin Rmarkdown-raportteina. Tässä kokeilen ja haen vielä oikeita asetuksia ympäristölle.

Omia muistiinpanoja:

ISO-8859-1 vai UTF8, doctype (raport vai article) jne.

Kuinka tarkasti on kuvattava datan muunnokset sopivaan R-muotoon?

Selvitettävä Rmarkdown -> LateX - yhdistelmä

Viitteet

[1] Eric J. Beh. Simple correspondence analysis: A bibliographic review. *International Statistical Review*, 72(2):257–284, 2004.

[2] Magne Flemmen and Johannes Hjellbrekke. Response: Not so fast: a comment on atkinson and deeming's ?class and cuisine in contemporary britain: the social space, the space of food and their homology? *The Sociological review*, 64(1):184–193, 2016.

[3] Michael Greenacre. Tying up the loose ends in simple correspondence analysis. 2001.

[4] Michael Greenacre. Canonical correspondence analysis in social science research. In Hermann Locarek-Junge and Claus Weihs, editors, *Classification as a Tool for Research*, pages 279–286, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ID: 10.1007/978-3-642-10745-$0_3$0.

[5] Michael Greenacre and Trevor Hastie. The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82(398):437–447, 06/01 1987. doi: 10.1080/01621459.1987.10478446.

[6] Michael Greenacre and Rafael Pardo. Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods Research*, 35(2):193–218, 11/01; 2018/02 2006. doi: 10.1177/0049124106290316; 06.

[7] Michael J. Greenacre. *Biplots in Practice*. Fundacion BBVA, Bilbao, Spain, 2010.

[8] Michael J. Greenacre. *Correspondence analysis in practice*. CRC Press, Boca Raton, Florida, third edition edition, 2017.

[9] Nina Kahma. Yhteiskuntaluokka ja maku, 2011-09-23.

[10] Nina Kahma and Arho Toikka. Cultural map of finland 2007: analysing cultural differences using multiple correspondence analysis. *Cultural Trends*, 21(2):113–131, 06/01 2012. doi: 10.1080/09548963.2012.674751.

[11] Seppo Mustonen. *Tilastolliset monimuuttujamenetelmät*. Survo Systems, Helsinki, 1995.

[12] Brigitte Le Roux and Henry Rouanet. *Geometric data analysis: from correspondence analysis to structured data analysis*. Kluwer Academic Publishers, Dordrecht, 2004.

[13] Antony Unwin. *Graphical data analysis with R*. CRC Press, Taylor Francis, Boca Raton, 2015.

[14] Kimmo Vehkalahti. *Kyselytutkimuksen mittarit ja menetelmät.* Tammi, Helsinki, 2008.