G Luku 1 Yksinkertainen korrespondenssianalyysi

Jussi Hirvonen

versio 1.04, tulostettu 2018-08-10

Sisältö

Kommentteja ja versionhallintaa:

- **edit:** oma kommentti, ei varsinaista tekstiä
- kirjastot/paketit ladataan jokaisessa Rmd-dokumentissa
- bib-formaatin viitetietokantaa tullaan kokeilemaan
- kuvasuhde (aspect ratio) edelleen epäselvä juttu! Mutta näyttää PDF-tulosteessa olevan ok.
- Datan käsittely ja hallinta +SPSS:n sallima kolme puuttuvan tiedon koodia saadaan mukaan read_spss-funktion (haven) parametrilla USER_NA = TRUE (mutta tarkistettava!) (25.4.18)
  - faktoreita ei ainakaan toistaiseksi muuteta ordinaaliasteikolle, CA ei tästä välitä
  - pidetään muuttujien ja tiedostojen nimeäminen selkeänä, tarkistetaan aika ajoin
- Taulukot: lisättiin riviprosentti- ja sarakeprosenttitaulut (25.4.18), kuva riviprofiileista puuttu vielä (15.5.2018)
- Datan esittelyssä on turhaa välitulostusta, ja samoin vähän muuallakin. Html on helpompi lukea, kun koodi on oletuksena piilossa
- PDF-tulosteessa koodi pääsääntöisesti näkyy toistaiseksi
- kokeiluja CA-karttojen tulostamiseen (a) suoraan koodilla ja (b) r-grafiikkaikkunasta tallennetun pdf-kuvan avlla. Paras toistaiseksi (a), jätin kokeilu näkyviin. Analyysit R:n grafiikkaikkunassa, jotta asp=1, ja tulkintaa varten voi tallentaa PDF-muodossa.
- rakenteeseen muutoksia (näkyvät sisällysluettelossa), ei erillistä teorialiitettä vaan sopivina annoksina. Lukuun 3 perusasiat, kaavat, määritelmät
- tehdään käsitetaulukko (kirjoittamista varten)
- 20.5.2018 (a) tulkita-osuuteen karttakuvia ja ca-tulokset (b) siistimpi taulukoiden tulostus löytyi (c) kaavaliite laajeni (dispo-haarassa)
- 23.5.2018 lisätään dataan toinen maa-muuttuja maa2, ikäluokkamuuttuja age_cat ja iän ja sukupuolen vuorovaikutusmuuttuja ga.
- 24.5.2018 lisättiin ca-kartta, jossa Saksan ja Belgian ositteet ja summarivit täydentävinä (passiivisina)

6.8.2018 versio 1.0

Siistitään tämä versio.

1 Data

edit tässä luvussa on paljon siistittävää, mutta data on ok. (13.5.2018)

Ladattavat paketit omana r-skriptinä (paketit.R), ei listata tilan säästämiseksi.

1.1 Luvun 1 tavoitteet

Datan esittely ja kuvailut - ehkä myös oma kappale muista mahdollisista lähestymistavoista aineiston analyysissä (deskriptiiviset, todennäköisyysteoriaan eksplisiittisesti perustuvat kuten MFA)

1. Eksploratiivinen ja graafinen menetelmä tarvitseen aineiston, hankalaa esitellä jollain synteettisellä esimerkkiaineistolla
2. CA (ja MCA) sopivat isojen moniulotteisten ja mutkikkaiden aineistojen analyysiin
3. Aineiston esittely, laajan kyselytutkimusaineiston tyypilliset ominaisuudet
4. Laadukkaan ja hyvin dokumentoidun aineiston edut
5. Tärkeä rajaus: CA sopii ja sitä on käytetty myös hyvin toisen tyyppisiin aineistoihin (ekologia ja biologia, arkeologia, kielen tutkimus)

1.2 Perhe ja muuttuvat sukupuoliroolit - ISSP:n kyselytutkimuksen data 2012

Hieman historiaa datasta, sosiaalisesti määräytyneen sukupuoliroolit (gender) tutkimusaiheena neljässä kansainvälisessä kyselytutkimuksessa.

ISSP Research Group (2016): International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012. GESIS Data Archive, Cologne. ZA5900 Data file Version 4.0.0, doi:10.4232/1.12661 **tämä doi-linkki ei toimi**

Aineistot 2012 **toimii - ja viitetieto tuossa edellä!**

Muuttujakuvaukset ja muut tiedot

Data ja dokumentit

Suomenkielinen lomake (ZA5900_q_fi-fi.pdf)

Käyttöehdot:

Tiedonkeruumenetelmä ja otoskoko: Viimeisin Portugali 29.06.2014 - 31.01.2015, ensimmäinen Bulgaria 16.08.2011 - 20.09.2011. Suurin osa muista 2012-13, kuten Suomi (21.09.2012 - 07.12.2012).

Havaintojen lukumäärät voi tarkistaa täältä .

edit: aineiston kuvailua voi ja kannattaakin jatkossa tarkentaa, ja laittaa se liitteeksi(?- tuskinpa). Dokumentointi on hyvin tarkka, tiedot löytyvät haastattelumenetelmista (parerilomake, tietokoneavusteinen haastattelu, jne), maakohtaisten taustamuuttujien harmonisoinnista maittain, otantamenetelmistä jne. Esittelen vain aineiston tärkeimmät rajaukset.

1.3 Aineiston rajaaminen

zxy Aineiston kuvailu omana osanaan (7.8.2018)

Ainestossa (jatkossa ISSP2012) on kyselytutkimukseen tulokset 41 maasta. Lisäksi aineistossa on runsaasti demografisia ja muita taustatietoja. R-koodista selviää käytetty versio (SPSS-tiedoston nimi) ja rajauksessa käytetyt muuttujat.

Rajaukset

zxy Miksi Espanja roikkuu mukana, ja Unkari? Pudoteaan ne elegantisti heti alussa, erilaisia kysymyksiä. **Aika iso ongelma datassa!**

1. Eurooppa ja samankaltaiset maat (28)

Bulgaria, Czech Republic, Denmark, Finnland,France, Germany, Great Britain, Ireland, Latvia, Lithuania, Norway, Poland, Sweden, Slovakia Slovenia, Spain, Switzerland, Australia, Austria, Canada, Croatia, Iceland, Russia, United States, Belgium, Hungary, Netherlands, Portugal

Pois jätettiin 13: Argentiina,Turkki, Venezuela, Etelä-Afrikka, Korea, Intia, Kiina, Taiwan, Filippiinit, Meksiko, Israel, Japani, Chile.

2. Maat joissa varsinaisissa tutkimuskysymyksissä on käytetty poikkeavia luokitituksia tms. Esimerkiksi Espanjan datassa on jätetty pois neutraali “en samaa enkä eri mieltä” - vaihtoehto, Unkarin datassa on omia versioita kysymyksistä jne. Espanja jätetään ainakin aluksi pois vertailukelpoisuuden vuoksi, Unkari ehkä myös.

3. kaikki havainnot, joissa on puuttuvia tietoja. Tämä rajaus on kyselytutkimuksessa ankara, tai oikeastaan kelvoton. Oikea menettely olisi imputoida jollain menetelmällä puuttuvat tiedot, mutta rajaan otantatutkimuksen menetelmät tutkielman ulkopuolelle (aiheesta löytyy artikkeleita...).Yksittäisten vastausten puuttuminen eli erävastauskato ohitetaan aluksi, mutta siihen palataan. Korrespondenssianalyysiin on helppo ottaa mukaan myös puuttuvat tiedot, sillä data on luokitteluasteikon dataa. Yksikkövastauskato eli otokseen poimitut joita ei ole tavoitettu ollenkaan on kansallisen tason ongelma, joka on ratkaistu vaihtelevin tavoin. Tiedot löytyvät aineiston dokumentaatiosta. Aineistossa on myös mukana painomuuttujat, mutta ne soveltuvat vain jokaisen maan omaan aineistoon. **zxy** Tärkein rajaus esimerkkianalyyseissä, ja voidaan esitellä CA:n käyttö puuttuvien vastausten analysoinnissa (Likert-asteikkolla).

edit: Tähän täsmennetään miten puuttuvia tietoja käsitellään.

4. Datan hallinta **liittyy reproducible research- periaatteeseen**

Aineistoa käsitellään ja muokataan niin, että jokaisen analyysin voi mahdollisman yksinkertaisesti toistaa suoraan alkuperäisestä datasta.

Aineiston muokkauksen (muuttujien ja havaintojen valikointi, muunnokset ja uusien muuttujien luonti jne.) dokumentoidaan r-koodiin.

```
# kolme maa-muuttujaa datassa. V3 erottelee joidenkin maiden alueita, V4 on koko maan
#two country code variables based on the ISO Code 3166. One identifies
#countries as a whole, the other one possible subsamples, such as East and West Germany. The cross
#tabulations shown in this Variable Report are based on a third, alphanumerical country code variable,
#which also identifies subsamples."
#V3 - Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)
# V3 erot valituissa maissa
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East
```

```
# 62001 PT-Portugal 2012: first fieldwork round (main sample)
# 62002 PT-Portugal 2012: second fieldwork round (complementary sample)
# Myös tämä on erikoinen, näyttää olevan vakio kun V4 = 826:
# 82601 GB-GBN-Great Britain
# Portugalissa ainestoa täydennettiin, koska siinä oli puutteita. Jako ei siis ole oleellinen,
# mutta muuut ovat. Tähdellä merkityt maat valitaan johdattelevaan esimerkkiin.
# Maat:
# 36 AU-Australia
# 40 AT-Austria
# 56 BE-Belgium*
# 100 BG-Bulgaria*
# 124 CA-Canada
# 191 HR-Croatia
# 203 CZ-Czech Republic
# 208 DK-Denmark*
# 246 FI-Finland*
# 250 FR-France
# 276 DE-Germany*
# 348 HU-Hungary*
# 352 IS-Iceland
# 372 IE-Ireland
# 428 LV-Latvia
# 440 LT-Lithuania
# 528 NL-Netherlands
# 578 NO-Norway
# 616 PL-Poland
# 620 PT-Portugal
# 643 RU-Russia
# 703 SK-Slovakia
# 705 SI-Slovenia
# 724 ES-Spain
# 752 SE-Sweden
# 756 CH-Switzerland
# 826 GB-Great Britain and/or United Kingdom
# 840 US-United States
#
# Belgian ja Saksan alueet:
#  V3
#  5601     BE-FLA-Belgium/ Flanders
#  5602     BE-WAL-Belgium/ Wallonia
#  5603     BE-BRU-Belgium/ Brussels
# 27601     DE-W-Germany-West
# 27602     DE-E-Germany-East

#valittavien maiden kolminumeroinen ISO 3166 - koodi vektoriin
incl_countries <- c(36, 40, 56,100, 124, 191, 203, 208, 246, 250, 276, 348, 352, 372, 428, 440,
                    528, 578, 616, 620, 643, 703, 705, 724, 752, 756, 826, 840)
ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav", user_na = TRUE)
#str(ISSP2012.data)
#
# lisäys 25.4.2018 user_na
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be # converted to NA"
```

```
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
#
#
#str(ISSP2012.data) #61754 obs. of  420 variables
ISSP2012jh1.data <- filter(ISSP2012.data, V4 %in% incl_countries)
#length((ISSP2012jh1.data))
#names(ISSP2012jh1.data)
#str(ISSP2012jh1.data) #37816 obs. of  420 variables
#
#EDIT: tiivistä, nämä ovat vain kokeiluja ja datan kaivelua (15.4.2018)
#
# V5 - V67 kysymyksiä, joillain mailla omat vastaukset joihinkin omina muuttujina, esim.  #  ES_V5 muut
#$ V5      :Class 'labelled'  atomic [1:37816] 5 1 2 2 1 NA 2 4 2 2 ...
#  .. ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
#  .. ..- attr(*, "format.spss")= chr "F1.0"
#  .. ..- attr(*, "labels")= Named num [1:8] 0 1 2 3 4 5 8 9
#  .. .. ..- attr(*, "names")= chr [1:8] "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree
# $ ES_V5   :Class 'labelled'  atomic [1:37816] NA NA NA NA NA NA NA NA NA NA ...
#  .. ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
#  .. ..- attr(*, "format.spss")= chr "F1.0"
#  .. ..- attr(*, "display_width")= int 4
#  .. ..- attr(*, "labels")= Named num [1:7] 0 1 2 3 4 8 9
#  .. .. ..- attr(*, "names")= chr [1:7] "NAP: other countries" "Strongly agree" "Agree" "Disagree" ...
#HU_V18
#V18$label
#attr(ISSP2012jh1.data$V6,'labels')
#attr(ISSP2012jh1.data$ES_V6,'labels')
```

Yllä esimerkiksi muuttujan V6 metatiedot. Perusvaihtoehdot ovat 1 - 5, ja joillain mailla on vaihtoehtona ollut myös “Can’t choose”, muilla taas on vain puuttuva tieto (No answer, 9).

Espanjan aineiston metatiedot muuttujulla ES_V6 taas ovat

```
attr(ISSP2012jh1.data$ES_V5,'labels')
```

```
## NAP: other countries       Strongly agree                Agree
##                    0                   1                   2
##             Disagree    Strongly disagree         Can't choose
##                    3                   4                   8
##            No answer
##                    9
```

```
temp1 <- ISSP2012jh1.data %>% filter(V4 == 724) %>% select(ES_V6, C_ALPHAN)
#str(temp1)
temp1$ES_V6 <- factor(temp1$ES_V6 )
summary(temp1)
```

```
##  ES_V6      C_ALPHAN
##  1: 195   Length:2595
##  2:1117   Class :character
##  3: 898   Mode  :character
##  4: 278
##  8:  91
##  9:  16
```

```
#typeof(ISSP2012jh1.data)
#class(ISSP2012jh1.data)
#storage.mode(ISSP2012jh1.data)
#attributes(ISSP2012jh1.data)
```

##Puuttuvat tiedot

zxy Perusasiat havaintojen puuttellisuudesta kyselytutkimusissa. Yksikkövastauskato (unit non-response), eräsvastauskato (item non-response). Mitä on raportoitava, kun käytetään valmista aineistoa? Erävastauskato on silti ongelma, vaihtelee kysymyksittäin, vaikka se ei kovin suuri olekaan.

Yksikkövastauskato on (onko?) otettu huomioon, kun kyselyn toteuttaja on editoinut ja tarkastanut datan.
Viittet

Aineistossa on tarkempi kolmen luokan koodaus puuttuvalle tiedolle, mutta toistaiseksi sitä ei käytetä.

Muiden kuin Espanjan vastaukset kysymykseen V6 jakautuvat näin:

```
temp2 <- ISSP2012jh1.data %>% filter(!(V4 == 724)) %>% select(V6, C_ALPHAN)

#str(temp1)
temp2$V5 <- factor(temp2$V6 )
temp2$maa <-  factor(temp2$C_ALPHAN)
#summary(temp2)
#str(taulu1)
taulu1 <- temp2 %>% tableX(V6,maa,type = "count")
taulu1a <- taulu1[ ,1:14]
knitr::kable(taulu1a,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset maittain")
```

Taulukko 1: Kysymyksen V6 vastaukset maittain

	AT	AU	BE	BG	CA	CH	CZ	DE	DK	FI	FR	GB-GBN	HR	HU
1	218	82	193	118	51	89	174	165	70	47	256	37	75	219
2	447	405	454	395	215	431	392	376	238	188	551	247	265	288
3	171	285	440	205	181	222	403	199	152	149	424	208	190	225
4	205	568	554	190	317	365	415	538	232	423	469	331	327	190
5	98	215	381	13	194	112	355	441	696	303	624	105	133	75
Missing	43	57	180	82	14	18	65	47	15	61	85	22	10	15
Total	1182	1612	2202	1003	972	1237	1804	1766	1403	1171	2409	950	1000	1012

```
taulu1b <- taulu1[ ,15:28]
knitr::kable(taulu1b,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset maittain")
```

Taulukko 2: Kysymyksen V6 vastaukset maittain

	IE	IS	LT	LV	NL	NO	PL	PT	RU	SE	SI	SK	US	Total
1	56	13	50	188	59	23	110	73	244	29	39	117	86	2881
2	250	138	438	395	296	186	395	495	542	124	272	246	350	9019
3	197	186	396	156	242	226	155	157	360	219	200	229	652	6829
4	478	552	220	209	445	579	365	215	254	276	365	298	196	9576
5	197	271	22	38	196	365	64	52	42	354	131	198	0	5675
Missing	37	12	61	14	77	65	26	9	83	58	27	40	18	1241
Total	1215	1172	1187	1000	1315	1444	1115	1001	1525	1060	1034	1128	1302	35221

	IE	IS	LT	LV	NL	NO	PL	PT	RU	SE	SI	SK	US	Total

Esimerkiksi Ruotsin puuttuviksi tiedoiksi koodatuista 29 havainnosta 19 valitsi “can’t choose”(8) ja 10 kieltäyti vastaamasta (9) tms. Dokumentti, s.12.

Tarkastellaan aineston puuttuvia havaintoja hieman tarkemmin. Puuttuvat tiedot on koodattu aineistoon näin: 0: Not applicapble (NAP), Not available (NAV) 7: (97,997, 9997,. . .): Refused 8: (98, 998, 9998,. . .): Don’t know 9: (99, 999, 9999,. . .): No answer

NAP ja NAV määritellään

"GESIS adds ‘Not applicable’(NAP) codes for questions that have filters. NAP indicates that only a subsample and not all of respondents were asked. Also in the case of country spesific variables, all the other countries are coded NAP.

GESIS adds ‘Not available’ for variables, which in singe countries may not have been conducted for whatever reason."

zxy Miten nämä tarkemmat tiedot (7, 8, 9 saadaan näkyviin?)

EDIT: Puuttuneisuuden lyhyttä kuvailua, ja rajausten vaikutus havaintojen lukumäärään muutamaan taulukkoon. Voi siirtää liitteisiin (25.4.2018)

Lyhyt taulukko, jossa maittain ja muuttujittain puuttuneiden tietojen osuus.

###Poikkeavat kysymykset

zxy Myös maakohtaiset erot, ja niiden vaikutus aineiston rajaamiseen

zxy yksi kappale: Aineitoa on harmonisoitu, kysymyksiä hiottu, vertailukelpoisuuteen on pontevasti pyritty. Silti eroja löytyy, osa ymmärrettäviä (lisäkysymykset jne) ja osa ei (Espanja!). Tällaista on kansainvälisen kyselytutkimuksen data.

edit: nämä merkinnät ovat muistiinpanoja, kun tarkemmin luin muuttujadokumenttia. Kysymyksissä on vaihtelua, ja tavallaan niin pitääkin olla kansainvälisessä kyselytutkimuksessa. Vastaajien on ymmärrettävä kysymyksen suurinpiirtein samalla tavalla. Kaikki on tarkasti dokumentoitu.

edit: täsmennettävä, periaatteessa vastaukset on harmonisoitu. Joistain maista joku tieto puuttuu, jos sitä ei ole kysytty. Joissain tapauksissa kysymysten vaihtoehdot poikkeavat standardista.

Aineistossa on ns. substanssimuuttujia 63 (V5 - V67). Suurin osa on kerätty jollain haastattelumenetelmällä, ja yleisin vastausvaihtoehto on viiden arvon Likert-skaala (1 = täysin samaa mieltä, samaa mieltä, en samaa enkä eri mieltä, eri mieltä, täysin eri mieltä =5). Eri maiden lomakkeissa on vaihtelua puuttuviksi tiedoiksi koodattujen muiden vastausten välillä. **Esimerkikisi Suomen lomakkeessa on kuudes vaihtoehto “en osaa sanoa”, ja lisäksi on koodattu vastaamisesta kieltäytyminen tai muuten puuttuva tieto.** Ensimmäisessä aineiston rajauksessa nämä kaikki jätetään pois, käytetään “yleistä” puuttuvan tiedon määritelmää (eli joku noista em.).

Espanjan lisäksi Unkarin osatutkimuksessa kysymyksen V18 V19 V20 vastausvaihtoehdot ovat poikkeavat siten, että keskimmäinen neutraali vaihtoehto on jätetty pois (em.dok, s. 48).

Islannissa kysymykseen V28 (Consider a couple who both work full-time and now have a new born child. One of them stops working for some time to care for their child. Do you think there should be paid leave available and, if so, for how long?) on tarjolla oma vastausvaihtoehto ((97) “Yes, but don’t know how many months”). Kysymyseen "V29 - Q9 Paid leave: Who should pay ja V30(Paid leave: How to divide between parents) Bulgarian kysely on poikkeava (0 NAP (code 0,98 in V28), s. 91).

Hollannin vastausvaihtoehdoissa kysymykseen V35 (Elderly people: Provider of domestic help) on oma variantti “5 Empoyers”, jonka kuitenkin on valinnut vain 6 vastajaa (0,5 %).

V39, V40, V41, V42, V43, V44, V45, V46, V47, V48, V50, V51, V52, V53, V54: paljon poikkeamia, aika vaikeaselkoisia kysymyksiä. Näitä ehkä pitää tutkailla... V55 (Life in general: How happy on the whole) ok.

V56-57 poikkeamia, V58 (Health status) ok V59 "ketjutettu kysymys", samoin V60-V64. s. 174 - puolison koulutus...

---

2 Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko

jäsennystä

Tässä esitellään yksinkertainen esimerkki, yksi kysymys (esim. V6) ja muutamia maita ristiintaulukoituna. Johdatteluna aiheeseen esitellään ca-käsitteet profiili, massa ja reunajakauma. Havainnollistetaan rivi- ja sarakeprofiilien vertailua vastaaviin keskiarvoprofiileihin.

Taulukoita kannattaa tarkastella ensin rivien (kuva puuttuu) ja sitten sarakkeiden suhteen. Miten ne poikkeavat keskiarvostaan, miten toisistaan saman kategorian profiilista. Usein taulukoissa muuttujilla on selvästi eri rooli, kuten tässä. Koitamme hahmottaa maiden (=aggregoituja yksilöitä) eroja ja yhtäläisyyksiä. Sarakkeiden vertailussa taas näemme, miten muuttujien profiilit poikkeavat keskiarvostaan. Monia riippuvuksia ja poikkeamia näyttäisi olevan. Klassinen ongelma, Pearson ja Fisher (ehkä turhaa tässä?).

Toiseksi riippumattomuushypoteesi ja χ^2 - riippumattomuustesti (pieni huomautus - on monta tapaa testata taulukon riippuvuuksia). Riippumattomuushypoteesi ehdollisena todennäköisyytenä reunajakauman suhteen.

χ^2 - etäisyys, yhteys hajontaan eli inertiaan ca-terminologiassa.

Dimensioiden vähentämisen idea ("the essence"), joka ei pienessä taulossa ole ihan ilmeinen. Toinen tavoite on visualisointi, yleensä kaksiulotteisena kuvana (karttana).

Yksinkertainen korrespondenssianalyysi on kahden luokitteluasteikon muuttujan riippuvuuksien geometrista analyysiä. Lähtökohta on kahden muuttujan ristiintaulukointi, alkuperäinen data voi olla muillakin asteikoilla mitattua. Menetelmän ydin on tarkastella molempien muuttujien – taulukon rivien ja sarakkeiden – riippuvuuksia kaksiulotteisena kuvana. Kuvaa kutsutaan myös kartaksi, ja tulkinnan ensimmäinen askel on kartan "koordinaatiston" tulkinta. Kaikki etäisyydet kuvassa ovat suhteellisia, vain rivi- ja sarakepisteiden etäisyydet kuvan origosta voidaan tulkita tarkasti. Koordinaatiston tulkinta aloitetaan "katsomalla mitä on oikealla ja vasemmalla, ja mitä on ylhäällä ja alhaalla" (viite LeRoux et.al, Bezecri-sitaatti). Vaikka pisteiden etäisyyksiä edes rivi- ja sarakepisteiden välillä ei voi tarkkaan tulkita (approksimaatioita), projektiossa kaukana toisistaan olevat pisteet ovat kaukana toisistaan myös alkuperäisessä "pistepilvessä".

Vanha lista:

1. Ensimmäinen taulukko: profiilit, massat, keskiarvoprofiilit, khii2 - riippumattomuustesti ja etäisyysmitta
2. Hyvin tiivis esitys CA:n perusideasta, mutta ilman aivan simppeleitä kolmiulotteisia kuvia (niitä on jo)
3. Ensimmäinen symmetrinen kartta, perustulkinta (mitä kuvasta voidaan sanoa, mitä ei)
4. Lyhyt viittaus graafisen esityksen tulkintapulmiin, jotka eivät ole kovin pahoja. Niihin palataan kaksoiskuva-jaksossa.
5. Tulkinnan syventäminen - CA-käsitteiden tarkempi esittely

Haaste: käsitteet ja niiden suhteet ovat abstraktien matemaattisten rakenteiden tuloksia (barycentric, sentroidi), ja ne pitää jotenkin johdonmukaisesti pala kerrallaan tuoda esimerkkien kautta tekstiin. Teen käsiteluettelon, ja kirjoitan kaavat yms. toistaiseksi omaan dokumenttiin (LaTeX).

Keskeiset lähteet: MG:n “Correspondence analysis in practice”, “Biplots in practice”, HY:n 2017 kurssin materiaali ja laskuharjoitukset. Näissä kaikissa on käytetty samaa dataa esimerkeissä. Lisäksi perusasioiden esittelyssä MG & Blasius artikkelikokoelma (“vihreä kirja”), joissain kohdin Lerouxin ja Romanetin teos.

Ensimmäinen symmetrinen kartta

Tulkinnat ja yksinkertaisimmat perussäännöt. Dimensiot ja kuinka paljon alkuperäisen taulukon inertiaa saadaan esitettyä kartalla. Sitten asian ydin, akseleiden tulkinta (“mitä on oikealla ja vasemmalla”). Jos pisteet ovat alkuperäisessä “pilvessä” kaukana toisistaan, ne ovat sitä myös projektiossa. Kartta, mutta etäisyyksillä ei suoraa tulkintaa paitsi eteisyyksinllä origoon. Rivipisteiden suhteelliset etäisyydet, samoin sarakepisteidet.

Varoitus virhetulkinnasta: ryhmien tunnistaminen rivi, jopa rivi- ja sarakepisteistä koostuvien ryhmien. **zxy** Ja silti tavallaan voi. Sarake- ja rivipisteiden etäisyyksille ei ole suoraa tulkintaa, mutta on “vetovoima” (attraktio) ja “työntövoima” (repulsio). Jos profiilissa sarakemuuttujan osuus on suuri (siis suurempi kuin keskiarvopisteessä, suhteellinen ero), se “ajautuu” lähelle sarekepistettä. MG: “loose ends” - paperi, symmetrinen kuva eräs suurin sekaannuksen lähde. Tätä koitetaan selventää myös MG:n JASA-artikkelissa.

zxy . Tarina: valitaan edellisessä luvussa esitetyn pohjalta osa muuttujista, perustellaan miksi työmarkkia-asenteen ovat kiinnostavia, valitaan esimerkkianalyyseihin **yksi** muuttuja ja kuusi maata.

2.1 Äiti työssä

zxy Perustellaan aineiston valinnan vaiheet. Esimerkiksi otetaan yksi kysymys.

zxy Suhde data-lukuun, siellä pitäisi esitellä aineisto sisällöllisesti. Tässä vain valitan esimerkkiä varten yksi kysymys ja kuusi maata.

zxy Muuttujien nimeäminen vaikuttaa (a) muuttujien faktorointiin ja (b) kuviin ja taulukoihin.

Aineisto muuttujat V5-V9 ovat vastauksia (1-5 Likert, täysin samaa mieltä - täysin eri mieltä) seuraaviin kysymyksiin (suomenkielinen lomake, kysymys 23):

(a) Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä

(b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä

(c) Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö

(d) On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat ensisijaisesti kodin ja lapsia

(e) Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen

```
#vähän hankalaa jos Rmd-tiedoston 'scope' vaatii aina kaiken ajamisen joka tiedostossa!
incl_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU)
ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav", user_na = TRUE) # Alkuperäinen data
#
# lisäys 25.4.2018 user_na
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be # converted to NA"
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
# str(ISSP2012.data)
#61754 obs. of  420 variables ja 61754 obs. of  420 variables 25.4.18
#
# kuusi maata
ISSP2012esim1.dat <- filter(ISSP2012.data, V4 %in% incl_esim1)
#str(ISSP2012esim1.dat) #8557 obs. of  420 variables
#
```

```r
# mukaan muuttujat, V3 jos halutaan jakaa Saksa ja Belgia
# SEX 1=male, 2=female AGE haastateltava ikä haastatteluhetkellä
#
ISSP2012esim1.dat <- select(ISSP2012esim1.dat, C_ALPHAN, V3,V4, V6, SEX, AGE)

#str(ISSP2012esim1.dat) #8557 obs. of  6 variables
#
#poistetaan havainnot, joissa puuttuvia tietoja
ISSP2012esim1.dat <- filter(ISSP2012esim1.dat, (!is.na(V6) & !is.na(SEX) & !is.na(AGE)))
#str(ISSP2012esim1.dat) #8143 obs. of  6 variables
#ISSP2012esim1.dat %>% table1(C_ALPHAN, splitby = V6) table1 tuottaa siitejä outputeja esim. #LateX-for
```

Tehdään aineistoon muutama muutos (eli faktoreiksi, mutta ei järjestystä), jotta sen käsittely on helpompaa.

zxy Faktoreita ei ordinaaliasteikolla, vaikka ilman puuttuvia näin voisi tehdä. Miten ordinaaliasteikon faktori toimii, jos on puuttuvia mukana? Tämä ohitus/ratkaisu pitää perustella.

```r
# muutetaan muuttujia faktoreiksi
#
# Luokittelumuuttujien tasoille labelit
#
# sp (sukupuoli) m = 1, f = 2
sp_labels <- c("m","f")
# 1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri, 4 = eri mieltä, 5 = täysin eri miel
vastaus_labels <- c("ts","s","ese","e","te")

# Faktoreiksi
ISSP2012esim1.dat$maa <- factor(ISSP2012esim1.dat$C_ALPHAN)
ISSP2012esim1.dat$sp <- factor(ISSP2012esim1.dat$SEX, labels = sp_labels)
ISSP2012esim1.dat$V6 <- factor(ISSP2012esim1.dat$V6, labels = vastaus_labels)
#
# toinen maa-muuttuja, jossa Saksan ja Belgian jako
#  V3
#  5601     BE-FLA-Belgium/ Flanders
#  5602     BE-WAL-Belgium/ Wallonia
#  5603     BE-BRU-Belgium/ Brussels
# 27601     DE-W-Germany-West
# 27602     DE-E-Germany-East
#
#tsekkauksia
#ISSP2012esim1.dat %>% tableX(maa,V6,type = "count")
#summary(ISSP2012esim1.dat$sp)
#
#Apuvälineitä - lisätietoa muuttujista
# kun faktoroidaan V6, niin metadata katoaa?
#
# typeof(ISSP2012esim1.dat$V6) # what is it?
# class(ISSP2012esim1.dat$V6) # what is it? (sorry)
# storage.mode(ISSP2012esim1.dat$V6) # what is it? (very sorry)
# length(ISSP2012esim1.dat$V6) # how long is it? What about two dimensional objects?
# attributes(ISSP2012esim1.dat$V6) # does it have any metadata?
# str(ISSP2012esim1.dat) #8143 obs. of  8 variables
```

Taulukot ja kuvat omina koodilohkoina

Frekvenssitaulukko

```
taulu2 <- ISSP2012esim1.dat %>% tableX(maa, V6, type = "count")
knitr::kable(taulu2,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset maittain")
```

Taulukko 3: Kysymyksen V6 vastaukset maittain

	ts	s	ese	e	te	Total
BE	191	451	438	552	381	2013
BG	118	395	205	190	13	921
DE	165	375	198	538	438	1714
DK	70	238	152	232	696	1388
FI	47	188	149	423	303	1110
HU	219	288	225	190	75	997
Total	810	1935	1367	2125	1906	8143

Riviprosentit

```
taulu3 <- ISSP2012esim1.dat %>% tableX(maa,V6,type = "row_perc")

knitr::kable(taulu3,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset, riviprosentit")
```

Taulukko 4: Kysymyksen V6 vastaukset, riviprosentit

	ts	s	ese	e	te	Total
BE	9.49	22.40	21.76	27.42	18.93	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00
DE	9.63	21.88	11.55	31.39	25.55	100.00
DK	5.04	17.15	10.95	16.71	50.14	100.00
FI	4.23	16.94	13.42	38.11	27.30	100.00
HU	21.97	28.89	22.57	19.06	7.52	100.00
All	9.95	23.76	16.79	26.10	23.41	100.00

Sarakeprosentit

```
taulu4 <- ISSP2012esim1.dat %>% tableX(maa,V6,type = "col_perc")

knitr::kable(taulu4,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset, sarakeprosentit")
```

Taulukko 5: Kysymyksen V6 vastaukset, sarakeprosentit

	ts	s	ese	e	te	All
BE	23.58	23.31	32.04	25.98	19.99	24.72
BG	14.57	20.41	15.00	8.94	0.68	11.31
DE	20.37	19.38	14.48	25.32	22.98	21.05
DK	8.64	12.30	11.12	10.92	36.52	17.05
FI	5.80	9.72	10.90	19.91	15.90	13.63
HU	27.04	14.88	16.46	8.94	3.93	12.24
Total	100.00	100.00	100.00	100.00	100.00	100.00

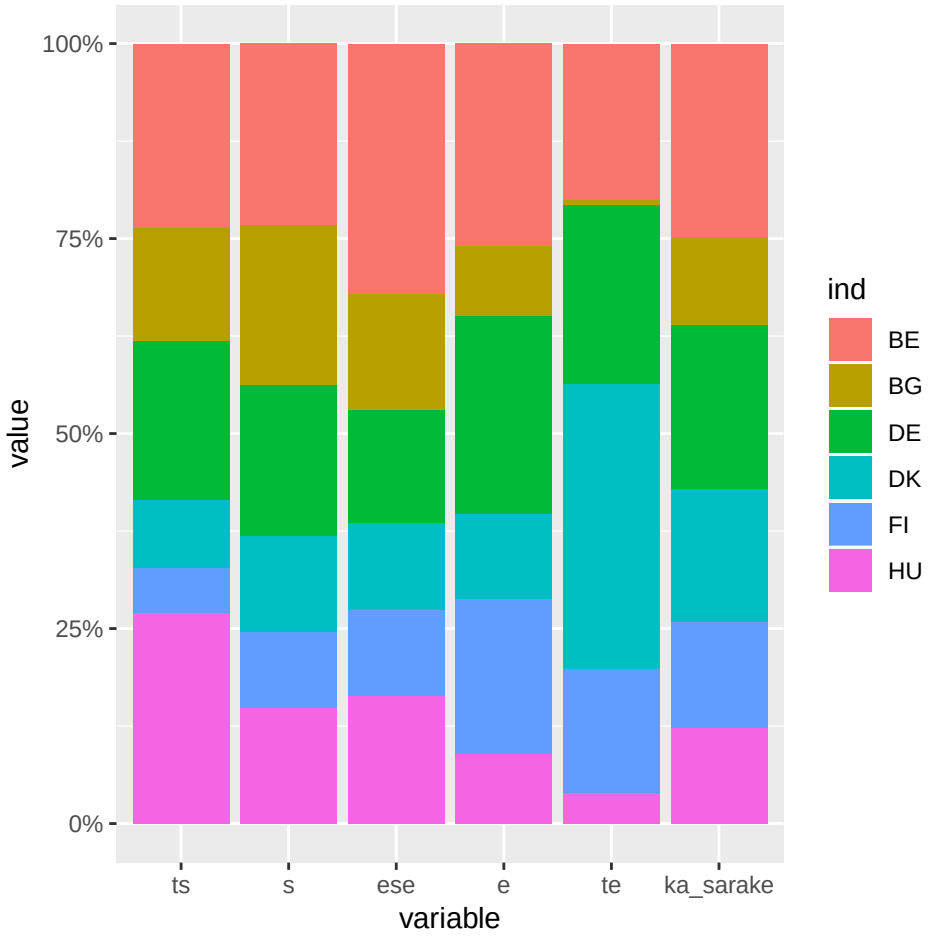
Taulukoissa on kuuden maan vastausten jakauma kysymykseen “Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä”. Taulukko on pieni, mutta havaintoja on melko paljon (N=8143). Alemman suhteellisten frekvenssien taulukon rivejä voi verrata toisiinsa ja alimpaan (“Total”“) keskimääräiseen riviin, sarakemuuttujien eli vastausvaihtoehtojen reunajakaumaan. Vastavasti sarakkeita voi verrata rivimuuttujien reunajakaumasarakkeeseen (”Total2). Eniten vastaajia on Belgiasta (25 %) ja Saksasta (21 %), vähiten Unkarista (12 %).

EDIT: Pienenkin taulukon pyörittely johdattelee hyvin, mihin korrespondenssianalyysiä tarvitaan. Näistähän riippuvuuden rakenteet näkee ilmankin, jos on tarpeeksi nokkela. Muiden pitää käyttää CA:ta.

```
simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat)
#tämä ajetaan jotta saadaan hieno kuva piirrettyä
```

edit: Riviprofiileista tarvitaan myös kuva, mutta hiotaan myöhemmin (13.5.2018)

```
#tauluG121 <- ISSP2012esim1.dat %>% tableX(maa, V6, type = "count")
#str(tauluG121)
#apu1 <- (tauluG121[-7, -6])
#str(apu1)
#apu1
#(rowSums(apu1))
#mutkikas kuvan piirto - sarakeprofiilit vertailussa
#ggplot vaatii df-rakenteen ja 'long data' - muotoon
##https://stackoverflow.com/questions/9563368/create-stacked-barplot-where-each-stack-is-scaled-to-sum-
#
# käytetään ca - tuloksia
apu1 <- (simpleCA1$N)
colnames(apu1) <- c("ts", "s", "ese", "e", "te")
rownames(apu1) <- c("BE", "BG", "DE", "DK", "FI", "HU")
apu1_df <- as.data.frame(apu1)
#lasketan rivien reunajakauma
apu1_df$ka_sarake <- rowSums(apu1_df)
#muokataan 'long data' - muotoon
apu1b_df <- melt(cbind(apu1_df, ind = rownames(apu1_df)), id.vars = c('ind'))

ggplot(apu1b_df, aes(x = variable, y = value, fill = ind)) +
         geom_bar(position = "fill", stat ="identity") +
         scale_y_continuous(labels = percent_format())
```



Riviprofiilien kuvat eivät onnistu ovat vielä tekemättä vailla valmiita (15.5.2018).

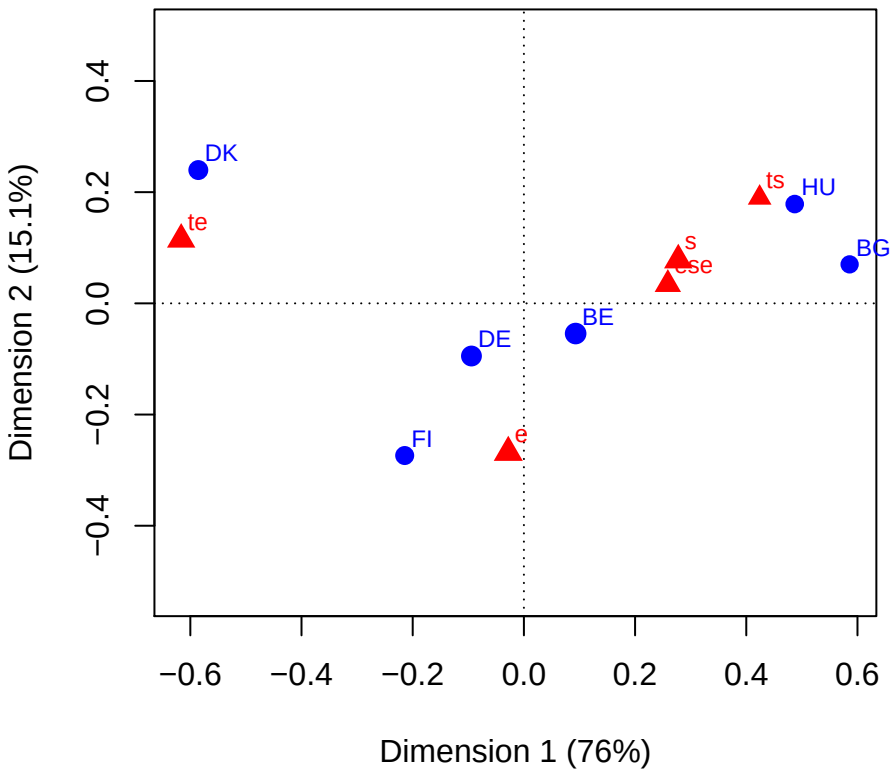
Ensimmäinen korrespondenssianalyysi - kokeiluja kuvasuhteen säätämiseksi output-dokumentissa. RStudiossa voi avata komentokehoitteessa grafiikka-ikkunan. Siitä käsin tallennettu pdf-kuva on ladattu alla Rmarkdownin omalla komennolla, kohdistus keskelle. Parhaiten näyttäisi toimivan knitrin funktio, mutta oletuskuvakoolla saa ca-kuvasta näköjään aika lähelle oikeanlaisen ilman mitään temppuja.

Lähtökohta: suhteelliset frekvenssit (korrespondenssimatriisi P)

```
taulu5 <-ISSP2012esim1.dat %>% tableX(maa,V6,type = "cell_perc")
knitr::kable(taulu5,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset maittain (%)")
```

Taulukko 6: Kysymyksen V6 vastaukset maittain (%)

	ts	s	ese	e	te	Total
BE	2.35	5.54	5.38	6.78	4.68	24.72
BG	1.45	4.85	2.52	2.33	0.16	11.31
DE	2.03	4.61	2.43	6.61	5.38	21.05
DK	0.86	2.92	1.87	2.85	8.55	17.05
FI	0.58	2.31	1.83	5.19	3.72	13.63
HU	2.69	3.54	2.76	2.33	0.92	12.24
Total	9.95	23.76	16.79	26.10	23.41	100.00



Kuva 1: V6: lapsi kärsii jos äiti on töissä

zxy Tätä ensimmäistä kuvaa on muistiinpanoissa kommentoitu (löytyy printattuna)

```
#simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat) suoritetaan ennen värikuvaa, tuloksia tarvitaan #siinä.
#symmetrinen kartta

plot(simpleCA1, map = "symmetric", mass = c(TRUE,TRUE))
```

```
#str(simpleCA1)
# 13.5.2018
# kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (X11()) ja sitten plot. Voi tallentaa pdf-muodossa
# grafiikkaikkunasta, ja ladata outputiin knitr-vaiheessa. Parempi tulostaa kuvatdsto pdf-ajurilla, jos
# näin tekemään.
# näitä kokeiln chunk-optioissa mutta ei toimineet (out.width = "6", out.hight = "6") (13.5.2018), vaan
# pandoc failed with error 43
#
```

Yritetään tuoda tähän pdf-muodossa kuvatiedosto, jossa kuvasuhde on oikea. Nämä toiminevat vain pdf-tulostuksessa.

Ja toinen tapa - kuvatiedoston lataaminen include_graphics - funktiolla. Ei esitetä tässä.

2.2 Korrespondenssianalyysin käsitteet

1. Profiilit
2. Massat
3. Profiilien etäisyydet (khii2)

zxy Ja tätä "triplettiä" täydentää neljä siitä johdettua käsitettä, en nyt (7.8.18) löydä MG-sitaattia.

---

3 Tulkinnan perusteita

Luvussa syvennetään esimerkin tulkinnan perusteita. Miksi symmetrinen kartta on yleensä paras vaihtoehto, siksi se oletusarvoisesti esitetäänkin. Milloin voi käyttää vaihtoehtoisia esitystapoja? **Ydinluku**.

Esimerkkiaineistossa tulee jo pohdittavaa, Guttman (arc, horseshoe) - efekti, ratkaisun dimensiot jne.

Asymmetrinen kartta, jossa riviprofiilit ovat pääkomponentti-koordinaateissa ja sarakeprofiilit standardikoordinaateissa.

(1) Sarakkeet ideaalipisteinä, edustavat kuvittellisia maita joissa kaikki ovat vastanneet vain yhdellä tavalla.

(2) Sarakepisteet kaukana origosta, koska skaalattu

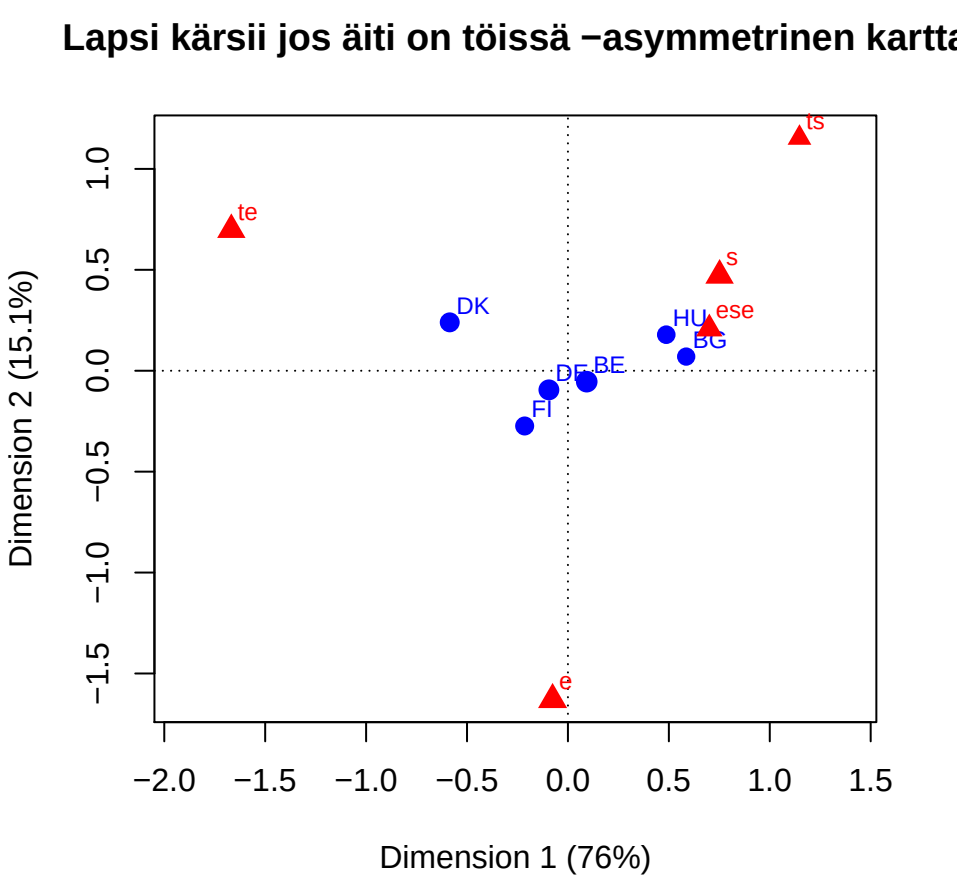
(3) Rivipisteet kasautuneet keskiarvopisteen ympärille

(4) Rivi-ja sarakepisteiden suhteelliset sijannit samat kuin symmetrisessä kuvassa

(5) Tässäkin kuvassa pisteen koko kuvaa sen massaa. Sarakkeista "täysin samaa mieltä" (ts) ja "ei samaa eikä eri mieltä" ovat massoiltaan pienimmät.

(6) Pisteiden koko kuvaa rivin tai sarakkeen massaa.

```
#asymmetrinen kartta - rivit pc ja sarakkeet sc
plot(simpleCA1, map = "rowprincipal",
     mass = c(TRUE,TRUE),
     main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta" )
```

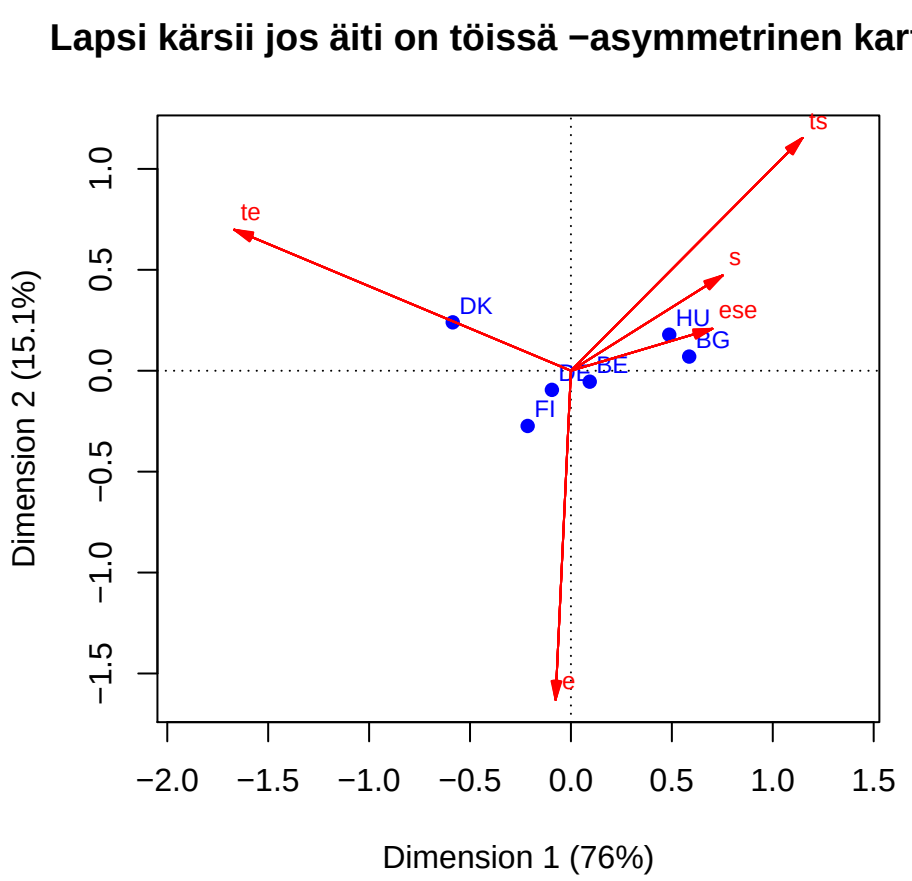


Tarinaa voi tarvittaessa jatkaa, tämä on CA:n hankalin asia. Kaksi koordinaatistoa, ja niiden yhteys.

(7) Asymmetrinen kuva ja akseleiden / dimensioiden tulkinta

Piirretään sama asymmetrinen kartta uudelleen, mutta yhdistetään sarakepisteet keskiarvopisteeseen (sentroidiin) suorilla. Mitä terävämpi on sarakesuoran (vektorin?) ja akselin kulma, sitä enemmän sarake määrittää tätä ulottuvuutta. Jos vektori on lähettä 45 asteen kulmaa, sarake määrittää yhtä paljon molempia ulottuvuuksia.

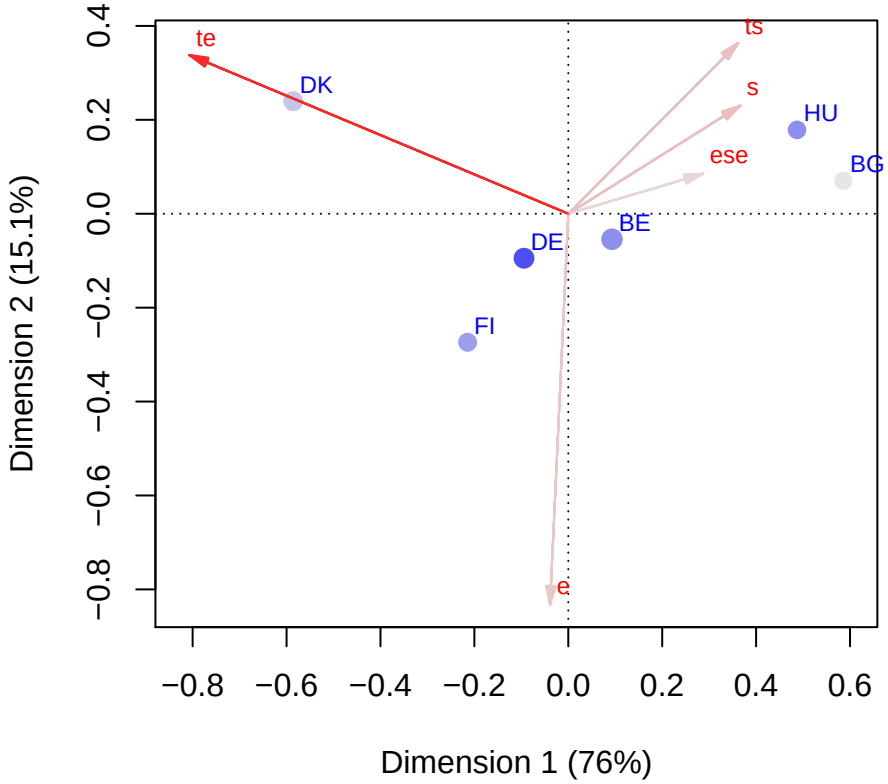
```
# asymmetrinen kartta - rivit pc ja sarakkeet sc
# sarakkeet vektorikuvina
plot(simpleCA1, map = "rowprincipal",
     arrows = c(FALSE,TRUE),
     main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta 1" )
```



Tärkein havainto on sarakkeen “Eri mieltä” (e) ja toisen ulottuvuuden yhteys. Myös sarake “täysin samaa mieltä” (ts) määrittää toista ulottuvuutta lähes yhtä paljon kuin ensimmäistä.

```
plot(simpleCA1, map = "rowgreen",
     contrib= c("absolute", "absolute"),
     mass = c(TRUE,TRUE),
     arrows = c(FALSE, TRUE),
     main = "Lapsi kärsii jos äiti on töissä - asymmetrinen kartta 2" )
```

Lapsi kärsii jos äiti on töissä − asymmetrinen kartta



Greenacre (2006, "loose ends -artikkeli") ehdotti asymmetrisessä kuvassa standardikoordinaattien skaalaamista niin, että ne kerrotaan massan neliöjuurella. Tämä skaalaus toimii hyvin pienen ja suuren inertian tapauksessa.

Asymmetrisessä kartassa 2 pisteiden koko on suhteessa niiden massaan, ja värisävy absoluuttiseen kontribuutioon (voi olla myös suhteellinen kontribuutio).

```
# CA:n numeeriset tulokset
summary(simpleCA1)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.136619  76.0  76.0  ********************
##  2      0.027089  15.1  91.1  ****
##  3      0.010054   5.6  96.7  *
##  4      0.005988   3.3 100.0  *
##         -------- -----
##  Total: 0.179751 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   BE |  247  465   34 |   93 347  16 |  -54 118  27 |
## 2 |   BG |  113  874  251 |  586 862 284 |   70  12  21 |
```

```
## 3 |   DE |  210  584   36 |  -94 291  14 |  -95 293  70 |
## 4 |   DK |  170  996  381 | -586 853 428 |  240 143 362 |
## 5 |   FI |  136 1000   92 | -214 380  46 | -274 620 377 |
## 6 |   HU |  122  889  206 |  487 783 213 |  179 105 144 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   ts |   99  784  152 |  424 653 131 |  190 131 132 |
## 2 |    s |  238  788  140 |  278 731 134 |   78  57  53 |
## 3 |  ese |  168  720   88 |  259 707  82 |   34  12   7 |
## 4 |    e |  261  982  108 |  -28  11   2 | -268 971 693 |
## 5 |   te |  234 1000  512 | -616 966 651 |  115  34 114 |
```

---

4 Yksinkertaisen korrespondenssianalyysin laajennuksia

Korrespondenssianalyysi sallii rivien tai sarakkeiden yhdistelyn tai "jakamisen". Tämä onnistuu esimerkkiaineistossa lisäämällä rivejä eli jakamalla eri maiden vastausksia useampaan ryhmään.

Sen avulla voi myös tarkastella ja vertailla erilaisia ryhmien välisiä tai ryhmien sisäisiä (within groups - between groups) eroja hieman. Teknisesti yksinkertaista korrespondenssianalyysiä sovelletaan muokattuun matriisiin. Datamatriisi rakennetaan useammasta alimatriisista, joko "pinoamalla" osamatriiseja (stacked matrices) tai muodostamalla symmetrinen lohkomatriisi (ABBA).

Lisätään esimerkkidataan uusia muuttujia, vastaajan luokitelut ikä ja sukupuoli.

** EDIT: ** Lisätäänkö muuttujat tässä, vai Data-luvussa? Lisätään aluksi tässä. Koitetaan aina pitää alkuperäinen data mahdollisimman "lähellä", luodaan siis kaikki uudestaan. Tarketeena .data jos koko aineisto ja .dat jos rajattu.

Toinen pulma: milloin laajennetaan dataa useampaan maahan?

```
#valittavien maiden kolminumeroinen ISO 3166 - koodi vektoriin - TÄSSÄ KAIKKI MAAT
#incl_countries <- c(36, 40, 56,100, 124, 191, 203, 208, 246, 250, 276, 348, 352, 372, 428, 440,
#                     528, 578, 616, 620, 643, 703, 705, 724, 752, 756, 826, 840)
#ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav", user_na = TRUE)
#
# lisäys 25.4.2018 user_na
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be # converted to NA"
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
#
#
#str(ISSP2012.data) #61754 obs. of  420 variables
#ISSP2012jh1.data <- filter(ISSP2012.data, V4 %in% incl_countries)
#
# Luodaan samanniminen data kuin edellisissä esimerkeissä, lisätään siihen uudet muuttujat.
incl_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU)
ISSP2012esim1.data <- read_spss("data/ZA5900_v4-0-0.sav", user_na = TRUE) # Alkuperäinen data
#
# lisäys 25.4.2018 user_na
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be # converted to NA"
```

```
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
# str(ISSP2012.data)
#61754 obs. of  420 variables ja 61754 obs. of  420 variables 25.4.18
#
# kuusi maata
ISSP2012esim1.dat <- filter(ISSP2012esim1.data, V4 %in% incl_esim1)
#str(ISSP2012esim1.dat) #8557 obs. of  420 variables
#
# mukaan muuttujat, V3 jos halutaan jakaa Saksa ja Belgia
# SEX 1=male, 2=female AGE haastateltava ikä haastatteluhetkellä
#
ISSP2012esim1.dat <- select(ISSP2012esim1.dat, C_ALPHAN, V3,V4, V6, SEX, AGE)

#str(ISSP2012esim1.dat) #8557 obs. of  6 variables
#
#poistetaan havainnot, joissa puuttuvia tietoja
ISSP2012esim1.dat <- filter(ISSP2012esim1.dat, (!is.na(V6) & !is.na(SEX) & !is.na(AGE)))
#str(ISSP2012esim1.dat) 8143 havaintoa, 6 muuttujaa
# sp (sukupuoli) m = 1, f = 2
sp_labels <- c("m","f")
# 1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri, 4 = eri mieltä, 5 = täysin eri miel
vastaus_labels <- c("ts","s","ese","e","te")

# Faktoreiksi
ISSP2012esim1.dat$maa <- factor(ISSP2012esim1.dat$C_ALPHAN)
ISSP2012esim1.dat$sp <- factor(ISSP2012esim1.dat$SEX, labels = sp_labels)
ISSP2012esim1.dat$V6 <- factor(ISSP2012esim1.dat$V6, labels = vastaus_labels)
#Tähän loppuu datan luonti
```

EDIT: Uudet muuttujat omana pätkänä

```
# 23.5.2018 maa2 - muuttuja
#
#  5601     BE-FLA-Belgium/ Flanders
#  5602     BE-WAL-Belgium/ Wallonia
#  5603     BE-BRU-Belgium/ Brussels
# 27601     DE-W-Germany-West
# 27602     DE-E-Germany-East
ISSP2012esim1.dat$maa2 <- factor(ISSP2012esim1.dat$V3,
                  levels = c("100","208","246","348","5601","5602","5603","27601","27602"),
                  labels = c("BG","DK","FI","HU","bF","bW","bB","dW","dE"))

#head(test6)
#str(test6$maa2)
taulu41 <- ISSP2012esim1.dat %>% tableX(maa,maa2,type = "count")
kable(taulu41,digits = 2, caption = "Uusi muuttuja maa2: Belgian ja Saksan ositus")
```

Taulukko 7: Uusi muuttuja maa2: Belgian ja Saksan ositus

	BG	DK	FI	HU	bF	bW	bB	dW	dE	Total
BE	0	0	0	0	1012	490	511	0	0	2013
BG	921	0	0	0	0	0	0	0	0	921
DE	0	0	0	0	0	0	0	1167	547	1714

	BG	DK	FI	HU	bF	bW	bB	dW	dE	Total
DK	0	1388	0	0	0	0	0	0	0	1388
FI	0	0	1110	0	0	0	0	0	0	1110
HU	0	0	0	997	0	0	0	0	0	997
Total	921	1388	1110	997	1012	490	511	1167	547	8143

4.1 Täydentävät muuttujat (supplementary points)

zxy Piste sinne piirretään, mutta muuttujassa on se tieto. "Täydentävät piste"" kuulostaa huonolta. Lisämuuttujat, havainnot?

Ref:CAip ss 89, HY2017_MCA

Aineistossa on havaintoja (rivejä) tai muuttujia (sarakkeita), joista voi olla hyötyä tulosten tulkinnassa. Nämä lisäpisteet voidaan sijoittaa kartalle, jos niitä voidaan jotenkin järkevästi vertailla kartan luomisessa käytettyihin profiileihin (riveihin ja sarakkeisiin).

EDIT Lisätään Belgian ja Saksan aluejako täydentäviksi riveiksi. Sopii tarinaan, dimensioiden tulkinta ei ollut esimerkissä kovin kirkas. Viite CAip:n lukuun, jossa vain todetaan että maita ei ole järkevää painottaa (massa) otoskoolla, vaan vakioidaan (jotenkin) sama (suhteellinen) massa kaikille. Samalla oikaistaan myös naisten yliedustus aineistossa.

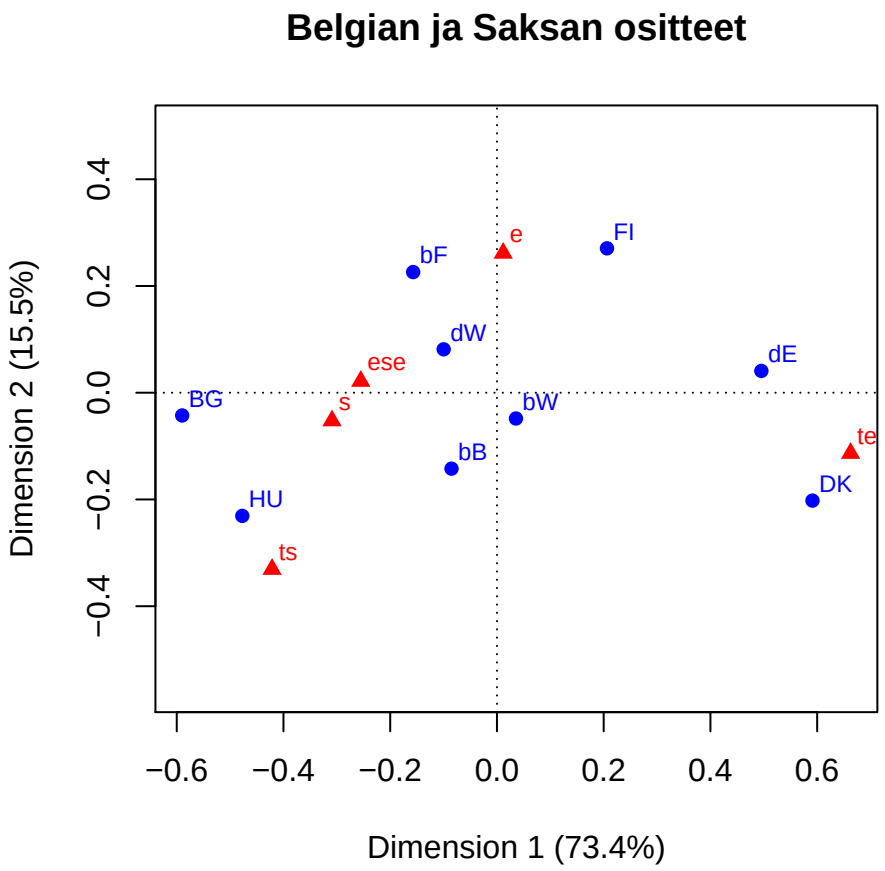
Active point, aktiivinen piste (aktiivinen havainto tai muuttuja).

Täydentävä piste (täydentävä havainto).

Täydentävien muuttujien kolme käyttötapaa:

- sisällöllisesti tutkimusongelman kannalta poikkeava tai erilainen rivi tai sarake
- outlayerit, poikkeava havainto jolla pieni massa (esimerkissä uusi sarakemuuttuja, jossa kovin vähän havaintoja)
- osaryhmät

```
#kömpelöä koodia, harjoitellaan taulukoiden yhdistelyä (CAtest1.Rmd)
# Belgian ja Sakasan jako lisäpisteinä 24.5.2018
#head(ISSP2012esim1.dat)
suppointCA1 <- ca(~maa2 + V6,ISSP2012esim1.dat)
plot(suppointCA1, main = "Belgian ja Saksan ositteet")
```



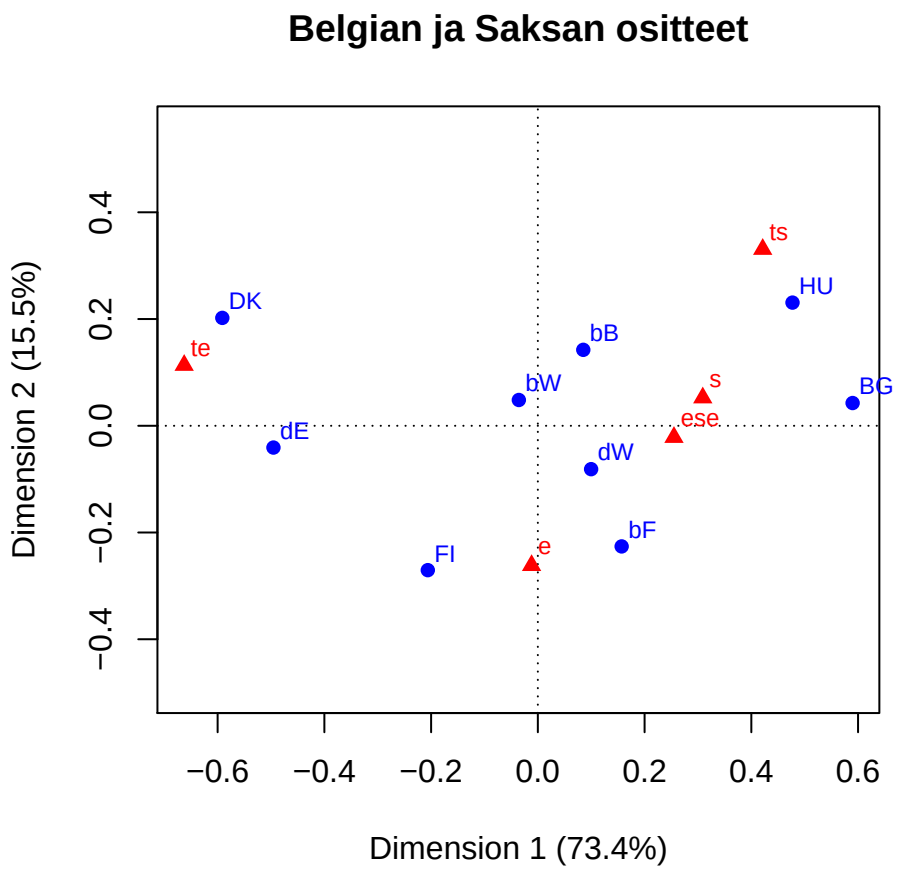

```
#kuva kääntyy ympäri, kerrotaan koordinaattivektorit luvulla -1
#summary(suppointCA1)
#print(suppointCA1)
#str(suppointCA1)
#
#Käännetään kuva

suppointCA1b <- suppointCA1
suppointCA1b$rowcoord <- suppointCA1b$rowcoord[,] * (-1)
suppointCA1b$colcoord <- suppointCA1b$colcoord[,] * (-1)
suppointCA1b$rowcoord
```

	Dim1	Dim2	Dim3	Dim4
BG	1.5024575	0.2364976	-1.5646535	1.2274009
DK	-1.5060223	1.1214678	-0.8891868	0.1996764
FI	-0.5252216	-1.5009862	0.5841156	0.1935193
HU	1.2154623	1.2803425	0.9947716	-0.9386679
bF	0.4000647	-1.2540425	-1.1182121	-1.6025782
bW	-0.0906315	0.2679979	0.0761877	-0.7901000
bB	0.2169124	0.7893585	1.3697862	-0.5617393
dW	0.2543232	-0.4511235	0.8757353	1.5124903
dE	-1.2620072	-0.2265947	0.7448562	-0.2844804

```
suppointCA1b$colcoord
```

	Dim1	Dim2	Dim3	Dim4
ts	1.0733103	1.8351327	2.1160478	-0.2360525
s	0.7872571	0.2909285	-0.9861563	1.2374779
ese	0.6497888	-0.1199336	-0.9123790	-1.9203632
e	-0.0298593	-1.4515479	0.8247769	0.2094281
te	-1.6881081	0.6291103	-0.1632819	-0.0121801

```
plot(suppointCA1b, main = "Belgian ja Saksan ositteet")
```




```
# Miten lisärivit? (24.5.2018)
# Luetaan data tauluksi - ei toimi, char-table
# yritetään uudestaan table-funktiolla
# data maa2-muuttujalla
suppoint1_df1 <- select(ISSP2012esim1.dat, maa2,V6)
str(suppoint1_df1)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    8143 obs. of  2 variables:
##  $ maa2: Factor w/ 9 levels "BG","DK","FI",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ V6  : Factor w/ 5 levels "ts","s","ese",..: 3 2 3 4 3 3 4 3 2 3 ...
##  - attr(*, "notes")= chr  "document Plan File: /Users/marcic/Desktop/old/GPS2011 sampling/ISSP2013.sa
```

```r
head(suppoint1_df1)
```

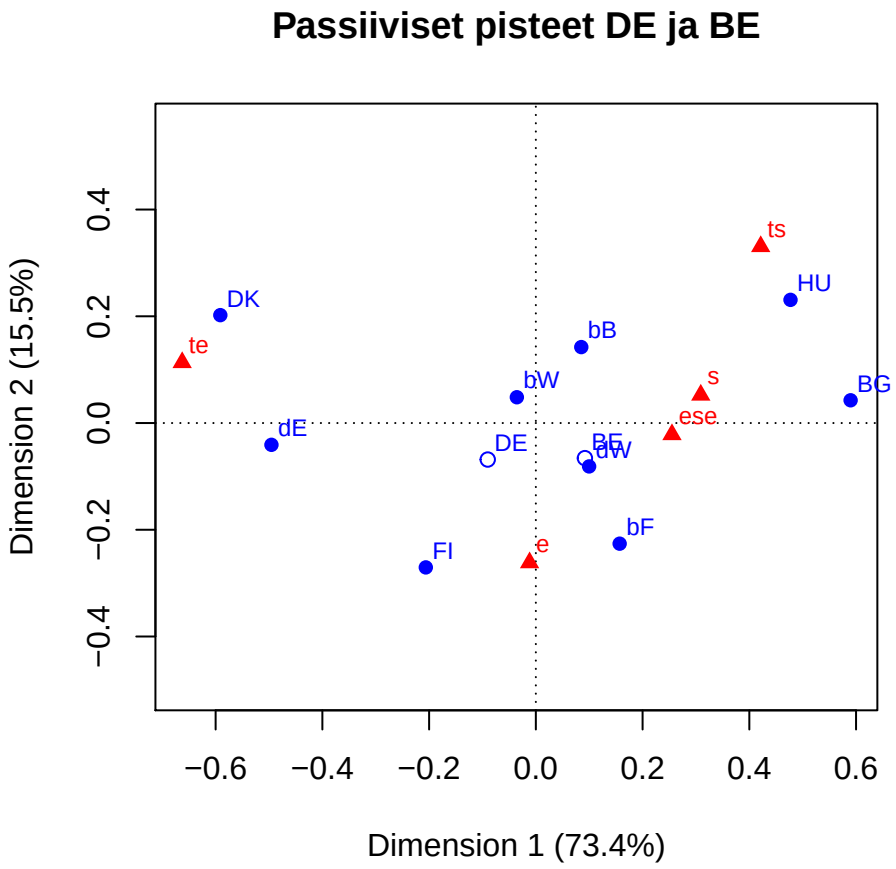
maa2	V6
BG	ese
BG	s
BG	ese
BG	e
BG	ese
BG	ese

```r
suppoint1_tab1 <- table(suppoint1_df1$maa2, suppoint1_df1$V6)
suppoint1_tab1
```

/	ts	s	ese	e	te
BG	118	395	205	190	13
DK	70	238	152	232	696
FI	47	188	149	423	303
HU	219	288	225	190	75
bF	51	241	262	312	146
bW	53	103	91	118	125
bB	87	107	85	122	110
dW	133	313	138	375	208
dE	32	62	60	163	230

```r
#plot(ca(~maa2 + V6, suppoint1_df1)) #toimii
#
# Saksan ja Belgian summarivit
#
suppoint2_df <- filter(ISSP2012esim1.dat, (maa == "BE" | maa == "DE"))
suppoint2_df <- select(suppoint2_df, maa, V6)
#head(suppoint2_df)
#tail(suppoint2_df)
#str(suppoint2_df)
#suppoint2_df
suppoint2_tab1 <- table(suppoint2_df$maa, suppoint2_df$V6)
#suppoint2_tab1
suppoint2_tab1 <- suppoint2_tab1[-2,]
# kömpelösti kolme kertaa
suppoint2_tab1 <- suppoint2_tab1[-3,]
suppoint2_tab1 <- suppoint2_tab1[-3,]
suppoint2_tab1 <- suppoint2_tab1[-3,]
#suppoint2_tab1

#lisätään rivit maa2-muuttujan taulukkoon

suppoint1_tab1 <- rbind(suppoint1_tab1, suppoint2_tab1)
#suppoint1_tab1
suppointCA2 <- ca(suppoint1_tab1[,1:5], suprow = 10:11)
#käännetään kuva
suppointCA2b <- suppointCA2
```

```
suppointCA2b$rowcoord <- suppointCA2b$rowcoord[,] * (-1)
suppointCA2b$colcoord <- suppointCA2b$colcoord[,] * (-1)

plot(suppointCA2b, main = "Passiiviset pisteet DE ja BE" )
```



Saksan ja Belgian summarivit ovat ositteiden painotettuja keskiarvoja (sentroideja), läntisen ja itäisen Saksan rivipisteiden välisellä janalla on koko maan summapiste DE.

4.2 Lisämuuttujat: ikäluokka ja sukupuoli

zxy Otsikkoa pitää harkita, CAip - kirjassa tämä on ensimmäinen esimerkki yksinkertaisen CA:n laajennuksesta. Otsikkona on “multiway tables”, ja tästä yhteisvaikutusmuuttujan (interactive coding) luominen on ensimmäinen esimerkki. Menetelmää taivutetaan sen jälkeen moneen suuntaan.

zxy “Taulokon käsite” vaatii oman kappaleen. Yhteys havaintojen/yksilöiden pilveen ja muuttujien pilveen.

Luodaan luokiteltu ikämuuttua age_cat, ja sen avulla iän ja sukupuolen interaktiomuuttuja ga. Maiden välillä on hieman eroja siinä, kuinka nuoria vastaajia on otettu tutkimuksen kohteeksi. Suomessa alaikäraja on 15 vuotta, monessa maassa se on hieman korkeampi. Ikäluokat ovat (1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 tai vanhempi). Vuorovaikutusmuuttuja ga koodataan f1,. . . , f6 ja m1,. . . ,m6. Muuttujien nimet kannattaa pitää mahdollisimman lyhyinä.

```
#age_cat
#AGE 1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 and older
#summary(ISSP2012esim1.dat$AGE)
#hist(ISSP2012esim1.dat$AGE)
ISSP2012esim1.dat <- mutate(ISSP2012esim1.dat, age_cat = ifelse(AGE %in% 15:25, "1",
                        ifelse(AGE %in% 26:35, "2",
                        ifelse(AGE %in% 36:45, "3",
                        ifelse(AGE %in% 46:55, "4",
                        ifelse(AGE %in% 56:65, "5", "6"))))))
ISSP2012esim1.dat$age_cat <- factor(ISSP2012esim1.dat$age_cat)

#test6 %>% tableX(AGE, age_cat, type = "count") aika iso taulukko, voi tarkistaa että muunnos ok.
taulu42 <- ISSP2012esim1.dat %>% tableX(maa,age_cat,type = "count")
kable(taulu42,digits = 2, caption = "Ikäluokka age_cat")
```

Taulukko 12: Ikäluokka age_cat

	1	2	3	4	5	6	Total
BE	208	333	336	375	368	393	2013
BG	77	115	159	148	198	224	921
DE	205	223	274	358	288	366	1714
DK	207	213	245	271	234	218	1388
FI	152	166	165	223	238	166	1110
HU	103	161	198	171	196	168	997
Total	952	1211	1377	1546	1522	1535	8143

```
taulu43 <- ISSP2012esim1.dat %>% tableX(maa,age_cat,type = "cell_perc")
kable(taulu43,digits = 2, caption = "age_cat: suhteelliset frekvenssit")
```

Taulukko 13: age_cat: suhteelliset frekvenssit

	1	2	3	4	5	6	Total
BE	2.55	4.09	4.13	4.61	4.52	4.83	24.72
BG	0.95	1.41	1.95	1.82	2.43	2.75	11.31
DE	2.52	2.74	3.36	4.40	3.54	4.49	21.05
DK	2.54	2.62	3.01	3.33	2.87	2.68	17.05
FI	1.87	2.04	2.03	2.74	2.92	2.04	13.63
HU	1.26	1.98	2.43	2.10	2.41	2.06	12.24
Total	11.69	14.87	16.91	18.99	18.69	18.85	100.00

Ikäjäkauma painottuu kaikissa maissa jonkinverran vanhempiin ikäluokkiin. Nuorempien ikäluokkien osuus on (alle 26-vuotiaan ja alle 26-35 - vuotiaat) varsinkin Bulgariassa (BG) ja Unkarissa (HU) pieni.

```
# ei kovin siisti

ISSP2012esim1.dat <- mutate(ISSP2012esim1.dat, ga = ifelse(((age_cat == "1")&(sp == "m")), "m1",
                        ifelse(((age_cat == "2")&(sp == "m")), "m2",
                        ifelse(((age_cat == "3")&(sp == "m")), "m3",
                        ifelse(((age_cat == "4")&(sp == "m")), "m4",
                        ifelse(((age_cat == "5")&(sp == "m")), "m5",
                        ifelse(((age_cat == "6")&(sp == "m")), "m6",
```

```
                            ifelse(((age_cat == "1")&(sp == "f")), "f1",
                            ifelse(((age_cat == "2")&(sp == "f")), "f2",
                            ifelse(((age_cat == "3")&(sp == "f")), "f3",
                            ifelse(((age_cat == "4")&(sp == "f")), "f4",
                            ifelse(((age_cat == "5")&(sp == "f")), "f5", "f6"))))))))))))
taulu44 <- ISSP2012esim1.dat %>% tableX(maa,ga,type = "count")
kable(taulu44,digits = 2, caption = "Ikäluokka ja sukupuoli ga")
```

Taulukko 14: Ikäluokka ja sukupuoli ga

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	116	198	174	199	186	185	92	135	162	176	182	208	2013
BG	40	64	94	85	114	149	37	51	65	63	84	75	921
DE	102	120	152	186	135	185	103	103	122	172	153	181	1714
DK	83	110	136	146	128	99	124	103	109	125	106	119	1388
FI	94	95	94	118	142	91	58	71	71	105	96	75	1110
HU	54	86	95	91	94	104	49	75	103	80	102	64	997
Total	489	673	745	825	799	813	463	538	632	721	723	722	8143

```
taulu45 <- ISSP2012esim1.dat %>% tableX(maa,ga,type = "cell_perc")
kable(taulu45,digits = 2, caption = "ga: suhteelliset frekvenssit")
```

Taulukko 15: ga: suhteelliset frekvenssit

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	1.42	2.43	2.14	2.44	2.28	2.27	1.13	1.66	1.99	2.16	2.24	2.55	24.72
BG	0.49	0.79	1.15	1.04	1.40	1.83	0.45	0.63	0.80	0.77	1.03	0.92	11.31
DE	1.25	1.47	1.87	2.28	1.66	2.27	1.26	1.26	1.50	2.11	1.88	2.22	21.05
DK	1.02	1.35	1.67	1.79	1.57	1.22	1.52	1.26	1.34	1.54	1.30	1.46	17.05
FI	1.15	1.17	1.15	1.45	1.74	1.12	0.71	0.87	0.87	1.29	1.18	0.92	13.63
HU	0.66	1.06	1.17	1.12	1.15	1.28	0.60	0.92	1.26	0.98	1.25	0.79	12.24
Total	6.01	8.26	9.15	10.13	9.81	9.98	5.69	6.61	7.76	8.85	8.88	8.87	100.00

zxy (7.8.2018) **DatanRukkailua2.R** parempi versio näistä muunnoksista.

4.3 Päällekkäiset matriisit (stacked matices)

Ref:CAip, CA_Week2.pdf (kalvot MCA-kurssilta 2017)

Concatenated tables (yhdistetyt taulut tai matriisit): (a) kaksi luokittelumuuttujaa (b) useita muuttujia stacked ("pinotaan").

MCA 2017 laskareissa ja kalvoissa esitetään, miten nämä saadaan kätevästi CA-paketin MJCA-funktion BURT-optiolla.

4.4 Matched matrices

Ref:CAip ss. 177, HY2017_MCA, Greenacre JAS 2013 (sovellus ISSP 1989, 4 kysymystä 'pitäisikö äidin olla kotona', 8 maata), tässä artikkelissa "SVD-based methods", joista yksi CA (muut biplots, PCA, compositional data/log ratios).

Edellisen menetelmän variantti, jossa ryhmien väliset ja sisäiset erot saadaan esiin. Inertian jakaminen. Samanlaisten rivien ja sarakkeiden kaksi samankokoista taulua, esimerkiksi sukupuolivaikutusten arviointi. Alkuperäinen taulukko jaetaan kahdeksi tauluksi sukupuolen mukaan. Matriisien yhdistäminen (concatenation) riveittäin tai sarakkeittain ei näytä optimaalisesti mm - matriisien eroja.

Ryhmien välisen ja ryhmien sisäinen inertian erottaminen, **ABBA** on yksi ratkaisu (ABBA matrix, teknisesti block circulanMat matrix).

Luokittelu voi olla myös kahden indikaattorimuuttujan avulla jako neljään taulukkoon (esim. miehet vs. naiset länsieuroopassa verratuna samaan asetelmaan itä-Euroopassa). Samaa ideaa laajennetaan.

Esimerkkinä “Attitudes to women working in 2012”.