

# Korrespondenssianalyysi - graafinen ja geometrinen data-analyysin menetelmä

Jussi Hirvonen

Versio 0.8, tulostettu 2020-11-22



# Sisällys



# Alkutoimia

```
# testaukseen 7.11.2020 virheilmoituksia varten arvo TRUE
options(tinytex.verbose = TRUE)
```

Tämä luku poistetaan kun tutkielma on valmis

Raportti yhtenä html-tiedostona ([https://hirjus.github.io/capaper/JH\\_capaper.html](https://hirjus.github.io/capaper/JH_capaper.html)), ja kaikki toimii. Tässä liitteet, lähdeviitteet ja koodilistaus.

**PDF-tulostus oikuttelee ja kaatuu, mutta pdf-syntyy.** MikTeX vaihdettu TinyTeX-engineen ja pdflatex -> xelatex (15.11.20). PDF-tulostus kaatuu **luultavasti** koodilistaukseen. Pandoc ei osaa rivittää koodia oikein tms. Koodilistauksen poistaminen ei auta, syystä tai toisesta.

**Suunnitelma: PDF html-raportista -> pdfExchange-ohjelmalla sivunumerot ja kansilehdet**

**22.11.2020 luonnos kansilehdeksi, sisällysluetteloksi ja tiivistelmän aihio**

(<https://hirjus.github.io/capaper/GraduKansilehti1.pdf>)      (<https://hirjus.github.io/capaper/wordTOC.pdf>)  
(<https://hirjus.github.io/capaper/GtiivisTesti1.pdf>)

RefWorksistä tuotu jhca2020.bib-tiedosto tarkistettu ja korjailtu virheet. Ei enään ladata uutta (21.11.2020).

Data-analyysi ja koodi omana projektina: (<https://hirjus.github.io/Galku>). Tarkempia tulostuksia, muunnosten tarkistuksia ja enemmän kuvia, myös koodi näkyy.

Gitbook-tulosteessa ei saa koodia “piilotettua”, asetus “code\_folding: hide” vaatii teeman (theme). `_output.yml` - tiedostoon lisätty `html_book` - formaatti, siinä voi tarvittaessa käyttää piilotusta.

Versiointi: 0.0n aloittelua, 0.n jäsentely koko paperille, 1.n.n valmiimpaa tekstiä.



# Luku 1

## Johdanto

**edit** Kirjoitetaan disposition pohjalta, keräillään kaikki yleiset ca-luonnehdinnat yhteen paikkaan eli johdantoon. Kirjoitetaan viimeiseksi, samoin yhteenvetoluku. Mahdollisimman lyhyt.

### johdannon ydinsisältö

1. CA:n periaatteet voi esitellä yksinkertaisen kahden luokittelumuuttujan taulukon analyysin avulla. Kun taulukko on pieni, on helppo vertailla CA:n karttoja dataan. Yksinkertainen kahden luokittelumuuttujan korrespondenssianalyysi antaa graafisen analyysin "...perussäännöt tulkinnalle. Kaikki muut korrespondenssianalyysin muodot ovat saman algoritmin soveltamista toisen tyyppiisiin datamatriiseihin, ja tulkintaa sovelletaan vastaavasti".((?)) s. 437). Tämä Greenacren ja Hastien artikkelissaan esittämä periaate helpottaa huomattavasti juuri graafisen data-analyysin eli korrespondenssianalyysin karttojen tulkinnan perussääntöjen oivaltamista. Juuri kartat, samassa kuvassa esitettävät havainnot ja muuttujat ovat CA:n tärkein menetelmä, mutta ei aivan helppo.

2. CA on kuitenkin vahvimmillaan isojen ja monimutkaisten aineistojen analyssissä. Siksi ns. monimuuttuja-korrespondenssianalyysin lyhyessä esittelyssä aineistoa laajennetaan, sillä "on erittäin vaikeaa osoittaa kotiakvaarioissa, että verkko on tehokas"((?)), s.15).

**loppuosa johdannosta vielä vanhoja tekstipätkiä (16.11.2020)**

### 1.1 Tutkielman tavoite

**k** Tässä kerrotaan, miksi tämä työ on kirjoitettu. Esitellään menetelmä käyttämällä oikeaa dataa. Täsmällisempi esitys sirotellaan esimerkkiaineiston analyysin tulosten esittelyn lomaan. Pitäisikö tässä tuoda esille ns. "ranskalaisen koulukunnan" matemaattisen perusteiden korostus, ja data-analyysin filosofia? Ehkä ei, koska sen pohdinta ei ole pääasia. Se tietysti mainitaan, ja asiaa pohditaan.

**ks** Esitellään korrespondenssianalyysin käsitteet ja graafisen analyysin periaatteet.

**k** -mitä ca on? - dimensioiden vähentäminen ja visualisointi

- mihin dataan se soveltuu: kahden luokittemuuttujan taulukon lukumäärädata (count data) tai suhdeasteikon muuttujia samassa mittayksikössä (esim. euroissa).
- määrittele graafinen, deskriptiivinen, eksploratiivinen data-analyysi
- yksinkertainen ca, useamman muuttujan ca

**zxy** Miksi eksploratiivinen (määrittele!) ja deskriptiivinen (määrittele!) menetelmä on esitettävä "in vivo", toiminnassa? Oppikirjoissa (viitteitä) erityisesti MG on havainnolistanut CA:n matemaattista ja geometristä taustaa synteettisillä aineistoilla. Turha kopioida tähän. Menetelmän ydin on yksinkertaisen graafisen esityksen –

kartan – avulla tulkita monimutkaisen empirisen aineiston muuttujien riippuvuuksia. Yhteyksiä ei tiivistetä todennäköisyyspäätelyn kriteereillä tilastolliseen malliin, vaan deskriptiivisen analyysin hengessä esitellään koko aineisto. Mallin sijaan vähennetään ulottuvuuksia, ja siinä menetetään informaatiota. Tavoitteena on säilyttää yleensä kaksilotteisessa kuvassa mahdollisimman suuri osa alkuperäisen datan vaihtelusta. Eksploratiivinen data-analyysi on vuoropuhelua aineiston kanssa. Analyysiä tarkennetaan, rajataan ja muokataan, kun aineisto paljastaa jotain kiinnostavaa tai yllättävää. Tästä saa jonkinlaisen aasinsillan matriisiyhtälöiden puolustukseksi. Saksan ja Belgian datan jakaminen on hyvä esimerkki, on “osattava tarttua” menetelmän tulosmatriiseihin.

**k** esitystavan perustelu

- kenelle kirjoitettu? Menetelmästä kiinnostuneelle tilastotieteen ja data-analyysin perusteet tuntevalle. R-ohjelmisto ei ole rajoitus, SPSS ja SAS sopivat (SPSS - MG:llä kriittinen huomio “loose ends - paperissa” tai CAip-teorialiitteessä).

## 1.2 Tärkeimmät lähteet ja ohjelmistot

Michael Greenacre luennoi lyhyen kurssin korrespondenssianalyysistä Helsingin yliopistossa keväällä 2017(?). Luennot ja laskuharjoitukset perehdyttivät minut ensimmäistä kertaa tähän menetelmään, ja kurssin materiaaleihin olen usein palannut. Michael Greenacren kärsivällisesti kirjoitettu “Correspondence Analysis in Practice” (jatkossa “CAIP”) (?) ja sen aikasemmat versiot ovat tehneet menetelmää laajasti tunnetuksi.“Biplots in Practice” (jatkossa “Biplots”) (?) esittää menetelman osana yleisempää kaksoiskuvien ideaa.

Ranskalaisten lähestymistä perusoppikirja(?) (GDA-kirja?) esittelee menetelmän matemaattiset perusteet. Lyhyt historiallinen katsaus ja menetelmä soveltamisen perusajatuksen esittely valaisevat ranskaa taitamattomalle data-analyysin koulukunnan ideoita. Kirjoittajat esittlevät perusteellisesti joitain empirisiä tutkimuksia, ja lyhyt mutta naseva matriisilaskennan kritiikki on hyvä panna merkille.

**edit** Hyvin lyhyesti, lause tai pari. On oma liite tekneisestä ympäristöstä.

**zxy** R, ca-paketti. löytyy myös muita paketteja. Rmarkdown(?), ja bookdown ((?) ja toinen viite (?)). Mikäs tuo jälkimmäinen on? PDF-lähdeluettelossa ei ole url-osoitteita.

**k** Helposti toistettavan tutkimukset periaatteet

1. Datastan perusmuunnokset ja muuttujatyypit tehdään kun data luetaan R-ohjelmistoon.
2. Koodi selkeää ja dokumentoitua. Tärkeä lähde (?)
3. R, LaTeX, pandoc - versiot dokumentoidaan

Tarkemmin liittessä.

## 1.3 Korrespondenssianalyysin historiaa

**k1** Tiivis esitys lähteineen. Historian voi aloittaa jo pari vuosikymmentä vallineesta tilanteesta. CA on yksi deskriptiivinen (ei-tn-teoriaan perustuvaa päätelyä) menetelmä muiden joukossa, eristyneisyys murtui hitaasti 80-luvun aikana.

**k2** Historialla on vain historiallista merkitystä. Kiinnostava juttu, mutta aika laaja ja lavea.

**k3** Peruskäsite monessa lähteessä (vihreä kirja, GDA-kirja jne.): synty ja kukoistus Ranskassa, loistava eristys (splendid isolation), pikkut hiljaa hyväksyntä.

Syiksi esitetään kaksoismuuria: abstrakti matemaattinen (“bourbakilainen”) perusta ja esitystapa ja kieli.

**k4** Mitä historiasta on hyvä tietää. 1. Matemaattinen perusta on “tosi”, mutta onko menetelmän soveltaminen riippuvainen siitä? Ei ole ollut.

2. Ristiriita data-analyyttisen/kuvalevan jne. lähestymistavan ja tilastollisen mallintamisen välillä - on läsnä edelleen mutta turha korostaa. Myös tilastollisen mallintamisen ja päätelyn sisällä on kiistoja, erilaisia näkemyksiä ja kuiluja.
3. "Esoteerinen tieteenfilosofia"? Kiinnostava aihe, ehkä. Murtag-sitaatti.



## Luku 2

# Data

Käytän tutkielmassa International Social Survey - projektin (ISSP) vuoden 2012 kyselytutkimusta "Perhe , työ ja sukupuolirootit"(International Social Survey Programme: Family and Changing Gender Roles IV). Tutkielman tärkeimmässä lähteissä (esim. CAIP, Biplots) tutkimuksen aikaisemmat versiot ovat esimerkkiaineistona, ~~samoin~~ Greenacren luennoilla 2017.



Länsi-Saksan ja USA:n tutkimuslaitosten yhteistyö vakiintui ISSP-organisaatioksi 1984 (<http://www.issp.org>). Vuonna 2015 neljän perustajajäsenen joukko oli kasvanut 49 maahan. Vertailevan tutkimuksen aineistoja on kerätty monista teemoista, perhearvoista ja naisten työmarkkina-asemasta neljä kertaa (1988, 1994, 2002, 2012). USA:n edustajana mukana ollut Tom W. Smith näkee aineistojen arvon juuri kansainvälisessä vertailevassa tutkimuksessa. ~~Järjestön julkaisuluettelossa~~ oli 2012 yli 5200, ja viime vuosina yli 400 vuodessa lisää. (?)



Saksalainen GESIS-tutkimuslaitos ylläpitää data-arkistoa, josta ~~suurin osa datasta ja dokumentaatiosta~~ on vapaasti saatavissa (<https://www.gesis.org/en/issp/home>). Suomessa tutkimuksen data ja dokumentaatio löytyvät [Tampereen yliopiston Aila-tietoarkistosta] ([https://services.fsd.uta.fi/catalogue/FSD2820?tab=summary&study\\_language=fi](https://services.fsd.uta.fi/catalogue/FSD2820?tab=summary&study_language=fi)).



GESIS-instituutin "datakatalogista" löytää kätevästi kaiken dokumentaation, mutta edes saksalaiset eivät voi estää www-sivustojen innokaita uudistajia. Kaikki linkit lähdeluettelossa vievät vain GESIS-arkistosivulle, josta löytää pitkän listan pdf-dokumentteja. Taulukossa ~~@Ref(tab:ISSPdocsTable)~~ on neljän tärkeimän dokumentin tiedostonimi.



Tutkimuksen viitetieto, linkki vie dokumenttiluetteloon koska hanke on päättynyt. Kaksi versiota, käytetään ensimmäistä (?) .

Data ja dokumentaatio (ml. käyttöehdot) löytyvät GESIS-instituutin datakatalogista (<https://zacad.gesis.org/>)? .

Tätä kirjoittaessa (10.11.2020) tutkimuksen aineisto löytyy osoitteesta [<https://zacad.gesis.org/webview/index.jsp?object=http://zacad.gesis.org/obj/fStudy/ZA5900>]. Datakatalogista löytyää tällä hetkellä (10.11.2020) dokumentit ja datan helpoiten.

Taulukko 2.1: ISSP 2012: tärkeimmät dokumentit

dokumentti	sisältö	tiedosto
Variable Report	Perusdokumentti, muuttujien kuvaukset ja taulukot	ZA5900_cdb.pdf
Study Monitoring Report	tiedokeruun toteutus eri maissa	ZA5900_mr.pdf
Basic Questionnaire	Maittain sovellettava kyselylomake	ZA5900_bq.pdf
Contents of ISSP 2012 module	substanssikysymykset taulukkona	ZA5900_overview.pdf
Questionnaire Development	kyselylomakkeen laatiminen	ssoar-2014-scholz_et_al-ISSP_2012

Taulukko 2.2: ISSP2012: Työelämä ja perhearvot - kysymykset

muuttuja	kysymyksen tunnus, lyhennetty kysymys
V5	Q1a Working mother can have warm relation with child
V6	Q1b Pre-school child suffers through working mother
V7	Q1c Family life suffers through working mother
V8	Q1d Women's preference: home and children
V9	Q1e Being housewife is satisfying
V10	Q2a Both should contribute to household income
V11	Q2b Men's job is earn money, women's job household
V12	Q3a Should women work: Child under school age
V13	Q3b Should women work: Youngest kid at school
SEX	Respondents age
AGE	Respondents gender
DEGREE	Highest completed degree of education: Categories for international comparison
MAINSTAT	Main status: work, unemployed, in education...
TOPBOT	Top-Bottom self-placement (10 pt scale)
HHCHILDR	How many children in household: children between [school age] and 17 years of age
MARITAL	Legal partnership status: married, civil partnership...
URBRURAL	Place of living: urban - rural

Koodikirjan (“Variable report”) (?) selostaa tarkasti tietosisällön. Tutkimuksen seurantaraportti (“Study Monitoring Report”) (?) kertoo miten tutkimus käytännössä toteutettiin. Kyselylomake (?) ja suomenkielinen versio (?) ja myös kaikki muut kielversiot voivat olla hyödyllisiä. Tiedonkeruun tarkoitus ja kyselyn suunnitelun ideat kerrotaan omassa raportissa (?).

## 2.1 Aineiston rajaaminen maat ja muuttujat

Olen valinnut laajasta aineistosta 25 maata ja joukon muuttujia. Maat on valittu niin, että ne ovat suhteellisen samankaltaisia ja valitut muuttujat ovat niissä samanlaisia. Kysymyksissä on jonkin verran pieniä eroja, mutta joissain tapauksissa ero on merkittävä. Esimerkiksi Espanja on jostain syystä jättänyt tässä käytetyistä muuttujista ns. neutraalin (“en samaa enkä eri mieltä”) vastausvaihtoehdon pois, joten Espanja jää pois.

Substanssimuuttujat ovat yksi ”kysymyspatteri”, jolla luodataan asenteita naisten roolista työmarkkinoilla. Aiheen pysyvää ajankohtaisuutta kuvaaa hyvin The Economist - lehden artikkeli Saksojen yhdistymisen 30-vuotispäivänä (3.10.2020, “A report...reveals the interplay between policy and attitudes that influences the decision to work.”). Artikkeli on maksumuurin takana mutta tutkimus on vapaasti luettavissa (DIW Weekly Report 38 / 2020, S. 403-410)



Taulukon 2.2 kysymysten lyhyet versiot ovat datassa mukana. Sarakkeessa ”muuttuja” on alkuperäisen aineiston muuttujanimi, kysymyksen tunnus on valittuun dataan luotu muuttujanimi. Vertailu muihin samalla aineistoilla tehtyihin tutkimuksiin on helpompaa.

☒ Kyselylomakkeilla kysymykset olivat hieman pidempiä, kuvassa 2.1 osa suomenkielistä lomaketta.

Valituista taustamuuttujista monet on kerätty haastattelulla. Tiedonkeruu, otantamenetelmät ja yksikkövaltauskadon huomioiminen on tehty joka maassa omalla tavallaan. Aineistoissa on mukana painot joilla tulokset voidaan korottaa perusjoukon tasolle, mutta kansainvälisiä vertailupainoja ei syystä tai toisesta ole. Taustamuuttujat kuten koulutustaso on harmonisoitu vertailukelpoisiksi.

Tutkimuksen kohdeperusjoukko on 18-vuotiaat tai sitä vanhemmat, poikkeuksina Suomi (15 - 74 vuotiaat), Islanti, Japani, Etelä-Afrikka ja Venezuela.

Seuraavaksi perheeseen, työhön ja kotitöihin liittyviä kysymyksiä.						
23. Mitä miettä olet seuraavista väittämistä? Rengasta jokaiselle... riviltä vain yksi vaihtoehto						
	Täysin samaa mieltä	Samaa mieltä	En samaa enkä eri mieltä	Eri mieltä	Täysin eri mieltä	En osaa sanoa
a) Työssäkävää äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä .....	1	2	3	4	5	8
b) Alle koululäisen lapsi todennäköisesti karsi, jos hänen äitiinsä käy työssä .....	1	2	3	4	5	8
c) Kaiken kaikian perhe-elämä karsi, kun naisella on kokopäivätöö .....	1	2	3	4	5	8
d) On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat ensisijaisesti kodin ja lapsia .....	1	2	3	4	5	8
e) Kotirouvana oleminen on aivan yhtä antoisa kuin ansioityön tekeminen .....	1	2	3	4	5	8

24. Mitä miettä olet seuraavista väittämistä? Rengasta kummallakin riviltä vain yksi vaihtoehto.						
	Täysin samaa mieltä	Samaa mieltä	En samaa enkä eri mieltä	Eri mieltä	Täysin eri mieltä	En osaa sanoa
a) Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen .....	1	2	3	4	5	8
b) Miehen tehtävä on ansaita raha; naisen tehtävä on huolehtia kodista ja perheestä .....	1	2	3	4	5	8

25. Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa? Rengasta kummallakin riviltä vain yksi vaihtoehto.						
Naisen tulisi...	käydä kokopäivätössä	käydä osa-alikäytössä	pysyä kotona	En osaa sanoa		
a) Kun perheessä on alle koululäinen lapsi .....	1	2	3	8		
b) Kun nuori lapsi on aloittanut koulunkäynnin .....	1	2	3	8		

Kuva 2.1: Suomenkielinen lomake

Jos ohitetaan pienet erot kysymyksissä ja vastausvaihtoehdissa jäljelle jää erävastauskato, kyselytutkimusten ominaisuus. Jostain syystä joihinkin kysymyksiin ei vastata. Esimerkiksi Ranskassa yli 20 prosenttia kieltyyti vastaanmasta lasten (HHCHILDR) lukumäärää kysyttäessä, ja aika moni myös muissa perherakenteeseen liittyvissä kysymyksissä. Tässä tutkielman monimuuttujakorrespondenssianalyysi käytetään tämän ongelman tai datan ominaisuuden analyysiin.

Poistin aineistosta havainnot, joissa tieto iästä tai sukupuolesta puuttuu ( $32969-32823 = 146$  havaintoa).

Aineiston luokittelu- ja järjestysasteikon muuttujat muunnetaan R-ohjelmiston factor-tietotyypiksi. Tein muunnokset useammassa vaiheessa heti kun data on luettu SPSS-tiedosta. Käsittelyssä koitan noudattaa helposti toisettavan tutkimuksen periaatteita (McNamara ja Horton (?)), koodi ei saisi olla kovin virhealtista ("haurasta") ja tarkistuksia tehdään paljon. Data-analysin ja ehkä erityisesti korrespondenssianalyysin idea on kuitenkin operoida matriiseilla, lisätä ja poistaa rivejä ja sarakkeita ja rakennella mutkikkaimpia matriiseja yksinkertaisemmissa. Analyysivaiheessa koodi muuttuu hauraammaksi.

### Aineisto ja korrespondenssianalyysi

Michael Greenacre on käyttänyt aineistoa eri vuosilta luentomateriaaleissa kuten Helsingissä 2017(?) ja ainakin kahdessa oppikirjassaan((?), (?)). ISSP - aineisto vuodelta 1989 on käytetty myös neljän "singuariarvohajoitelman perustuvan menetelmän" vertailuun(?). Blasius ja Thiessen ((?)) arvioivat aineiston laatua ja ja maiden vertailtavuutta vuoden 1994 aineistolla.

Substanssitutkimusta en käsitlele, näiden esimerkkien lisäksi ISSP:n ja GESIS-instituutin www-palveluista löytyy paljon muitakin.

Sukupuoliroolien (gender roles) ja niihin liittyvien asenteiden vertailevaa kansainvälistä (cross-cultural) tutkimusta on tehty paljon. Tutkimusongelman sisällöllisten ja teoreettisen kysymysten nykytilaa kuvaavat tuore artikkeli (?).

Toisessa esimerkissä (?) tutkitaan ensin 18 OECD-maan perhepolitiikan muutoksia kolmen viime vuosikymmenen ajalta. Näkökulma on työllisyyspolitiikka ja menetelmänä monimuuttuja-korrespondenssianalyysi (MCA). Havaittuille kehityssuunnille etsitään toisessa vaiheessa selityksiä. Aineistona on viisi kansainvälineen vertailuun soveltuvala aineistoa, yhtenä niistä ISSP:n data kolmelta kierrokselta (1988,1994,2002).



## Luku 3

# Yksinkertainen korrespondenssianalyysi

Korrespondenssianalyysin peruskäsitteet ja muuttujien yhteyden graafisen analyysin periaatteet voi esittää kahden luokitelumuuttujan ristiintaulukoinnin eli kontigenssitaulun analyysin avulla. Kyse ei ole pelkästään helposta esimerkistä, vaan peruskäsitteet ja geometrisiin perusteisiin nojaava graafinen analyysi ovat oleellisilta osin samat myös monimutkaisemmissa menetelmän sovelluksissa. MG&Hastie korostivat tätä, ja MG:n oppikirjat ovat hyvä esimerkki perusteellisesta yksinkertaisen taulukon analyysin esitystavasta. GDA-kirja (viite) korostaa ranskalaisten perinteiden mukaisesti matemaattista teoriaperustaa, mutta myös siinä menetelmä peruskäsitteet ja tulkinnat esitellään yksinkertaisella esimerkillä (Fisherin Cairness-data, Mustonen (viite) käyttää samaa dataa).

Esitän tässä jaksossa korrespondenssianalyysin peruskäsitteet intuitiivisesti, matemaattiset yksityiskohdat löytyvät liitteestä 1. Esitystavan etu on taulukon pieni koko, johtopäätökset voi helposti tarkastaa datasta. Nämä päästävät nopeasti pääasiaan, graafiseen analyysiin.

Tämä ei ole ainoa mahdollinen näkökulma, korrespondenssianalyysillä on useita vaihtoehtoisia tulkintoja. Se voidaan ymmärtää myös moniulotteisen varianssianalyysin kaltaiseksi hajonnan dekomponoinnin menetelmäksi. GDA-kirjassa numeeriset tulokset ovat tulkinnan lähtökohtana ja vasta sitten katsotaan graafista esitystä. CAIP (viite) pitää numeerisia tuloksia tärkeinä, niiden avulla varmistetaan kuvaan tulkinnan pätevyys.

### 3.1 Äiti töissä - kärsiikö lapsi?

Aineisto on kuuden maan vastaukset kysymykseen Q1b: "Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä". Kysymys on voimakkaasti muotoiltu (~~edit~~MG on jossain jutussa havainnut, että kyselyn kahdessa kysymyksessä on sana "äiti" ja niiden vastausten jakaumat poikkeavat. Jos löydän lähteensä lisäänsen tähän). Kysymykset on suunniteltu kokonaisuudeksi, ja niitä analysoidaan yhdessä luvussa 7, yhden taulukon analyysi esittelee menetelmän. (alaviite:Tämän havaitsi eräs lastensuojelun ammattilainen: kysymys on irrallisenä monimerkityksellinen (vähintään ns. "double - barreled"), eikä siihen voi oikein järjevästi vastata. Pitäisi tietää missä lapsi on, mitä hän tekee). Havainnot joissa tieto vastauksesta puuttuu on poistettu aineistosta. Taustamuuttuja ovat vastaajan sukupuoli ja ikä.

Frekvenssitalukossa 3.1 on esitetty vastausten suhteellinen jakauma, lukumäärät on jaettu havaintojen lukumäärällä (8143). Korrespondenssianalyysissä kaikki on suhteellista, ja analyysi perustuu tähän taulukkoon. Taulukon reunajakaumat kertovat jokaisen maan ja jokaisen vastausvaihtoehdon suhteellisen osuuden. Näitä suhteellisia osuuksia kutsutaan korrespondenssianalyysissä *rivi- ja sarakemassoiksi*.

Muuttujien luonne on usein erilainen. Tähän aineistoon sopii riviprosenttientaulukko, vertaillaan vastausten jakaumia maiden välillä. Taulukon sarakkeet ovat muuttuja ja rivit havaintoja. Rivit on saatu summaamalla (aggregoimalla) vastaukset maittain. (viite: MG kutsuu näitä rivejä termillä samples, osajoukot).

Sarakeprosentit antavat toisen näkökulmaan samaan dataan.

### Taulukko 3.1: Kysymyksen Q1b vastaukset maittain, suhteelliset frekvenssit

	S	s	?	e	E	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

#### Taulukko 3.2: Kysymyksen Q1b vastaukset, riviprosentit

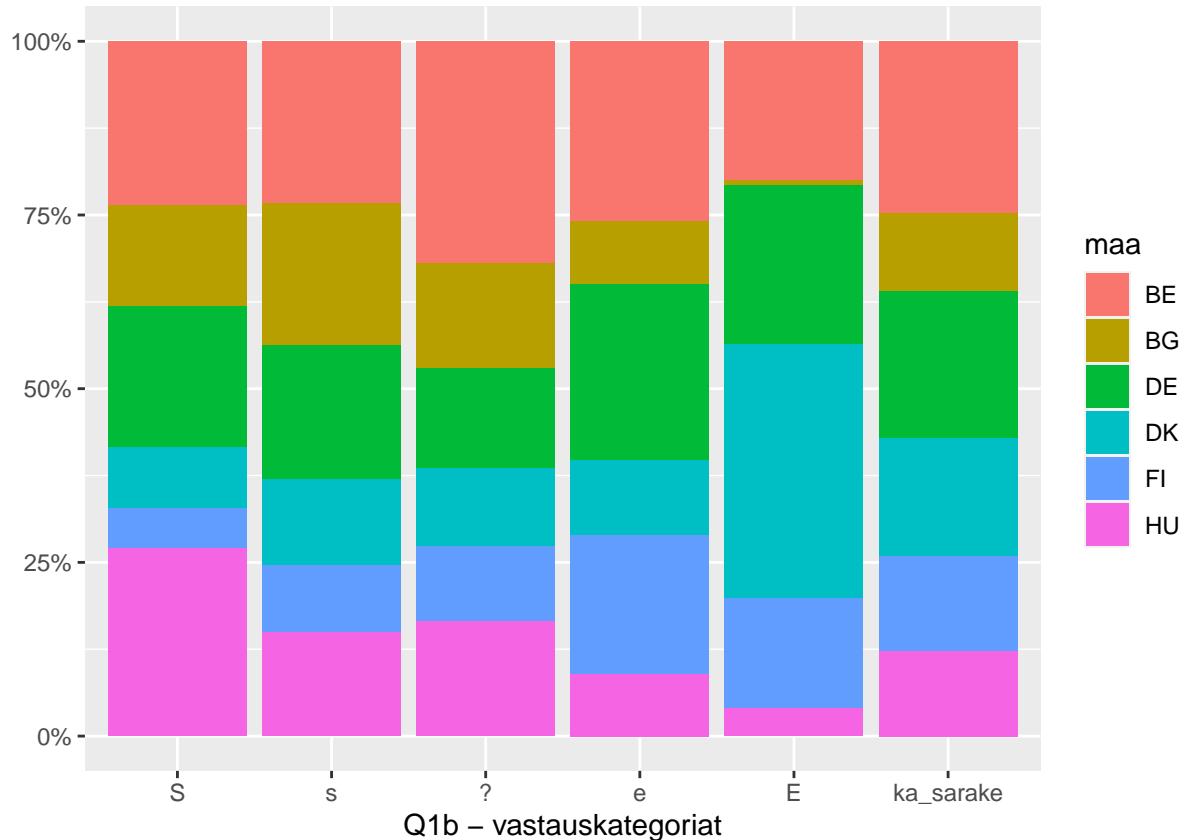
	S	s	?	e	E	Total
BE	9.49	22.40	21.76	27.42	18.93	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00
DE	9.63	21.88	11.55	31.39	25.55	100.00
DK	5.04	17.15	10.95	16.71	50.14	100.00
FI	4.23	16.94	13.42	38.11	27.30	100.00
HU	21.97	28.89	22.57	19.06	7.52	100.00
All	9.95	23.76	16.79	26.10	23.41	100.00

### Taulukko 3.3: Kysymyksen Q1b vastaukset, sarakeprosentit

Tavoitteena on korrespondenssianalyysin kartta, jossa rivi- ja sarakepisteet on esitetty samassa kuvassa. Sarakeprosenttien taulukossa on esitetty sarakkeiden suhteelliset jakaumat. Näitä suhteellisia rivejä ja sarakkeita kutsutaan korrespondenssianalyysissä *rivi- ja sarakeprofileiksi*.

**k** Rivist on saatu alkuperäisestä aineistosta osajoukkojen summina. MG:n terminologialla “samples”.

Korrespondenssianalyysin perusidea on analysoida rivien ja sarakkeiden yhteyttä (korrespondenssia) rivi- tai sarakeprofilien hajonnan avulla. Hajontaa mitataan poikkeamilla kesiarvorivistä tai sarakteesta, ja massat otetaan huomioon, kun hajonnat lasketaan yhteen.



Kuva 3.1: Q1b:Sarakeprofilit ja kesiarvoprofiili

Kuvasta 3.2 3.2 esimerkiksi näkee, että Tanska (DK) näyttäisi poikkeava kesiarvorivistä paljon, samoin Bulgaria. Bulgarian massa on kuitenkin aineiston pienin (11,31 %), Tanskan taas kohtalainen (17 %). Sarakeprofiluvassa 3.1 täysin eri mieltä - vastaus (E) on selvästi erilainen ja sen massa on suuri (23%). Kaikki luvut ovat suhteellisia, havaintojen lukumäärä ei vaikuta tulkintaan periaatteessa mitenkään.

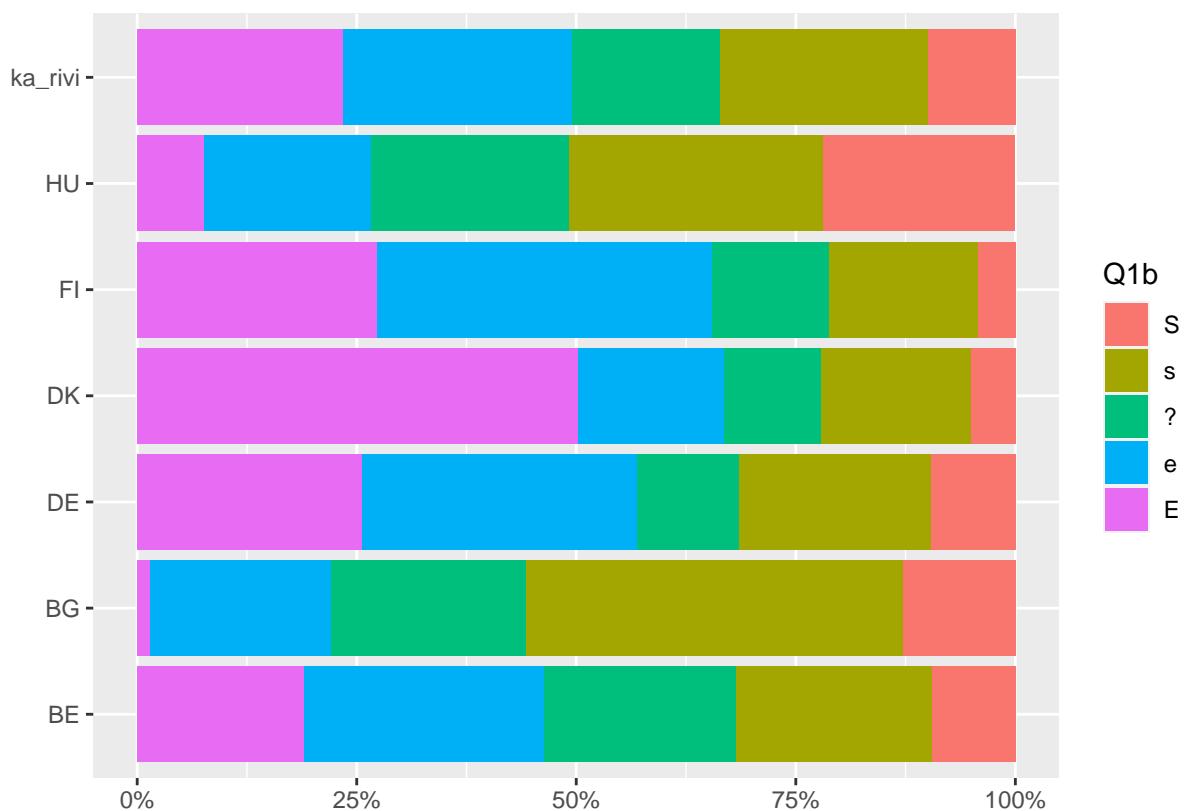
Mikä on rivien (havaintojen) ja sarakkeiden (muuttujien) yhteys?

Kahden luokittelumuuttujan riippuvuutta voidaan testata  $\chi^2$  - testillä. Riippumattomuushypoteesin mukainen odotettu solufrekvenssi on taulukon 3.1 reunajakaumien alkioiden tulo.

Testisuure saadaan laskemalla yhden jokaisen solun havaittujen ja odotetettujen frekvenssien erotukset muodossa

$$\chi^2 = \frac{(havaittu - odotettu)^2}{odotettu} \quad (3.1)$$

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



Kuva 3.2: Q1b: riviprofililit ja keskiarvorivi

$$\text{rivisumma} \times \frac{(havaittu riviprofiili - odotettu riviprofiili)^2}{odotettu riviprofiili} \quad (3.2)$$

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

$$\frac{\chi^2}{n} = \phi^2 \quad (3.3)$$

Jakajassa ei ole vapausastekorjausta ( $n-1$ ), korrespondenssianalyysi on deskriptiivistä data-analyysiä.

Tunnusluku  $\phi^2$  on korrespondenssianalyysisä kokonaisinertia (total inertia). Se kuvaa, kuinka paljon varianssia taulukossa on ja on riippumaton havaintojen lukumäärästä. Tilastotieteessä tunnusluvulla on useita vaihtoehtisia nimiä (esim. mean square contingency coefficient), ja sen neljöjuurta kutsutaan  $\phi$  - kertoimeksi.

Korrespondenssianalyysin algoritmit operoivat suhteellisten frekvenssien taulukkolla. Kaavojen (3.1) ja (3.2) yhteyden pitäisi olla selkeä.

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

$$\sqrt{(p_{11} - p_{31})^2 + (p_{12} - p_{32})^2 + (p_{13} - p_{33})^2 + (p_{14} - p_{34})^2 + (p_{15} - p_{35})^2} \quad (3.4)$$

Rivien  $\chi^2$ -etäisyys on painotettueuklidinen etäisyys, jossa painoina ovat riviprofilin odotetut arvot. Ne ovat riippumattomuushypoteesin mukaisesti riviprofilien keskiarvoprofilin vastaavat alkioit  $r_i$ .

$$\sqrt{\frac{(p_{11} - p_{31})^2}{r_1} + \dots + \frac{(p_{15} - p_{35})^2}{r_5}} \quad (3.5)$$

Inertia voidaa esittää rivien ja keskiarvorivin (sentroidin)  $\chi^2$ -etäisyysien neliöiden painotettuna summana, jossa painoina ovat rivien massat  $m_i$  ja summa lasketaan yli rivien  $i$ .

$$\phi^2 = \sum_i (\text{massa } m_i) \times (\text{profiili } i \chi^2 - \text{etaisyys sentroidista})^2 \quad (3.6)$$

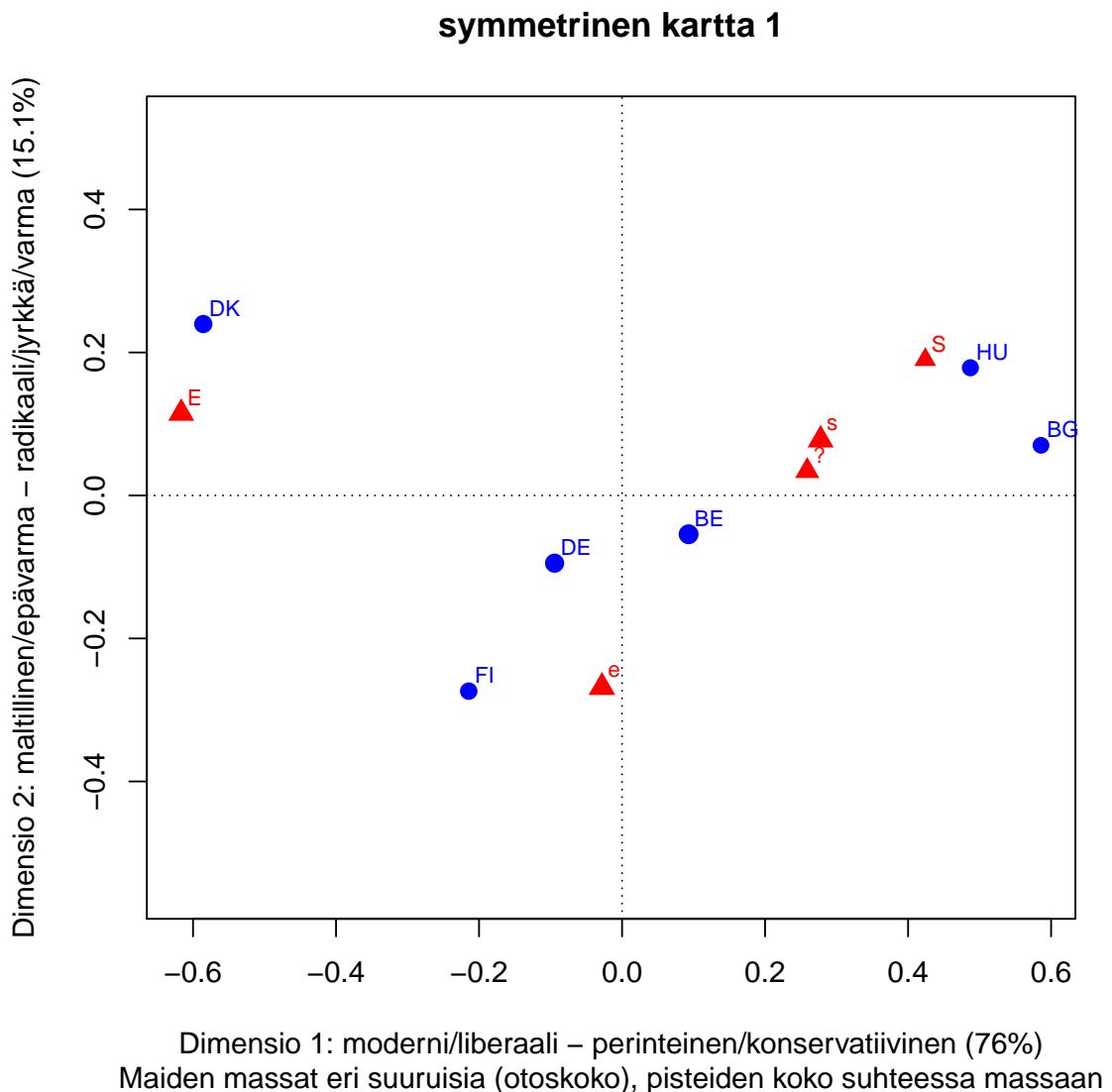
**edit** Tässä esitystavassa viite on CAiP, teorialitteessä tarkemmin. Tarkoitus on esittää yksinkertaisesti taulukon datan analyysin käsittein ja CA:n peruskäsittein profili, massa ja  $\chi^2$ -etäisyys

Rivi- ja sarakeprofilien taulukoista näkee helposti, että keskiarvoprofilien alkiot ovat massoja. Rivien keskiarvoprofilin alkiot ovat sarakemassoja, ja sama pätee sarakkeille. Tämä rivi- ja sarakeongelmien duaalisuus on yksinkertaisen korrespondenssianalyysin keskeinen idea. Greenacren perusoppikirjassa ensimmäiset luvut esittävät lähes kaikki menetelmän perusideat eri näkökulmilla tällaiseen taulukkoon (CAiP).

**k** ca-ratkaisu: rivi- ja sarekepilvien dimensio on sarakkeiden tai rivien lukumäärä vähenettynä yhdellä, pienempi kahdesta vaihtoehdosta. Tämä seuraa yksikertaisesti rivi- ja sarakeprofilien suhteellisuudesta, niiden summat ovat 1.

**k** Etsitään kaksiulotteinen ratkaisu (taso), joka minimoii pisteen etäisyyksien poikkeamien summan eli on mahdollisimman lähellä pistettä.

**K** rivi- ja sarakeongelma ratkaisu johtaa saamaan lopputulokseen. Tämä duaalisuus on korrespondenssianalyysin perusominaisuutta.



## 3.2 Symmetrinen kartta

Kartassa (symmetrisessä kartassa) on jo nimetty molemmat akselit, mutta tulkinta aloitetaan prosenteista. Ne kertovat, kuinka paljon aineiston inertiesta eli hajonnasta on kaksiulotteisessa projektiossa saatu kuvattua akseleille.

Ensimmäinen akseli saa aina suurimman osan inertiesta, tässä 76 prosenttia. Kun toinen akseli kuvaaa 15 prosenttia koko inertiesta, on kartalla esitetty 91 prosenttia aineiston hajonnasta. Loput 9 prosenttia jää 3. ja 4. dimensiolle. Nämä "selitysosuudet" ovat samantapainen laskelma kuin perinteisen regressiomallin "selitetty" vaihtelu ja "jäännösvaihtelu".

*Kontrastit* määrittävät akselien tulkinnan. Benzacrin ohjeen mukaan (1992, teoksessa GDA s. 49) katsotaan mitä on oikealla ja mitä vasemmalla. Akselien tulkinta perustuu siihen, mitä mitä yhteistä on kaikilla elementeillä jotka ovat origon vasemmalla puolella ja vastaavasti origon oikealla puolella. Samalla tavalla tulkitaan toinen akseli, mitä on ylhällä ja alhaalla.

Tässä tapauksessa taulukon rivit ovat havaintoja ("samples") ja sarakkeet muuttujia, ja akselien tulkinta tehdään muuttujien avulla. Vasemmalla on E ja oikealla puolella samanmielistä vastaukset s ja S. Neutraali "?" on s-vastausten vasemmalla puolella. Kun erot ovat suhteellisia, ei kuvan perusteella voi sanoa kuinka paljon.

Sarakkeet ovat oikeassa järjestyksessä, mutta niiden koordinaatit x-akselilla eivät ole tasaväliä. Jos muuttuja jostain syystä halutaan esittää välimatka- tai suhdeasteikon muuttujana koordinaatti ensimmäisellä dimensiolla on hyvä vaihtoehto.

Ensimmäisen dimensioon tulkinta on aika selkeä. Toinen akseli on kontrasti lievemmän tai maltillisemman erimielisyyden ja muiden vastausten kanssa. Se on 1. dimension suuntaan kaikkein lähimpänä origoa. Hieman varovaisemmin akselin voi tulkita maltillisesta ja jyrkemmin tai varmemmasta mielipiteestä kontrastiksi.

Maiden vertailu tehdään näiden akselien suuntaan. Sekä sarekepisteiden että rivipisteiden keskinäiset välimatkat approksimoivat optimaaliseksi niiden (khii2) etäisyyksiä. Sarake- ja rivipisteiden välisillä etäisyyksillä ei ole mitään suoraa tulkintaa. Pisteiden etäisyydet samassa pistepilvessä ovat suhteellisia, Saksa on konseptiivisempi kuin Suomi mutta emme tiedä kuinka paljon. Maiden järjestys oikealta vasemmalle on selkeä, Tanska on vasemmalla liberaalina "ääripääänä", oikealla taas Unkari ja Bulgaria. Pystyakselin suuntaan nähdään, että kaikkein "maltillisin" mutta kuitenkin liberaali on Suomi, jyrkimmät mielipiteet löytyvät Unkarista ja Tanskasta.

Näitä tulkintoja voi vertailla edellä esitettyihin kahteen kuvaan rivi- ja sarakeprofileista. Kartta kertoo aika paljon enemmän. Kartta on approksimaatio neliuolitteisen pistepilven hajonnalle. Vain origo on siinä tarkasti esitetty, se on koko aineiston keskiarvopiste, ja pisteiden hajonta sen ympärillä kuvaa poikkeamia riippumattomuushypoteesista.

Tärkeä geometrinen periaate on se, että kaukana on kaukana myös alkuperäisessä pistepilvessä, mutta kartalla lähellä olevat pisteet eivät välttämättä ole lähellä. Projektio kutistaa pisteiden etäisyyksiä.

Approksimaation laatu selviää korrespondenssianalyysin numeerisista tuloksista, samoin se miten rivi- ja sarakepisteiden määrittävät akselit.

Kartoissa tärkein tekninen yksityiskohta on kuva- tai muotosuhde (aspect ratio). Akseleiden mittayksikön pitää olla sama eli muotosuhteen yksi. Jos kuvia tulostetaan useassa formaatissa kannattaa olla tarkkana. Kuvien on jo analyysivaiheessa oltava lukukelpoisia, ja symbolien kokoa joutuu isoissa aineistoissa säätämään. Tulosten esittäminen lopullisessa muodossa vaatii jo paljon vaivannäköä, tässä tutkielman esitetään vain datan analysoinnin valikoituja kuvia. Graafinen data-analyysi on vaivatonta vasta sitten kun se tehty. En jatkossa esitä kuvalievia akseleiden nimiä kuvissa, akseleiden nimeäminen on kuvan tulkinnan toinen askel.

**k** Kuva tai kartta - käytän termejä synonyymeinä - on se taso, joka parhaiten "selittää" neliuolitteisen pisteparven hajontaa suhteessa koko aineiston keskiarvopisteesseen eli sentroidiin. Matemaattisesti ratkaisu saadaan soveltamalla singulaariarvohojitelmaa, ja tulokseksi saadaa taso joka on lähimpänä pistepilviä. Etäisyyttä mitataan massoilla painotetulla khii2-etäisyysmitalla.

**k** Intuitiivisesti idea on aivan sama kuin pääkomponenttianalyysissä (PCA, principal component analysis). Ratkaisu löydetään akseli kerrallaan. Ensi pistepilvestä etsitään akseli, jolle ortogonaalisesti projisoitujen pisteiden hajonta on suurin. Sitten etsitään sille kohtisuora toinen akseli samalla säännöllä, ja näin jatketaan kunnes koko

pilven hajota on jaettu näille uusille akseleille. tavoitteena on muutaman dimension approksimaatio moniulotteiselle datalle, yleensä kaksiulottoinen kartta.

**k** CA on painotettu PCA

### 3.3 Korrespondenssianalyysin peruskäsitteet

**edit** Sulava kuvaus tulkinnasta, painotus kuvien tulkinnassa. CA:n numeeriset tulokset vasta seuraavassa luvussa. Tässä ”mitä kuvasta näkee”, ei muuta (paitsi varoitukset - mitä ei näe). Idea koko ajan taulukon sarakkeiden ja riveien yhteyksien visualisointi.

**edit** Tärkeää selkeä kuvaus pääkoordinaattien ja standardikoordinaattien suhteesta. Tarkemmin teorialiitteessä, tässä heuristikesti jotta kuvia osaa tulkita.

Korrespondenssianalyysille on vakiintunut oma käsitteistö, joista tärkeimmät on jo mainittu. Kun tulkinta perustuu ”ääripäihin”, puhutaan kontrasteista ja distinktiosta. Luokittelumuuttujan arvot taas ovat modaliteetteja. Tärkein periaate on se, että kaikki on suhteellista. Ydinkäsitteitä ovat *korrespondenssianalyysin ”tripletti”*: *khii2-etäisyys, massat ja profilit*.

Kolmikko täydentää ”kvartetti”, neljä siitä johdettua käsittettä: *inertia* eli (painotettu) varianssi, *sentroidi* (painotettu keskiarvo, barysentrinen periaate), *aliavaruuus* ja *projektio*. (CAiP, s. 49).

*k* rivi- ja sarakeratkaisun duaalisuus: viite CAiP, jossa käydään läpi perusteellisesti. Rivi- ja sarakeratkaisut liittyvät tiivisti toisiinsa, kts. teorialiite.

*k* khii2-etäisyys on profilien painotettu euklidinen etäisyys (ja toki neliöjuuri!) jossa painoina ovat keskiarvo-profilin elementtien käänteisluvut eli elementtien etäisyyden neliö jaetaan keskiarvoprofilin alkiolla.

**edit:** khii2-etäisyydestä ehkä teorialiitteeneseen?

*k* khii2-testin oletukset eivät välittämättä ole voimassa kaikissa aineistoissa, mutta etäisyyssmittaa käytetään silti, sen perustelu on paljon yleisempi.

*k* khii2-etäisyys on ainoa etäisyyssmitta, joka toteuttaa distributional equivalence - periaatteen, CA:n ”tärkein juttu” (Benzecri), avain kaikkiin CA:n ominaisuuksiin. (Viite:CAip epilogi)

*k* normalisointi, samaan tapaan kuin PCA:ssa. Jos lukumäärätaulukko, Poisson-jakauman hajonta on sama kuin odotusarvo eli jaetaan poikkeama keskiarvosta hajonnalla. Poisson-jakaumassa odostusarvo ja hajonta ovat sama parametri. Tämä tulkinta khii2-etäisyydelle ei kuitenkaan saisi hämärtää massojen kaksoisroolia: ne ovat profilien painoja ja samalla standardoivat khii2-etäisyyden.

*k* CAiP epilogi: khii2 on yhteys Mahalanobis-etäisyyteen ja multinomijakaumaan, jonka realisaatioiksi profilit voidaan tulkita. (s. 301).

**Millaisista dataa?**

Korrespondenssianalyysin sovelletaan yleisimmin frekvenssitaulujen analyysiin, lukumäärädataan (count data). Periaatteessa mikä tahansa data sopii, kunhan se voidaan järkevästi esittää suhteelisina lukumäärinä (relative amounts), siis suhdeasteikon (ratio scale) muuttujana. Tässä oleellista on tulkittavuus tutkimusongelman näkökulmasta. Välttämätön ehto on sama mittayksikkö: lukumäärä, rahayksikkö, pituusmitta kelpaavat. (CAiP s. 15). Toinen ehto on ei-negatiivisuus (nolla tai nollaa suurempi luku), kts. kuitenkin GDA (tietyillä ehdoilla myös neg. arvot. Rajat ovat melko joustavat, kun mukaan otetaan erilaiset uudelleenskaalaukset ja transformaatiot. Tämä oli menetelmän perusidea jo Benzecrillä(CAiP ch 26, s. 201).

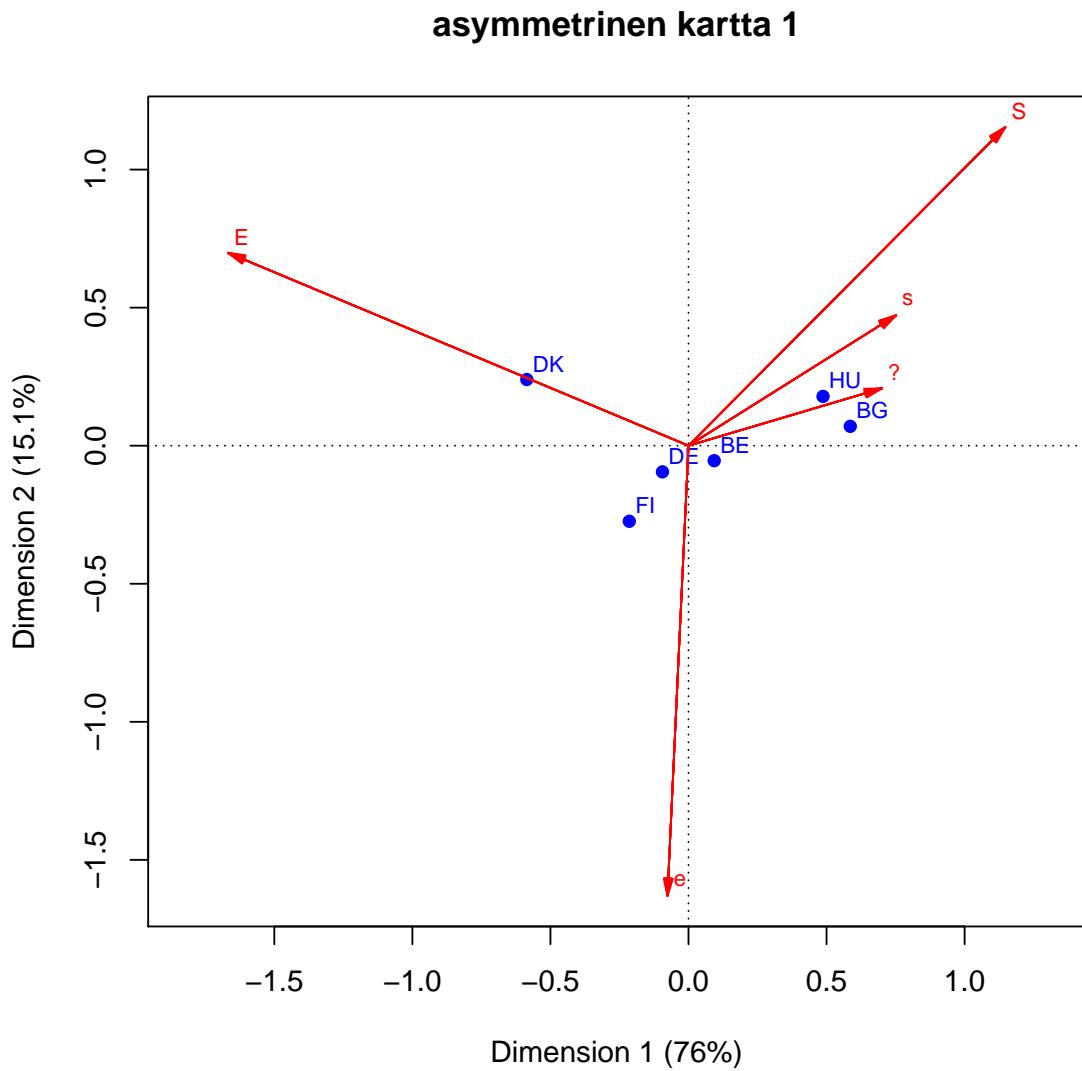
Menetelmää sovelletaan profileihin jotka painotetaan massoilla, ja profilien etäisyyksiä mitataan khii2-etäisyyssitalla. Jos datan voi esittää tässä muodossa, menetelmää voi käyttää.

#### 3.3.1 Asymmetrisen kartta ja ideaalipisteet

Symmetrisen kartta 3.3 on peruskuvaa ja esimerkiksi tässä käytetyn R-paketin ”ca” oletus. Siinä molemmat pisteparvet on esitetty pääkoordinaateissa (principal coordinates) ikääntkuin pääallekkäin, samassa kuvassa.

Toinen vaihtoehto on asymmetrinen kartta. Sarakeet ovat aineistossa muuttujia, joten ne voi esittää ns. standardikoordinateissa ja rivipisteet pääkoordinateissa.

Sarakepisteitä kutsutaan ideaalipisteiksi, ne edustavat kuviteellisia maita joissa kaikki vastaukset ovat samoja. Matemaattisesti kartalle projisoidut ideaalipisteet ovat (tässä esimerkissä) neliuolotteisen avaruuden verteksin (monikulmion) kärkipisteitä. Rivipisteet ovat tämän verteksin sisällä.



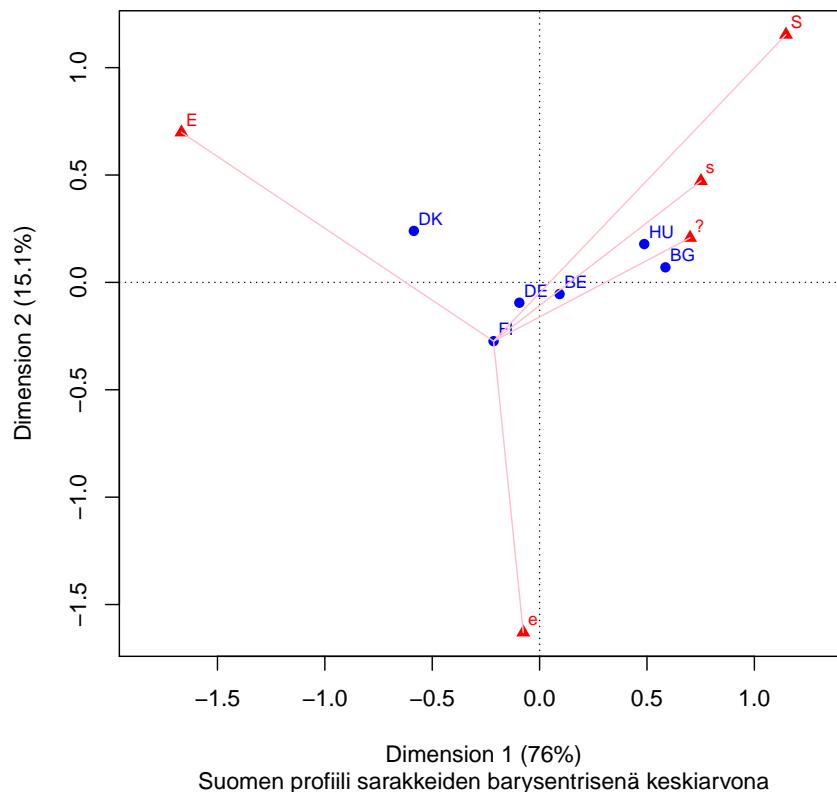
Kuva 3.4: Q1b: lapsi kärsii jos äiti on töissä

Sarakepisteet kuvaavat maksimi-inertiaa, ja rivipisteiden paljon pienempi hajonta kuvaaa niiden poikkeamaa tästä hypoteettisesta tilanteesta. Sarakepisteet skaalautuvat origosta ulospäin. Asymmetrisessä kartassa rivi- ja sarakepisteiden etäisyydellä on tulkinta, samoin rivipisteiden välisellä etäisyydellä. Sarakepisteiden välisillä etäisyyksillä ei ole tulkintaa. Sarakepisteet on skaalattu ja mittakaavan ero symmetriseen karttaan näkyi selvästi.

### 3.3.2 Barysentrinen periaate

Rivipisteet ja sarakepisteet yhdistää *barysentrinen periaate*. Jokainen rivipiste on ideaalipisteiden painotettu keskiarvo, painoina sarakkeiden käänneinen osuus riviprofilissa.

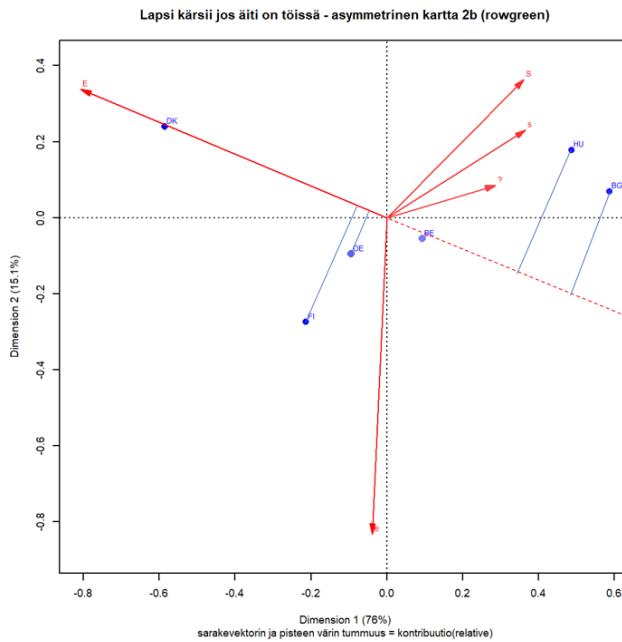
**edit** kuva ei ehkä tarpeen? Tehdään vähän pienempi (out.width = 60%, muuten 90%).



Kuva 3.5: Q1b: lapsi kärsii jos äiti on töissä

Suomen profili on kaukana S-sarakkeesta ja lähellä ? - saraketta, S-vastausten osuus on siis pieni ja ? - vastausten suuri.

Ideaalipisteiden tulkinnan voi varmistaa sarake kerrallaan, projisjoimalla rivipisteet origon kautta piirrettylle janalle. Kuvassa @ref(fig:G1\_3\_asymmtulk2) nähdään mikä on maiden järjestys E-vastausvaihtoehdossa.



Asymmetrinen kartta antaa kaksi uutta näkökulmaa rivien ja sarakkeiden suhteeseen. Sen huono puoli on ideaalipisteiden karkaaminen kauas origosta ja rivipisteiden pakkautuminen pieneksi parveksi. Jos rivipisteiden hajonta on suuri, kuva on käytännöllinen. Kyselytutkimusaineistoissa näin ei yleensä ole.

### 3.3.3 Kontribuutiot kartalla

Analyyseissä käytetty r-paketti “ca” esittää kartoilla myös pisteiden massat pisteen symbolin kokona, mutta tässä aineistossa eroja on vaikea nähdä. Tärkeämpi on pisteiden *kontribuutioiden* esittämien värisävynä.

Kun kartalla pistejoukon inertia kuvataan akseleille, on jokaisella pisteellä oma osuutensa akseleiden kuvamasta inertiesta. Absoluuttinen kontribuutio kertoo rivin tai sarakkeen osuuden akselin inertiesta. Vaikutuksessa on mukana pisteen massa.

Suhteellinen kontribuutio taas kertoo akselin osuuden pisteen inertiesta. Tämä tunnusluku kuvailee pisteen projektio laatuia, kuinka hyvin se on kartalla esitetty.

Kontribuutiokartta on asymmetrinen kartta, jossa sarakevektorit on skaalattu (kerrottu) massojen nelioillä. Nämä sarakevektorit “kutistuvat” kohti origoa mutta vektorin pituus kertoo edelleen sen suhteellisen massan. Kartta sopii niin pieniin kuin suureen inertian tilanteisiin (kts. esim. (?))

#### Absoluuttiset kontribuutiot

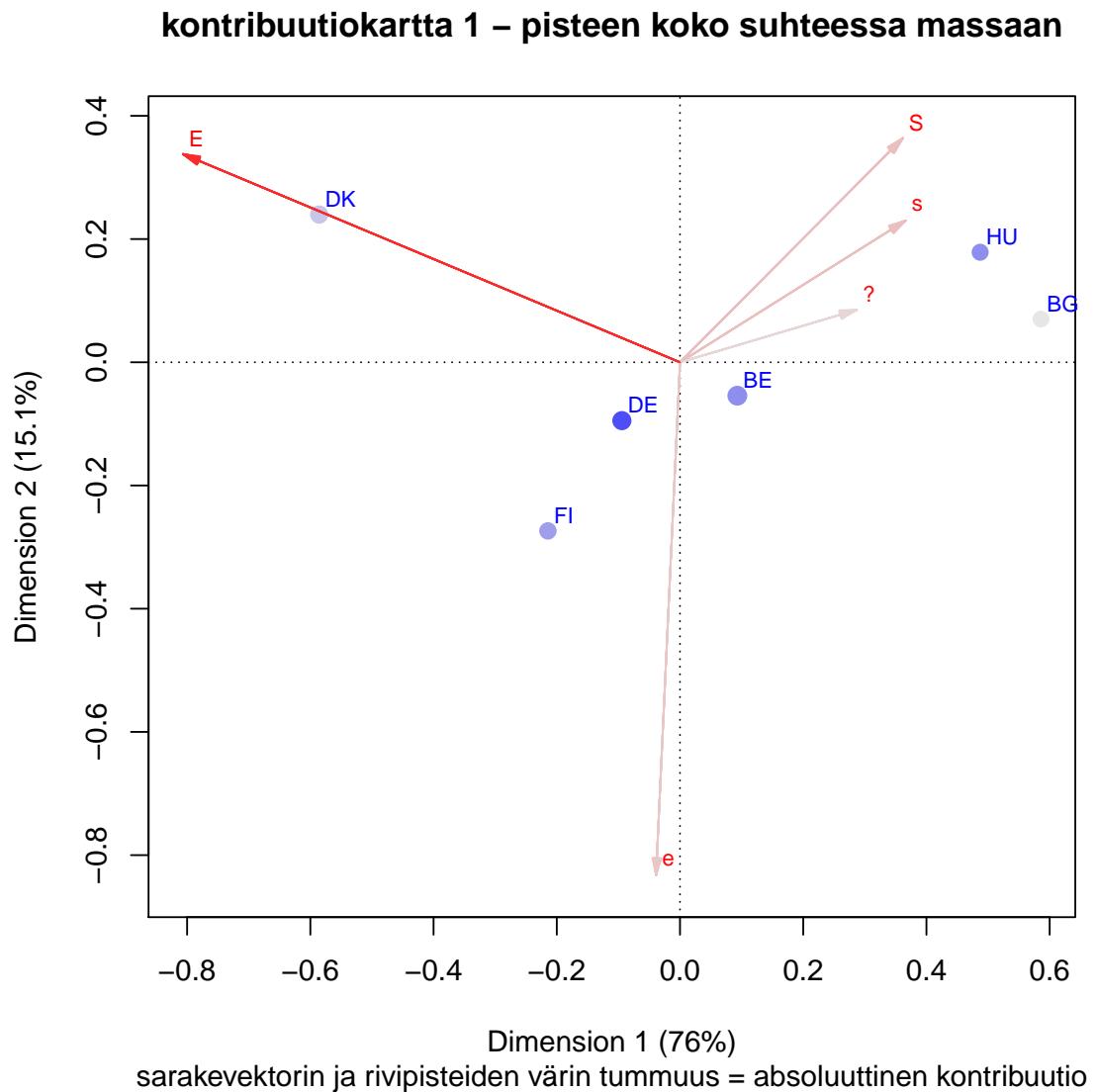
Absoluuttisten kontribuutioiden jakautumista akseleille voi varovaisesti päätellä sarakevektorin ja akseleiden välisistä kulmista. Mitä lähempänä sarakevektori on akselia, sitä suurempi on sen osuus akselin inertiesta. Samanlaisia päätelmiä voi tehdä myös rivipisteistä hahmottamalla janan niistä origoon.

Käsitteisiin palataan tarkemmin seuraavissa luvuissa ja teorialiitteessä, ja liian tarkkaan karttaa ei kannata tutkia. Numeeriset tulokset ovat yksityiskohdissa selkeämpia.

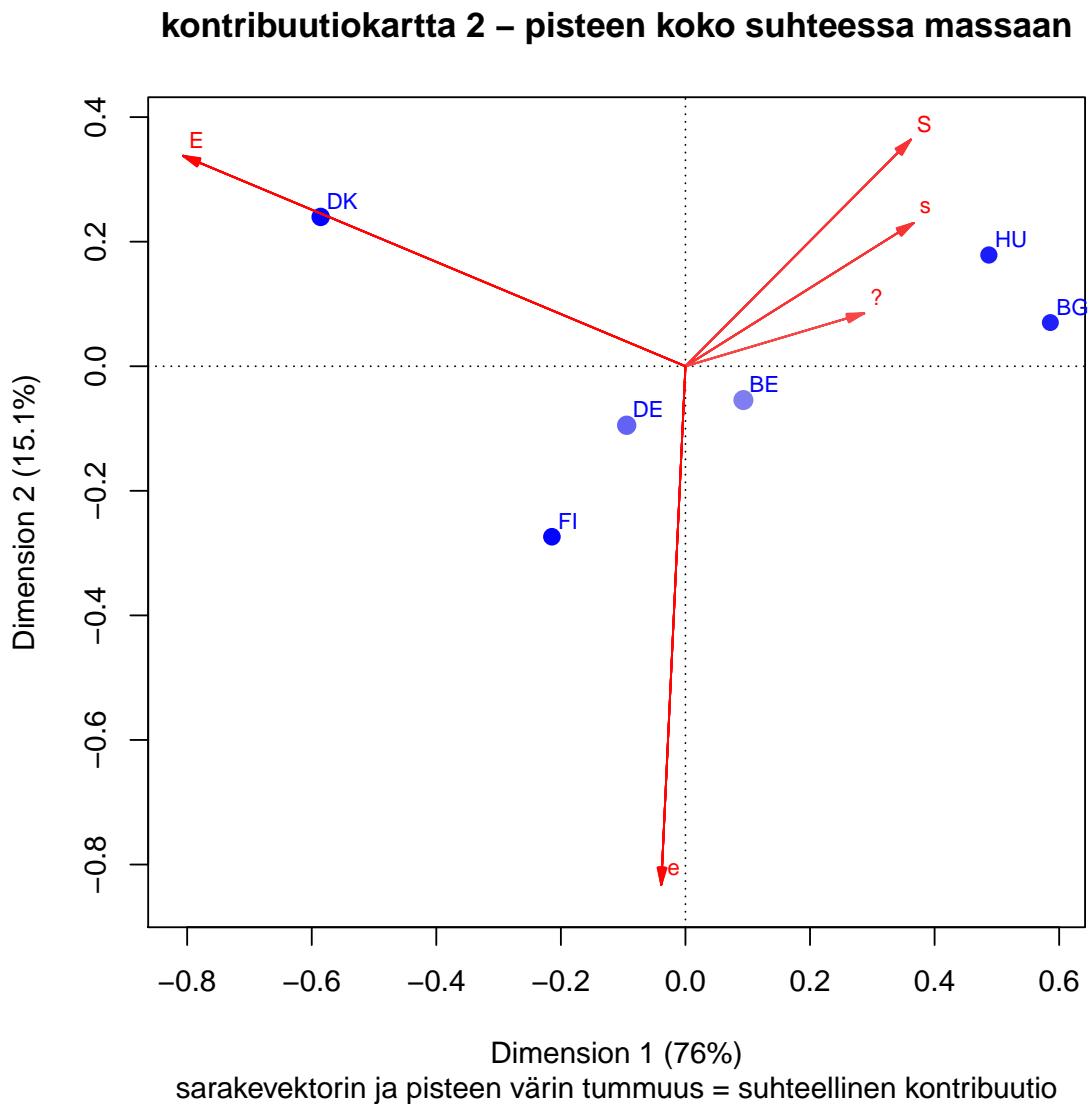
**edit** käytää termiä “vektori” vain kuvaan piirretyn “nuolen” nimityksenä.

Sarakkeista ratkaisuun vaikuttaa selvästi eniten E, ja juuri ensimmäiseen dimensioon. Toista dimensiota määrittää vahviten e, mutta myös kaikki muut sarakkeet x-akselin yläpuolella. Samaa mieltä olevien (S ja s) vaikutus näyttäisi jakautuvan selvimmin molemmille dimensioille.

Vaikka massojen suhteellisia eroja ei kovin helposti pistekoosta erota, se näkyy epäsuorasti Saksan melko vahimpana kontribuutiona. Bulgarian vähäisin kontribuutio näyttäisi olevan ensimmäiselle dimensiolle.



Kuva 3.6: Q1b: lapsi kärsii jos äiti on töissä



Kuva 3.7: Q1b: lapsi kärsii jos äiti on töissä

### Suhteelliset kontribuutiot

Kaikki edellä esitettyt pääätelyt perustuvat tietysti kaksiulotteiden projektioon. Jos pistet on esitetty hyvin eli niiden inertiasta (poikkeamasta keskiarvosta) suuri osa on kuvattu kartalle, rivipiste on sitä lähempänä ideaalipistettä mitä suurempi ideaalipisteen osuus on sen profilissa.

Sarakkeiden laatu näyttäisi olevan hyvä, mutta rivipisteistä Saksa ja erityisesti Belgia erottuvat hieman heikommin esitettyinä.

#### 3.3.4 Massat

**edit** Onko vakioitujen massojen kartta liian aikaisin? Tämä ei ole pääasia, vaan selvennys. Miksi tässä? Perusteltava, miksi en vakioi massoja maille, sukupuolle jne. (a) perusteltua kun tarkempi tutkimusongelma, esim. erottelut maiden ja sukupuolten välillä. Varianssianalyysin tapaan varianssin hajoittaminen ryhmisen sisäiseen ja ryhmien väliseen. Kts. teorialiitteestä esim. ABBA. (b) CA “perusmuodossa”, massa on yksi kolmesta tärkeimmästä käsitteestä. (c) on aika työlästä!

**edit** Galkussa verrattu molempien painotusten khii2-etäisyysksiä, jos tarpeen niin teoria-liitteeseen.

Massat ovat korrespondenssianalyysin keskeinen käsite, ja niiden kaksoisrooli on menetelmän ytimessä. Massat ovat normalisoiva muunnon khii2-etäisyysmitalle ja profiilien painoja. Tässä jälkimmäisessä roolissa massat liittyvät tutkimusongelmaan, mitä halutaan vertailla? Kun vertaillaan eri maita, ei ole kovin perusteltua käyttää massoina eri maiden otoskokoja. Jos taas halutaan vertailla vaikkapa miesten ja naisten vastauksia on luonnollista normalisoida miesten ja naisten massat yhtä suuriksi. Rivi- ja sarakemassat ovat verrannollisia taulukon rivi- ja sarakesummiin, frekvenssitaulukon reunajakaumiin. Ne voidaan tutkimusongelmaan sopivalla tavalla skaalata uudelleen. CAIP(s. 23) esimerkissä viiden koulutustaso-ryhmän massat skaalataan verrannollisiksi niiden väestötason osuuksiin, ei otoksen osuuksiin. Tällainen datan esikäsittely on normaali osa korrespondenssianalyysin soveltamista.

Jos massat halutaan vakioida yhtä suuriksi osajoukoissa, ratkaisu on yksinkertainen. Korrespondenssianalyysin taulukoksi otetaan riviprofiltaulukko, jossa rivien summat ovat yksi.

Kuvassa 3.8 on tehty näin, ja kartta eroaa hämmästyttävän vähän maiden otoskokoja massoina käyttävästä kartasta.

Pienimpien otosten maat (Bulgaria, Unkari) liikahtavat hieman origoa kohti, Bulgaria hieman enemmän kohti matallista puolta x-akselia.

Kontribuutiokarttakaan ei eroa edellä esitetystä kartasta. **edit** Tämä kuva on ehkä tarpeeton?

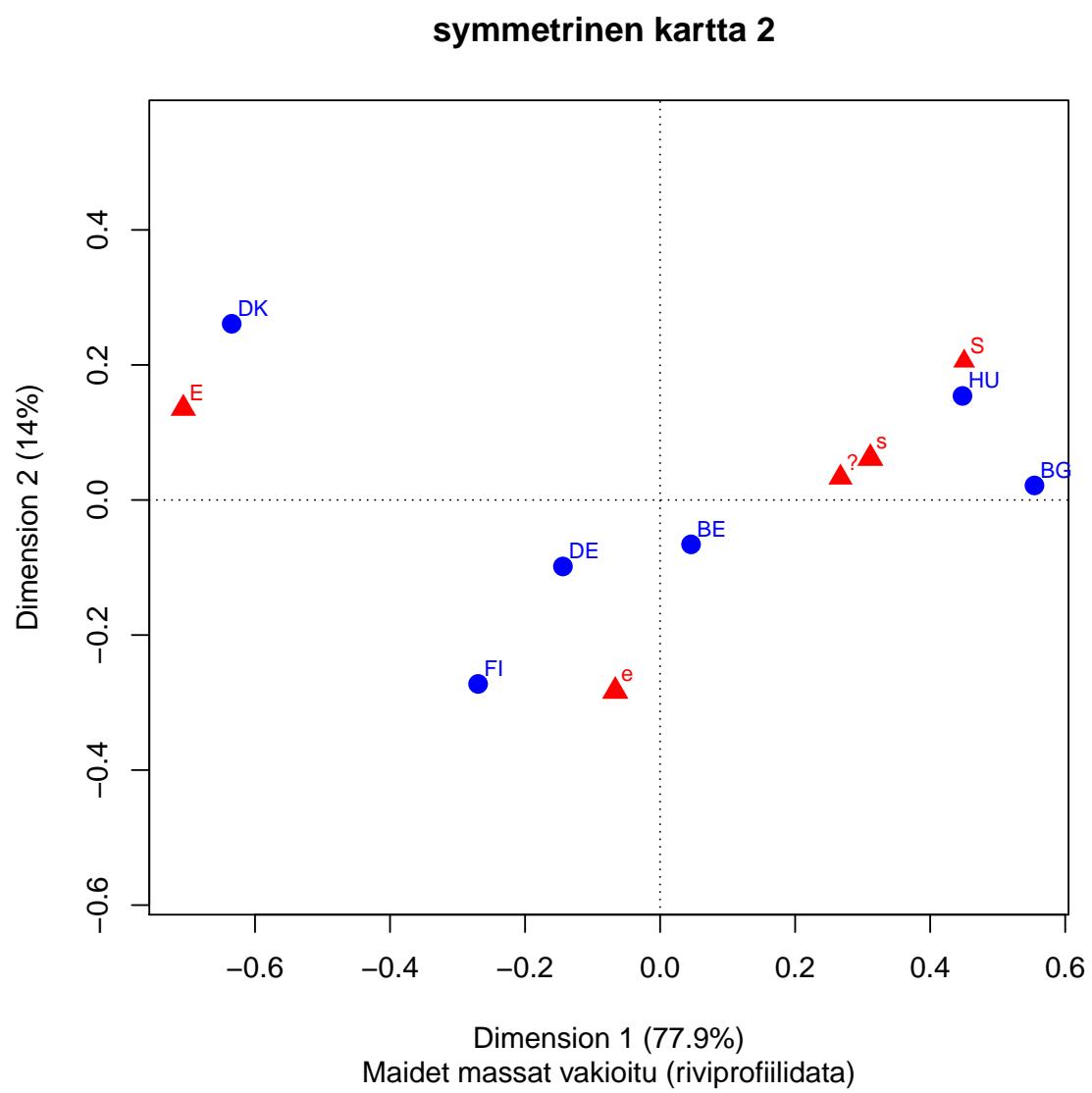
En ole vakioinut vertailtavien ryhmien (tässä maat) suhteellisia osuuksia. Syy on yksinkertainen: esittelen menetelmää sen perusmuodossa ilman kovin täsmällisiä tutkimusongelmia. Oikeiden tutkimuskymysten vastausia pitää tietysti etsiä järkevillä massojen skaalausella. Korrespondenssianalyysi on inertian eli kokonaishajonnaan dekomponointia, jakamista osiin.

#### 3.3.5 Karttojen erot

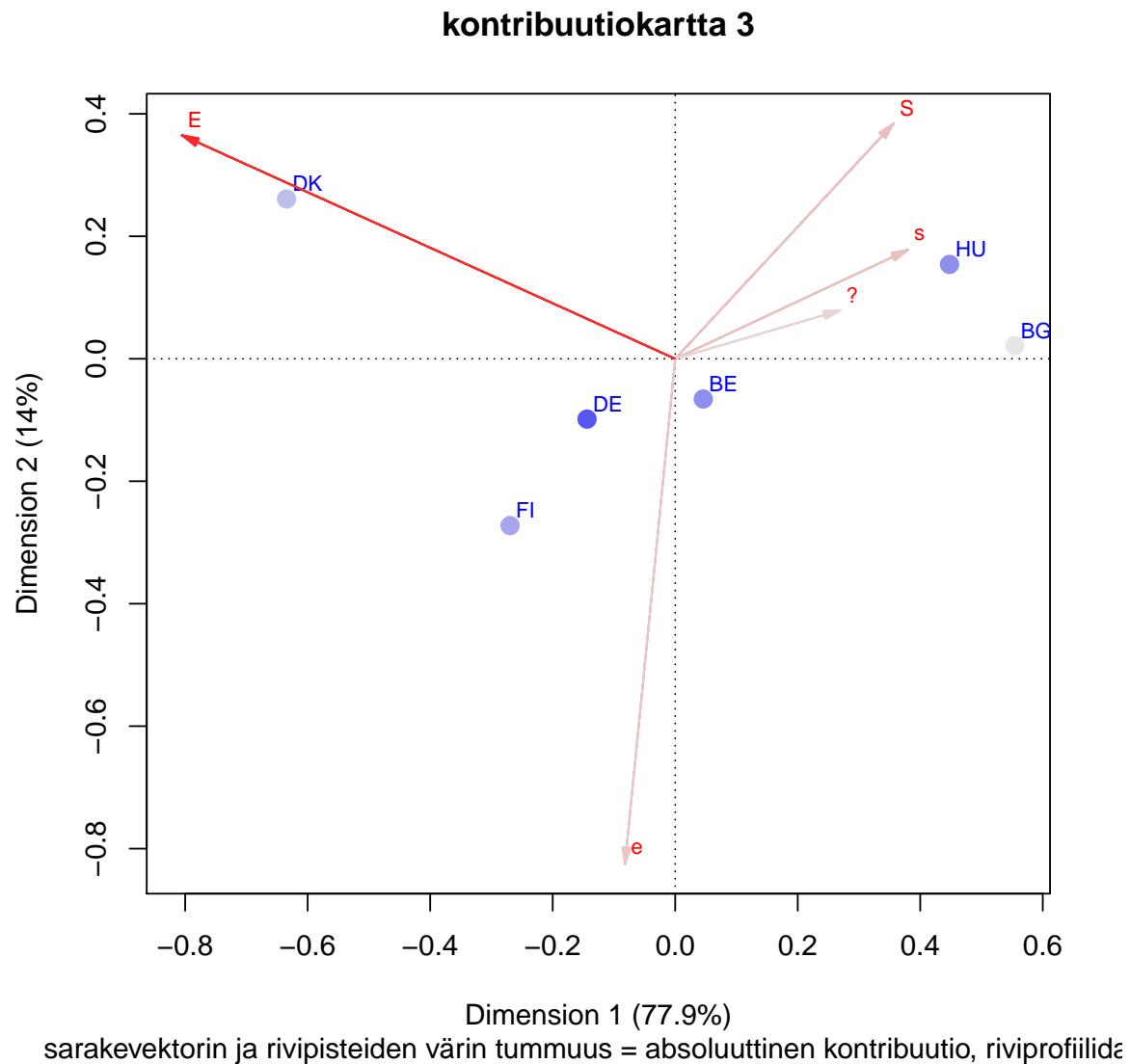
Yksinkertaisen korrespondenssianalyysin peruskuvan on symmetrinen kartta. Ehkä yllättäen sen “...tulkinta on edelleen menetelmän kaikkein kiistanalaisin aspekti.” (?) (s.295),(?) vrt. myös, Johdanto, SPSS - kritiikki.

Sarake- ja rivisteet esitetään siinä ikään kuin päällekkäin, samassa koordinaatistossa. Niiden pääkoordinaatit ovat kuitenkin eri pistejoukoista tai avaruuksista. Asymmetrisessä kartassa pistet ovat samassa avaruudessa, ja ero on Greenacren mukaan vain skaalaus. Asymmetrisessä kartassa standardikoordinaateissa esitettyt ideaalipisteet skaalataan pääakselien suunnassa vastaavilla pääakselien inertioiden neliöjuurilla. Siten pisteen suuntavektorit niin pääkoordinaateissa kuin standardikoordinaateissa ovat lähes samat kun akselien inertioiden (principal inertias) neliöjuuret eivät ole liian erisuuria.

Jos pääinertioiden neliöjuuret ovat hyvin eri suuruisia, tulkintaongelmaa voi tulla, mutta niillä ei käytännössä ole merkitystä. Siksi hän pitää skaalausdebattia akateemisena kiistana, käytännön sovelluksissa sillä ei ole



Kuva 3.8: Q1b: lapsi kärsii jos äiti on töissä



Kuva 3.9: Q1b: lapsi kärsii jos äiti on töissä

merkitystä. Kiista on ollut aika sitkeä (esimerkiksi 1989 Greenacren kommentoi skaalausta perusteellisesti (?)), mutta lienee laantunut.

Symmetrisen kartta hyvä vaihtoehto, sillä asymmetrisessä skaalaus vie ideaalipisteet usein kauas pääkoordinaateissa esitettyt pisteet pakkautuvat kuvan keskelle. Toisaalta jos dataa tulkitaan ”asymmetrisesti” kontribuutiokartta on hyvä vaihtoehto. Silloin rivipisteiden etäisyydet esitetään optimaalisesti, sarakkeiden suuntavektoreille projisoiduilla pistellä on kaksoiskuva-tulkinta (biplot) ja niiden pituudetkin kertovat jotain.

Greenacren mukaan kartoilla voi tavoitella kolmea eri asiaa, joista vain kaksi voi totetua yhtä aikaa. Kuvassa voi esittää rivipisteiden etäisyydet, sarakepisteiden etäisyydet tai rivi- ja sarakepisteiden etäisyydet. Jäkimäinen on kaksoiskuvien (biplot) ns. skalaritulo-ominaisuus. Rivi- ja sarakepisteiden skalaritulo ”palauttaa” alkuperäisen datan, tässä tapauksessa taulukon solun.

Näistä vain kaksi voidaan optimaalisesti esittää yhtä aikaa.

*Symmetrisessä kartassa* khii2-etaisyydet rivipisteiden välillä ja sarakepisteiden välillä esitetään optimaalisesti. Rivi- ja sarakepisteiden välisiä etäisyyksiä ei esitetä optimaalisesti, mutta ne voidaan tulkita kohtalaisen hyvin jos pääakselien inertioiden neliöjuuret eivät ole liian erisuuria.

*Asymmetrisessä kartassa* pääkoordinaateissa esitetyn pistejoukon etäisyydet kuvataan optimaalisesti, standardikoordinaateissa esitettyt pisteet ovat ”ääripofileja”, verteksin kulmapisteitä. Rivi- ja sarakepisteiden etäisyydet esitetään optimaalisesti, mutta sarakepisteiden etäisyyksillä ei ole suoraa tulkintaa,

*Kontribuutiokartta* on muunnelma asymmetristä kartasta. ”Ääriprofilit” vedetään kohti origoa kertomalla ne massojen neliöjuurilla. Nämä kuva selkenee, ja ”kutistetun” pisteen etäisyys origosta (“vektori”) kertoo sen kontribuution pääakselille. Näiden pisteiden välisillä etäisyyksillä ei ole suoraa tulkintaa.

Jako standardi- ja pääkoordinaatteihin on suora seuraus korrespondenssianalyysin matemaattisesta ratkaisusta. Greenacre esittelee kaksoiskuvia käsitlevää kirjassaan (?) selkeästi koordinaattien yhteyden ratkaisualgoritmien, singulaariarvohajotelmaan.

Koordinaattien yhteys voidaan esittää kahtena yksinkertaistettuna kaavana ( ?, s.174 ):

$$pkoordinaatit = standardikoordinatit \times \sqrt{pakselien inertiat} \quad (3.7)$$

$$kontribuutiokoordinaatit = \sqrt{massat} \times standardikoordinatit \quad (3.8)$$



## Luku 4

# Täydentävät pisteet

Kartat ovat analyysin väline, ja usein on hyödyllistä esittää kuvassa lisäinformaatiota tulkinnan avuksi. Täydentävät pisteet (supplementary points, CAIP s. 89-) ovat rivejä tai sarakkeita jotka lisätään karttaan. Mikä tahansa rivi tai sarake voidaan voidaan lisätä kuvaan, jos se on järkevästi vertailukelpoinen kartan määttäneiden profiilien kanssa.

Tällainen piste on kartan laskennassa *passiivinen*, sillä on sijainti kartalla mutta ei massaa eikä vaikutusta inertiaan. Passiivisilla pisteillä ei ole vaikutusta (kontribuutiota) kartan pääakseleihin.

Täydentävillä pisteillä on kolme yleistä käyttötarkoitusta. Kartalle voidaan lisätä profili, joka on jollain lailla sisällöllisesti erilainen kuin muut. Esimerkkiaineistossa kartalle voisi lisätä joitain Euroopan ulkopuolisia maita. Vaikka nämä riviprofilit eivät vaikuta kartan akseleiden määräytymiseen, ne voidaan esittää kuuden maan määrittämässä “avaruudessa”. Projektion laatu (suhteelliset kontribuutiot) voidaan myös esittää.

Toinen käyttötapaus on pienen massan profili. Tällaisella pisteellä voi olla iso vaikutus ratkaisuun, mutta passiivisena pisteenä se sijoitetaan muiden pisteiden määrittämälle kartalle. Jo sisällöllisistä syistä pienen massan pisteiden esitystä kannattaa harkita, ne sijaitsevat kaukana origosta ja huonontavat kuvan laatua. Esimerkkiaineistossa puuttuvat vastaukset voisi ottaa mukaan täydentävänä pisteenä.

Kolmas mahdollisuus on jakaa pistejoukkoja osajoukkoihin ja esittää niiden summaprofili täydentävänä pisteenä. Summaprofili on osiensa painotettu (barysentrisen) keskiarvo. Kun se esitetään passiivisena pisteenä, havaintoja ei oteta ratkaisuun kahta kertaa. Profilien yhdistämiseen liittyy korrespondenssianalyysin tärkein periaate, jakaumaekvivalenssi (*distributional equivalence*). Profileiltaan samanlaiset rivit voidaan yhdistää, analyysin tulokset eivät muutu. Khii2-etäisyysmitta taas on ainoa etäisyysmitta joka touttuaa tämän periaatteen. En esittele täitä ydinkäsittettä tämän enempää (kts. esim. CAIP tai perusteellinen matemaattinen esitys (?)).

Täydentävien profiilien lisääminen vaatii jo yksinkertaisia matriisioperaatioita. Korrespondenssianalyysi on käytännössä matriisien muokkausta tutkimusongelman tarpeisiin.

### 4.1 Saksan ja Belgian alueet

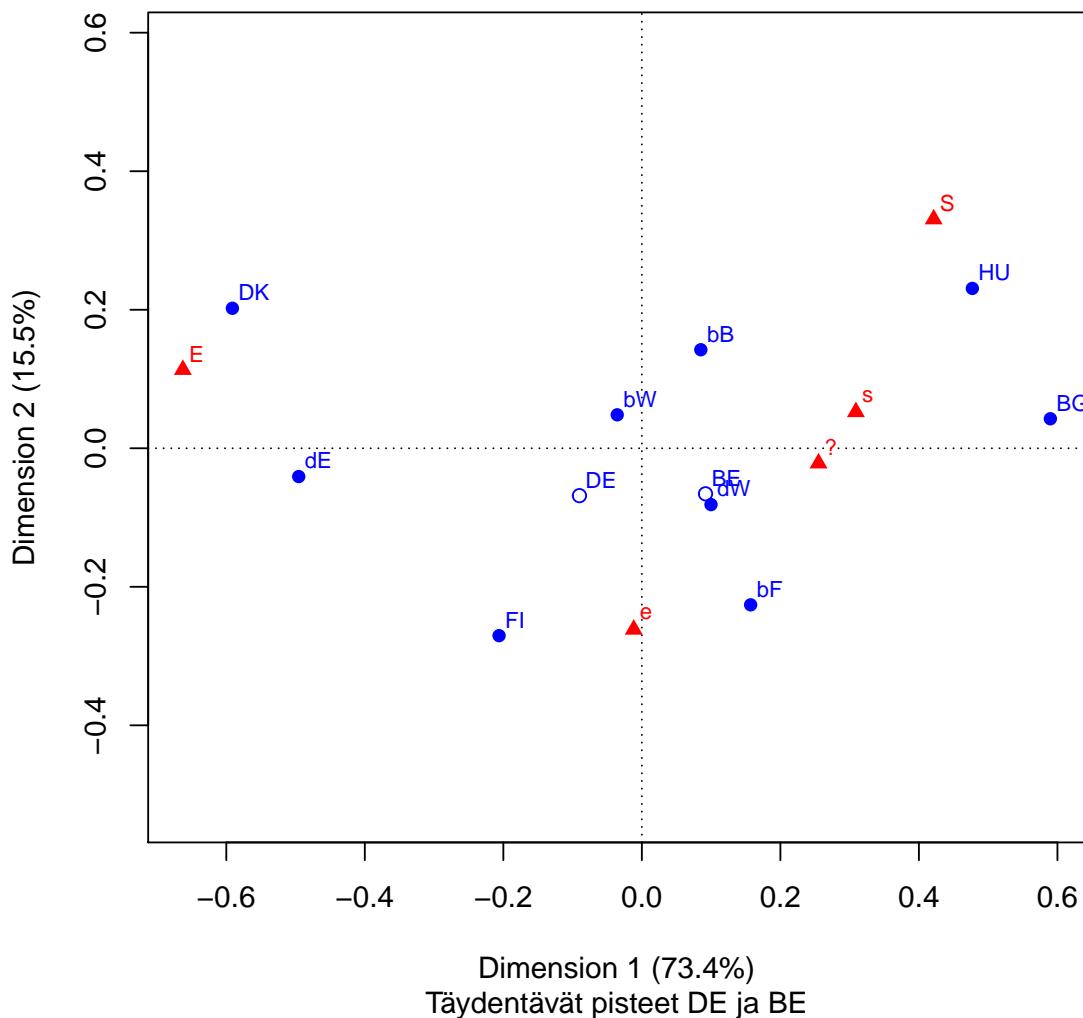
Saksan ja Belgian aineistossa on mukana aluejako: entiset itä- ja länsi-Saksa (dE,dW), Flanders (bF), Wallonia (bW) ja Bryssel (bB).

Aineistoon lisätään passiivisina riveinä Saksan ja Belgian maaprofilit (DE, BE). Maiden massoja ei skaalta yhtä suuriksi, otoskoot vaikuttavat ratkaisuun.

Saksan ja Belgian täydentävät pisteet ovat osiensa barysentrisiä keskiarvoja, etäisyys on sitä pienempi mitä suurempi on osuus. Saksan piste sijaitsee siksi lähempänä länsi-Saksan pistettä. Karttaa kannattaa verrata kuvaan jossa aluejako ei ole ??fig:simpleCA1map1), mutta Saksan ja Belgian osien sijoittuminen on kiinnostava. Itäinen Saksa on selvästi liberaalilla puolella, ensimmäisellä dimensiolla lähinnä Tanskaa. Läntinen Saksa on ensimmäisellä dimensiolla konservatiivisella puolella Belgian maapisteen tasolla. Belgian alueista Wallonia

Taulukko 4.1: Q1b vastaukset, Saksan ja Belgian alueet

	S	s	?	e	E	Total
bF	5.04	23.81	25.89	30.83	14.43	100.00
bW	10.82	21.02	18.57	24.08	25.51	100.00
bB	17.03	20.94	16.63	23.87	21.53	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00
dW	11.40	26.82	11.83	32.13	17.82	100.00
dE	5.85	11.33	10.97	29.80	42.05	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

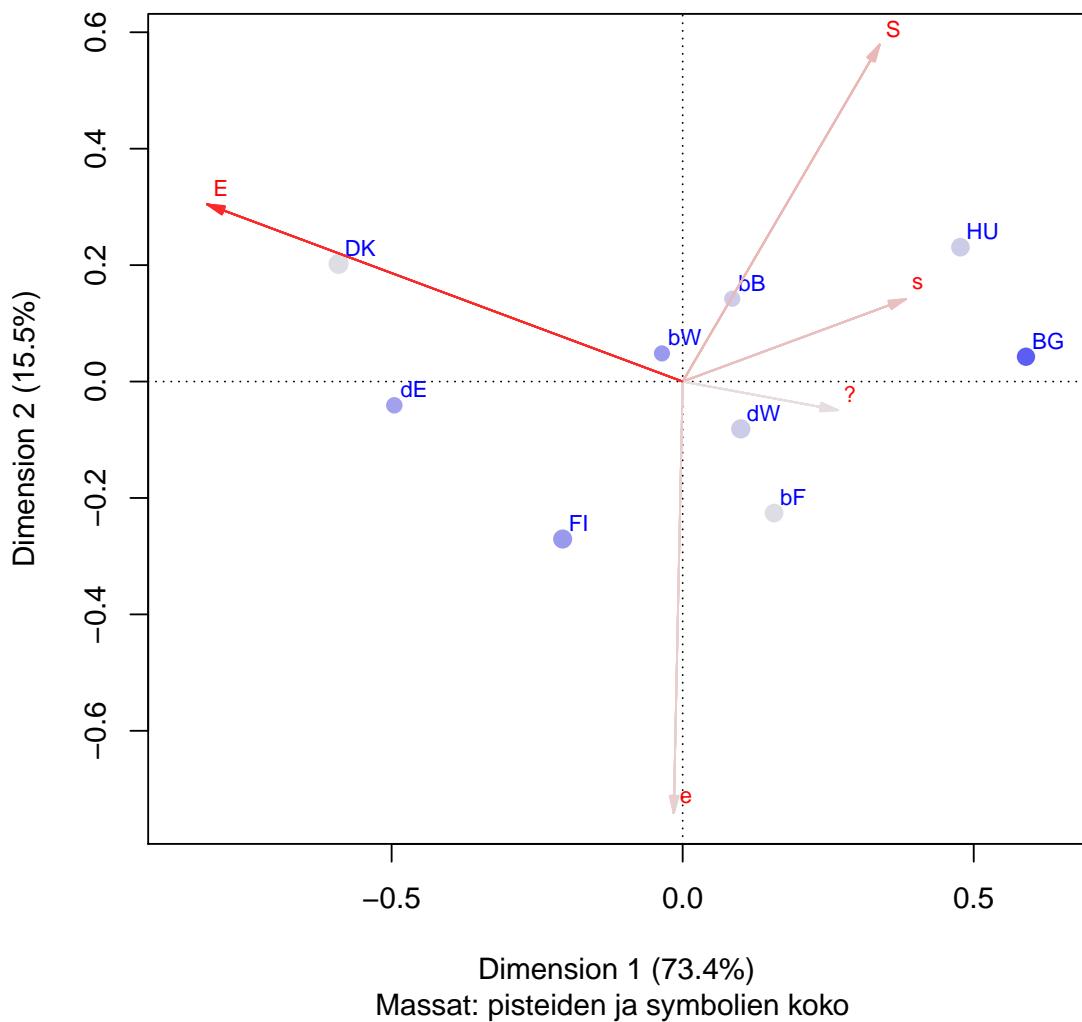
**Symmetrinen kartta 1**

Kuva 4.1: Q1b: Saksan ja Belgian aluejako

(bW) on liberaalilla puolella mutta kaikkein eniten oikealla. Bryssel ja Flander ovat konservatiivisella puolella, toinen länsi-Saksaa liberaalimpi ja toinen konservatiivisempi. Belgian osat hajoavat toiseen suuntaan kuin Saksaan, liberaalein Flanders on myös kaikkein maltillisin ja Bryssel vastaavasti tiukempien mielipiteiden puolella. Sarakepisteiden suhteelliset sijainnit toisiinsa nähden eivät oleellisesti muutu.

Bryssel ja Wallonia näyttävät olevan hyvin lievästi U-muotoisen maapisteiden parven sisällä. Tämä kaariefekti tai *Guttman-efekti* on kartoissa yleinen. Se on tavallaan seuraus ratkaisun geometriasta. Rivipisteiden pilvi on sarakkeiden ideaalipisteiden virittämän verteksin sisällä, ja ainoa reitti verteksin kulmasta toiseen kulkee tasolla kaarveasti (CAIP, s. 127). Voi myös sanoa, että kaariefekti taustalla on järjestysasteikon muuttujan korrelaatio (viite: LeRoux). Kaaren sisäpisteet ovat usein polarisoituneita ensimmäisen dimension "ääripää-vastausten" välillä. Tässä vaikutus on heikko, taulukossa 4.1 ei mitää selvää polariaatiota näy.

### kontribuutiokartta 1 – absoluuttiset kontribuutiot



Kuva 4.2: Q1b: Saksan ja Belgian aluejako

Kontribuutiokartasta täydentävät pisteet on jätetty pois, ne eivät vaikuta ratkaisuun. Pisteiden koko auttaa hahmottamaan niiden massojen eroja, sarakkeiden massoja ei juuri tässä kuvassa erota. Sarakkeiden kontribuutiot ovat samantapaiset kuin alkuperäisessä kartassa 3.6. Rivipisteiden kontribuutioista osa on selvästi pienempiä, erityisesti länsi-Saksaa kaksi Belgian aluetta (bB, bF). Unkarin ja Bulgarian kontribuutiot muuttuvat eri suuntiin, Unkarin pienenee ja Bulgarian kasvaa.

## 4.2 Korrespondenssianalyysin numeeriset tulokset

Korrespondenssianalyysin numeeriset tulokset ovat tärkeitä tulkinnan varmistamiselle ja antavat tarkemman kuvaan ratkaisusta. Nämä tulokset ovat erilaisia kokonaisinertiaan dekomponointeja. Kokonaisinertia (total inertia) profiilien ja keskiarvoprofilin kiih2-etaisyyksien massoilla painotettu summa ((3.6). Se kuvaaa profilipisteiden hajontaa ideaalipisteiden verteksin sisällä. Maksimi-inertia saavutetaan kun profililit ovat verteksiin kärkipisteissä, jokaisessa profilissa on vain yksi luokittelumuuttajan arvo. Inertia on sama kuin ratkaisun dimensio, tässä esimerkissä 4(sarakkeiden lukumäärä - 1). Tärkein lähde on CAiP:n luku 11 ja liitte B.

R-paketti "ca" (versio 0.71.1) listaa numeeriset tulokset suppeasti (print) ja laajemmin (summary), laajempi tulostus on alla. Paketin pdf-dokumentissa s.20 on kirjoitusvirhe, tiiviimmässä tulosteessa ei ole pääkoordinaattaja (principal coordinates) vaan standardikoordinaatit.

**k** Lyhyt selostus - nämä aika selkeitä

Ensimmäisenä on listattu kokonaisinertia pääakseleittain. Tässä suhteelliset luvut on esitetty prosentteina. Muut luvut on luettavuuden vuoksi skaalattu, joko kerrottu tuhannella tai esitetty "permills" (summa on 1000).

```
summary(suppointCA2)
```

```
##  
## Principal inertias (eigenvalues):  
##  
##   dim      value      %    cum%   scree plot  
##  1      0.154101  73.4  73.4 ****  
##  2      0.032489  15.5  88.9 ***  
##  3      0.014294   6.8  95.7 **  
##  4      0.008944   4.3 100.0 *  
##  
##  
## Total: 0.209828 100.0  
##  
##  
## Rows:  
##  
##   name   mass   qlt   inr   k=1 cor   ctr   k=2 cor   ctr  
##  1 | bF    124  650    69 | 157 212   20 | -226 438  195 |  
##  2 | bW     60  388     3 | -36 137    0 |   48 252    4 |  
##  3 | bB     63  481    17 |  85 127    3 | 142 354   39 |  
##  4 | BG    113  878   215 | 590 874   255 |   43  5    6 |  
##  5 | dW    143  345    33 | 100 208    9 | -81 138   29 |  
##  6 | dE     67  966    82 | -495 960   107 |  -41  7    3 |  
##  7 | DK    170  971   327 | -591 869   387 | 202 102   214 |  
##  8 | FI    136  957    79 | -206 352   38 | -271 605   307 |  
##  9 | HU    122  927   177 | 477 751   181 | 231 176   201 |  
## 10 | (*)BE | <NA> 512 <NA> |  92 338 <NA> | -66 173 <NA> |  
## 11 | (*)DE | <NA> 418 <NA> | -90 265 <NA> | -68 153 <NA> |  
##  
## Columns:  
##  
##   name   mass   qlt   inr   k=1 cor   ctr   k=2 cor   ctr  
##  1 | S    99  816   167 | 421 505 115 | 331 311 335 |  
##  2 | s    238 781   143 | 309 759 147 |  52 22 20 |  
##  3 |       168 594    88 | 255 589  71 | -22  4  2 |  
##  4 | e    261 871    98 | -12   2  0 | -262 870 550 |  
##  5 | E    234 999   505 | -663 971 667 | 113 28 93 |
```

Rivi- ja sarakeprofileista esitetään samat tiedot. Ensimmäisessä kolmen sarakkeen joukkossa kerrotaan pisteen massa, laatu (qlt) ja inertiantribuutio.

Inertiakontribuutio on suhteellinen osuus kokonaisinertiaasta. Aktiivisia rivejä on 9, joten tasaisesti jaettu ineria oli noin 110. Tanska, Bulgaria ja Unkari ”selittäväät” suurimman osan inertiaasta. Belgian ja Saksan alueiden kontribuutiot ovat pieniä. Nämä inertiaosuudet liittyvät kokonaisinertiaan alkuperäisessä neljässä ulottuvuudessa.

Laatu kertoo miten hyvin piste on esitetty kartalla, miten suuri osa sen inertiaasta on esitetty kartalla. Kaksiulotteinen kartta kuten tässä on yleisin valinta, laatu kerrotaan valitulle dimensioiden määälle. Laatu ei riipu massasta, vaan pisteen ja kartan akseleiden välisistä kulmista (kts. teorialiite). Saksan osien ero laadussa on iso, itä-Saksalla erittäin hyvä ja länsi-Saksalla huono. Belgian alueista Wallonia on kehtoinen esitetty, ja vain Flandersin laatu on kohtuullisen hyvä. Kovin hyvä ei ole täydentävien maapisteidenkään laatu.

Kaksi seuraavaa lohkoaa kertovat tulokset valituille dimensiolle eli ratkaisulle. Molempien dimensioiden (“ $k=1$ ”, “ $k=2$ ”) pääkoordinaattien ( $x$  1000) lisäksi raportoidaan dimension *suhteellinen kontribuutio* pisteen inertiaan (“cor”). Nämä tunnusluvut summautuvat laaduksi (qlt), ja ne voidaan tulkita korrelaation neliöiksi (kts. teorialiite). Erityisesti Belgian alueiden projektiön laatu on huonompi ensimmäisellä dimensiolla. Itä-Saksa ja Bulgaria taas ovat hyvin esittyjä vain ensimmäisellä dimensiolla eivätkä juuri ollenkaan korreloivat toisen dimension kanssa.

Pisteen *absoluuttinen kontribuutio* kertoo sen osuuden dimension inertiaasta (summa 1000). Jos katsotaan sarakkeita, nähdään E-sarake ”selittää” ensimmäisen dimension inertiaasta lähes 70 prosenttia, ja dimensio saman verran kokonaisinertiaasta.

**k** Tulosten käsitteiden esittely - tavoite kuvan laadun varmistus, akselien tulkinnan tarkistus. Tarkemmin teorialiitteessä. Tästä pitäisi nähdä, miksi seuraavat kartat ovat sellaisia kuin ovat. Nämä kolme sitaattia tekstin tarkistuksen tueksi, eivät tule lopulliseen versioon (18.11.20)

**k1 Contributions.** The contribution of point to axis is a statistic that depends both on the distance from the point to the origin point along the axis and on the weight of the point. The contributions of points to axes are the main aid to interpretation.

The contribution of a point to an axis is equal to the relative weight multiplied by the squared coordinate and divided by the eigenvalue.

Note on relative contributions. Both the contribution of a point to an axis (Ctr) and the quality of representation ( $\cos^2$ ) are relative contributions, since both are obtained by dividing the amount of variance of axis due to the point, by the variance of axis (Ctr) and by the amount of the overall variance due to the point ( $\cos^2$ ), respectively. Tästä kuva teorialiitteessä.

**k2 Varmuuden vuoksi:** CAiP-laskentaliitteestä (s.263):

mass: masses (1000) of the respective row and column points;

qlt: quality of representation (out of 1000) of the point in the solution of chosen dimensionality, in this case two-dimensional

inr: part of total inertia (out of 1000) of the point in the full space of the rows or columns

$k=1$  and  $k=2$ : principal coordinates on first two dimensions, multiplied by 1000

cor: relative contributions (out of 1000) of each dimension to the inertia of individual points. These are also interpreted as squared correlations ( $_1000$ )

ctr: contributions (out of 1000) of each point to the principal inertia of a dimension

**k3 Kontribuutiot:** yleisesti high contribution of a point to the inertia of the axis -> high relative contribution of the axis to the inertia of the point. Ei päde käänneen. ”Point ‘secretaries’ on the first axis is extremely well represented, but its contribution to the axis is minimal.

## 4.3 Esimerkki 3d- kartasta - Saksan ja Belgian dimensiot

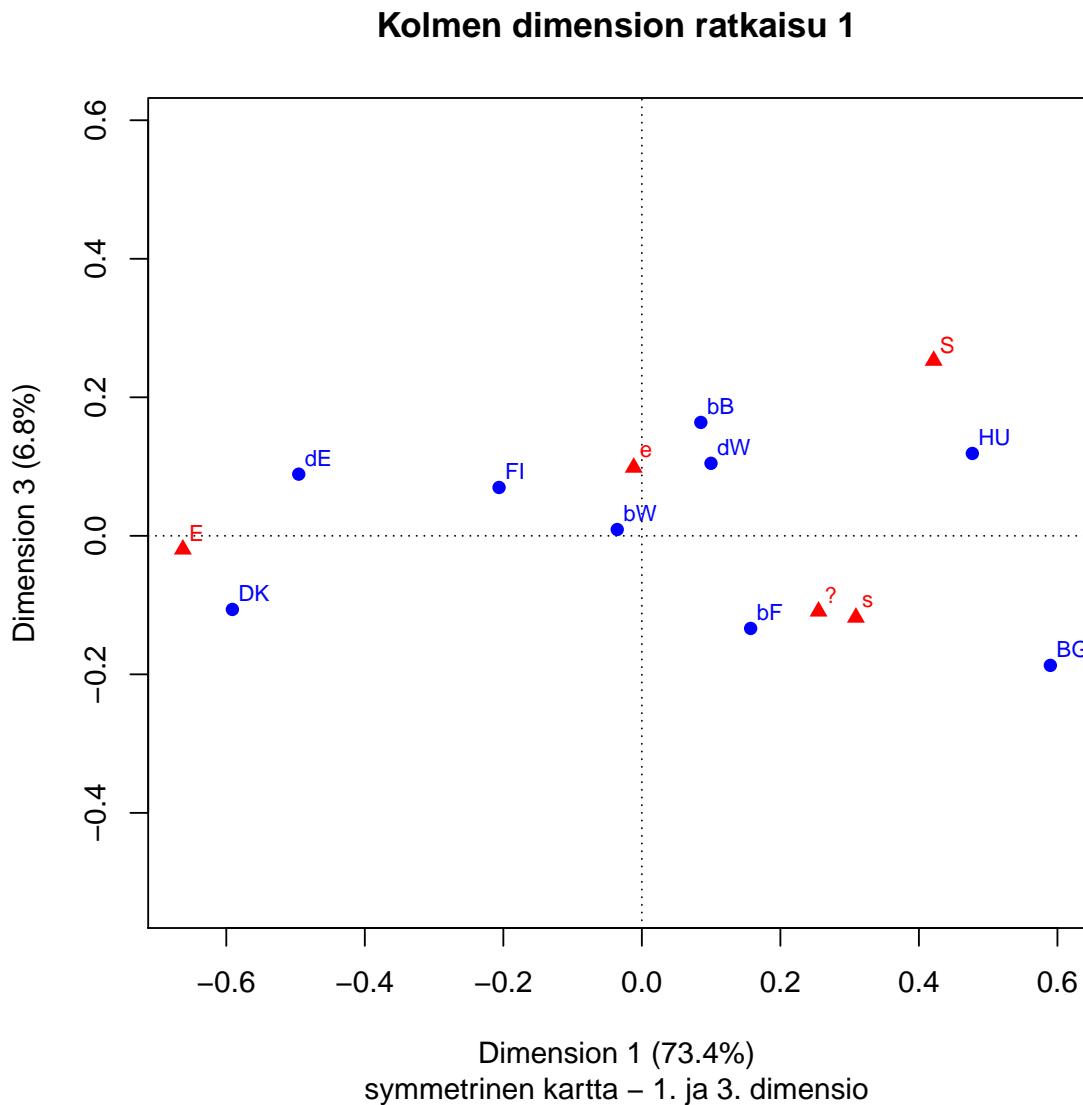
**k** Ei kovin hyviä kuvia, mutta periaate on tärkeä. Kartta on approksimaatio, pitää päättää milloin se on tarpeeksi hyvä. Tai mille pisteille hyvä, mille huonompi.

**edit 26.10.2020** summary-funktio ei toimi, kun dimensioita CA-ratkaisussa kolme. Numeeriset tulokset voisi laskea "käsityönä". Kehno kvalitetti 2d-ratkaisussa saa kuvissa selityksen.

**Kaksi karttaa - edit** 2d-ratkaisu esitetty, nyt 3d. ca-ratkaisun akselit ovat "nested"/sisäkkäisiä. **edit** "Kolmisormisääntö" auttaa tulkitsemaan kaksilotteisia "oikean käden" kuvia. Ensimmäinen dimensio on oikean käden peukalo, toinen etusormi ja kolmas keskisormi.

Esimerkki kolmiulotteisen ratkaisun tarkasta tulkinnasta ( ?, s.365), Ranskan poliitiikan dimensiot ("French political space") 1990-luvun lopulla.

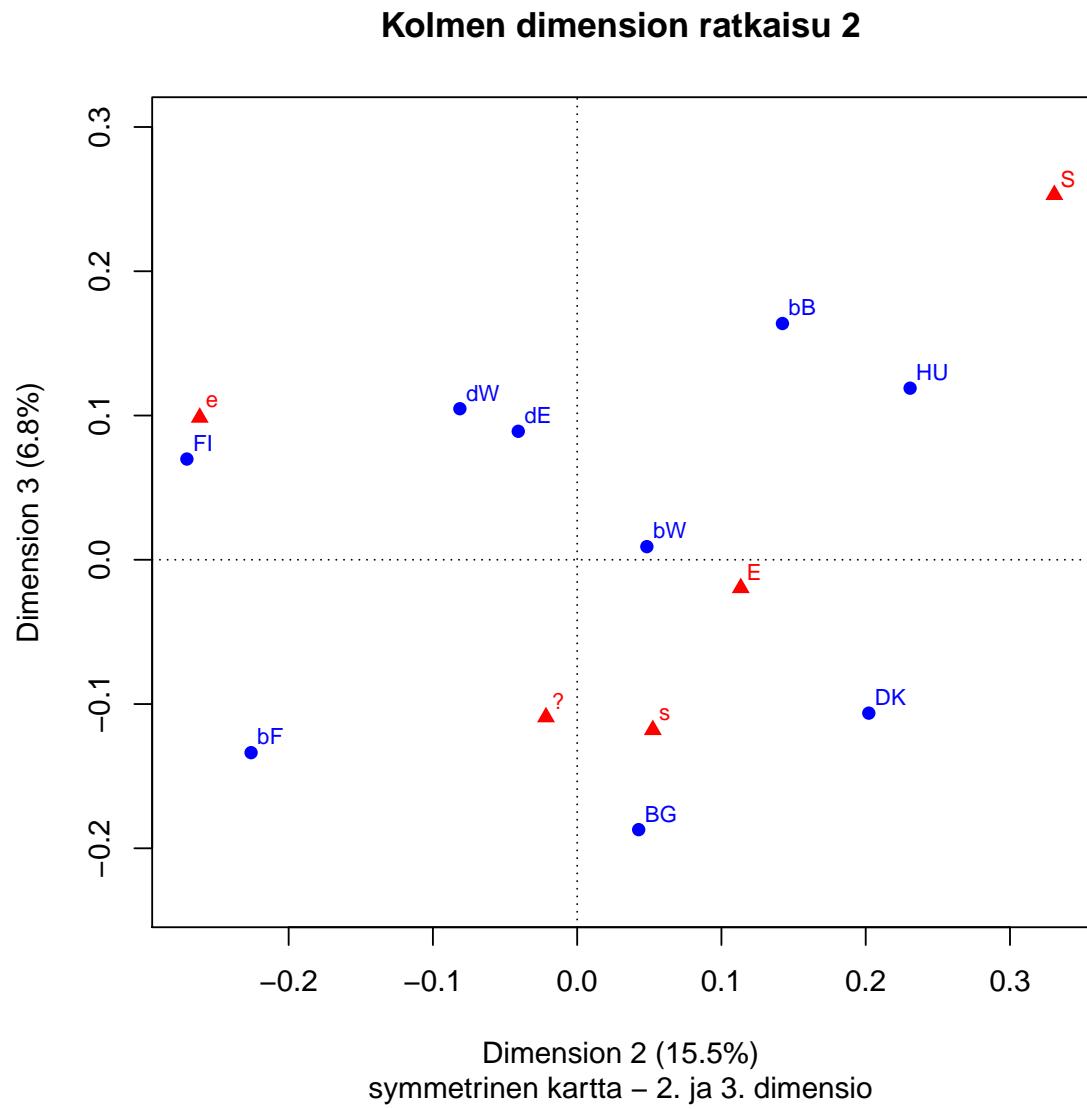
Ensimmäisen ja kolmannen dimesion kuvassa näkyy pisteparven hajonta tärkeimän dimension ympäillä. Sarakepisteiden järjestys säilyy samana, samoin maapisteiden oikealta vasemmalle.



Kuva 4.3: Q1b: Saksan ja Belgian aluejako

Toisen ja kolmannen dimension kartalla on esitetty noin viidesosa kokonaisinertiastä. Tässä Belgian pisteet ovat kuvan diagonaalilla.

**k** Tulkinta on aika hankalaa, ehkä riittää että toteaa selvän kolmiulotteisen rakenteen jossa Belgian alueiden ero näkyy.



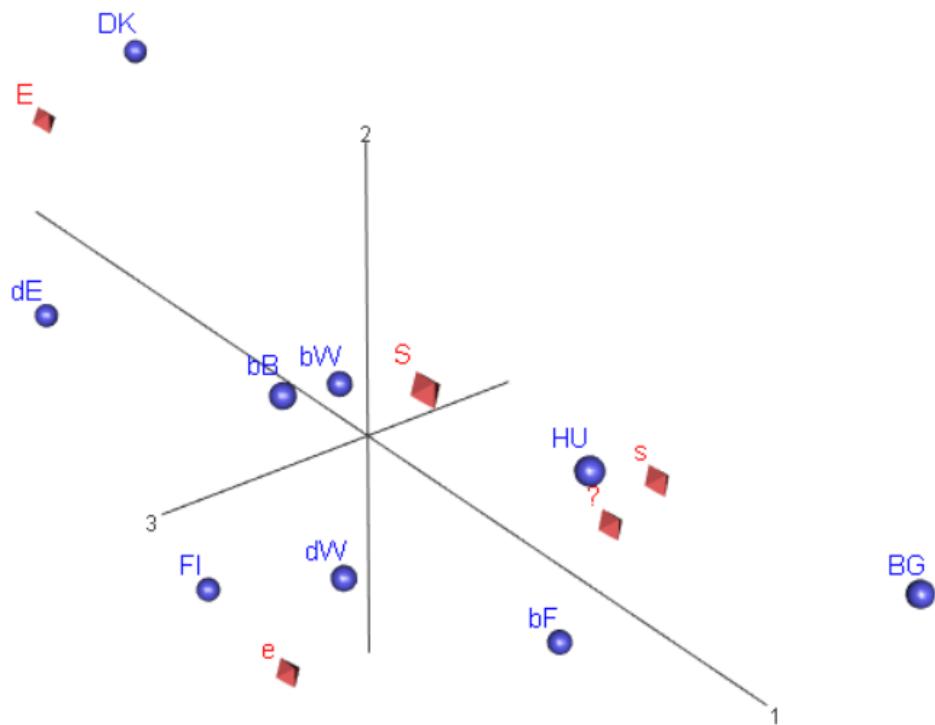
Kuva 4.4: Q1b: Saksan ja Belgian aluejako

Kahdesta projektiossa näkee kolmannen dimension suuntaiset suurimmat poikkeamat, niitä voi vertailla rivi- ja sarakepisteiden laatuun kaksiulotteisella kartalla.

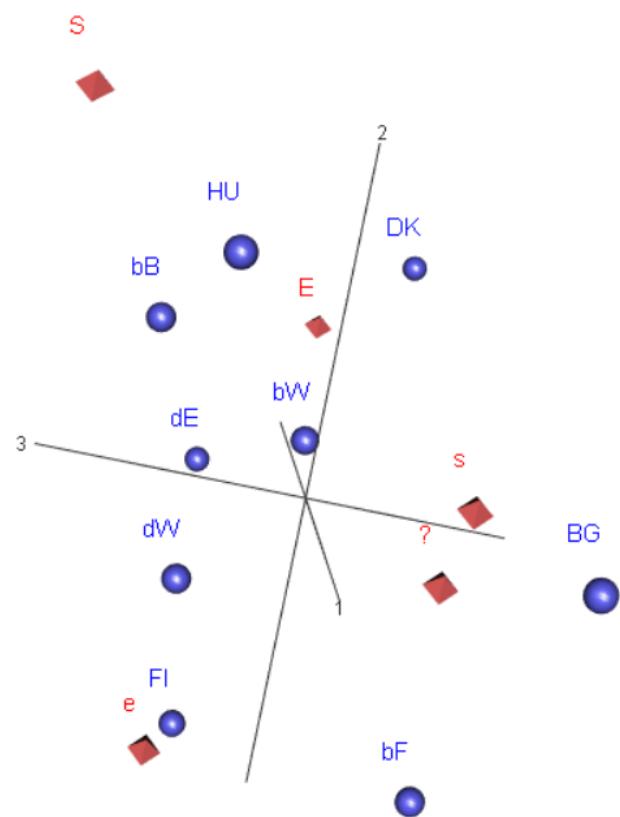
**k** Kolmiulotteisen kuvan tulkinta pitäisi aloittaa alusta, rivi- ja sarakepisteet hajaantuvat ulos tasosta. Ratkaisu on kuitenkin "sisäkkäinen" / "nested", kaksiulotteisen kartan pisteet vain siirtyvät kolmannen hieman ulos tasosta. Kolmannen dimension osuus kokonaisinertiaasta on noin seitsemän prosenttia. Belgian alueiden ero näkyy erityisesti kartassa 4.4 ja 4.6, samoin E-sarakkeen kohtalainen muutos.

**edit** - toinen ehkä riittää?

**k** Kuvakaappaukset ca-paketin kolmiulotteisista kuvista eivät kovin hyvin näйтä eroja. Dynaamisella kuvalla pistepilvien rakenteen hahmottaisi helpommin, tämä onnistuu useissa R-ympäristöissä.



Kuva 4.5: Saksan ja Belgian aluejako - 3d-kuval



Kuva 4.6: Saksan ja Belgian aluejako - 3d-kuva2



# Luku 5

## Yhteisvaikutusmuuttujat

Yksinkertaisin tapa tutkia taustamuuttujien yhteisvaikuksia on yhdistää kaksi muuttujaan uudeksi luokitelmumuuttujaksi (“interactive coding”). Miehet ja naiset on luokiteltu kuuteen ikäluokkaan (1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 tai vanhempi).

Poikkileikkausaineistossa vastaajan ikä kertoo myös ikäluokan (kohortin). Vastaajat ovat kokeneet kaksi suurta mullistusten vuotta elämänsä eri vaiheissa. Kaksin nuorinta ikäluokka on ollut 1990 alle 14-vuotiaita ja vanhin ikäluokka yli 44-vuotiaita. Finannsikriisin vuonna 2008 toiseksi nuorin ikäluokka on ollut 22-31 vuotiaita, ja kaksi vanhinta yli 51-vuotiaita. Pelkän ikävaikutuksen analyysi edellyttäisi vähintään kahden aineiston yhdistämistä.

Jatkan esimerkkiä kolmen muuttujan yhteisvaikutusmuuttjalla, jossa otetaan mukaan myös maa. Käytännössä kolmen luokittelumuuttujan yhdistäminen tekee taulukosta jo hieman huteron, joissain soluissa havaintojen määrä pienenee. Tässä kaikissa soluissa on sentää viisi havaintoa tai enemmän. Pienten massojen ja harvinaisten luokkien vaikutukset on kuitenkin arvioitava, ne voivat joskus mutta onneksi harvoin määrittää sitä liikaa.

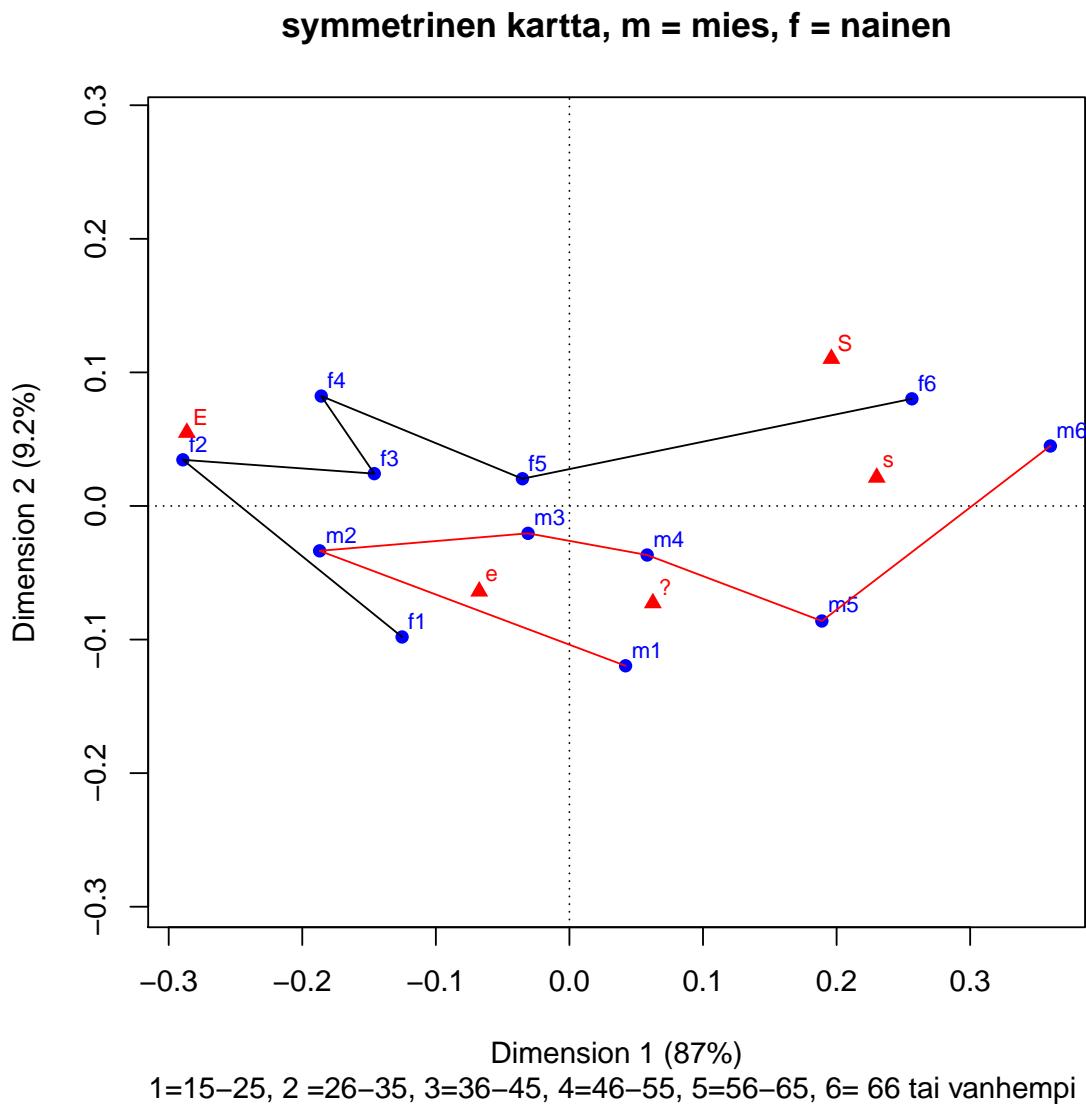
### 5.1 Ikä ja sukupuoli

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

Ikäluokilla on luonnollinen järjestys, niiden pistet voi yhdistää nuorimmasta vanhimpaan.

Ratkaisu on melko yksilöllinen, ensimmäinen dimensio kuvaa 87 prosenttia kokonaisinertiastä. Dimensioiden tulkinta on suurinpiirtein sama kuin kuin edellisissä kartoissa, mutta S-sarake on kiusallisesti s-sarakkeen vasemmalla puolella. Numeerisista tuloksista näkee, että sarakkeiden s ja E osuus kokonaisinertiastä (sarake irr) on 768. Niiden kontribuutio x-akselille on yhteensä vielä suurempi (849). Muut sarakkeet taas kontribuoivat y-akselin inertiaan, mutta sen osuus kokonaisinertiastä on vain 9 prosenttia. Kun sarakkeet kuitenkin ovat aika hyvin esitettyjä (qlt), voidaan x- akseli tulkinta hieman karkeammin samaa mieltä - eri mieltä - tasolla samaksi liberaalien ja konservatiivisten asenteiden ulottuvuudeksi. Toinen dimensio kuvaa tiukempaa samanmielisyyttä (S), kontrastina neutraali (?) ja maltillinen erimielisyys (s).

```
##  
## Principal inertias (eigenvalues):  
##  
##   dim    value      %   cum%   scree plot  
##   1     0.037448  87.0  87.0  ****  
##   2     0.003977   9.2  96.2  **  
##   3     0.001041   2.4  98.6  *  
##   4     0.000590   1.4 100.0  
##   ----- -----
```



Kuva 5.1: Q1b: ikäluokka ja sukupuoli

```
## Total: 0.043055 100.0
##
##
## Rows:
##      name   mass   qlt   inr    k=1 cor ctr    k=2 cor ctr
## 1 | f1 | 60 990 36 | -125 614 25 | -98 376 145 |
## 2 | f2 | 83 997 163 | -289 983 185 | 35 14 25 |
## 3 | f3 | 91 984 47 | -146 958 52 | 24 26 13 |
## 4 | f4 | 101 1000 97 | -186 836 93 | 82 164 172 |
## 5 | f5 | 98 879 4 | -35 658 3 | 20 221 10 |
## 6 | f6 | 100 951 176 | 256 866 175 | 80 85 162 |
## 7 | m1 | 57 659 32 | 42 72 3 | -120 587 205 |
## 8 | m2 | 66 977 57 | -187 946 62 | -34 30 19 |
## 9 | m3 | 78 457 5 | -31 318 2 | -20 139 8 |
## 10 | m4 | 89 674 14 | 58 482 8 | -37 192 30 |
## 11 | m5 | 89 988 90 | 189 818 85 | -86 170 166 |
## 12 | m6 | 89 978 277 | 360 963 307 | 45 15 45 |
##
## Columns:
##      name   mass   qlt   inr    k=1 cor ctr    k=2 cor ctr
## 1 | S | 99 915 128 | 196 695 102 | 110 220 304 |
## 2 | s | 238 969 304 | 230 961 336 | 21 8 27 |
## 3 | | 168 777 46 | 62 330 17 | -73 447 223 |
## 4 | e | 261 897 58 | -68 473 32 | -64 424 268 |
## 5 | E | 234 997 464 | -286 962 513 | 55 35 177 |
```

Rivien massat ovat yllättävän saman kokoisia, mutta kolmen ryhmän (f2, f6 ja m6) osuus kokonaisinertiaista on 616 ja niiden kontribuutio ensimmäiselle dimensiolle on 567. Vain 36-45-vuotiaiden miesten (m3) piste on huonosti esitetty (qlt = 457). Tulkinta on hankalaa miesten ja naisten nuorimman ryhmän osalta, vaikka efekti kartalla on iso. Molempien osuus kokonaisinertiaista on pieni (inr). Nuoret naiset (f1) on kuvattu kartalla erittäin hyvin. Nuorten miesten (m1) esityksen laatu on heikompi, ja kaikista suurin kontribuutio on vain y-akselille. Kun muut ikäryhmät (paitsi f3) ovat ikäjärjestyksessä vasemmalta oikealle, voi nuorimpien ja vanhimpien ikäryhmien sijainnin tulkita osittain toisen dimension (varma mielipide - epävarma mielipide) avulla.

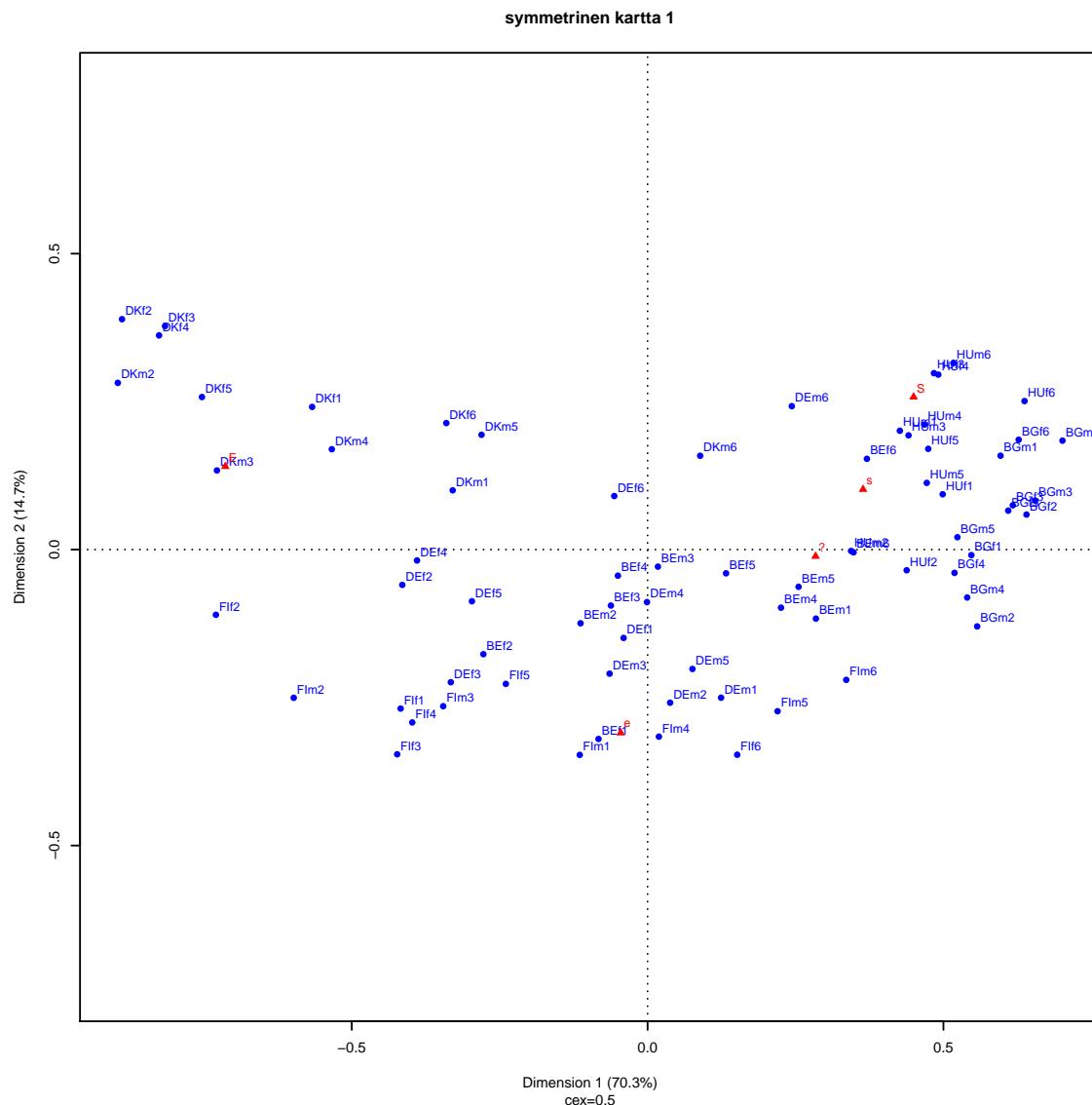
Selvästi kaikissa ikäluokissa miehet ovat konservatiivisempia kuin naiset. Nuorin ikäluokka on hieman vähemmän varma mielipiteistään kuin vanhin. Yksi mahdollinen selitys kartan tulkinnan ongelmilelle on se, että maiden välistet erot mielipiteissä ovat paljon suurempia kuin sukupuolten välistet maiden sisällä (ISSP 1994 aineisto, CAIP, s.126).

## 5.2 Ikä, sukupuoli ja maa

**edit** Yksi vaikeaselkoinen kartta täynnä pisteitä, tihrustellaan. Kommentoidaan numeeristen tulosten perusteella lyhyesti, oleelliset asiat.

```
ISSP2012esim1b.dat <- mutate(ISSP2012esim1b.dat,
                                maaga = paste(maa, ga, sep = ""))
# tarkistus, muunnos ok
# ISSP2012esim1b.dat %>% tableX(maa, maaga)
# head(ISSP2012esim2.dat)
# str(ISSP2012esim2.dat)
```

Kuvasta saa jotenkin selvää, kun sen suurentaa mutta pisteitä on selvästi liikaa. Joitain muuttujien nimiä voisi



Kuva 5.2: Q1b: ikäluokka ja sukupuoli maittain

lyhentää, kuva-alaa voisi rajata joihinkin osiin mutta osajoukon korrespondenssianalyysi tarjoaa pätevimmän vaihtoehdon.

**edit: Numeeristen tulosten listaus?** Numeeriset tulokset vievät paljon tilaa, mutta toisaalta niistä voi "lukea" miten kartta on rakentunut. Ne voisi ehkä jättää näkyviin ainakin seuraavassa luvussa?

Sarakkeiden järjestys vasemmalta oikealle ja ylhäältä alas on sama kuin edellisissä kartoissa. Dimensioiden tulokinta on sama, osuus inertista pienenee x-akselilla noin 6 prosenttiyksikköä. Pisteiden järjestys liberaalista konservatiiviseen alkaa Tanskan ja Suomen pistestä, sitten tulee Saksan ja Belgian pistetä ja konservatiisimpia ovat oikeassa laidassa Unkarin ja Bulgarian osajoukot. Toisella aksellilla maltillisia ja neutraaleja ovat hyvin karkeasti Suomen pistet ja lähes kaikki Saksan ja Belgian pistet. Eri maiden osajoukkojen suhteita on hankalampi hahmottaa, erityisesti kartan oikealla laidalla.

```
# Vilkaistaan numeerisia tuloksia, kopioidaan tekstiin jos on tarpeellista
```

```
# print(maagaCA1)
```

```
summary(maagaCA1)
```

```
##  
## Principal inertias (eigenvalues):  
##  
##   dim      value      %    cum%    scree plot  
##  1      0.184895 70.3 70.3 ****  
##  2      0.038751 14.7 85.0 ***  
##  3      0.024006  9.1 94.1 **  
##  4      0.015502  5.9 100.0 *  
##  
##  
## Total: 0.263154 100.0  
##  
##  
## Rows:  
##  
##   name   mass   qlt   inr   k=1 cor ctr   k=2 cor ctr  
##  1 | BEf1   | 14  678   9 | -83  43  1 | -320 635 38 |  
##  2 | BEf2   | 24  914  11 | -278 650 10 | -177 264 20 |  
##  3 | BEf3   | 21  320   3 | -62  96  0 | -95 224 5 |  
##  4 | BEf4   | 24  164   3 | -50  92  0 | -44 71 1 |  
##  5 | BEf5   | 23  332   5 | 133 304 2 | -40 28 1 |  
##  6 | BEf6   | 23  832  17 | 371 710 17 | 153 121 14 |  
##  7 | BEm1   | 11  429   9 | 284 367 5 | -117 62 4 |  
##  8 | BEm2   | 17  372   5 | -113 169 1 | -125 203 7 |  
##  9 | BEm3   | 20  108   1 | 17 29 0 | -29 79 0 |  
## 10 | BEm4   | 22  966   5 | 225 812 6 | -98 154 5 |  
## 11 | BEm5   | 22  728   8 | 255 686 8 | -63 42 2 |  
## 12 | BEm6   | 26  788  15 | 348 788 17 | -5 0 0 |  
## 13 | BGf1   | 5   531  11 | 547 531 8 | -9 0 0 |  
## 14 | BGf2   | 8   860  14 | 640 853 17 | 59 7 1 |  
## 15 | BGf3   | 12  815  21 | 617 804 24 | 75 12 2 |  
## 16 | BGf4   | 10  932  12 | 519 927 15 | -39 5 0 |  
## 17 | BGf5   | 14  880  23 | 609 870 28 | 66 10 2 |  
## 18 | BGf6   | 18  921  32 | 627 846 39 | 186 74 16 |  
## 19 | BGm1   | 5   940   7 | 596 878 9 | 159 62 3 |  
## 20 | BGm2   | 6   830   9 | 557 788 11 | -130 43 3 |  
## 21 | BGm3   | 8   709  19 | 655 698 19 | 83 11 1 |  
## 22 | BGm4   | 8   771  11 | 540 754 12 | -81 17 1 |  
## 23 | BGm5   | 10  979  11 | 524 977 15 | 21 2 0 |
```

```

## 24 | BGm6 | 9 692 27 | 701 647 24 | 184 45 8 |
## 25 | DEF1 | 13 425 3 | -41 29 0 | -149 395 7 |
## 26 | DEF2 | 15 938 10 | -415 919 14 | -60 19 1 |
## 27 | DEF3 | 19 846 13 | -333 582 11 | -224 264 24 |
## 28 | DEF4 | 23 985 13 | -390 982 19 | -18 2 0 |
## 29 | DEF5 | 17 839 7 | -297 772 8 | -87 67 3 |
## 30 | DEF6 | 23 116 8 | -56 32 0 | 90 84 5 |
## 31 | DEM1 | 13 912 4 | 124 180 1 | -250 732 20 |
## 32 | DEM2 | 13 766 4 | 38 16 0 | -259 749 22 |
## 33 | DEM3 | 15 737 4 | -64 63 0 | -210 674 17 |
## 34 | DEM4 | 21 137 5 | -1 0 0 | -89 137 4 |
## 35 | DEM5 | 19 603 5 | 76 75 1 | -202 529 20 |
## 36 | DEM6 | 22 849 12 | 244 427 7 | 242 422 34 |
## 37 | DKf1 | 10 991 15 | -567 839 18 | 241 152 15 |
## 38 | DKf2 | 14 991 49 | -888 831 58 | 389 160 53 |
## 39 | DKf3 | 17 963 53 | -816 793 60 | 377 170 61 |
## 40 | DKf4 | 18 977 57 | -826 820 66 | 362 157 61 |
## 41 | DKf5 | 16 998 38 | -753 894 48 | 258 105 27 |
## 42 | DKf6 | 12 808 9 | -340 579 8 | 214 229 14 |
## 43 | DKm1 | 15 981 7 | -329 898 9 | 100 83 4 |
## 44 | DKm2 | 13 989 43 | -895 900 55 | 282 89 26 |
## 45 | DKm3 | 13 982 28 | -728 950 38 | 134 32 6 |
## 46 | DKm4 | 15 941 19 | -534 855 24 | 170 86 11 |
## 47 | DKm5 | 13 643 9 | -281 435 6 | 194 208 13 |
## 48 | DKm6 | 15 355 5 | 89 85 1 | 158 270 9 |
## 49 | FIIf1 | 12 980 11 | -417 693 11 | -269 287 21 |
## 50 | FIIf2 | 12 927 26 | -730 907 34 | -110 21 4 |
## 51 | FIIf3 | 12 984 13 | -423 590 11 | -346 394 36 |
## 52 | FIIf4 | 14 991 14 | -398 644 12 | -292 347 32 |
## 53 | FIIf5 | 17 952 8 | -240 502 5 | -227 450 23 |
## 54 | FIIf6 | 11 835 7 | 151 134 1 | -347 701 35 |
## 55 | FIIm1 | 7 787 5 | -115 78 1 | -347 710 22 |
## 56 | FIIm2 | 9 977 14 | -598 832 17 | -250 146 14 |
## 57 | FIIm3 | 9 998 6 | -345 629 6 | -265 369 16 |
## 58 | FIIm4 | 13 837 6 | 19 3 0 | -316 834 33 |
## 59 | FIIm5 | 12 734 7 | 220 289 3 | -273 446 23 |
## 60 | FIIm6 | 9 911 6 | 336 637 6 | -220 274 12 |
## 61 | HUf1 | 7 723 9 | 499 698 9 | 93 25 1 |
## 62 | HUf2 | 11 689 11 | 438 685 11 | -35 4 0 |
## 63 | HUf3 | 12 808 18 | 484 586 15 | 298 222 27 |
## 64 | HUf4 | 11 768 18 | 491 564 15 | 296 204 25 |
## 65 | HUf5 | 12 850 13 | 474 753 14 | 170 97 9 |
## 66 | HUf6 | 13 671 34 | 637 581 28 | 251 90 21 |
## 67 | HUm1 | 6 935 5 | 426 766 6 | 201 170 6 |
## 68 | HUm2 | 9 381 11 | 344 381 6 | -2 0 0 |
## 69 | HUm3 | 13 957 12 | 441 803 13 | 193 154 12 |
## 70 | HUm4 | 10 999 10 | 468 830 12 | 211 169 11 |
## 71 | HUm5 | 13 942 12 | 472 891 15 | 113 51 4 |
## 72 | HUm6 | 8 726 15 | 517 529 11 | 315 197 20 |

##
## Columns:
##   name mass qlt inr k=1 cor ctr   k=2 cor ctr
## 1 | S | 99 653 155 | 450 492 109 | 258 162 171 |
## 2 | s | 238 741 174 | 364 687 170 | 102 54 63 |
## 3 | | 168 535 96 | 284 534 73 | -11 1 1 |

```

```
## 4 |   e | 261 941 103 | -45 20   3 | -310 921 646 |
## 5 |   E | 234 1000 471 | -714 962 645 | 141 37 119 |
```

Sarakkeet on kohtailaisen hyvin esitetty, heikoinmin neutraali vaihtoehto (535). Kun sen suhteellinen kontribuutio (cor) on vain 1 toisella dimensiolla jää loppuosa x-akselille. Maltillisuuden dimensiota määrittää e-sarakke (ctr = 646), ja vain sitä. Ensimmäistä dimensiota määrittää vahvimmin E-sarake (ctr = 645) liberaaliin ja samaa mieltä olevien sarakkeet (s, S) konsertatiiviseen suuntaan.

Kun aineistossa on 72 riviä on inertian suhteellisen kontribuution keskiarvo noin 14. Tämän ylittäviä kontribuutiota on Bulgaria naisilla (BGf2, BGf3, BGf5 ja BGf6) kaikilla konservatiiviseen suuntaan. Sama pätee Unkarin naisille, muuten naisten ikäluokat kontribuoivat yleensä liberaaliin suuntaan. Suomen pisteiden absoluuttiset kontribuutiot lähes pelkästään toiselle dimensiolle maltillisseen suuntaan. Tanska taas kontribui vahvasti jyrkempien mielipiteiden suuntaan.

Kaikissa taulukon soluissa on vähintään viisi havaintoa, muutama pieni massan havainto kontribuoii kohtuullisen paljon. Kuvan laadun takia ryhmiä pitäisi yhdistellä.

### 5.3 Stabiilisuus

Tarkastelen tässä vain ratkaisustabiiliutta (solution stability). Siinä data on annettu, ja ratkaisun numeerisista tuloksista nähdään miten pisteet määritättävät akselit. Ulkoinen stabiilius on laajempi käsite, mikä on datan suhde esimerkiksi johonkin perusjoukkoon (CAiP, s. 225). Ratkaisu on stabiili niiden pisteiden suhteen jotka eivät vaikuta siihen.

Korresponduenssianalyysiä ja erityisesti khii2- etäisyysmittaa on arvosteltu siitä, että se on liian herkkä harvinaisille luokittelumuuttujan arvoille. Yhteenvetotoartikkeli MG (Loose ends) tarttuu ”vaikuttavien poikkeavien havaintojen myytiin”, ja pitää sitä lähes aina perusteettomana.

Harvinaiset kategoriat ovat usein kartalla kaukana origosta, mutta jokaisella pisteellä on massa ja näillä ”out-layerilla” se on pieni. Niinpä niiden vaikutuskin on vaatimaton.

Harvinaisen kategoroiden vaikutus voi olla suuri, joten numeerisista tuloksita on tarkistettava onko hyvin pienien massan pisteillä suuri kontribuutio ratkaisuun. Käytännössä näin voi käydä esimerkiksi silloin, kun jonkun harvinaisen luokittelumuuttujan arvon havainnot ovat keskittyneet muutamaan profiliin joissa niiden osuuus on suuri (CAiP, s 298). Luvussa 7 nähdään, miten melko vähäinen määärä puuttuvia vastuaksia kasaantuu samaan vastaajien osajoukkoon ja mitä seurauksia sillä on.

Stabiiliutta voi helposti kokeilla määrittelemällä joitain pisteitä täydentäviksi pisteiksi.

Edellisen karten numeerisista tuloksista ei löydy pienien massan pisteitä joilla on merkittävä kontribuutio akselihin.



## Luku 6

# Osajoukon korrespondenssianalyysi

**k** Osajoukon korrespondenssianalyysin tarve: kuvat menevät tukkoon. Kun muuttujia on paljon kartat ovat täynnä pisteitä ja niitä on vaikea lukea. Usein myös havaitaan, että ratkaisun päädimensiot kertovat yllätyksettömän ilmeisen tarinan, ja kiinnostavammat yhteydet ovat piilossa ylemmissä dimensioissa (MG ja Pardo, "Vihreä kirja" ss. 198).

**k** Teoreettinen idea: inertian dekomponointi (taas). Lasketaan ensin ca-ratkaisu ja rajataan aineisto. Yksinkertainen toteuttaa. Massat ja reunajakaumat säilyvät, ja siksi aineiston ("pilven") kokonaisinertia voidaan jakaa osiin osajoukkojen inertiaksi. Uusia yhteyksiä näkyviin, tarkempi kuva dataasta.

Kuvan tukkoisuuteen voi toki vaikuttaa muillakin keinoilla. Muuttujien arvoja voi yhdistellä, kuvasta voi jättää pois pisteitä joiden kontribuutio on pieni tai kuva voi yksinkertaisesti suurentaa analyysivaiheessa näytöllä.

Kuva-aluetta voi myös rajata johonkin osaan kuvaa. Kaikki pisteet ovat kuitenkin mukana määräämässä ratkaisun geometriaa.

Rivejä tai sarakkeita voi myös muuttaa passiivisiksi, mutta silloin ei saavuteta käytännöllistä inertian dekomponointi-ominaisuutta. MCA-jaksossa osajoukon analysointi on pääasia ja mahdollisia vaihtoehtoja käsitellään hieman lisää.

Lähteet MG ja Prado (2006)(?), "vihreä kirja" (?), CAIP (?).

**k** Simple CA subset

**k1** Täydentävät pisteet voi ottaa mukaan jos ne eivät ole rajatusta "pilvestä". Esim. rivien osajoukon analyysiin voi ottaa sarakepisteitä täydeäntvinä pisteinä normaalisti. Osajoukon profilit muuttuvat, niiden summa ei enää ole yksi, barysentristä ominaisuutta ei voi suoraan käyttää täydentävien rivipisteiden koordinaattien laskemiseen.

**k2** Kuvien luettavuus ja pistekoko, hankala juttu

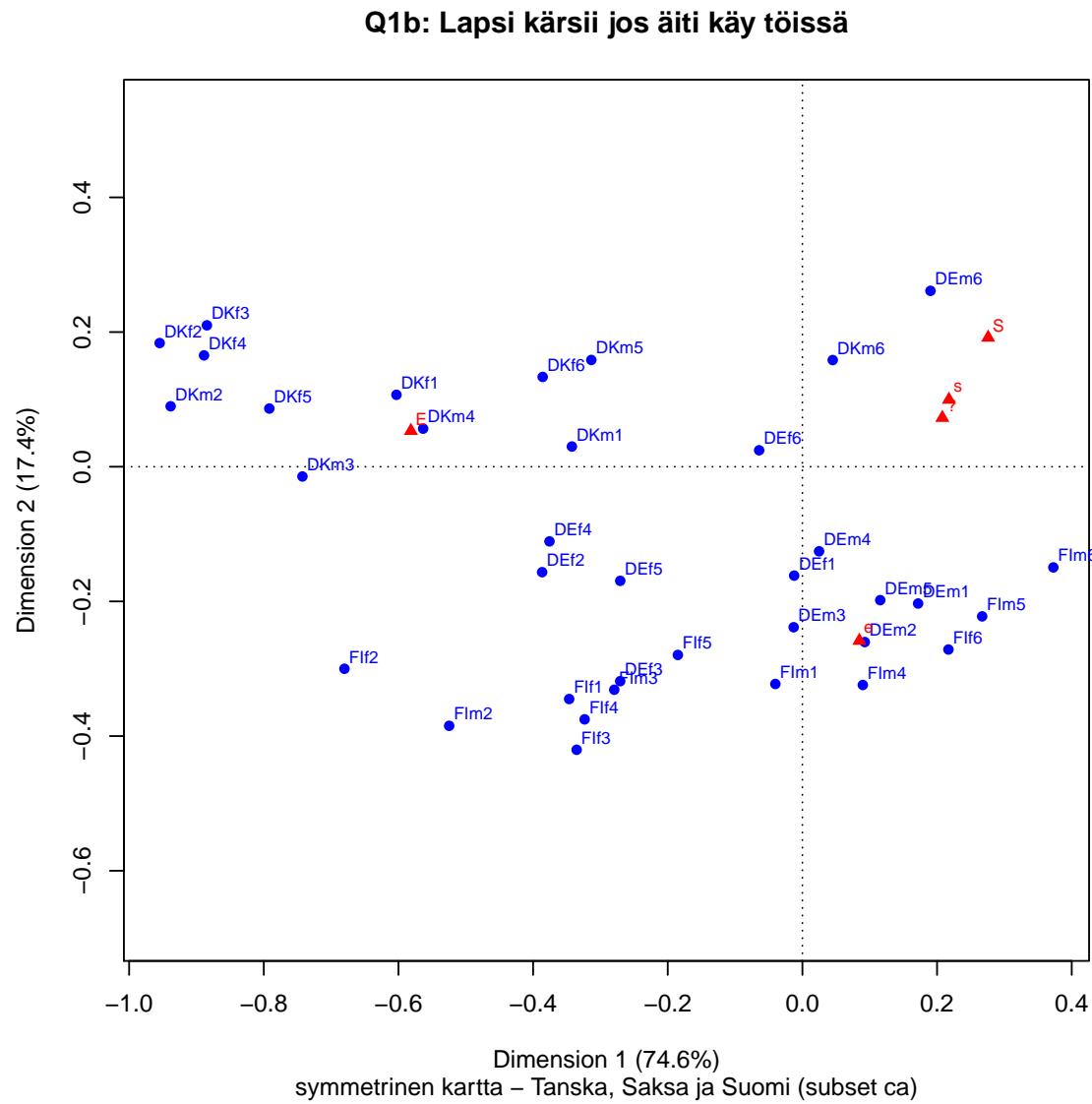
**k** Kuvien järjestykseen perustelu. 1. symmetrinen, voidaan arvioida rivipisteiden ja sarakepisteiden etäisyyksiä. Akselien tulkinta.

2. Kontribuutikartta: laatu, tulkinnan tarkentaminen
3. Rivipisteiden etäisyydet - siksak

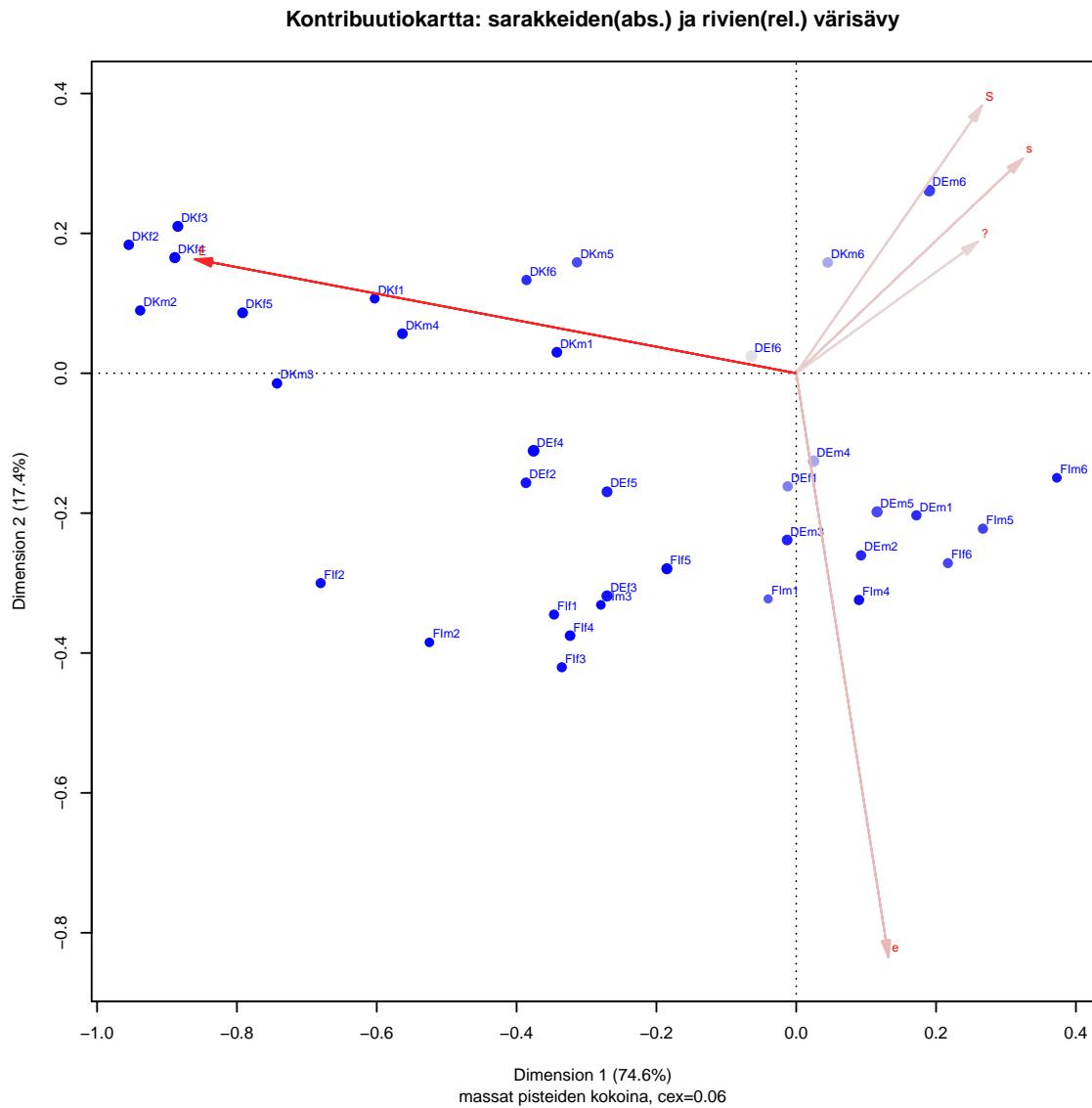
**k** Kontribuutiokarttan tulkinta

**k** Akseleiden tulkinta: 1. dimensio on kontrasti vahvasti eri mieltä olevien ja kaikkien muiden vastausten välillä. Toinen dimensio kontrasti e-vastauksen ja kaikkien muiden vastausten kanssa, erityesti S, s, ja myös neutraali ?.

**k** Ei tulkintaa sarakkeiden etäisyksille, mutta kontribuutiolle on.

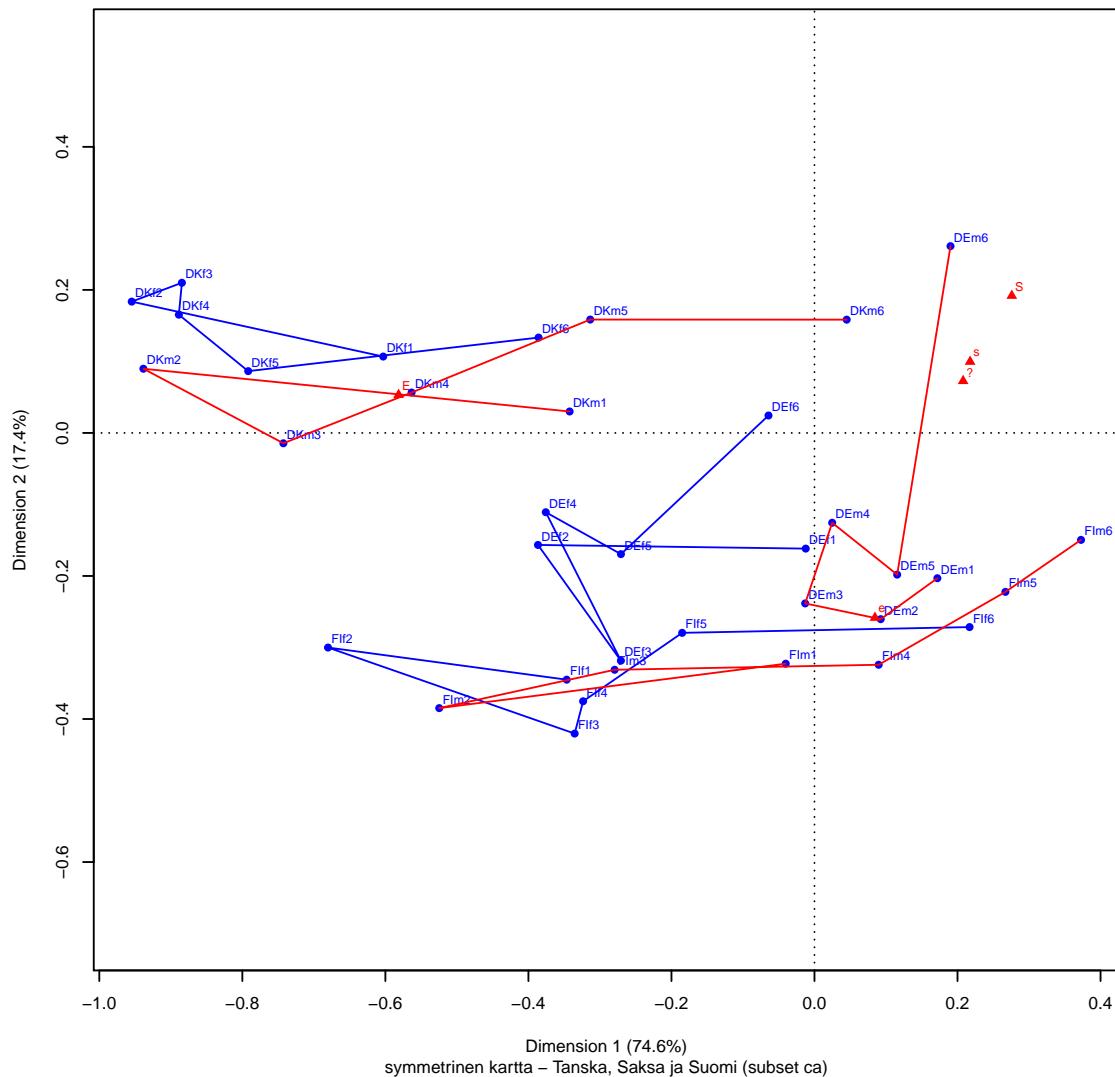


Kuva 6.1: Ikä, sukupuoli ja maa:Tanska-Saksa-Suomi



Kuva 6.2: Ikä, sukupuoli ja maa:Tanska-Saksa-Suomi

Karttoja 6.1 ja 6.2 joutuu katsomaan aika tarkkaan ennnekkuin uskoo, että akseleiden skaalaus on akateeminen pulmma vailla käytännön merkitystä (kts. jakso 3.3.4). Dataa analysoidaan graafisesti, ja kuvat näyttävät erilaisilta. Pääakselien inertioiden neliöjuuret ovat 0.327 ja 0.158, sarakkeet voisi siis tulkita myös kontribuutio-kuvista.



Kuva 6.3: Ikä, sukupuoli ja maa:Tanska-Saksa-Suomi

#### k Rivipisteiden tulkinta

```
##
## Principal inertias (eigenvalues):
##
##   dim   value      % cum%  scree plot
## 1    0.107090 74.6 74.6 ****
## 2    0.024985 17.4 92.0 ****
## 3    0.006594  4.6 96.6 *
## 4    0.004882  3.4 100.0 *
##
```

```

## Total: 0.143551 100.0
##
##
## Rows:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | DEf1   13  467     5  -12   3   0  -162 464 13 |
## 2 | DEf2   15  930    19 -387 799  21 -157 131 14 |
## 3 | DEf3   19  919    25 -271 385  13 -318 533 76 |
## 4 | DEf4   23  993    25 -376 913  30 -111 80  11 |
## 5 | DEf5   17  893    13 -271 641  11 -169 252 19 |
## 6 | DEf6   23    48    15  -64   42   1   24   6   1 |
## 7 | DEM1   13  827     8 172 345   3 -203 482 21 |
## 8 | DEM2   13  855     8   93  96   1 -260 759 34 |
## 9 | DEM3   15  874     7  -13   3   0 -238 871 34 |
## 10 | DEM4   21  285     8   25  11   0 -126 274 13 |
## 11 | DEM5   19  684    10 116 174   2 -198 510 30 |
## 12 | DEM6   22  750    22 190 260   8 261 490 61 |
## 13 | DKf1   10  979    27 -603 949  35 107 30  5 |
## 14 | DKf2   14  996    89 -955 960 115 184 36 18 |
## 15 | DKf3   17  985    98 -885 933 122 210 53 29 |
## 16 | DKf4   18  983   104 -889 950 132 165 33 20 |
## 17 | DKf5   16 1000    69 -792 988  92  86 12  5 |
## 18 | DKf6   12  834    17 -386 745  17 133 89  9 |
## 19 | DKm1   15  978    13 -342 971  17  30  7  1 |
## 20 | DKm2   13  997    79 -938 988 104  90  9  4 |
## 21 | DKm3   13  989    52 -743 989  69 -14  0  0 |
## 22 | DKm4   15  962    36 -563 952  45  57 10  2 |
## 23 | DKm5   13  682    16 -314 543  12 159 139 13 |
## 24 | DKm6   15  291     9   45  22   0 158 269 15 |
## 25 | FIf1   12  951    20 -346 478  13 -345 474 55 |
## 26 | FIf2   12  941    48 -680 788  50 -300 153 42 |
## 27 | FIf3   12  952    24 -335 370  12 -420 582 82 |
## 28 | FIf4   14  999    25 -323 426  14 -375 573 82 |
## 29 | FIf5   17  982    14 -185 299   6 -280 683 55 |
## 30 | FIf6   11  704    13  217 274   5 -271 430 33 |
## 31 | FIm1    7  624     8  -40  10   0 -323 614 30 |
## 32 | FIm2    9  984    26 -525 640  22 -385 344 52 |
## 33 | FIm3    9  990    12 -279 412   6 -331 578 38 |
## 34 | FIm4   13  944    11   90  67   1 -324 877 54 |
## 35 | FIm5   12  722    14  267 426   8 -222 295 23 |
## 36 | FIm6    9  911    11  373 785  12 -150 126  8 |

##
## Columns:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | S     99  731 107  276 493  71  192 238 147 |
## 2 | s     238 832 114  218 688 105 100 144  94 |
## 3 |       168 647  88  208 576  68  73  70  35 |
## 4 | e     261 992 135   85  96  17 -258 896 697 |
## 5 | E     234 1000 556 -582 992 739   53   8  27 |

```

## Tulkintaa

Kolmen maan osajoukon ratkaisussa 2. dimensiolla (maltillinen liberaali?) on inertiaasta 17 prosenttia, edellä ollut paljon yksilottaisempia ratkaisuja. Huono kvaliteetti on DEf1 (467) ja DEf6 (48!), DEM4 (285). Tanskan havainnoista vanhimmat miehet (DKm6,291) ovat kaikkein huonoimmin esitettyjä ratkaisussa, ja hieman nuoremmatkin (DKm5, 682). Suomen aineistossa vain nuoret miehet (FIm1, 624) on esitetty kartalla huonosti.

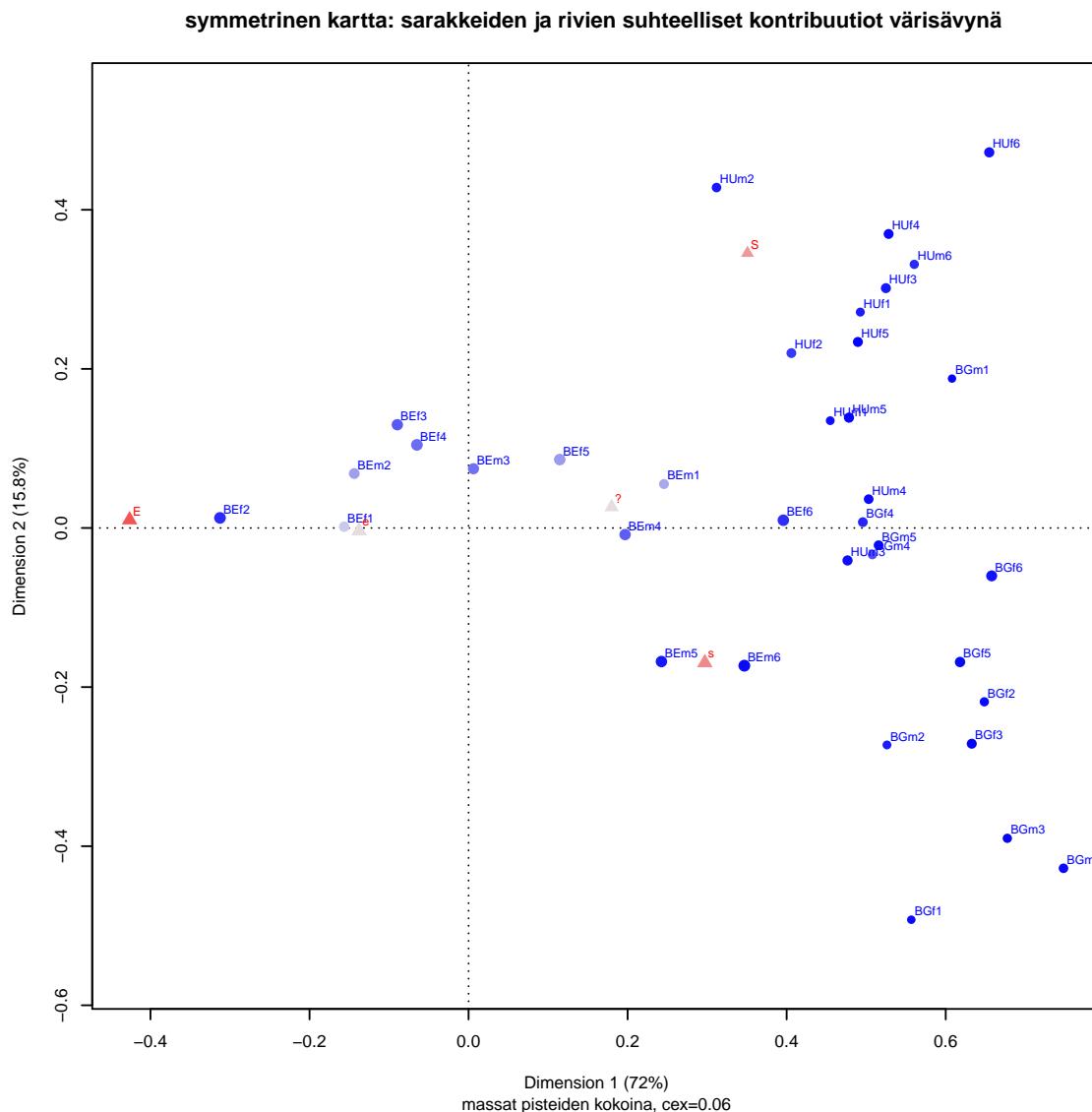
Kaksi dimensiot selittävät osajoukon kokonaishajonnasta 92%, mutta muutaman ryhmän hajonta on muissa dimensioissa. Saksan naisten iäkkään ikäluokka (DEff6) ja keski-ikäisen miehet (DEm4) vain näyttävät olevan lähekkääin origon tuntumassa, samoin muutama muu huonosti tasoon sijoitettu piste.

Huonosti kuvatuista pisteistä ei kuva ei siis kerro oikeastaan mitään muuta.

Sarakkeet on esitetty kohtalaisen hyvin, ja symmetrisessä kartassa tärkeimmälle dimensioille projisodut sarakepisteet ovat odotetussa järjestyksessä.

Kontribuutiokartasta nähdään, että tärkein kontrasti on tiukan erimielisyyden (E) ja kaikkien muiden vastausvaihtoehtojen välillä. Epävarmojen tai maltillisten (e) kontrasti hallitsee toista dimensiaa, erityisesti S- ja s-kategoroiden kanssa. Samalla kuvasta näkee (ja numeerisist tuloksista voi vahvistaa), että S-piste on on lähempänä (kulma on pienempi) pystyakselia. Kontribuutio on suurempi (147 vs. 71 x-akselille). Toisaalta x-akseli selittää selvästi suurimman osan kaikkien muiden sarakepisteiden inertiesta, ja y-akseli taas lähes täysin e-pisteen inertian.

Kontribuutiokartalla ei voi tulkita sarakkeiden etäisyyksiä, joten ensin symmetrisinen kartta.



Kuva 6.4: Ikä, sukupuoli ja maa: Belgia-Bulgaria - Unkari 1

Ensimmäisen dimension tulkinta pysyy samana, mutta nyt molemmat erimieliset (E, e) vastauskategoriat ovat selvästi oikealla. Ne ovat lähes x-akselin pääällä, kun ensimmäisen osajoukon kartalla e-sarake oli oikealla ja alhaalla kontrastina S- ja s- vastauksille ja myös neutraalille vaihtoehdolle. Kartan toinen dimensio erottelee nyt tiukasti ja lievemmin samaa mieltä olevat, neutraali vaihtoehto jää väliin.

Belgian nuoremmat ikäluokat ovat liberaalilla puolella, ja kiinnostavasti kaksi vanhinta miesten ryhmää on pystysuunnassa kaikkein maltillisimpia. Bulgarian ja Unkarin pisteen ovat x-akselilla tiukasti konservatiivisa. Vaihtelua on maltillisemman ja jyrkemmän konservatiivisuuden välillä pystysuuntaan. Toisen dimension kontrasti on myös hieman yllättäen Bulgarian nuorimpien naisten (BGf1) Unkarin vanhimpiten naisten (HUF6) välillä.

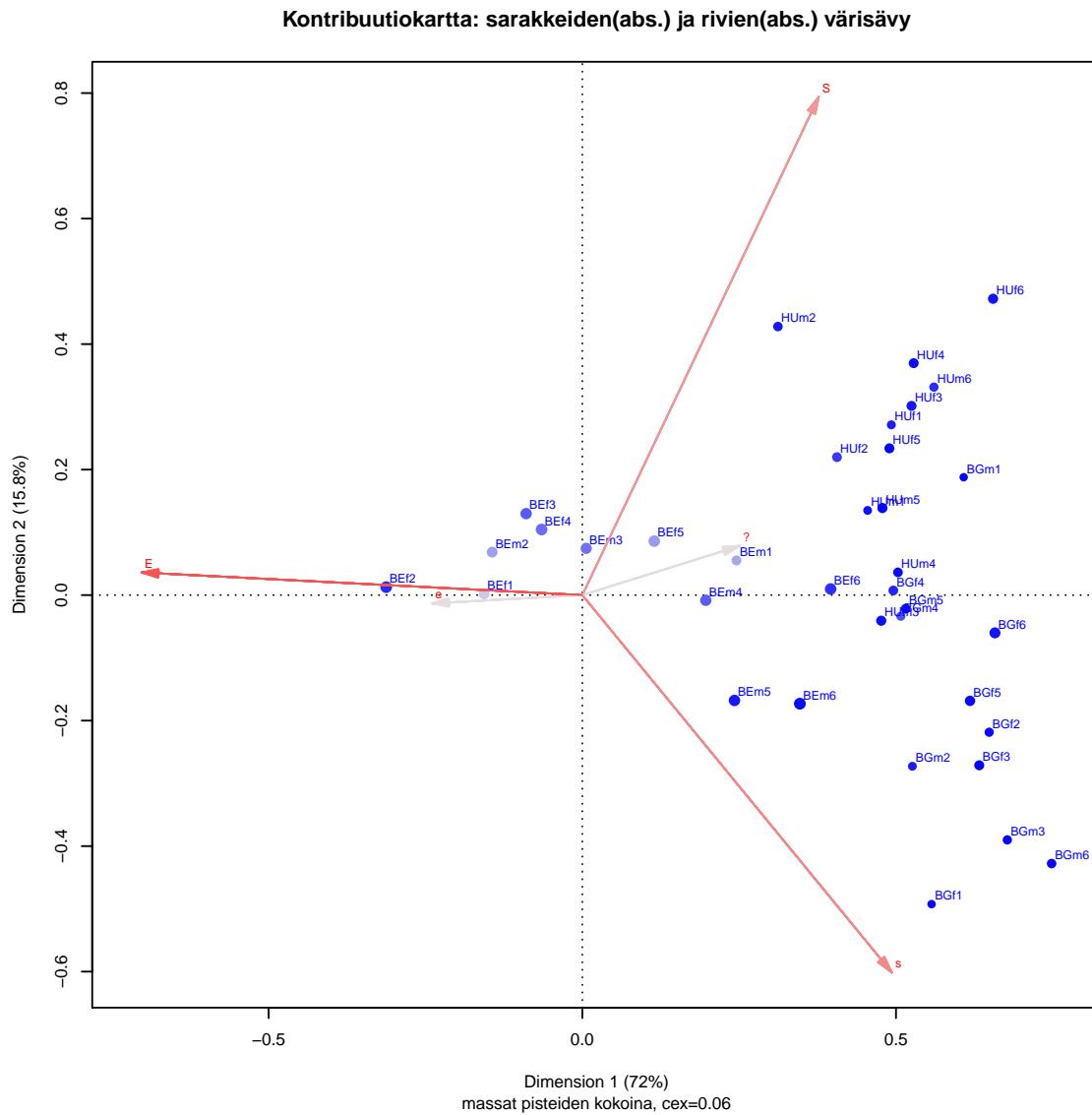
Kuvan 6.3 tapaan ei Bulgarian ja Unkarin ikäluokkia kannata yhdistää. Järjestys toki löytyy, mutta ei ollenkaan niin selkeä. Saksan naisten ikäluokkakuva alkaa erkaantua hieman Suomen ja Taskan hyvin samanalaistista kuvioista. Saksan miehillä on jo eroja paljon toisen dimension suuntaan, Unkarin ja Bulgarian osajoukkojen erot ovat lähes pelkästään pystysuoria.

Suhteellinen kontribuutio eli pisteen laatu (numeerisissa tuloksissa "cor") on esitetty värisävynä. Sarakkeista e ja "?" on esitetty huonosti, riveistä Belgian nuorimmat miehet ja naiset.

Kontribuutiokartta ??fig:maagaCA2sub3map2) eroaa kartasta ??fig:maagaCA2sub2map3) kolmen akselin (E, S ja s) tasapainosemmassa vaikutuksella ratkaisuun. Konservatiiviset sarakepisteet ovat vaikuttavampia kuin E, maltillinen liberaali (s) ja neutraali vaihtoehto vaikuttavat vähemmän.

**k** Yksityiskohdat ovat kiinnostavia, mutta graafisen analyysin päätavoite on yleiskaava. Tässä ekspolatiivinen data-analyysi kuitenkin kulkee hieman eri polkuja kuin tavaniomainen tilastollisten mallien analyysi. Seurataan minne data kuljettaa, ettiään uusia näkökulmia. Iän ja sukupuolen yhteys vastauksiin on rakenteeltaan erilainen, se ei ole ongelma vaan datan ominaisuus.

```
##
## Principal inertias (eigenvalues):
##
##   dim      value      %    cum%    scree plot
## 1    0.086111  72.0  72.0 ****
## 2    0.018841  15.8  87.8 ***
## 3    0.011172   9.3  97.1 **
## 4    0.003477   2.9 100.0 *
##
## Total: 0.119602 100.0
##
##
## Rows:
##
##   name    mass   qlt   inr   k=1 cor ctr   k=2 cor ctr
## 1 | BEf1 | 14 152   19 | -156 152   4 | 2   0   0 |
## 2 | BEf2 | 24 826   24 | -313 824   28 | 13  1   0 |
## 3 | BEf3 | 21 623    7 | -90 201   2 | 130 422  19 |
## 4 | BEf4 | 24 556    6 | -65 155   1 | 105 401  14 |
## 5 | BEf5 | 23 355   11 | 115 227   3 | 86 128   9 |
## 6 | BEf6 | 23 810   37 | 396 810   41 | 10  0   0 |
## 7 | BEm1 | 11 288   21 | 246 274   8 | 55 14   2 |
## 8 | BEm2 | 17 333   11 | -144 271   4 | 68 61   4 |
## 9 | BEm3 | 20 531    2 | 6   4   0 | 75 528   6 |
## 10 | BEm4 | 22 620   11 | 197 618   10 | -8  1   0 |
## 11 | BEm5 | 22 917   18 | 243 620   15 | -168 297  33 |
## 12 | BEm6 | 26 977   33 | 347 782   36 | -173 195  41 |
## 13 | BGf1 | 5 979   23 | 557 549   18 | -492 430  63 |
## 14 | BGf2 | 8 974   32 | 649 875   38 | -219 99   20 |
## 15 | BGf3 | 12 1000  46 | 633 844   54 | -271 155  45 |
## 16 | BGf4 | 10 847   25 | 496 847   30 | 7   0   0 |
```



Kuva 6.5: Ikä, sukupuoli ja maa: Belgia-Bulgaria - Unkari 2

```

## 17 | BGf5 | 14 961 50 | 618 894 62 | -168 66 21 |
## 18 | BGf6 | 18 939 71 | 658 931 92 | -60 8 4 |
## 19 | BGm1 | 5 999 15 | 608 912 19 | 188 87 9 |
## 20 | BGm2 | 6 892 21 | 526 703 20 | -273 189 25 |
## 21 | BGm3 | 8 994 41 | 677 746 43 | -390 247 64 |
## 22 | BGm4 | 8 669 25 | 508 666 23 | -34 3 0 |
## 23 | BGm5 | 10 949 24 | 516 947 32 | -22 2 0 |
## 24 | BGm6 | 9 978 58 | 748 737 60 | -428 241 89 |
## 25 | HUf1 | 7 888 20 | 493 681 19 | 271 207 26 |
## 26 | HUf2 | 11 762 25 | 406 589 20 | 220 173 27 |
## 27 | HUf3 | 12 916 39 | 525 688 37 | 301 227 56 |
## 28 | HUf4 | 11 970 40 | 528 651 36 | 370 319 81 |
## 29 | HUf5 | 12 985 29 | 490 802 32 | 234 183 34 |
## 30 | HUf6 | 13 933 75 | 655 614 64 | 472 319 151 |
## 31 | HUm1 | 6 948 12 | 455 871 14 | 135 77 6 |
## 32 | HUm2 | 9 902 24 | 312 313 10 | 428 589 90 |
## 33 | HUm3 | 13 945 26 | 477 938 33 | -41 7 1 |
## 34 | HUm4 | 10 965 22 | 503 960 29 | 36 5 1 |
## 35 | HUm5 | 13 993 26 | 478 916 33 | 139 77 13 |
## 36 | HUm6 | 8 839 33 | 560 622 29 | 331 217 46 |

##
## Columns:
##      name   mass   qlt   inr   k=1 cor ctr     k=2 cor ctr
## 1 |    S | 99 944 214 | 351 479 142 | 346 465 630 |
## 2 |    s | 238 942 247 | 297 711 244 | -169 231 362 |
## 3 |    | 168 435 107 | 180 426 63 | 26 9 6 |
## 4 |    e | 261 640 65 | -138 639 57 | -4 0 0 |
## 5 |    E | 234 966 368 | -426 965 494 | 10 1 1 |

```

Kahden osajoukon inertioiden summa on sama kuin koko aineiston ( $0.144 + 0.12 = 0.263$ ), Selitysasteet nousevat hieman, mutta vielä oleellisempaa on kahden aika erilaisen ryhmän havaitseminen.

Belgian jokun pisteet: huono kvaliteetti (BEf1, BEf5,BEm1, BEm2). Bulgaria ja Unkari hyvin esitetty. Belgia on pulmallinen tapaus, ehkä taas omissa dimensioissaan.



## Luku 7

# Monimuuttuja-korrespondenssianalyysi (MCA)

**k** MG:n jäsentely CA:n tutkimusasetelmista (within - between), kokoaan myös edeltävät analyysit “samana katon alle”.

**k** Idea: taulukoiden yhdistely “supertaulukoksi” tai matriisiksi, analysoidaan kahden muuttujan ristiintaulukoinnin laajennuksena useita kahden muuttujan ristiintaulukointeja.

**k** MCA stacked&concat - analyysin erikoistapauksena.

**k** CA SVD:tä soveltavana ratkaisualgoritmmina, jolle “syötetään” sopiva matriisi. Benzecri: löydettävä vain matriisi joka dekomponoidaan.

**k1** kahden muuttujuajoukon väliset yhteydet

**k2** yhden muuttujuajoukon muuttujien väliset yhteydet

**k3** wrt samples ja wrt variables

**k4** Yksinkertaisesta monimutkaisempaan 0. Täydentävät pistet tulkinnan apuna, yleistyökaluna (“jack of all trades but master of none”) esitelly ensimmäisenä. Tärkeä lisä kaikissa analyyseissä.

1. yhteisvaikutusmuuttujat **K4-1** -max kolmen muuttujan yhtesivaikutusmuuttuja - taulukko harsoontuu

2. Yksi “selittävä” muuttuja, useampi “selittävä” **K4-2**”

- pinotut taulukot (käytännössä matriisit, helpompi pinota R:ssä, Burtin taulun käyttö)

3. Taulukoiden yhdistäminen yleisemmin **K4-3**

- monta “selittää” ja “selittää”: taulukoita pääallekkäin ja rinnakkain, lohkomatriisi
- ABBA: voidaan esim. jakaa taulukko (viite) osatauluiksi
- tästä hieman tarkemmin teorialitteessä; symmetriset taulut (esim. vastaukset eri vuosina, isän ja pojan ammatti), matched matrices

4. MCA **K4-4**

### K-n

tulkinnan tärkeimmät periaatteet

- kaikki erot suhteellisia eroja
- ca-ratkaisu approksimaatio; voi olla hyvä, mutta myös huono -CA on hajonnan (intertian) dekomponoinnin menetelmä.
- Toiset (esim. MG, "ranskalaiset") painottavat graafista menetelmää ja kuvien oikeaa tulkintaa), voi ymärtää myös moniulotteisenä skaalausena (MDS: etsi skaalaus, joka maksimoi korrelaation) tai moniuotteisena varianssianalyysinä (ANOVA)
- tulkinnan tarkennus molemmissa alaluvuissa

\*\*edit\* Tässä jaksossa korostuu data-analyykin eksplotatiivinen luonne, datan kaivelu ja näkökulmien muuntelu. Missä dimensiossa yhteydet ja riippuvuudet piileksivät? Kartta on tässä mielessä kartta (vaikka hieman outo), osoittaa yhdessä numeeristen tulosten kanssa suuntaa.

## 7.1 Pinotut ja yhdistetyt taulukot (stacked and concatenated tables)

Pinotut taulut: perusidea

```
knitr::include_graphics('img/stacked1.png')
```

kaksi selittävää - kolme selittäjää		
	Q1b: S,s, ?, e, E	Q1c: S,s, ?, e, E
maa		
kä-sukupuoli (ga)		
koulutustaso (edu)		

Kuva 7.1: Pinotut ja yhdistetyt taulut - periaate

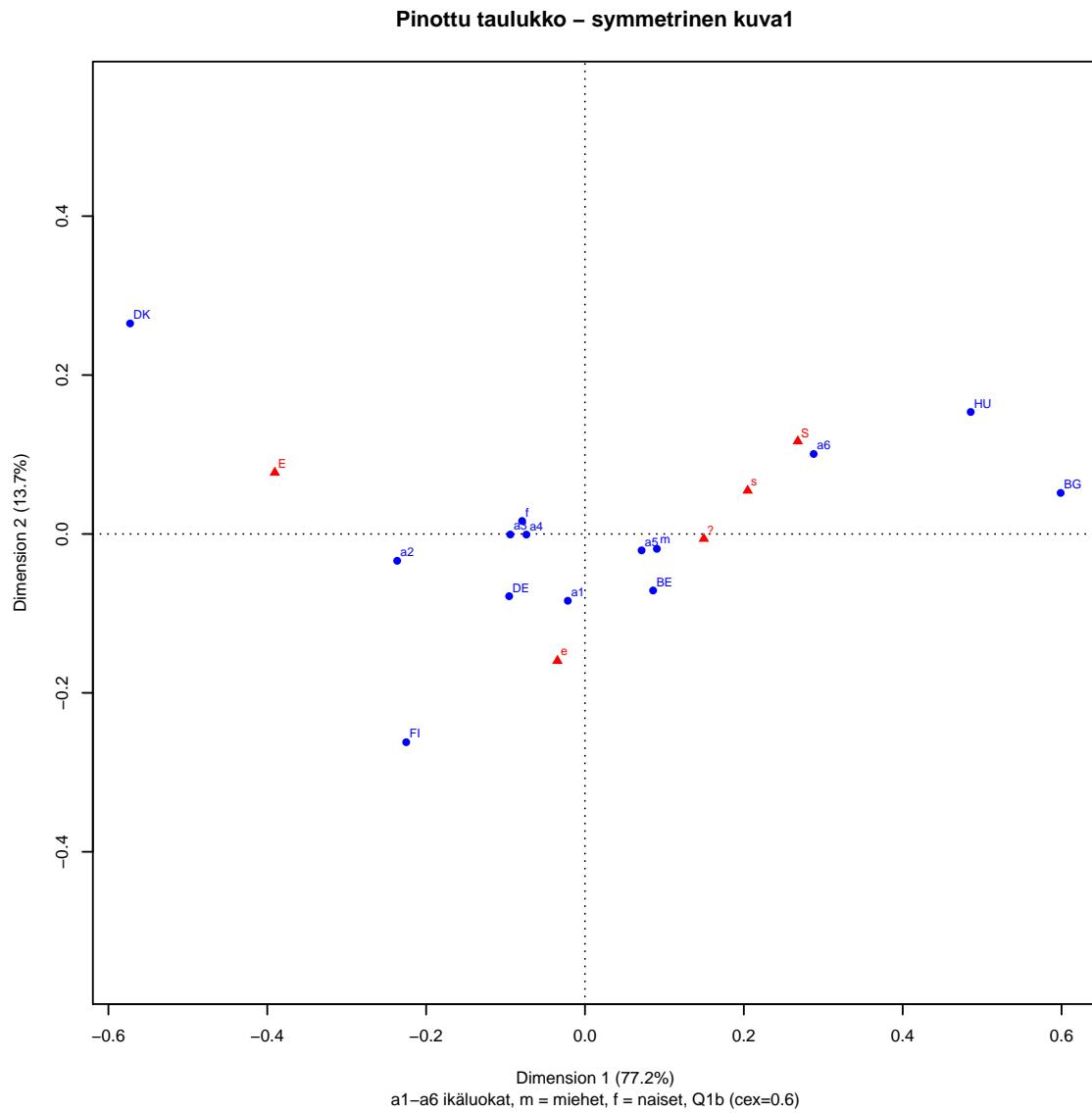
Lisättiin kuva

Kun yhdistetty taulu rakennetaan matriiseista, karttojen muuttujanimiä joudutaan siistimään. Kuvat voivat myös herkästi käentyä akselien ympäri, ne kannattaa kääntää vertailun helpottamiseksi samanlaisiksi kuin muut.

### Kartan tulkinta; miten eroaa yhteisvaikutusmuuttujan analyysistä?

Perustulkinta akseleille ei muudu, eikä maapisteiden sijoittuminenkaan.

Mikä on maapisteiden ja kahden selittävän (eksogenisen) muuttujan pisteiden yhteys sarakepisteisiin?



Kuva 7.2: Q1b: Lapsi kärsii jos äiti käy työssä

Koko aineiston kartassa ikäluokkapisteet ja sukupuolipisteet ovat pakkautuneet maapisteitä tiiviimmin origon ympärille. Ikäluokkapisteiden (koko aineiston keskiarvot) selvä kontrasti on vanhimman (a6) ja toiseksi nuorimman välillä 1. dimenision suuntaan.

Ikäluokkapisteet ovat koko aineiston keskiarvopisteitä, niiden sijantia voi tulkita pistejoukko kerrallaan kuuden maapisteidenkin. Mitään yhteisvaikutuksia ei analysoida eksplisiittisesti. Karttaa voi verrata sukupuoli-ikäluokka yhteisvaikutusmuuttujan analyysin aiemmin. Naispiste on tiukassa nipussa ikäluokkien a3 ja a4 kanssa aivan origon vasemmalla puolella. Miesten keskiarvopiste on hieman origosta oikealle, yhdessä ikäluokan a5 kanssa.

Taustamuuttujat: numeeristen tulosten tarkastelua

Lisäpisteet on hyvin esitetty, niiden etäisyysistä voi luotettavasti arvioida kuvasta. Poikkeus on nuorin ikäluokka (a1, qlt = 501). Inertian osuudet (inr) ovat yhtä vaativammia kuin Belgian (28) ja Saksan (29), (m = 20, f = 17, a2 = 40, a6 = 83), samoin kontribuutiot akseleiden inertiaan. 1. dimension kontribuutio (ctr) on suuri (>800) kaikilla paitsi nuorimmalla ikäryhmällä (a1) jolla 2. dimension selittää lähes puolet sen inertiesta (470).

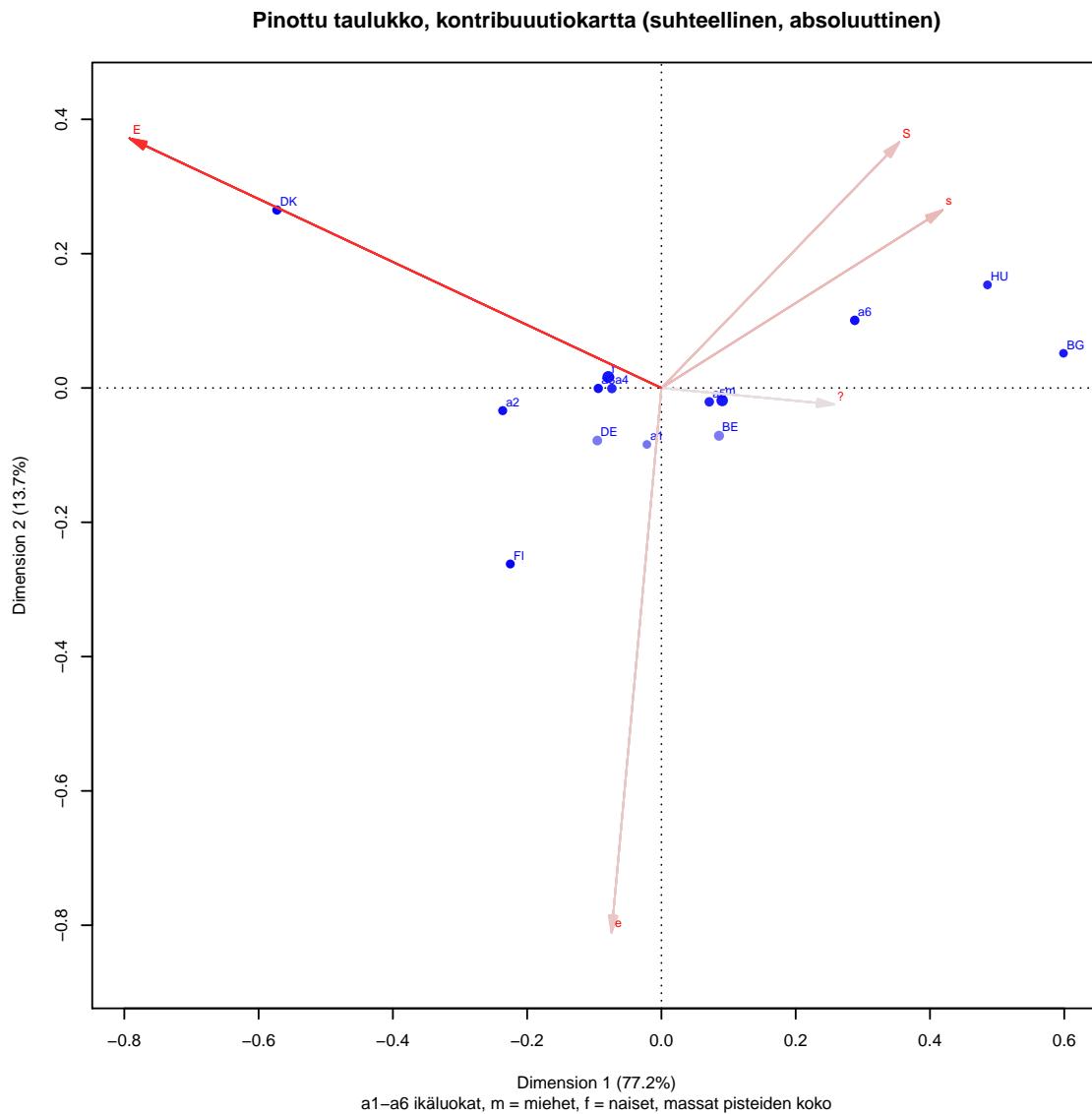
Vilkaistaan taas numeerisia tuloksia, varmistetaan tulkinta. Tämän voi poistaa lopullisesta versiosta.

```
summary(Concat1jh.CA1)
```

```
##
## Principal inertias (eigenvalues):
##
##   dim    value      %   cum%   scree plot
##  1     0.056877  77.2  77.2 ****
##  2     0.010116  13.7  91.0 ***
##  3     0.003923   5.3  96.3 *
##  4     0.002711   3.7 100.0 *
##   -----
## Total: 0.073628 100.0
##
##
##
## Rows:
##   name   mass  qlt  inr   k=1 cor ctr   k=2 cor ctr
##  1 | BE    82   498   28 |  86 295 11 | -71 203 41 |
##  2 | BG    38   907   204 | 599 901 238 |  52  7 10 |
##  3 | DE    70   498   29 | -95 298 11 | -78 200 43 |
##  4 | DK    57   990   310 | -573 816 328 | 265 175 394 |
##  5 | FI    45   987   75 | -225 419 40 | -262 568 309 |
##  6 | HU    41   856   168 | 486 778 169 | 153 78 95 |
##  7 | m    156   910   20 |  91 873 22 | -19 37 5 |
##  8 | f    178   910   17 | -79 873 20 |  16 37 5 |
##  9 | a1   39   501    8 | -22 31 0 | -84 470 27 |
## 10 | a2   50   958   40 | -236 939 49 | -34 19 6 |
## 11 | a3   56   958    7 | -94 958 9 | -1 0 0 |
## 12 | a4   63   841    6 | -74 841 6 | -1 0 0 |
## 13 | a5   62   868    5 |  71 801 6 | -21 67 3 |
## 14 | a6   63   957   83 | 288 852 92 | 101 104 63 |
##
## Columns:
##   name   mass  qlt  inr   k=1 cor ctr   k=2 cor ctr
##  1 | S    99   786   147 | 268 661 126 | 117 125 134 |
##  2 | s    238  843   172 | 205 787 175 |  55  56 70 |
##  3 |       168  640   80 | 150 639 66 | -6  1 1 |
##  4 | e    261  970   97 | -35 44 6 | -160 926 657 |
##  5 | E    234 1000   504 | -390 962 628 |  77  38 138 |
```

# 14 riviä, inertiaakontribuution keskiarvo  
# 1000/14 = 71

**edit** Galkussa myös subsetCA-kartta, josta Unkari ja Bulgaria on jätetty pois. Ei mitään oleellista lisätietoa, jotken ei esitetä tässä. Ensimmäisen dimension osuus interitasta laskee ja toisen kasvaa. Kartan selkeys ei parane.



Kuva 7.3: Q1b: Lapsi kärsii jos äiti käy työssä

Sarekkeista E-sarake (täysin eri mieltä) määrittää akseleita vahvasti, kontrastina kaksi konservatiivista vastausta (S ja s) ja myös neutraali vaihtoehto (e). Numeerista tuloksista nähdään, että ikäluokat vaikuttavat juuri ensimmäiseen tärkeimpään dimensioon.

Belgian ja Saksan pisteet on esitetty kartassa huonosti, samoin nuorin ikäluokka. Muiden pisteiden sijaintia voidaan arvioida myös sarakkeiden ja rivipisteiden välillä. Ikäluokkien kontrasti on selvä toiseksi nuorimman (a2) ja vanhimman (a6) välillä.

**edit** Ikäluokkipisteet asymmetrisessä kontribuutiokartassa sarakkeiden keskiarvona, samoin maapisteet. Esimerkiksi vanhimmassa ikäluokassa on suhteellisesti paljon vastauksia samaa mieltä olevia ja neutraaleja ja vähän erimielisiä.

Hyvin yksinkertainen esimerkki.

**edit** Mutta tässä esimerkkiaineisto, jossa ei puuttuvia tietoja. Ne olisivatkin aika pulmallisia, varianssin dekomponointi ei onnistu jos reunajakaumat ovat alitaluiissa erilaisia.

**Matched matrices edit** Ei omaa lukua, vaan pari kappaletta matriisien pinoamisesta ja yhdistelystä.

**k** lyhyesti matriisien yhdistelystä hyvin yleisenä analyysin menetelmänä. Voidaan tarkastella luonteeltaan erilaisten muuttujaryhmien tai muuttujien yhteyksiä kuten edellisessä esimerkissä.

**kIdea:** matriisien yhdistämisellä saadaan ote monenlaiseen tutkimusongelmaan. Benzecri: data-analyysissä on vain löydettävä oikea matriisi joka diagonalisoidaan.

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). (?)

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

Edellisen menetelmän variantti, jossa ryhmien väliset ja sisäiset erot saadaan esiin. Inertian jakaminen.

Samanlaisten rivien ja sarakkeiden kaksi samankokoista taulua, esimerkiksi sukupuolivaikutusten arvointi. Al-kuperäinen taulukko jaetaan kahdeksi taulukiksi sukupuolen mukaan. Matriisien yhdistäminen (concatenation) riveittäin tai sarakkeittain ei näytä optimaalisesti mm - matriisien eroja.

Luokittelut 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”.

**ABBA** “We consider the joint analysis of two matched matrices which have common rows and columns, for example multivariate data observed at two time points or split according to a dichotomous variable. Methods of interest include principal components analysis for interval-scaled data, correspondence analysis for frequency data, log-ratio analysis of compositional data and linear biplots in general, all of which depend on the singular value decomposition. A simple result in matrix algebra shows that by setting up two matched matrices in a particular block format, matrix sum and difference components can be analysed using a single application of the singular value decomposition algorithm. The methodology is applied to data from the International Social Survey Program comparing male and female attitudes on working wives across eight countries. The resulting biplots optimally display the overall cross-cultural differences as well as the male-female differences. The case of more than two matched matrices is also discussed.”

## 7.2 MCA - monimuuttujakorrespondenssianalyysi

**k** Terminologiasta: monta muuttuja on jo ollut käytössä. MCA on monimuuttujamenetelmä samassa mielessä kuin faktorianalyysi. Analysoidaan usean statukseltaan samanlaisen muuttujan välistä suhteita, ja myös niiden yhteyksiä tutkimusongelman kannalta “eksogenisiin” taustamuuttuihin tai “selittäjiin”. Surveytutkimuksen kyselylomakkeen kysymyspatterit luotaavat tietoa joistain taustalla olevista asenteista.

**edit** yksi kappale, jossa tutkimusasetelmaa verrataan tilastollisten mallien asetelmaan? Jako “selittäjiin” ja selittävään, moniyhtälömallit? Faktorianalyysi tässä selkein vertailukohde

**k** Matemaattisesti kaikki muuttuu paljon mutkikkaammaksi, ja yksinkertaisen perustapauksen selkeät tulkinnot eivät toimi. Tärkeä asia: CA:n skaalausominaisuudet ja visuaalinen tulkinta pätevät edelleen.

**edit** Teorialiitteessä tästä enemmän, ranskalaiset hieman eri mieltä **Viitteitä**

MG ja MG&Prado ovat tutkineet ISSP-datalla tätä ongelmaa. Biplots In practice - kirja (?), ss.142- “dimensions of middleness”. Mg ja Prado artikkelissaan (?) ja samasta teemasta laajentaen kokoomateoksessa (?), ss. 197-217. Aineistona jälkimmäisessä tutkimuksessa on sama kysymyssarja kuin tässä 1994 datasta. Kysymyksissä on jonkin verran eroja. Ensimmäisessä artikkelissa he komentoivat vertaisarvioitsijoiden ehdotuksia. Kuvia voisi

selkiyttää esittämällä vain osan pisteistä sarjassa karttoja, mutta ratkaisu perustuisi siinäkin koko dataan ja sitä nimenomaan ei haluta. Toinen ehdotus on yhdistää ne vastausvaihtoehdot jotka eivät ole tutkimuksen kohteena yhdeksi yhdistetyksi kategoriaksi. Tällöin osajoukon korrespondenssianalyysin kätevä inertia dekomponointi osajoukille ei toimisi. Ratkaisu ei toimi ollenkaan yleisimmissä osajoukoissa.

#### **k Data**

**k** Taustamuuttujien taulukoissa on yllättävän isoja eroja, jotkut taulukoiden luokat ovat nollia tai hyvin vähän havaintoja. Luokka pitäisi yhdistellä, jo pelkästään ”kuvaroskan” takia. Ei tehdä.

**k** Puuttuvat havainnot, muutama numero vain mitä “listwise delete” saa aikaan.

#### *#Puuttuvien tietojen yleiskuva*

*# Puuttuvat tiedot aineistossa – viite datan dokumentointiin jossa taulukot.*

*# Vaihtelee maittain ja muuttujittain, paljon.*

*# Koko data (G1\_1\_data2.Rmd – skriptissä valitut muuttujat ja 25 maata)*

*#*

*#sum(!complete.cases(ISSP2012jh1d.dat)) = 9455*

*#dim(ISSP2012jh1d.dat) = 32823*

*#9455/32823 = 0.2880602*

#### *# Puuttuvat tiedot valitussa MCA-aineistossa*

*#missingMCAvars1 <- c("Q1a", "Q1b", "Q1c", "Q1d", "Q1e", "Q2a", "Q2b", "edu",*

*#                      "sosta", "urbru", "maa", "ika", "sp" )*

*#missingTestMCA1.dat <- ISSP2012jh1d.dat %>% select(all\_of(missingMCAvars1))*

*#sum(!complete.cases(missingTestMCA1.dat)) = 6101*

*#dim(missingTestMCA1.dat) = 32823*

*#6101/32823 = 0.1858758 Puuttelisten havaintojen osuus.*

Koko tähän tutkimukseen valitussa aineistossa (25 maata ja muuttujat, poistettu havainnot joissa ikä tai sukupuoli puuttuu) 71% havainnoista on kaikki tiedot.

MCA-analyyseihin valitun  $7 + 3 = 10$  muuttujan aineiston havainnoista 81% on vailla puuttuvia tietoja. Jos puuttuvat tiedot poistetaan (ns. “listwise delete”, poistetaan jos yksi tai useampi tieto puuttuu) viidesosa datasta jää pois.

**k** lyhyt kappale - puuttuvien tietojen käsittely on laaja aihe. Otantamenetelmissä ja survey-aineistojen (kyselytutkimusten) oppaissa käsitelty monipuolisesti.

**k** Miten liittyy tähän? CA on koko aineiston kuvalevaa analyysiä, ei päättelyä otoksesta perusjoukon tasolle. Miten puuttuvia vastauksia voisi kuvalla? Toisaalta edellisen jakson menetelmien soveltaminen on mahdotonta, jos jotain ei tehdä (reunajakaumat oltava samoja).

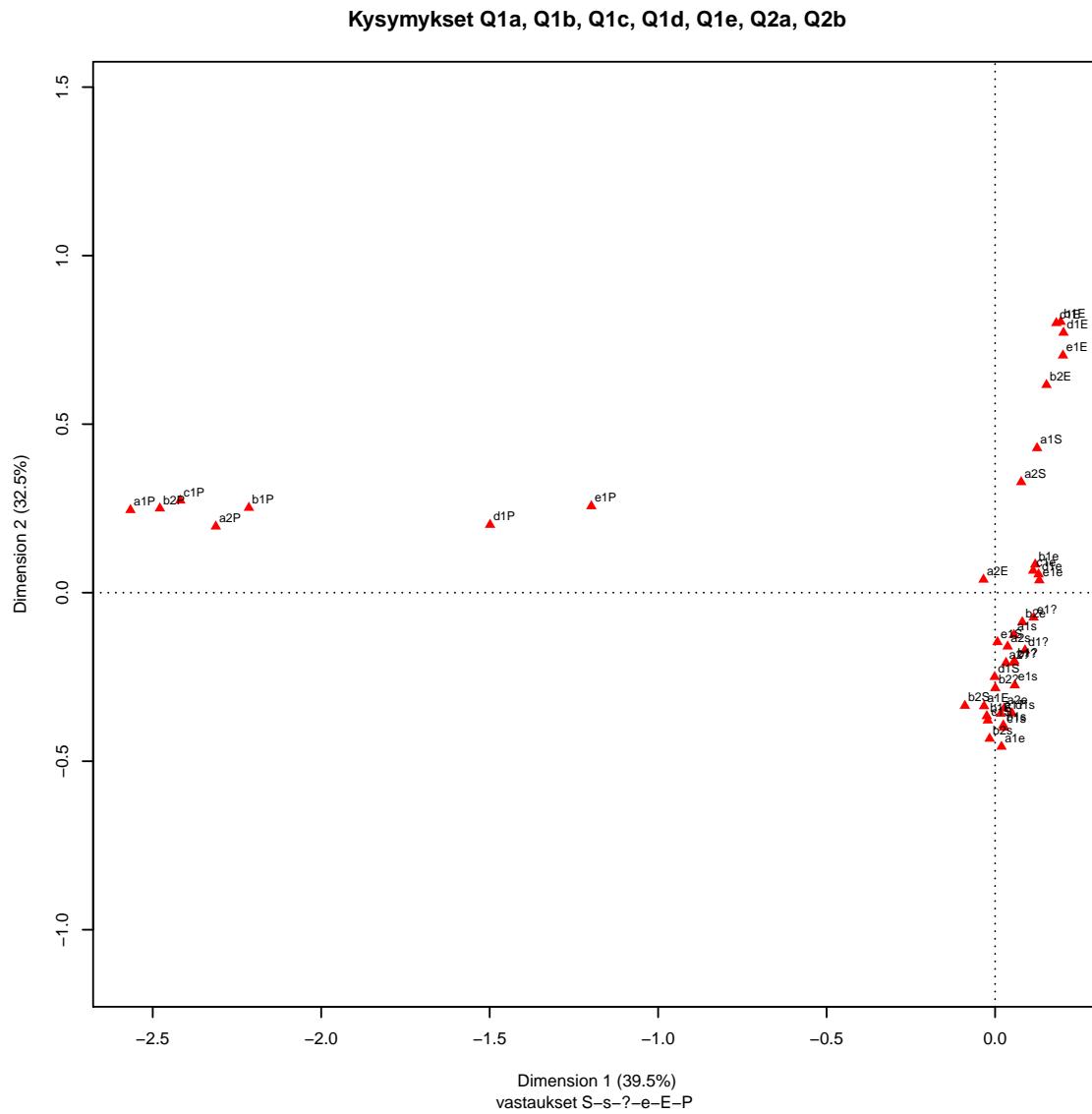
MCA1 - seitsemän kysymystä (jokaisessa viisi vaihtoehtoa)

Aineistossa on 32 823 havaintoa ja seitsemän muuttuja.

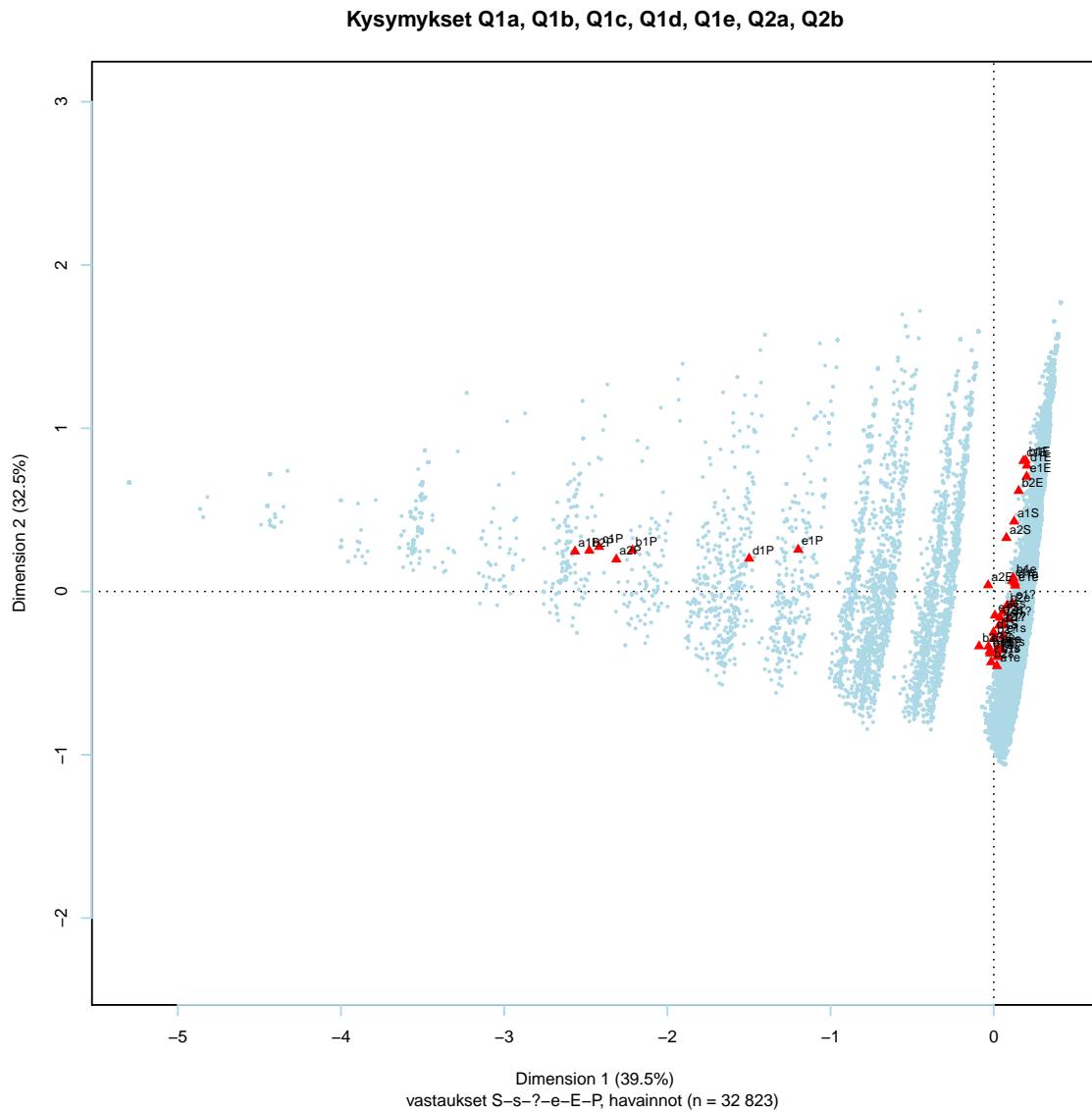
**k** Perustulkinta: Inertian selitysosuudet ovat paljon pienempiä, ja ratkaisu on selvästi kaksiluottein. Puuttuvat vastaukset erottuvat omana ryhmänä, ja varsinaiset vastaukset ovat pakkautuneet y-akselin oikealle puolelle. Niiden erot näkyvät vain toisessa dimensiossa. Ensimmäinen dimensio kuvailee vastaamattomuutta (syystä tai toisesta) vs. kaikkia vastauksia.

Pystyakselin suuntaan kontrasti näyttäisi olevan konservatiiviset ylhäällä, modernit ja liberaalimmat alhaalla. Pisteitä on vaikea erottaa toisistaan.

Karttaa voi parantaa lisäämällä siihen vastaajien ( $n = 32\ 823$ ) pisteen (viite MG - ekologiakirja, jossa vanhempi ISSP-data).



Kuva 7.4: MCA: Seitsemän kysymystä - 25 maata, kartta 1



Kuva 7.5: MCA: Seitsemän kysymystä - 25 maata, kartta 2

Jokainen havainto on sarakevektoreiden keskiarvopiste. Sarakevektoreita ei voi tulkita yhtä selkeästi kuin yksinkertaisessa korrespondenssianalyysissä. Ne eivät edusta kysymystä vaan kysymyksen yhtä vastauskategoriaa, pitkä jono nollia ja ykkösiä. Rivipiste on vastauksiaan vastaavien sarakepisteiden keskiarvopiste. Jos vastaanaja on valinnut vaihtoehdot (a1S, a2s,..., b2?) se on näiden pisteiden keskiarvopiste.

Oleellista MCA-kartan tulkinnassa on kuitenkin yleiskuvan muodostaminen, geometrinen tulkinta on hieman hankala. Skaalausominaisuudet MCA kuitenkin säilyttää (**Viite**, teorialiitteessä hieman tarkemmin).

Myös riviprofilit ovat samanlaisia, jokaisella rivilla on kuudessa kategoriassa (vastaukset ja puuttuva tieto) nollia ja yksi ykkönen. Rivien välistä etäisyyttä määrittelevä ainoastaan "erimielisydyt", ja GDA-kirjan (**viite**) ohjeen mukaan MCA-karttojen tulkinta pitäisi aloittaa yksilöiden pilven ääripäistä.

Pistepilven muoto kertoo , kuinka pienenevä joukko vastaanajia lähestyy kiilana puuttivien tietojen pisteitä. Kaikkiin kysymyksiin vastanneet ovat massana kuvan oikeassa laidassa. Pistepilvet oikealta vasemmalle kuvaavat kuinka moneen kysymykseen on jätetty vastaamatta.

(MG-vastaava kuva (?), luku 14. "dimensions of middleness".) Pieni joukko määräää koko kartan koordinaatiston.

### Osajoukon MCA

**k** Osajoukon MCA ratkaisee ongelman tyylikkäästi.

**k** Tulkinta: kontrasti "ääripäiden" välillä, vahvat mielipiteet (S ja E) hallitsevat vasenta alakulmaa ja oikeaa laitaa x-akselin tuntumassa. Maltilliset vastaukset ja neutraali vaihtoehto ovat ylhällä vasemmalla.

**k** Johtopäätös: seitsemän kysymystä erottlee hyvin vastaanajat liberaali - konservatiivi - aksellilla. Maltilliset ja neutraalit vastaukset sijaitsevat ääripäiden välissä. Karttaan voi hahmotella diagonaalisen akselin vahvojen mielipiteiden ryppäiden välille. Muut vastaukset ovat ns. kaariefektiin mukaisesti näiden välisellä U-muotoisella linjalla. Tämä kuuluusa Guttan efekti kertoo järjestysasteikon muuttujien korrelatiosta, hyvä asia kun kysymyksillä luodataan asenteita naisten työssäkäyntiin. Kaariefekti on myös "geometrinen vältämättömyys", tarkemmin teorialiitteessä.

Kun kartalle lisätään havainnot, nähdään selvästi kuinka suuri hajonta on havaintojen pilvessä verrattuna vastauskategorioiden pilveen.

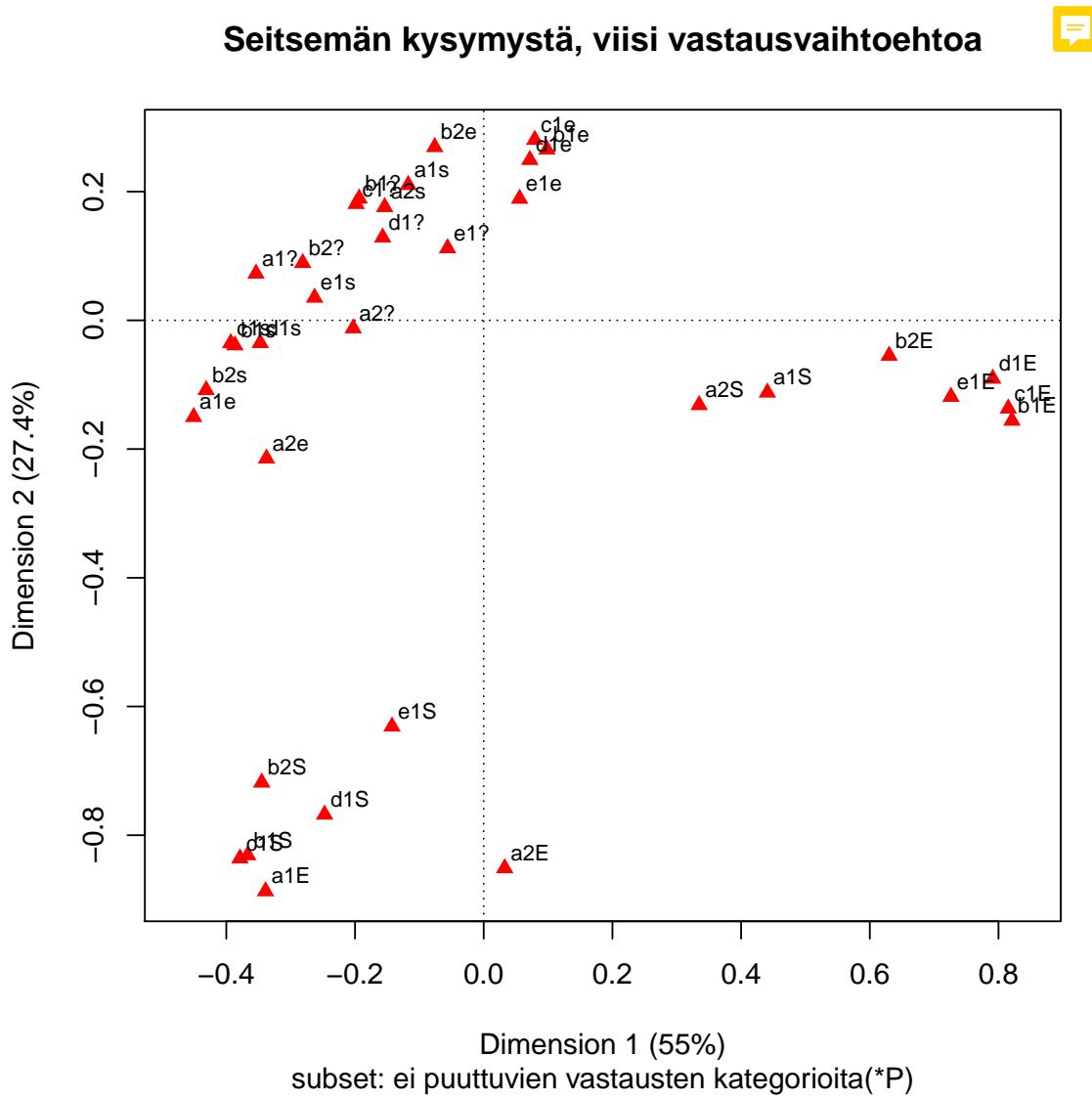
Asymmetrisen kartta näyttäisi olevan paras vaihtoehto, vastausvaihtoehdot erottuvat selvimmin.

**edit1** Jostain syystä en saa toimimaan mjca-funktiossa subsecat - parametrin kanssa täydentäviä muuttuja supcol.

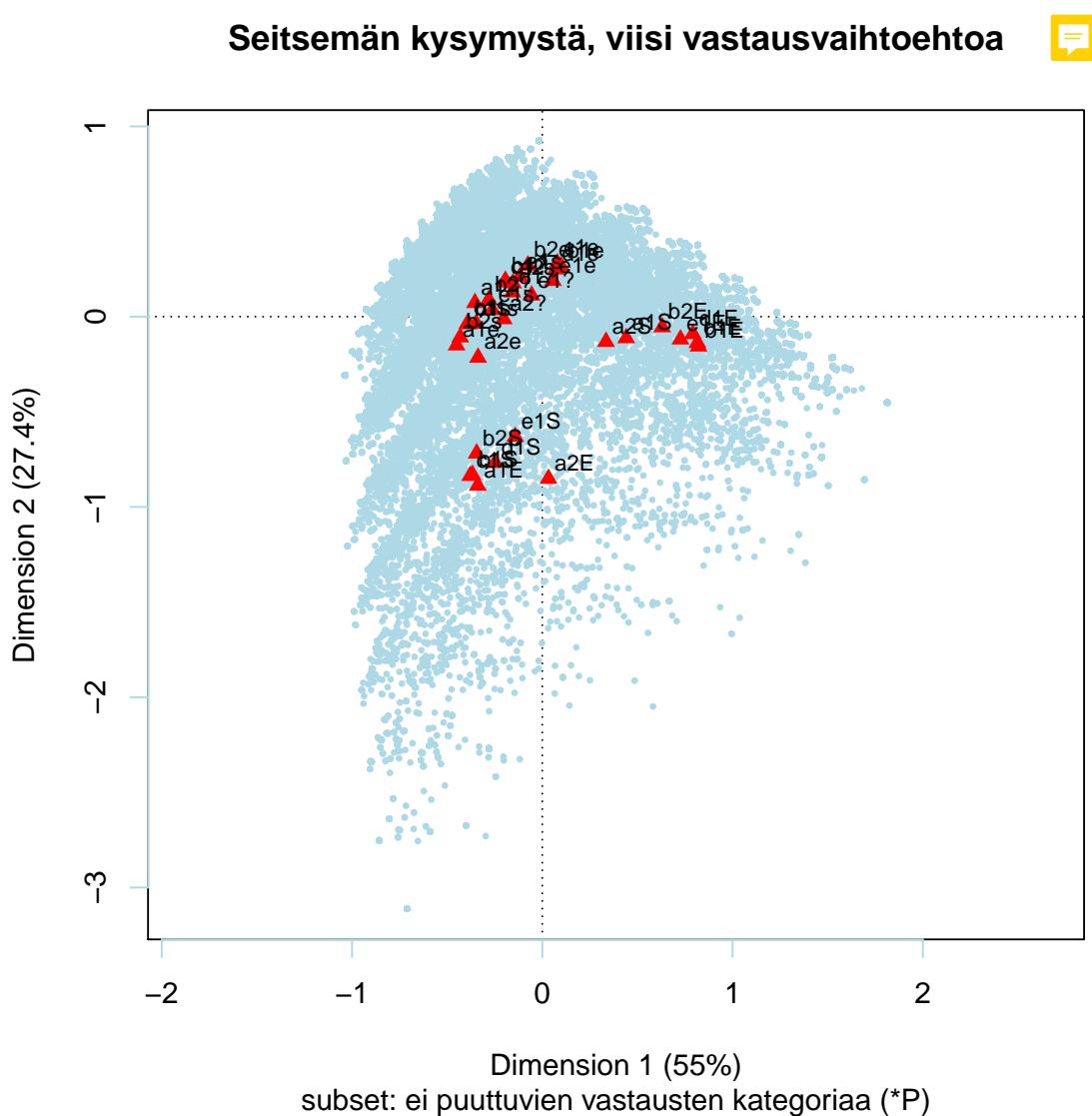
### edit2 MCA - numeeriset tulokset

```
summary(Qmuuttujat2.mca)
```

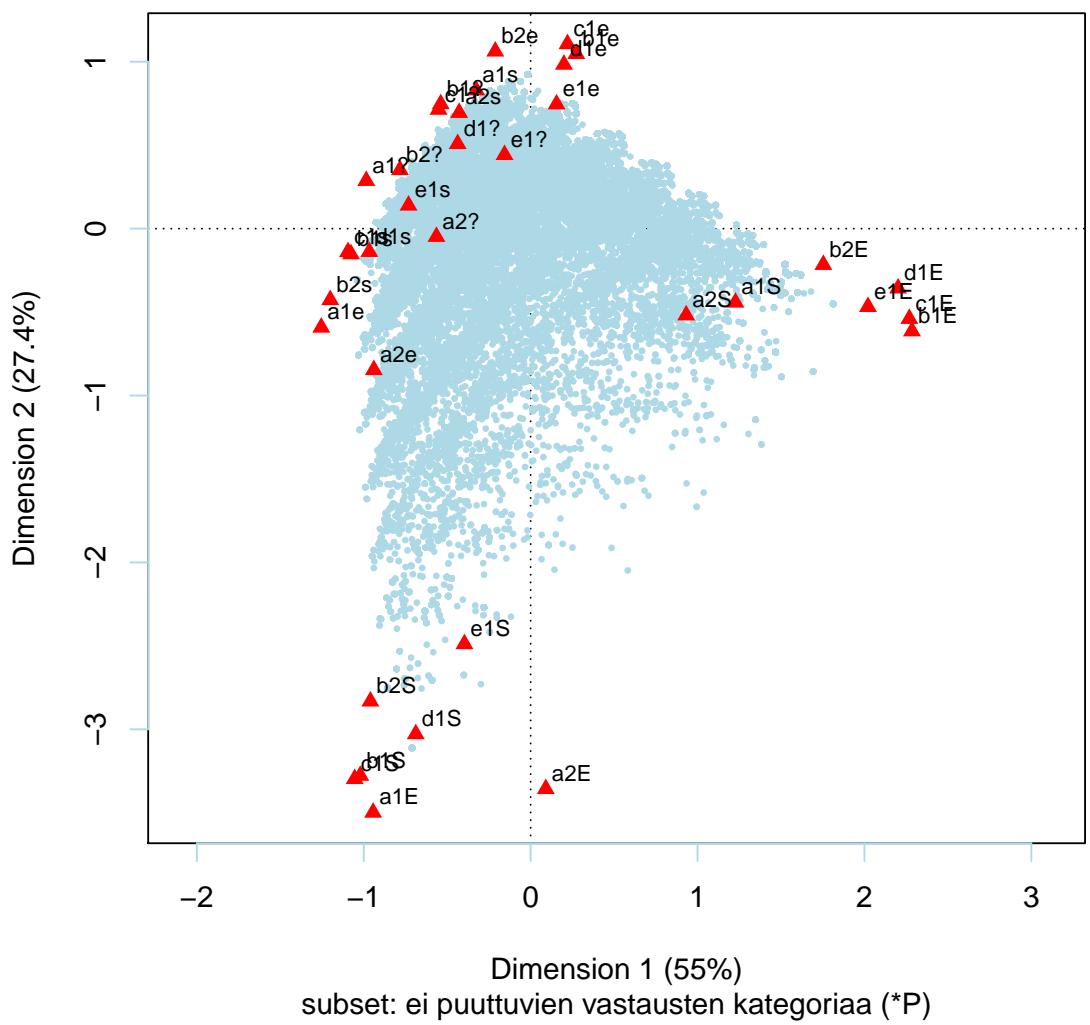
```
##  
## Principal inertias (eigenvalues):  
##  
##   dim   value      %   cum%   scree plot  
##  1    0.129087  55.0  55.0  ****  
##  2    0.064296  27.4  82.4  *****  
##  3    0.014807   6.3  88.7  **  
##  4    0.008871   3.8  92.5  *  
##  5    0.000538   0.2  92.7  
##  6    0.000221   0.1  92.8  
##  7    0.000156   0.1  92.9  
##  8    6.9e-050   0.0  92.9  
##  9    1e-06000   0.0  92.9  
##  
##   -----  
## Total: 0.234728  
##  
##
```



Kuva 7.6: MCA: Seitsemän kysymystä - 25 maata, kartta 3



Kuva 7.7: MCA: Seitsemän kysymystä - 25 maata, kartta 4

**Seitsemän kysymystä, viisi vastausvaihtoehtoa – asymmetrinen kartta**

Kuva 7.8: MCA: Seitsemän kysymystä - 25 maata, kartta 5

```
## Columns:
##   name mass qlt inr   k=1 cor ctr   k=2 cor ctr
## 1 a1S  48  975  29 | 441 916 73 | -112 59  9 |
## 2 a1s  54  869  20 | -117 206  6 | 210 663 37 |
## 3 a1?  15  653  27 | -354 626 14 | 73  26  1 |
## 4 a1e  18  788  29 | -451 709 28 | -150 79  6 |
## 5 a1E  5   884  30 | -339 113  4 | -887 771 56 |
## 6 b1S  12  818  37 | -367 133 12 | -831 685 128|
## 7 b1s  37  696  28 | -387 689 42 | -39  7  1 |
## 8 b1?  26  527  25 | -194 269  8 | 189 258 14 |
## 9 b1e  39  555  25 |   98 67  3 | 266 488 43 |
## 10 b1E 24   889  43 | 821 858 126| -156 31  9 |
## 11 c1S  12  826  38 | -379 141 14 | -836 685 134|
## 12 c1s  36  698  29 | -393 692 43 | -35  6  1 |
## 13 c1?  26  517  25 | -198 282  8 | 181 235 13 |
## 14 c1e  38  550  26 |   79 41  2 | 281 509 46 |
## 15 c1E  26  890  43 | 815 865 134| -137 24  8 |
## 16 d1S  12  851  34 | -247 80  6 | -768 771 112|
## 17 d1s  33  751  26 | -347 744 31 | -35  8  1 |
## 18 d1?  32  561  23 | -157 336  6 | 129 226  8 |
## 19 d1e  34  620  24 |   72 47  1 | 249 572 33 |
## 20 d1E  22  944  38 | 791 932 106| -91 12  3 |
## 21 e1S  15  910  31 | -142 44  2 | -631 866 90 |
## 22 e1s  36  752  23 | -263 738 19 |   35 13  1 |
## 23 e1?  34  350  22 |   -56 70  1 | 112 279  7 |
## 24 e1e  32  597  23 |   56 48  1 | 189 549 18 |
## 25 e1E  15  1012 32 | 726 985 62 | -119 26  3 |
## 26 a2S  49  1005 24 | 335 870 43 | -132 134 13 |
## 27 a2s  59  989  19 | -154 428 11 | 176 561 28 |
## 28 a2?  22  584  24 | -203 582  7 | -12  2  0 |
## 29 a2e  8   794  26 | -338 565  7 | -215 228  6 |
## 30 a2E  2   870  27 |   33 1  0 | -851 869 20 |
## 31 b2S  12  922  33 | -345 173 11 | -718 750 94 |
## 32 b2s  22  783  28 | -431 736 32 | -108 46  4 |
## 33 b2?  27  640  25 | -281 582 16 |   89 58  3 |
## 34 b2e  40  711  24 |   -76 53  2 | 269 658 45 |
## 35 b2E 39   939  37 | 630 932 119| -55  7  2 |
```

Vilkaistaan 3d-ratkaisua - onko kolmella keskimmäisellä vaihtoehdolla "dimension of middleness"?

edit tämä poistetaan lopullisesta versiosta.



```
Qmuuttujat2d3 <- mjca(mcaDat11jh.dat, ps="", nd = 3,subsetcat=eiPvastaukset)
summary(Qmuuttujat2d3)
```

```
##
## Principal inertias (eigenvalues):
##
##   dim   value    %  cum%  scree plot
## 1   0.129087 55.0 55.0 ****
## 2   0.064296 27.4 82.4 ****
## 3   0.014807  6.3 88.7 **
## 4   0.008871  3.8 92.5 *
## 5   0.000538  0.2 92.7
## 6   0.000221  0.1 92.8
```

```

## 7      0.000156   0.1  92.9
## 8      6.9e-050   0.0  92.9
## 9      1e-06000   0.0  92.9
## -----
## Total: 0.234728
##
##
## Columns:
##   name mass qlt inr   k=1 cor ctr   k=2 cor ctr   k=3 cor ctr
## 1 | a1S | 48  976 29 | 441 916 73 | -112 59 9 | -18 1 1 |
## 2 | a1s | 54  966 20 | -117 206 6 | 210 663 37 | 81 98 24 |
## 3 | a1? | 15  786 27 | -354 626 14 | 73 26 1 | -163 133 26 |
## 4 | a1e | 18  837 29 | -451 709 28 | -150 79 6 | -119 49 17 |
## 5 | a1E | 5   935 30 | -339 113 4 | -887 771 56 | 229 51 16 |
## 6 | b1S | 12  869 37 | -367 133 12 | -831 685 128 | 226 51 41 |
## 7 | b1s | 37  798 28 | -387 689 42 | -39 7 1 | -149 102 55 |
## 8 | b1? | 26  573 25 | -194 269 8 | 189 258 14 | -80 46 11 |
## 9 | b1e | 39  836 25 | 98 67 3 | 266 488 43 | 201 281 107 |
## 10 | b1E | 24  910 43 | 821 858 126 | -156 31 9 | -129 21 27 |
## 11 | c1S | 12  876 38 | -379 141 14 | -836 685 134 | 226 50 43 |
## 12 | c1s | 36  805 29 | -393 692 43 | -35 6 1 | -155 107 58 |
## 13 | c1? | 26  571 25 | -198 282 8 | 181 235 13 | -87 54 13 |
## 14 | c1e | 38  843 26 | 79 41 2 | 281 509 46 | 213 293 116 |
## 15 | c1E | 26  908 43 | 815 865 134 | -137 24 8 | -119 18 25 |
## 16 | d1S | 12  902 34 | -247 80 6 | -768 771 112 | 197 51 32 |
## 17 | d1s | 33  828 26 | -347 744 31 | -35 8 1 | -112 77 28 |
## 18 | d1? | 32  647 23 | -157 336 6 | 129 226 8 | -80 86 14 |
## 19 | d1e | 34  896 24 | 72 47 1 | 249 572 33 | 173 276 69 |
## 20 | d1E | 22  956 38 | 791 932 106 | -91 12 3 | -90 12 12 |
## 21 | e1S | 15  949 31 | -142 44 2 | -631 866 90 | 135 40 18 |
## 22 | e1s | 36  793 23 | -263 738 19 | 35 13 1 | -62 42 10 |
## 23 | e1? | 34  475 22 | -56 70 1 | 112 279 7 | -75 126 13 |
## 24 | e1e | 32  847 23 | 56 48 1 | 189 549 18 | 128 250 35 |
## 25 | e1E | 15  1023 32 | 726 985 62 | -119 26 3 | -78 11 6 |
## 26 | a2S | 49  1005 24 | 335 870 43 | -132 134 13 | -2 0 0 |
## 27 | a2s | 59  1011 19 | -154 428 11 | 176 561 28 | 35 22 5 |
## 28 | a2? | 22  770 24 | -203 582 7 | -12 2 0 | -115 186 19 |
## 29 | a2e | 8   796 26 | -338 565 7 | -215 228 6 | 23 3 0 |
## 30 | a2E | 2   920 27 | 33 1 0 | -851 869 20 | 202 49 5 |
## 31 | b2S | 12  961 33 | -345 173 11 | -718 750 94 | 162 38 21 |
## 32 | b2s | 22  873 28 | -431 736 32 | -108 46 4 | -151 91 35 |
## 33 | b2? | 27  741 25 | -281 582 16 | 89 58 3 | -117 101 25 |
## 34 | b2e | 40  939 24 | -76 53 2 | 269 658 45 | 159 228 68 |
## 35 | b2E | 39  944 37 | 630 932 119 | -55 7 2 | -45 5 5 |

```





## Luku 8

# Yhteenveto

Jäsennysdokumentissa on muutama ajatus, ja viite. Kirjoitetaan tämä viimeiseksi.

**k** Onko maiden vertailu järkevä? Blasius ja Thiessen "This paper provides empirically-based criteria for selecting Items and countries to develop measures of an underlying construct of interest that are comparable in cross-national research. Using data from the 1994 International Social Survey Program and applying multiple correspondence analysis to a set of common items in each of the 24 participating countries, we show that both the quality of the data, as well as its underlying structure - and therefore meaning - vary considerably between countries. The approach we use for screening countries and items is especially useful in situations where the psychometric properties of the items have not been well established in previous research." (?)

**k** voiko järjestysasteikon muuttujilla tehdä vertailuja maiden välillä?

"Surullinen totuus onnellisuustutkimuksesta" ? (ilman sulkua) ja suluilla (?)

**k** Eksploratiivinen data-analyysi ja todennäköisyysteoreettinen päätely

Gifi-nimimerkillä kirjoittavat Jan De Leeuw jatkavat verkkokirjassaan (?) keskustelua konfirmatorisen ja data-analyttisen eksploratiivisen lähestymistavan eroista. He ovat tiukasti eksploratiivisen linjan kannattajia, ja korostavat konfliktin pitkää historiaa. Nyt historia on heidän mielestään loppunut: "We shall not pay much attention any more to these turf and culture wars, because basically they are over. Data analysis, in its multitude of disguises and appearances, is the winner. Classical statistics departments are gone, or on their way out. They may not have changed their name, but their curricula and hiring practices are very different from what they were 20 or even 10 years ago."

**k** Miksi ei molempia?

**k1** Visualisointi on tehokasta tapa tutkia aineiston rakenetta, yhteyksiä muuttujien välillä ja eri havaintojoukkojen eroja. Ei automaattisesti helppoa, mutta kahden luokitelumuuttujan taulukko on ehkä yleisin tapa esittää mitä tahansa dataa. CA on aika pätevä väline taulukon riippuvuuksien hahmottamiseen yhdellä kartalla.

**k2** Jo oppikirjoista näkee, että tarvitaan monta menetelmää ja näkökulmaa. "jack of all trades but master of none", sellainen on data-analytikko.

\*\*k21\*\* MCA-esimerkki, eikö ole mainio lähtökohta faktorianalyysille?

**k** Tulevaisuus?

"The applicability of a dimension-reduction technique on very large categorical data sets or on categorical data streams is limited due to the required singular value decomposition (SVD) of properly transformed data. The application of SVD to large and high-dimensional data is unfeasible because of the very large computational time and because it requires the whole data to be stored in memory (no data flows can be analysed). The aim of the present paper is to integrate an incremental SVD procedure in a multiple correspondence analysis (MCA)-like procedure in order to obtain a dimensionality reduction technique feasible for the application on very large categorical data or even on categorical data streams"(?).



# Lähteet



# Liite 1: Korrespondenssianalyysin teoriaa

## Korrespondenssianalyysin perusyhtälöt ja kaavat

Tässä lähteenä Greenacren kirja(?) (ca in practice) ja sen liite “Theory of CA”.

**edit** Muistiinpanoja löytyy, joissa viitataan myös Biplots in practice - kirjaan. Kevään 2017 kurssin luentokalvoja on myös käytetty. Lisäillään vielä käsitteitä LeRouxin ja Rouanetin kirjasta, jos on tarvis.

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

Korrespondenssimatriisi  $P$  saadaan jakamalla matriisin  $N$  alkiot niiden summalla  $n$ .

Merkitään matriisin  $P$  rivisummien vektoria  $r (= (r_1, \dots, r_I))$  ja sarakesummien vektoria  $c (= (c_1, \dots, c_J))$ . Niitäh vastaavat diagonaalimatriisit ovat  $D_r$  ja  $D_c$ .

Korrespondenssianalyysin ratkaistaan singulaariarvohajoitelman avulla. Hyvin yleinen tulos, jostain syystä tilastotieteessä tullu tunnetuksi melko myöhään (Mustonen 1985).

Singulaariarvohajoitelmalla (singular value decomposition) tuottaa ratkaisun kun sitä sovelletaan standardoitun residuaalimatriisiin  $S$ .

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

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

$$D_r^{-1} P D_c^{-1} = \left( \frac{p_{ij}}{r_i c_j} \right) \quad (2)$$

$$S = D_r^{1/2} (D_r^{-1} P D_c^{-1} - 11^T) D_c^{-1/2} . \quad (3)$$

Toinen esitystapa on hyödyllinen, kun tarkastellaan CA:n yhteyksiä muihin läheisiin menetelmiin. Näitä menehtelmiä kuten myös korrespondenssianalyysiä kutsutaan monilla nimillä: “suhteellisten osuuksien datan” analyysi (log ratio analysis of compositiona data), moniulotteinen skaalaus, lineaarinen diskriminantianalyysi, kanoninen korrelatioanalyysi, pääkomponettianalyysi, kaksoiskuvat ja muita SVD-hajoitelman perustuvat dimensioiden vähentämisen menetelmät.

Samat kaavat voi esittää myös alkiomuodossa:

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \quad (4)$$

ja toinen

$$s_{ij} = \sqrt{r_i} \left( \frac{p_{ij}}{r_i c_j} \right) \sqrt{c_j} . \quad (5)$$

Alkimuodosasta esitetyistä kaavoista näkee intuitivisesti rivi- ja sarakeratkaisujen sidoksen. Ratkaisujen duaalisuus on teoreettinen tulos, jonka voi perustella täsmällisesti algebrallisen geometrian avulla. Käytännössä rivi- ja sarekeongelman duaalisuus tarkoittaa sitä, että vain toinen ongelma on ratkaistava.

Singulaariarvohajoitelta (singular value decomposition, SVD) matriisille  $S$  on

$$S = U D_\alpha V^T \quad (6)$$

missä  $D_\alpha$  on diagonaalimatriisi, jonka alkiot ovat singulaariarvot suuruusjärjestyksessä  $\alpha_1 \geq \alpha_2 \geq \dots$ .

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

$$S_{(m)} = U_{(m)} D_{\alpha(m)} V_{(m)}^T \quad (7)$$

Korrespondenssianalyysin ratkaisualgoritmissa tästä 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

$$\Phi = D_r^{-\frac{1}{2}} U \quad (8)$$

Sarakkeiden standardikoordinaatit

$$\Gamma = D_c^{-\frac{1}{2}} V \quad (9)$$

Rivien pääkoordinaatit

$$F = D_r^{-\frac{1}{2}} U D_\alpha = \Phi D_\alpha \quad (10)$$

Sarakkeiden pääkoordinaatit

$$G = D_c^{-\frac{1}{2}} V D_\alpha = \Gamma D_\alpha \quad (11)$$

Pääakseleiden inertiat (principal inertias)  $\lambda_k$

$$\lambda_k = \alpha_k^2, k = 1, \dots, K, K = \min\{I-1, J-1\} \quad (12)$$

Ratkaisun dimenisio on myös maksimi-inertia. Tässä aineistossa ja vastaavissa kyselytutkimusdatoissa inertia on yleensä paljon maksimia pienempi. Asymmetrisissä kartoissa ideaalipisteet ovat kaukana origon lähelle pakkauteesta havaintojen pilvestä.

Korespondenssianalyysi ratkaisun akseleiden inertiaa kutsutaan usein ominaisarvoksi, mutta periaatteessa SVD-ratkaisulla saadaan singulaariarvot, ja niiden neliöt ovat akseleiden inertioita. Ominaisarvojen ja sigulaariarvojen yhteyts on läheinen ja riippuu diagonalisoitavan matriisin ominaisuuksista.

Korrespondenssimatriisi  $P$  voidaan esittää matriisi- ja alkamuodossa ns. palautuskaavana (reconstitution formula).

$$P = D_r \left( 11^T + \Phi D_{\lambda}^{\frac{1}{2}} \Gamma^T \right) D_c \quad (13)$$

$$p_{ij} = r_i c_j \left( 1 + \sum_{k=1}^K \sqrt{\lambda_k} \phi_{ik} \gamma_{jk} \right) \quad (14)$$

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. barysentrisen ominaisuus - barycentric relationships)

$$F = D_r^{-1} P \Gamma \quad (15)$$

$$G = D_c^{-1} P^T \Phi \quad (16)$$

Pääkoordinaatit pääkoordinaattien funktointa:

$$F = D_r^{-1} P G D_{\lambda}^{-\frac{1}{2}} \quad (17)$$

$$G = D_c^{-1} P^T F D_{\lambda}^{-\frac{1}{2}} \quad (18)$$

Yhtälöt (15) ja (16) esittävät profilipisteet ideaalipisteiden (vertex points) painotettuna keskiarvoina, painoina profilin elementit. Asymmetriset kartat (rivien tai sarakkeiden suhteen) perustuvat näihin yhtälöihin. Yhtälöiden (17) ja (18) kahdet pääkoordinaatit ovat perusta symmetrisille karttoille. Myös niitä yhdistää barisentrinen painotetun keskiarvon riippuvuus, mutta mukana ovat skaalaustekijät  $\frac{1}{\sqrt{\lambda_i}}$ . Skaalaustekijä on jokaisessa dimensiossa sen inertia, suurimmasta pienimpään.

### **Pisteet ja projektio aliavaruuteen**

Kuva on kurssimateriaaleista(?).

```
knitr::include_graphics('img/CAquality.png')
```

Kuvassa on esitetty korrespondenssianalyysin ratkaisun minimoointiongelma. Pisteen projektio on sitä parempi mitä pienempi kulma on sentroidista pisteeseen piirtetyn janan ja pisteen projektiota välillä. COR - tunnusluku ca-funktion numeerisissa tuloksissa tämän kulman kosinin neliö. Pisteen kuvauslaatu (qlt) ca-tuloksissa on valitun approksimaation akseleiden kvaliteettien (COR) summa.

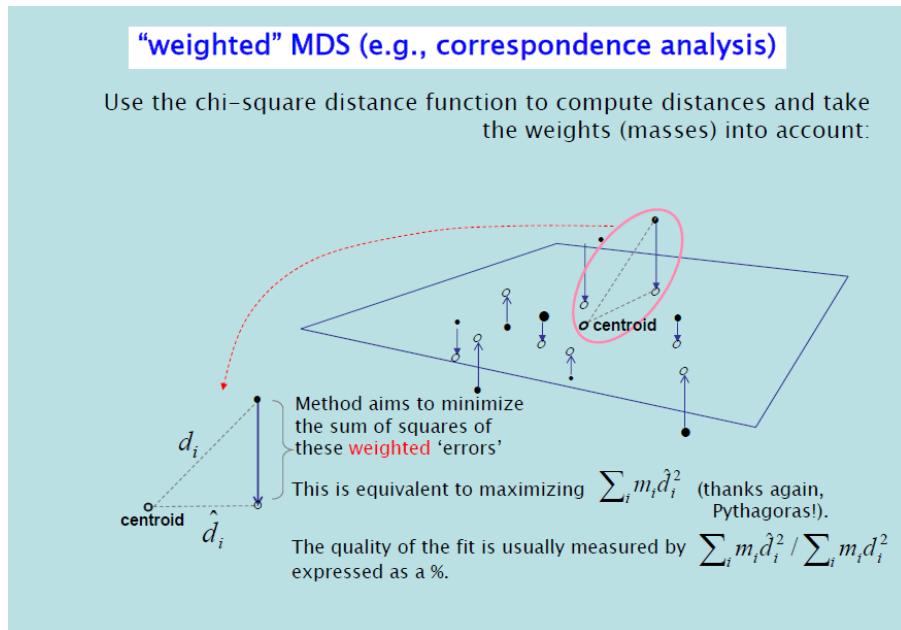
Kuvasta voi myös hahmottaa sen periaatteen, että projektiossa kaukana olevat pisteet ovat kaukana myös alkuperäisessä avaruudessa. Projektiossa lähekkäin olevat pisteet voivat olla alkuperäisessä avaruudessa kaukana toisistaan, jos niiden projektiota laatu on huono.

## **Matriisit ja niiden havainnollistaminen**

**edit** Nämä tässä varalla, jos matriisiä tarvitaan lisää. Ehkä ei?

Korrespondenssianalyysin sovelluksissa tutkimusongelman ratkaisu on usein sopivan matriisin rakentaminen.

**edit: kaavaesimerkkejä**



Kuva 1: Pisteen projektio aliavaruuteen

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & \ddots & & \\ \vdots & & & \\ a_{n1} & \dots & \dots & a_{nk} \end{bmatrix} \quad (19)$$

Ehkäpä ABBA onnistuu paremmin tällä notaatiolla?

$$A = \begin{bmatrix} A_{11} & B_{12} \\ B_{21} & A_{22} \end{bmatrix} \quad (20)$$

$$A = \begin{bmatrix} A_{maa,Q1a} & A_{maa,Q1b} \\ B_{gage,Q1a} & B_{gage,Q1b} \end{bmatrix} \quad (21)$$

### Pinotut tai yhdistetyt matriisit (“stacked and concatenated matrices”)

Yksinkertainen korresponduenssianalyysi on kahden luokittelumuuttujan määrittämän taulukon analyysiä, mutta sitä voi soveltaa myös usean muuttujan analyysiin. Menetelmän matemaattinen perusta ja ratkaisualgoritmi (SVD) toimivat, tulkinta vain muuttuu.

Yksinkertaisin laajennus on lisätä alkuperäisen taulukon alle toinen taulukko. Rivist ovat esimerkissä maittan summattuja vastauksia, ja niiden alle voidaan lisätä joku toinen luokittelumuuttuja. Havaintojen määrä yhdityssä (“pinotussa”) taulussa kaksinkertaistuu.

Taulukoiden yhdistämisen idea on inertian dekomponointi. Yhdistetyn matriisin inertia voidaan eri tavoin esittää alimatriisien inertian summana. Tällöin jokaisen alimatriisin reunajakauman tulee olla sama, ja puuttuvat tiedot vääristäävät tuloksia.

Merkitään edellisten analyysien kuuden maan ja viiden vastausvaihtoehdon taulukkoa matriisilla  $A_{IJ}$ , missä  $I$  on rivien ja  $J$  sarakkeiden lukumäärä. Taulukoidaan ikäluokan (1 - 6) ja sukupuolen ( $f$  = nainen,  $m$  = mies) vuorovaikutusmuuttuja ( $f_1, \dots, f_6$  ja  $m_1, \dots, m_6$ ) samojen vastausvaihtoehojen kanssa. Jos tästä taulukkoaa merkitään matriisilla  $B_{I'J}$ , voimme muodostaa yhdistetyn matriisin

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

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

### **Monimuuttuja-korrespondenssianalyysi MCA**

Usean muuttujan korrespondenssianalyysissä tutkitaan usean muuttujan välistä yhteyksiä. Kartan tulkinnan apuna siihen voidaan lisätä havaintojen sijaan niiden keskiarvopisteitä ja niille simuloituja luottamusellipsejä. Kuvien pääongelma on liian suuri määärä pisteitä, ja analyysin lopputulos on usein mahdollisimman yksinkertainen kartta.

Usean muuttujan analyysissä kohteena on joko indikaattorimatriisi  $Z$  tai Burtin matriisi  $B$

Indikaattorimatriisissa rivit ovat havaintoja ja sarakkeet luokittelumuuttujan arvoja. Havaintoa vastaa rivi nollia ja arvo 1 valitun vastausvaihtoehdon kohdalla. Tästä seuraa, että vain erilaiset vastaukset määrittävät rivien etäisyyksiä.

Burtin matriisi on erikoistapaus yhdistetyistä matriiseista. Siihen on koottu kaikki tutkittavien muuttujien pa-reittan muodostetut taulukot. Diagonaallilla ovat muuttujien ristiintaulukoinnit itsensä kanssa. Ratkaisu riippuu vain näistä parittaisista taulukoista.

Burtin matriisi on kätevä välivaihe matriisien yhdistelyssä.

Molemmat matriisit paisuttavat keinotekoisesti kokonaisinertiaa, ja esimerkiksi kaksiulotteisen kartan selitetyn inertian osuudet jäävät melko pieniksi. Ratkaisuna on inertian oikaisu tai korjaus (adjusted inertia), jossa mm. poistetaan kokonaisinertialaskelmista Burtin matriisin diagonaallilla olevat alimatriisit. Näillä korjauksilla ei ole vaikutusta kartan pisteen sijaintiin. Tämä menetelmä on ca-paketin mjca-funktion oletus.

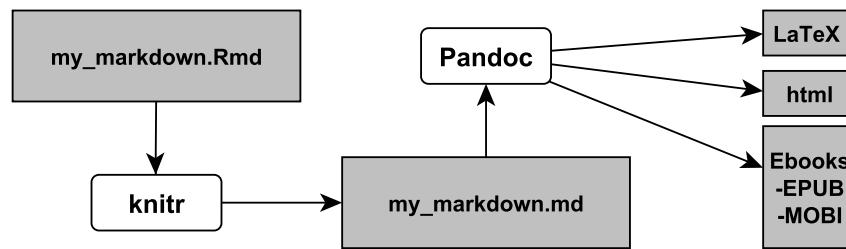
Kolmas vaihtoehto on ns. yhdistetty korrespondenssianalyysi (joint ca).



## Liite 2: Tekninen ympäristö ja Bookdown-paketti

Muokataan tiiviimpi pätä esimerkkireposta bookdown-testi1. Tämä kuva kertoo vain julkaisuteknikan ympäristön.

```
knitr:::include_graphics('img/BookdownProc.png')
```



Kuva 2: Tulostiedoston prosessointi

```
# pois out.width='50%',
```

