

Korrespondenssianalyysi - graafinen ja geometrinen data-analyysin menetelmä

Jussi Hirvonen

Versio 0.9, tulostettu 2020-11-25

Sisällys

Alkutoimia

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

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

Tämä luku poistetaan kun tutkielma on valmis

Raportti yhtenä html-tiedostona (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>)

Tiivistelmästä löytyy uudempi pohja, jota käytetään.

RefWorksistä tuotu jhca2020.bib-tiedosto tarkistettu ja korjailtu virheet. Ei enään ladata uutta (21.11.2020), packages.bib (r-pakettien viitekanta) tarkistettu ja korjattu erikoismerkit. Ei päivitetä enään (24.11.20)

Data-analyysi: (<https://hirjus.github.io/Galku>).

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 analyysissä. Siksi ns. monimuuttujakorrespondenssianalyysin lyhyessä esittelyssä aineistoa laajennetaan, sillä “on erittäin vaikeaa osoittaa kotia-kvaarioissa, 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 luokittelumuuttujan taulukon lukumäärädata (count data) tai suhdeasteikon muuttujia samassa mittayksikössä (esim. euroissa).
- määrittele graafinen, deskriptiivinen, eksploraatiivinen data-analyysi
- yksinkertainen ca, useamman muuttujan ca

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

kartan – avulla tulkita monimutkaisen empiirisen aineiston muuttujien riippuvuuksia. Yhteyksiä ei tiivistetä todennäköisyyspäättelyn kriteereillä tilastolliseen malliin, vaan deskriptiivisen analyysin hengessä esitellään koko aineisto. Mallin sijaan vähennetään ulottuvuuksia, ja siinä menetetään informaatiota. Tavoitteena on säilyttää yleensä kaksiulotteisessa kuvassa mahdollisimman suuri osa alkuperäisen datan vaihtelusta. Eksploraatiivinen 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 tulostuloksiin.

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 aikaisemmat versiot ovat tehneet menetelmää laajasti tunnetuksi. “Biplots in Practice” (jatkossa “Biplots”) (?) esittää menetelmän osana yleisempää kaksoiskuvien ideaa.

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

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 liitessä.

1.3 Korrespondenssianalyysin historiaa

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

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

k3 Peruskäsitys monessa lähteessä (vihreä kirja, GDA-kirja jne.): syntyy ja kukoistuu Ranskassa, loistava eristys (splendid isolation), pikku 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-analyttisen/kuvailevan jne. lähestymistavan ja tilastollisen mallintamisen välillä - on läsnä edelleen mutta turha korostaa. Myös tilastollisen mallintamisen ja päättelyn sisällä on kiistoja, erilaisia näkemyksiä ja kuiluja.
3. “Esoteerinen tieteenfilosofia”? Kiinnostava aihe, ehkä. Murgtag-sitaatti.

Luku 2

Data

Käytän tutkielmassa International Social Survey -projektin (ISSP) vuoden 2012 kyselytutkimusta “Perhe , työ ja sukupuoliroolit”(International Social Survey Programme: Family and Changing Gender Roles IV). Tutkimuksen aikasempien toteutusten dataa on käytetty tutkielman tärkeimmissä lähteissä esimerkkidatana.

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 julkaisua. Viime vuosina luetteloon on lisätty noin 400 julkaisua vuodessa (?).

Data ja dokumetaatio on vapaasti saatavilla saksalaisen GESIS-tutkimuslaitoksen ylläpitämästä data-arkistosta (<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).

GESIS-instituutin ”datakatalogista”(<https://zcat.gesis.org>) löytää kätevästi kaiken dokumentaation(?), mutta edes saksalaiset eivät voi estää www-sivustojen innokaita uudistajia. Monet linkit lähdeluettelossa vievät GESIS-arkistosivulle, josta löytyy pitkä lista pdf-dokumentteja (?).

Taulukkoon 2.1 on koottu neljän tärkeimmän dokumentin tiedostonimet ja lyhyt kuvaus.

Tätä kirjoittaessa (10.11.2020) ISSP 2012 - aineisto löytyy osoitteesta [<https://zcat.gesis.org/webview/index.jsp?object=http://zcat.gesis.org/obj/fStudy/ZA5900>].

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

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

2.1 Aineiston rajaaminen maat ja muuttajat

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 kuvaa hyvin The Economist - lehden artikkeli Saksojen jälleenyhdistymisen 30-vuotispäivänä (3.10.2020, "A report...reveals the interplay between policy and attitudes that influences the decision to work."). Artikkelin 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.

Kyselylomakkeilla kysymykset olivat hieman pidempiä. Kuva 2.1 on osa suomenkielistä lomaketta.

Valituista taustamuuttujista monet on kerätty haastattelulla. Tiedonkeruu, otantamenetelmät ja yksikkövaustuskadon (unit non-response, otokseen valitulta ei saada mitään tietoja) 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.

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

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

Seuraavaksi perheeseen, työhön ja kotiin liittyviä kysymyksiä.

23. Mitä mieltä olet seuraavista väittämistä?
Rengasta jokaiseltä... luvulta vain yksi vaihtoehto

	Täysin samaa mieltä	Samaa mieltä	En samaa enää eri mieltä	Eri mieltä	Täysin eri mieltä	En osaa sanoa
a) Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä.....	1	2	3	4	5	8
b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä.....	1	2	3	4	5	8
c) Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö.....	1	2	3	4	5	8
d) On hyvä käydä töissä mutta tosiasiasa useimmat naiset haluavat ensisijaisesti kodin ja lapsia.....	1	2	3	4	5	8
e) Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen.....	1	2	3	4	5	8

24. Mitä mieltä olet seuraavista väittämistä?
Rengasta kummaltakin riviltä vain yksi vaihtoehto.

	Täysin samaa mieltä	Samaa mieltä	En samaa enää 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 rahaa; naisen tehtävä on huolehtia kodista ja perheestä.....	1	2	3	4	5	8

25. Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?
Rengasta kummaltakin riviltä vain yksi vaihtoehto.

Naisen tulisi...	käydä koko- päivätyössä	käydä osa- aikatyössä	pysyä kotona	En osaa sanoa
a) Kun perheessä on alle kouluikäinen lapsi.....	1	2	3	8
b) Kun nuorin lapsi on aloittanut koulunkäynnin.....	1	2	3	8

Kuva 2.1: Suomenkielinen lomake

Aineiston luokittelu- ja järjestysasteikon muuttujat muunnetaan R-ohjelmiston factor-tietotyyppiä. Teen muunnokset useammassa vaiheessa heti kun data on luettu SPSS-tiedosta. Käsittelyssä koitan noudattaa helposti toistettavan tutkimuksen periaatteita (?). koodi ei saisi olla kovin virheeltistä ("haurasta") ja tarkistuksia tehdään paljon. Data-analyysin ja ehkä erityisesti korrespondenssianalyysin idea on kuitenkin operoida matriiseilla, lisätä ja poistaa rivejä ja sarakkeita ja rakennella mutkikkaampia matriiseja yksinkertaisemmista. 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 "singuaariarvohajoitelmaan perustuvan menetelmän" vertailuun(?). Blasius ja Thiessen ((?)) arvioivat aineiston laatua ja ja maiden vertailtavuutta vuoden 1994 aineistolla.

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

? tutkivat ensin 18 OECD-maan perhepolitiikan muutoksia kolmen viime vuosikymmenen ajalta. Näkökulma on työllisyyspolitiikka ja menetelmänä monimuuttuja-korrespondenssianalyysi (MCA). Havaituille kehityssuunnille etsitään toisessa vaiheessa selityksiä. Aineistona on viisi kansainväliseen vertailuun soveltuvaa 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 kontingenssitaulun analyysin avulla. Kyse ei ole pelkästään helposta esimerkistä, vaan peruskäsitteet ja geometrisiin perusteisiin nojaava graafinen analyysi ovat oleellisilta osin samat myös monimutkaisemmissa menetelmän sovelluksissa. (?)

Greenacren oppikirjat ovat hyvä esimerkki perusteellisesta yksinkertaisen taulukon analyysin esitystavasta. ? esittelivät menetelmän ranskalaisen perinteen mukaisesti korostaen matemaattista teoriaperustaa, mutta myös siinä menetelmä peruskäsitteet ja tulkinnot esitellään yksinkertaisella esimerkillä. ? käyttää samaa Fisherin Cairness-aineistoa korrespondenssianalyysin esittelyyn

Esitän tässä jaksossa korrespondenssianalyysin peruskäsitteet intuitiivisesti, matemaattiset yksityiskohdat löytyvät liitteestä 1. Esitystavan etu on taulukon pieni koko, johtopäätökset voi helposti tarkastaa datasta. Datan analyysin tärkein väline on kuva, yleensä kaksiulotteinen kartta. Tulkinta ja erityisesti väärin johtopäätösten välttäminen vaati kartan tulkinnan varmistamista ratkaisun numeerista tuloksista. Kun analysoitava taulukko, sen rivien ja sarakkeiden riippuvuuksia kuvaava kartta ja kartan perustana olevat numeeriset tulokset esitetään yhdessä on helpompi ymmärtää tulkinnan perussäännöt.

Greenacren oppikirjojen ja ? esitystavassa on pieni ero. Molemmille juuri graafinen analyysi on oleellista, mutta ranskalaiset tutkijat korostavat numeeristen tulosten ensisijaisuutta. Analyysi pitää aloittaa tutkimalla numeerisen ratkaisun ominaisuuksia. Greenacren mielestä numeerisia tuloksia tarvitaan johtopäätösten varmistamiseen, ensin katsotaan karttaa. Eroa ei kannata liioitella, molempia tarvitaan. Eksploratiivisessa data-analyysissä näkökulmaa muutetaan kun datan ominaisuudet tai omituisuudet havaitaan. Kun kartat ovat aina approksimaatiota, numeerisia tuloksia tarvitaan.

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. Eräs lastensuojelun ammattilainen piti vastaamista mahdottomana; pitää tietää missä lapsi on , mitä hän tekee. Kysymykset on kuitenkin suunniteltu kokonaisuudeksi, ja niitä analysoidaan yhdessä luvussa 7. Yhden taulukon analyysi esittelee menetelmän, oikeassa tutkimuksessa pitää käyttää vähintään koko kysymyssarjaa.

Havainnot joissa tieto vastauksesta puuttuu on poistettu aineistosta. Taustamuuttujia ovat vastaajan sukupuoli ja ikä. Taulukoissa vastausvaihtoehtojen tunnuksina käytetään samoja symboleja kuin kuvissa (E = täysin eri mieltä, e = eri mieltä ? = ei samaa eikä eri mieltä, s = samaa mieltä, S = täysin samaa mieltä).

Frekvenssitaulukossa 3.1 on esitetty vastausten suhteellinen jakauma, lukumäärät on jaettu havaintojen lukumäärällä (8143). Korrespondenssianalyysissä kaikki on suhteellista, ja analyysi perustuu tähän taulukkoon.

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

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 muuttujia ja rivit havaintoja. Rivit on saatu summaamalla (aggregoimalla) vastaukset maittain. Greenacre käyttää näistä yksittäisten vastausten (havaintojen) summari-veistä termiä “sample”.

Sarakeprosentit antavat toisen näkökulmaan samaan dataan.

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*.

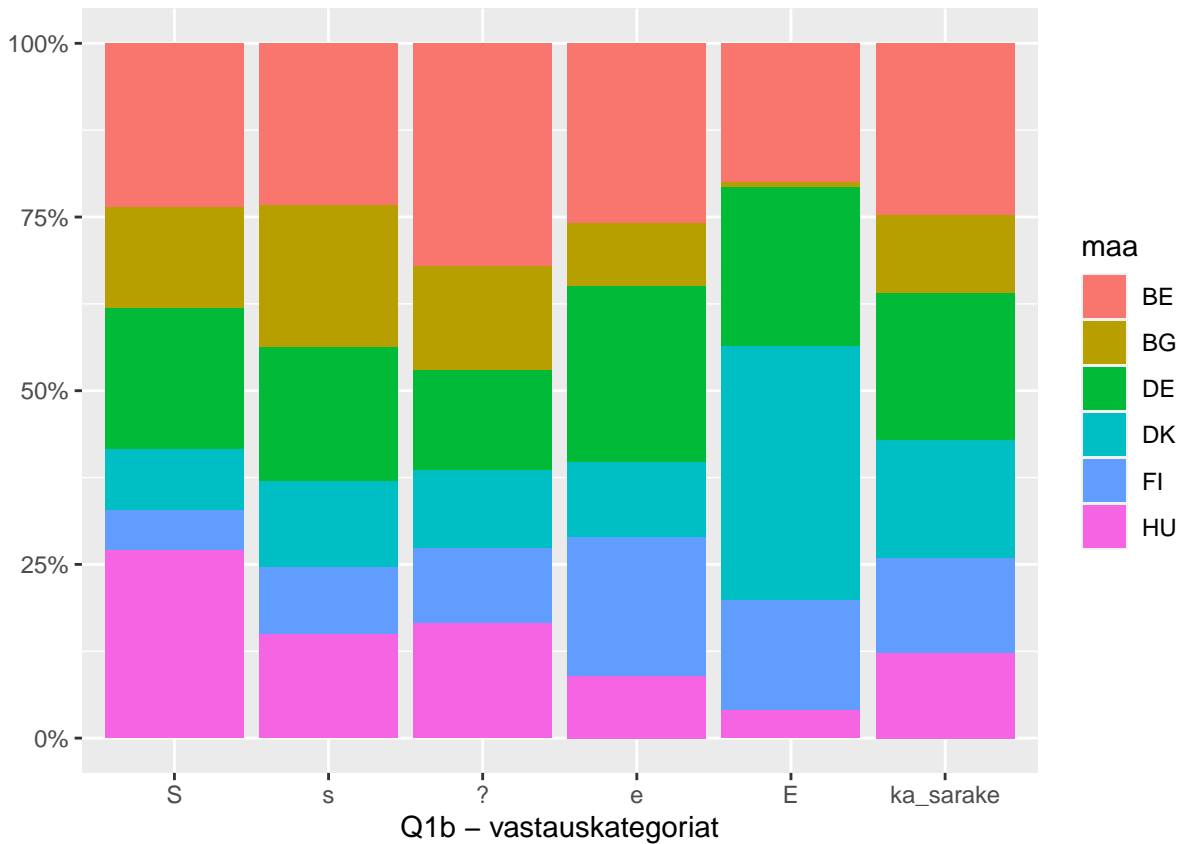
k Rivit on saatu alkuperäisestä aineistosta osajoukkojen summina. MG:n terminologialla “samples”.

Korrespondenssianalyysin perusidea on analysoida rivien ja sarakkeiden yhteyttä (korrespondenssia) rivi- tai sarakeprofiilien hajonnan avulla. Hajontaa mitataan poikkeamilla keskiarvorivistä tai sarakkeesta, ja massat

Taulukko 3.3: Kysymyksen Q1b vastaukset, sarakeprosentit

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

otetaan huomioon, kun hajonnat lasketaan yhteen.



Kuva 3.1: Q1b:Sarakeprofiilit ja keskiarvoprofiili

Kuvasta 3.2 3.2 esimerkiksi näkee, että Tanska (DK) näyttäisi poikkeava keskiarvorivistä paljon, samoin Bulgaria. Bulgarian massa on kuitenkin aineiston pienin (11,31 %), Tanskan taas kohtalainen (17 %). Sarakeprofiilikuvassa 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 ja sarakkeiden yhteys?

Kahden luokittelumuuttujan riippuvuutta voidaan testata χ^2 - testillä. Riippumattomuushypoteesin mukainen odotettu solufrekvenssi on taulukon 3.1 reunajakaumien alkioden tulo.

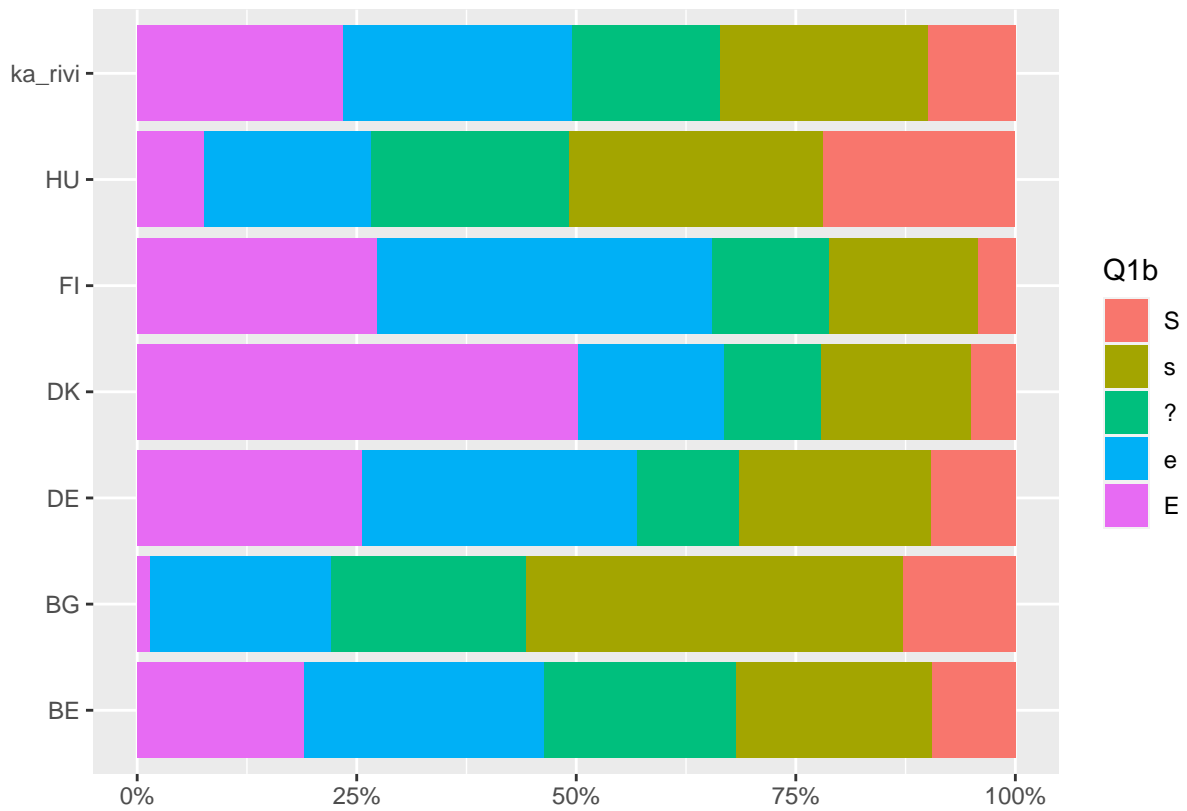
Testisuure saadaan laskemalla yhten jokaisen solun havaittujen ja odotettujen frekvenssien erotukset muodossa

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

Tämä voidaan esittää korrespondenssianalyysin esittelyyn sopivammalla tavalla parilla muunnoksella, jolloin saamme riveittäin vastaavat termit rivisummalla painotettuna.

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

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



Kuva 3.2: Q1b: riviprofiilit ja keskiarvorivi

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

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

Tunnusluku ϕ^2 on korrespondenssianalyysissä *kokonaisinertia* (total inertia). Se kuvaa, kuinka paljon varianssia taulukossa on ja on riippumaton havaintojen lukumäärästä. Tilastotieteessä tunnusluvulla on useita vaihtoehtoisia nimiä (esim. mean square contingency coefficient) ja sen neliöjuurta kutsutaan ϕ - kertoimeksi.

Korrespondenssianalyysin ratkaisussa käytetään suhteellisten frekvenssien taulukkoa.

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 χ^2 - etäisyys on painotettu euklidinen 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 voidaan esittää rivien ja keskiarvorivin (sentroidin) χ^2 -etäisyyksien neliöiden painotettuna summana, jossa painoina ovat rivien massat m_i ja summa lasketaan yli rivien i .

$$\phi^2 = \sum_i (massa\ m_i) \times (profiilin\ i\ \chi^2 - etäisyys\ sentroidista)^2 \quad (3.6)$$

Korrespondenssianalyysin kolmen peruskäsitteen ”tripletti” – *profiili*, *massa* ja χ^2 - *etäisyys* – on esitely tarkemmin liitteessä 1.

Rivi- ja sarakeprofilien taulukoista huomaa, että keskiarvoprofilien alkioit ovat massoja. Rivien keskiarvoprofilin alkioit ovat sarakemassoja, ja sama pätee sarakkeille. Tämä rivi- ja sarakeongelmien duaalisuus on yksinkertaisen korrespondenssianalyysin keskeinen idea (CAiP, s. 57). Rivi- tai sarakeongelman ratkaisu tuottaa saman tuloksen.

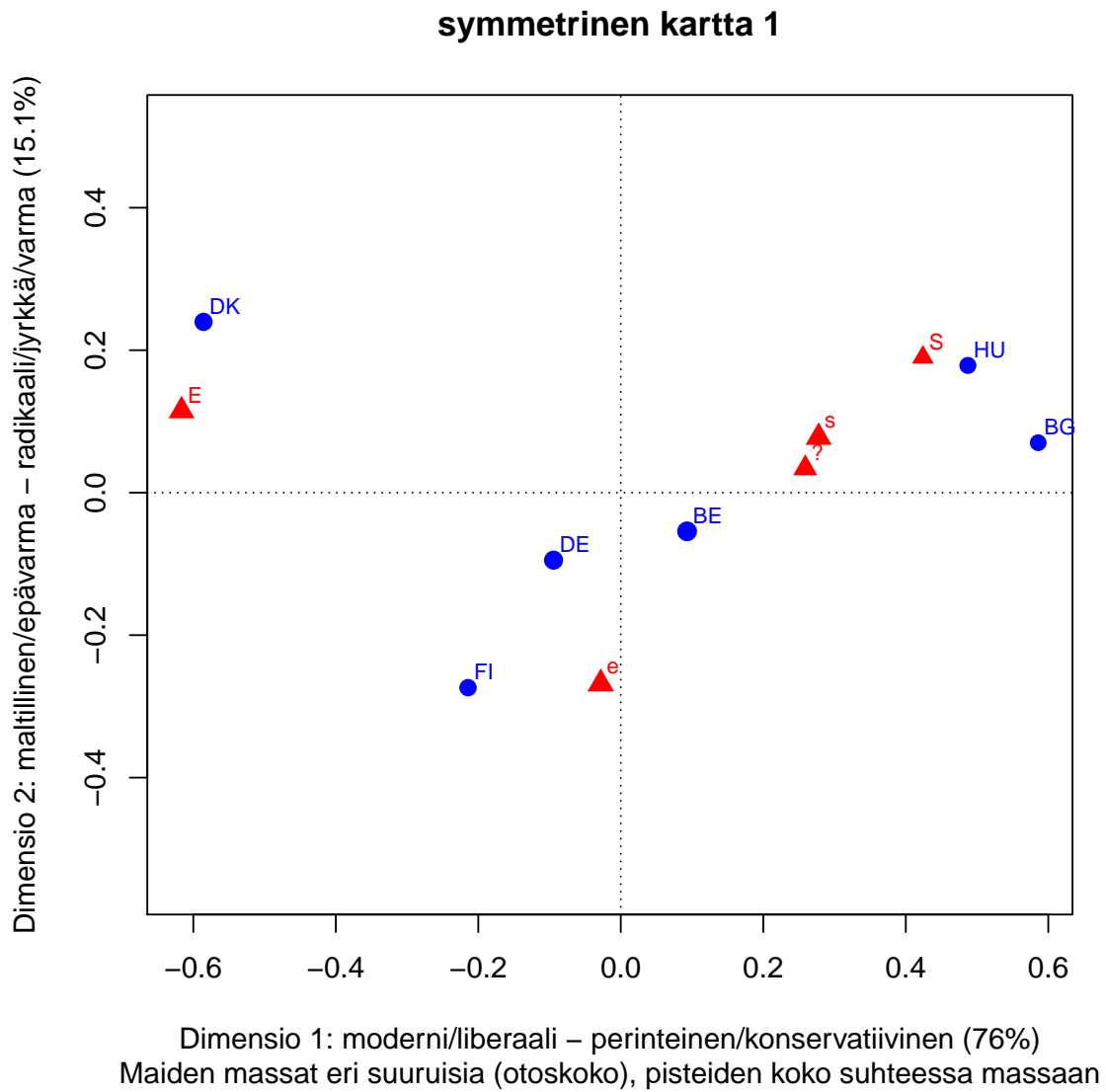
Ratkaisun dimensio on sarakkeiden tai rivien lukumäärä vähennettynä yhdellä, pienempi kahdesta vaihtoehdosta. Se on myös kokonaisinertian teoreettinen maksimi.

Korrespondenssianalyysi on läheistä sukua pääkomponenttianalyysille. Etäisyysmitta on khii2-etäisyys (käytän tekstissä tätä kirjoitusasua) ja mukana ovat massat painoina. Ratkaisussa etsitään haluttu yleensä kaksiulotteinen ratkaisu (taso), joka minimoi pisteiden khii2-etäisyyksien poikkeamien summan eli on mahdollisimman lähellä pisteitä. Alkuperäisen täyden dimension (full space) data projisoidaan tälle tasolle.

3.2 Symmetrinen kartta

Kartassa on jo nimetty molemmat akselit, mutta tuloksin voi aloittaa akseleille merkityistä prosentista. Ne kertovat, kuinka paljon aineiston inertiaasta eli hajonnasta on kaksiulotteisessa projektiossa saatu kuvattua akseleille.

Akselit ovat sisäkkäisiä (nested) Ensimmäinen akseli saa aina suurimman osan inertiaasta, tässä 76 prosenttia. Kun toinen akseli kuvaa 15 prosenttia koko inertiaasta, 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”.



Kuva 3.3: Q1b: lapsi kärsii jos äiti on töissä

Kontrastit määrittävät akselien tulkinnan. Benzacrin ohjeen mukaan (1992, teoksessa (?), 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 ylhäällä ja alhaalla. Tulkinta tehdään akseleiden suuntaan.

Kun taulukon rivit ovat havaintoja ja sarakkeet muuttujia, akselien tulkinta tehdään muuttujien avulla. Vasemmalla on E ja oikealla puolella samanmieliset vastaukset s ja S. Neutraali ”?” on s-vastausten vasemmalla puolella. Kaikki erot ovat suhteellisia, kuvan perusteella voi sanoa kuinka paljon.

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

Ensimmäisen dimensioin tulkinta on aika selkeä. Toinen akseli on kontrasti lievemman tai maltillisemman erimielisyyden ja muiden vastausten kanssa. Se on 1. dimension suuntaan kaikkein lähimpänä origoa. Hieman varovaisemmin akselin voi tulkita maltillisen ja jyrkemmän tai varmemman mielipiteen kontrastiksi.

Maiden vertailu tehdään näiden akselien suuntaan. Sekä sarekepisteiden että rivipisteiden keskinäiset välimatkat approksimoivat optimaalisesti niiden (khii2)etäisyyksiä. Sarake- ja rivipisteiden välillä etäisyyksillä ei ole mitään suoraa tulkintaa. Pisteiden etäisyydet samassa pistepilvessä ovat suhteellisia, Saksa on konservatiivisempi kuin Suomi mutta emme tiedä kuinka paljon. Maiden järjestys oikealta vasemmalle on selkeä, Tanska on vasemmalla liberaalina ”ääripäänä”, 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 neliulotteisen pistepilven hajonnalle. Vain origo on siinä tarkasti esitetty, se on koko aineiston keskiarvopiste, ja pisteiden hajonta sen ympärillä kuvaa poikkeamaa 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 tutkielmassa esitetään vain datan analysoinnin valikoituja kuvia. Graafinen data-analyysi on vaivatonta vasta sitten kun se tehty. En jatkossa esitä kuvailevia akseleiden nimiä kuvissa, akseleiden nimeäminen on kuvan tulkinnan toinen askel.

k Kuva tai kartta - käytän termejä synonyymeinä - on se taso, joka parhaiten ”selittää” neliulotteisen pisteparven hajontaa suhteessa koko aineiston keskiarvopisteeseen eli sentroidiin. Matemaattisesti ratkaisu saadaan soveltamalla singulaariarvohajotelmaa, ja tulokseksi saadaan taso joka on lähimpänä pistepilviä. Etäisyyttä mitataan massoilla painotetulla khii2-etäisyysmitalla.

k Intuiitiivisesti 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 hajonta on jaettu näille uusille akseleille. tavoitteena on muutaman dimension approksimaatio moniulotteiselle datalle, yleensä kaksiulotteinen 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 heuristisesti 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 modaliitteja. Tärkein periaate on se, että kaikki on suhteellista.

Ydinkäsitteitä ovat *korrespondenssianalyysin ”tripletti”*: *khii2-etäisyys*, *massat* ja *profiilit*.

Kolmikkoo täydentää ”kvartetti”, neljä siitä johdettua käsitettä: *inertia* eli (painotettu) varianssi, *sentroidi* (painotettu keskiarvo, barysentriinen periaate), *aliavaruus* ja *projektio*. (CAiP, s. 49).

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

k khii2-etäisyys on profiilien painotettu euklidinen etäisyys (ja toki neliöjuuri!) jossa painoina ovat keskiarvo-profiilin elementtien käänteisluvut eli elementtien etäisyyden neliö jaetaan keskiarvoprofiilin alkiolla.

edit: khii2-etäisyydestä ehkä teorialiitteeseen?

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

k khii2-etäisyys on ainoa etäisyysmitta, 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 odotusarvo ja hajonta ovat sama parametri. Tämä tulkinta khii2-etäisyydelle ei kuitenkaan saisi hämärtää massojen kaksoisroolia: ne ovat profiilien painoja ja samalla standardoivat khii2-etäisyyden.

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

Millaista dataa?

Korrespondenssianalyysin sovelletaan yleisimmin frekvenssitaulujen analyysiin, lukumäärädataan (count data). Periaatteessa mikä tahansa data sopii, kunhan se voidaan järkevästi esittää suhteelisinä lukumäärinä (relative amounts), siis suhteasteikon (ratio scale) muuttujana. Tässä oleellista on tulkittavuus tutkimusongelman näkökulmasta. Välttämätön ehto on sama mittayksikkö: lukumäärä, rahayksikkö, pituusmitta kelpaavat. (CAiP s. 15). Taulukon lukujen on oltava ei-negatiivisia (positiivisia, nolla sallittu).

Rajat ovat joustavia, 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 profiilien etäisyyksiä mitataan khii2-etäisyysmitalla. Jos datan voi esittää tässä muodossa, menetelmää voi käyttää.

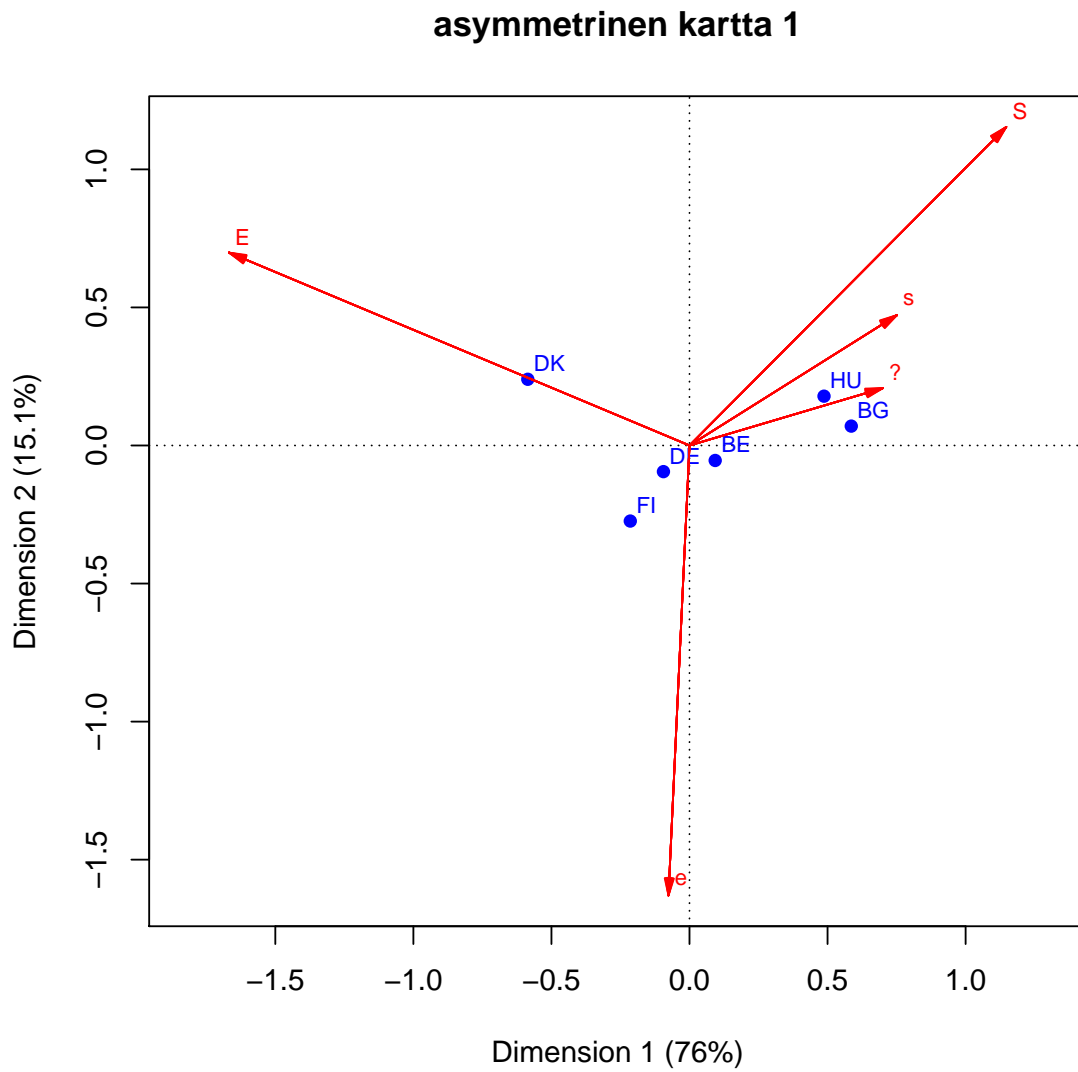
3.3.1 Asymmetrinen kartta ja ideaalipisteet

Symmetrinen kartta 3.3 on peruskuva ja esimerkiksi tässä käytetyn R-paketin ”ca” oletus. Siinä molemmat pisteparvet on esitetty pääkoordinaateissa (prinipal coordinates) ikäänkuin päällekkäin, samassa kuvassa.

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

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

Sarakepisteet kuvaavat maksimi-inertiaa, ja rivipisteiden paljon pienempi hajonta kuvaa niiden poikkeamaa tästä hypoteettisesta tilanteesta. Sarakepisteet skaalautuvat origosta ulospäin, Asymmetrisessä kartassa rivi-



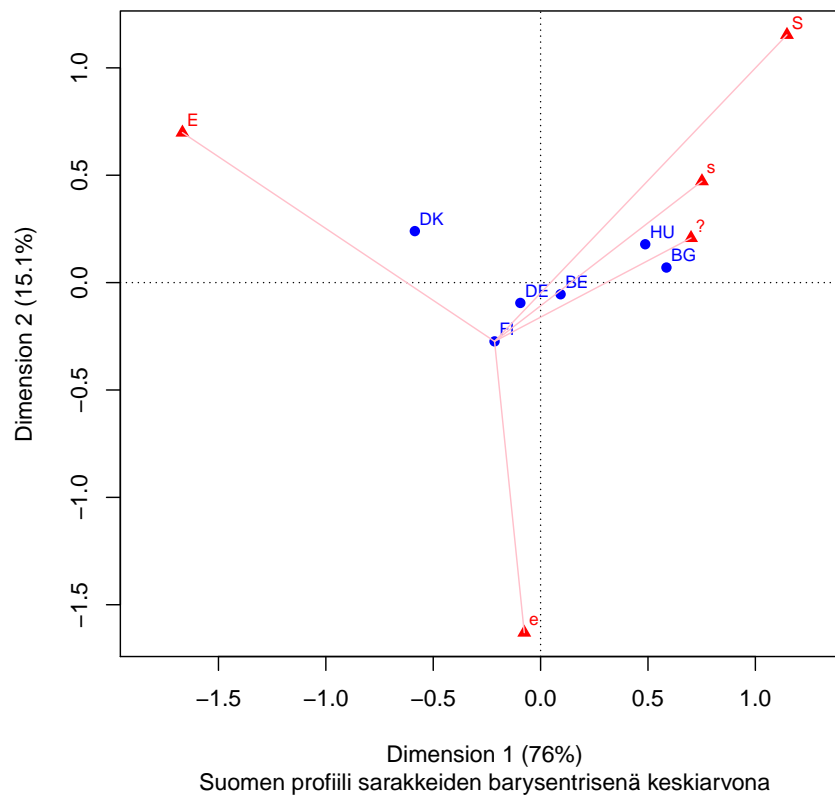
Kuva 3.4: Q1b: lapsi kärsii jos äiti on töissä

ja sarakepisteiden etäisyydellä on tulkinta, samoin rivipisteiden välisellä etäisyydellä. Sarakepisteiden välisillä etäisyyksillä ei ole tulkintaa. Sarakepisteet on skaalattu ja mittakaavan ero symmetriseen karttaan näkyy selvästi.

3.3.2 Barysentrinen periaate

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

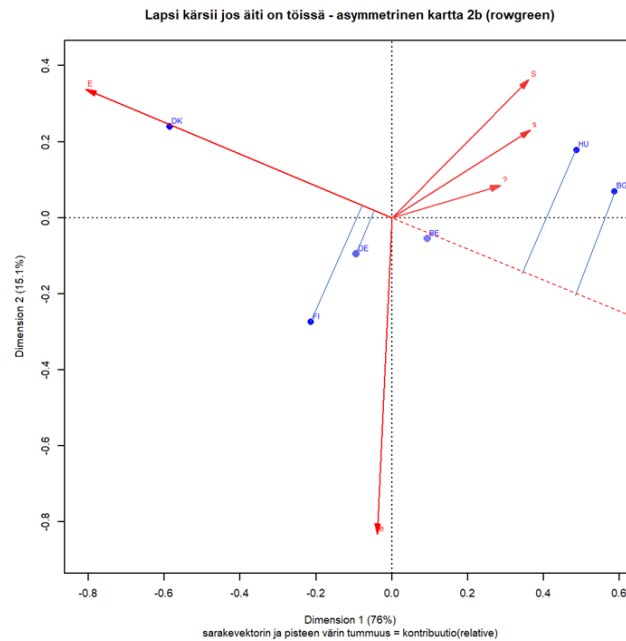
edit kuva ei ehkä tarpeen? Tehdään vähän pienempi (out.width = 60%, muuten 90%).



Kuva 3.5: Q1b: lapsi kärsii jos äiti on töissä

Suomen profiili 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 piirretylle 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 akselien kuvaamasta inertiaasta. Absoluuttinen kontribuutio kertoo rivin tai sarakkeen osuuden akselin inertiaasta. Vaikutuksessa on mukana pisteen massa.

Suhteellinen kontribuutio taas kertoo akselin osuuden pisteen inertiaasta. Tämä tunnusluku kuvaa pisteen projektion laatua, kuinka hyvin se on kartalla esitetty.

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

Absoluuttiset kontribuutiot

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

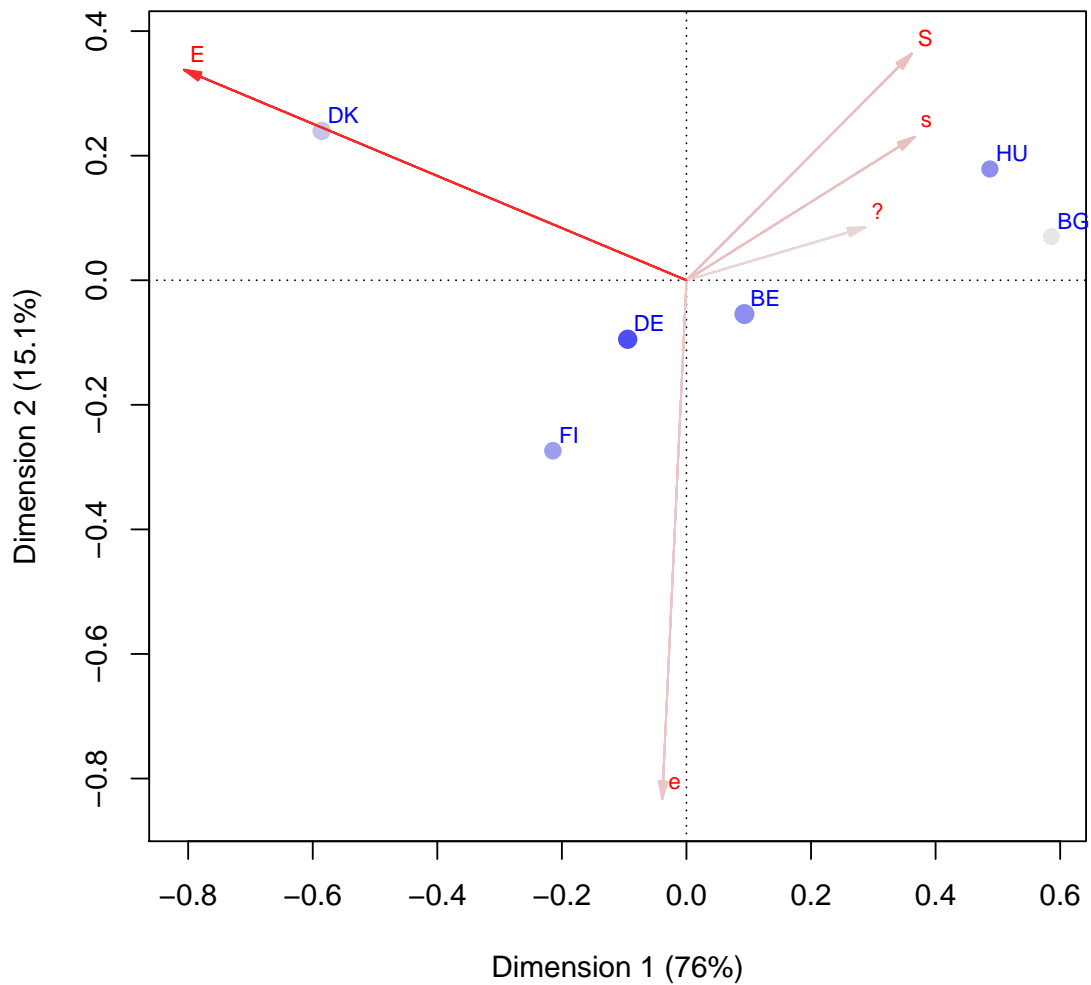
Käsitteisiin palataan tarkemmin seuraavissa luvuissa ja teorialiitteessä, ja liian tarkkaan karttaa ei kannata tutkia. Numeeriset tulokset ovat yksityiskohdissa selkeämpiä.

edit käytän termiä “vektori” vain kuvaan piirretyn “nuolen” nimityksenä.

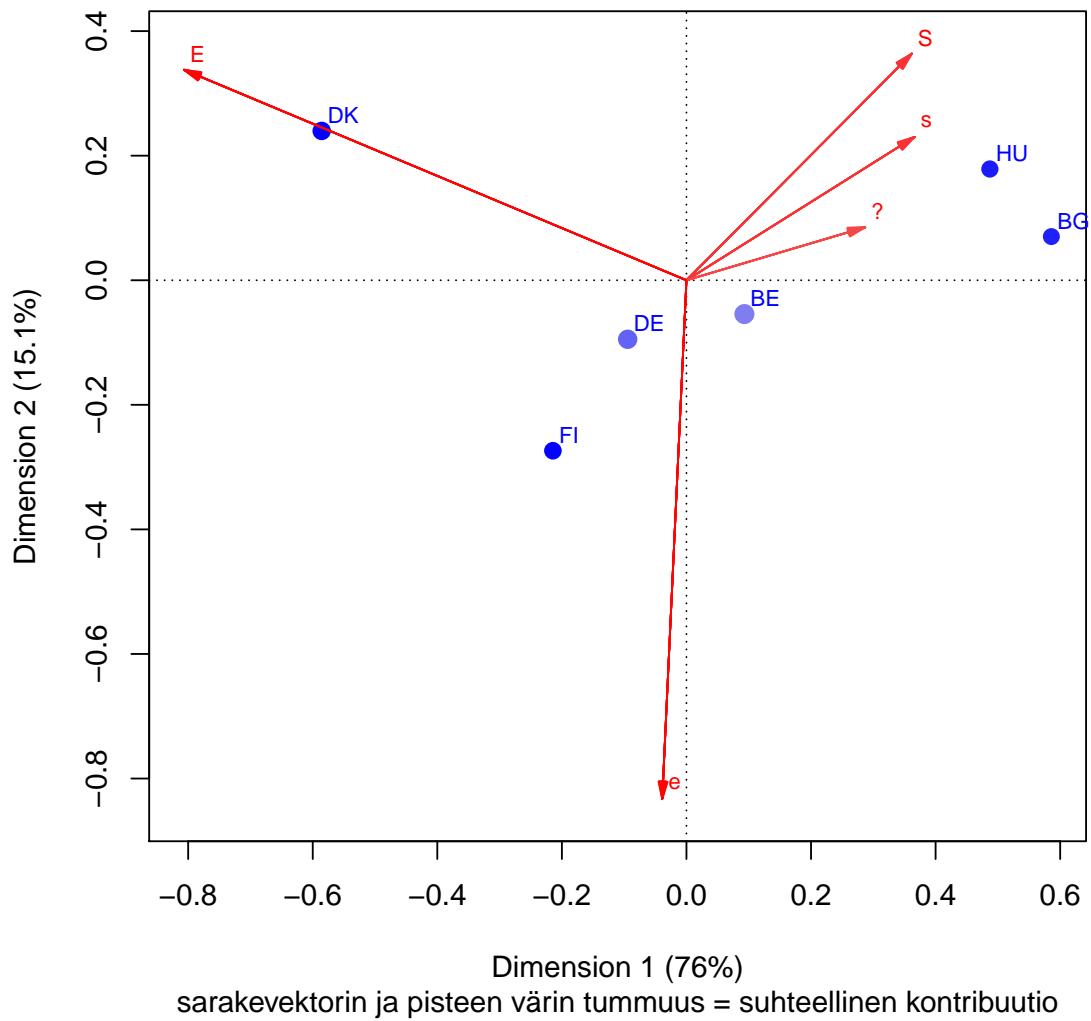
Sarakkeista ratkaisuun vaikuttaa selvästi eniten E, ja juuri ensimmäiseen dimensioon. Toista dimensiota määrittää vahviten e, mutta myös kaikki muut sarakkeet x-akselin yläpuolella. Samaa mieltä olevien (S ja s) vaikutus näyttäisi jakautuvan selvimmin molemmille dimensioille.

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

kontribuutiokartta 1 – pisteen koko suhteessa massa



Kuva 3.6: Q1b: lapsi kärsii jos äiti on töissä

kontribuutiokartta 2 – pisteen koko suhteessa massa

Kuva 3.7: Q1b: lapsi kärsii jos äiti on töissä

Suhteelliset kontribuutiot

Kaikki edellä esitetyt päättelyt perustuvat tietysti kaksiulotteiseen projektioon. Jos pisteet on esitetty hyvin eli niiden inertiaasta (poikkeamasta keskiarvosta) suuri osa on kuvattu kartalle, rivipiste on sitä lähempänä ideaalipistettä mitä suurempi ideaalipisteen osuus on sen profiilissa.

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

3.3.4 Massat

Massat ovat korrespondenssianalyysin keskeinen käsite, ja niiden kaksoisrooli on menetelmän ytimessä. Massat ovat normalisoiva muunnos 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 vakioda yhtä suuriksi osajoukoissa, ratkaisu on yksinkertainen. Korrespondenssianalyysin taulukoksi otetaan riviprofiilitaulukko, 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 maltillista 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 skaalauksella. Korrespondenssianalyysi on inertian eli kokonaishajonnan dekomponointia, jakamista osiin.

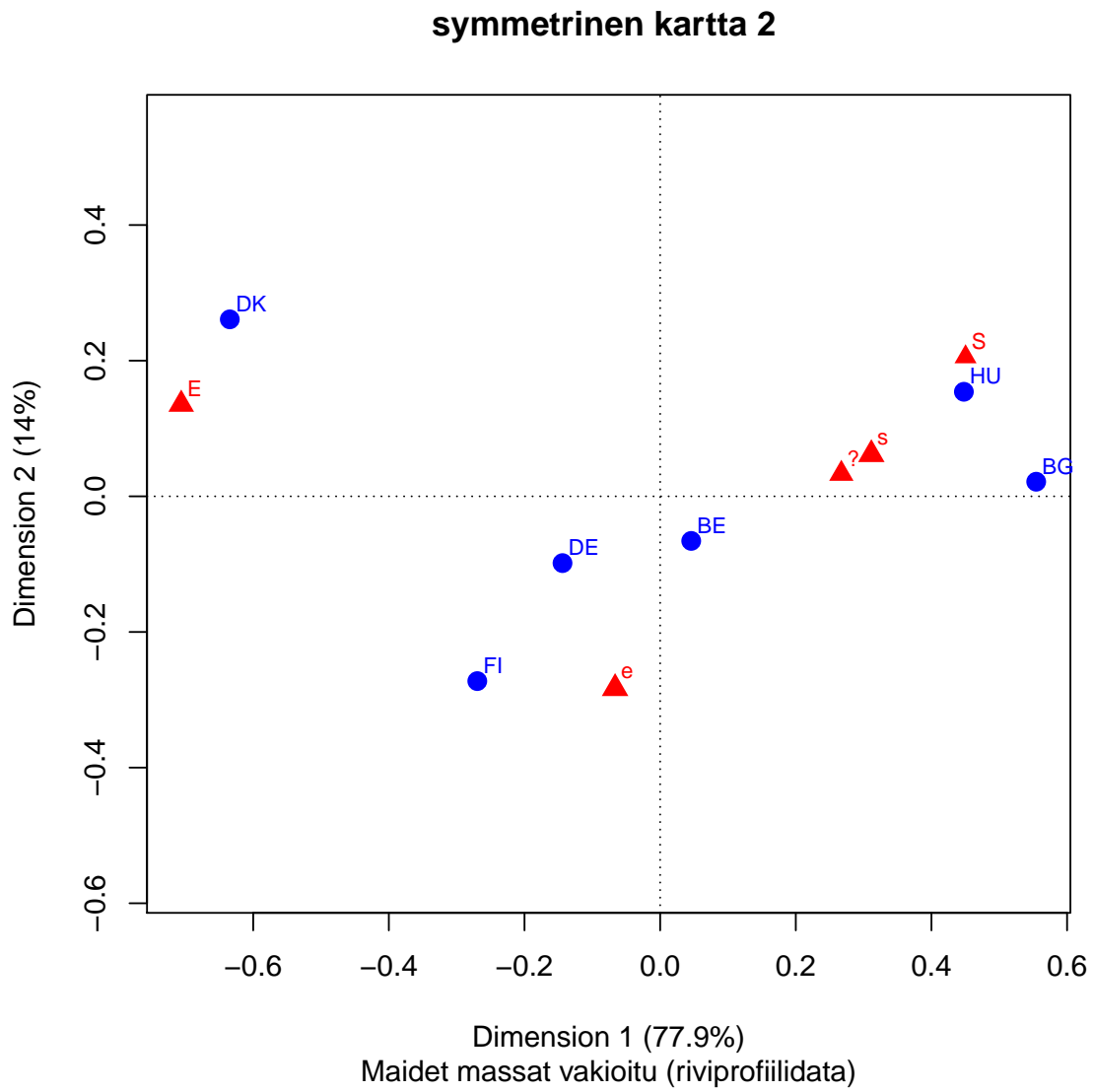
3.3.5 Karttojen erot

Yksinkertaisen korrespondenssianalyysin peruskuvana on symmetrinen kartta. Ehkä yllättäen sen “...tulkinta on edelleen menetelmän kaikkein kiistanalaisin aspekti.” (CAiP s.295, ?.)

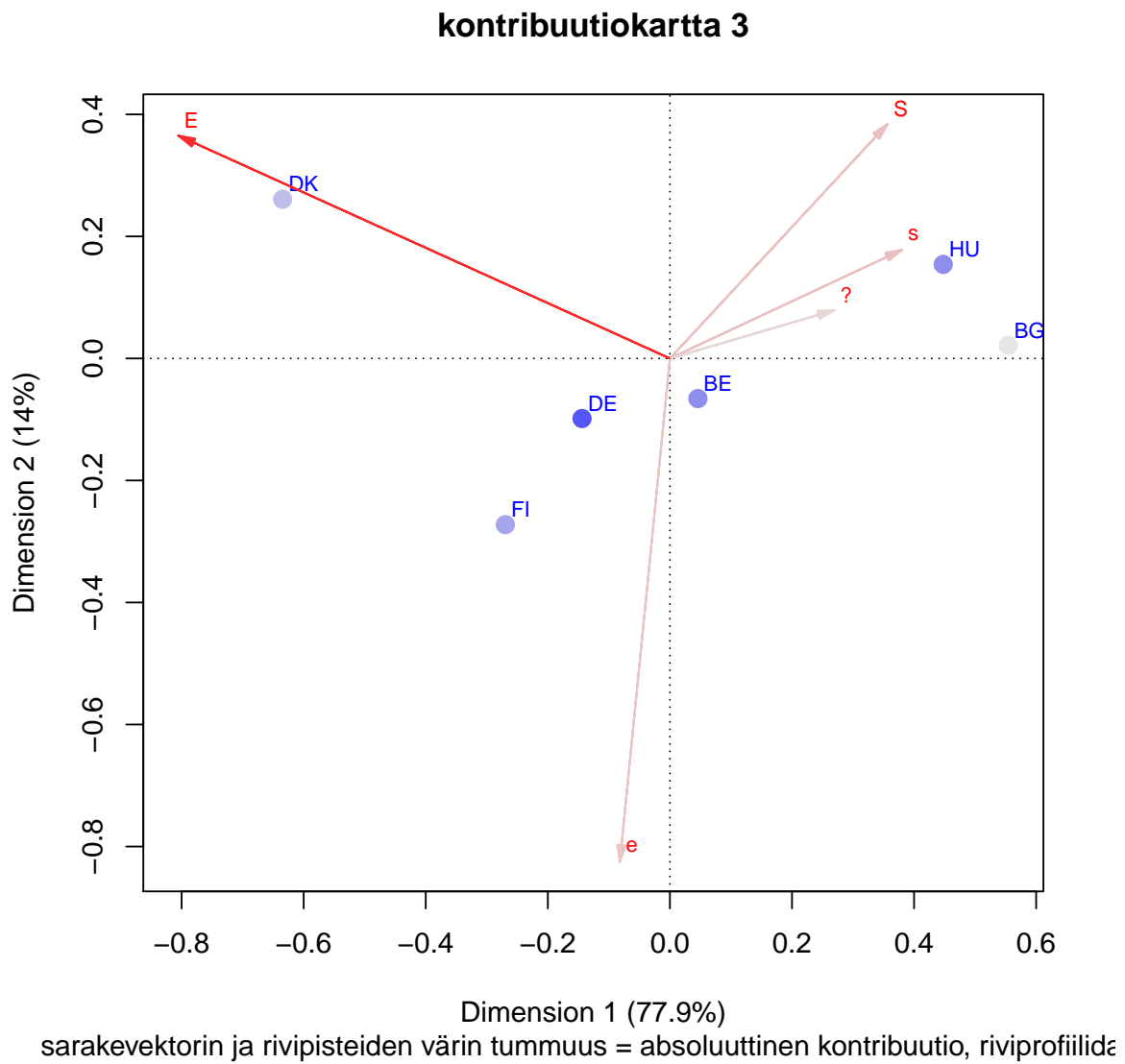
Sarake- ja rivisteet esitetään siinä ikään kuin päällekkäin, samassa koordinaatistossa. Niiden pääkoordinaatit ovat kuitenkin eri pistejoukoista tai avaruuksista. Asymmetrisessä kartassa pisteet ovat samassa avaruudessa, ja ero on Greenacren mukaan vain skaalaus. Asymmetrisessä kartassa standardikoordinaateissa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

Jos pääinertioiden neliöjuuret ovat hyvin eri suuruisia, tulkintaongelmia voi tulla, mutta niillä ei käytännössä ole merkitystä. Siksi hän pitää skaalausdebatia akateemisena kiistanä, käytännön sovelluksissa sillä ei ole merkitystä. Kiista on ollut aika sitkeä (esimerkiksi 1989 Greenacren kommentoi skaalausta perusteellisesti (?)), mutta lienee laantunut.

Symmetrinen kartta hyvä vaihtoehto, sillä asymmetrisessä skaalaus vie ideaalipisteet usein kauas pääkoordinaateissa esitetyt 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.



Kuva 3.8: Q1b: lapsi kärsii jos äiti on töissä



Kuva 3.9: Q1b: lapsi kärsii jos äiti on töissä

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. skalaaritulo-ominaisuus. Rivi- ja sarakepisteen skalaaritulo ”palauttaa” alkuperäisen datan, tässä tapauksessa taulukon solun.

Näistä vain kaksi voidaan optimaalisesti esittää yhtä aikaa.

Symmetrisessä kartassa khii2-etäisyydet 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 erisuuruisia.

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

Kontribuutiokartta on muunnelmä asymmetrisestä kartasta. ”Ääriprofilit” vedetään kohti origoa kertomalla ne massojen neliöjuurilla. Näin kuva selkenee, ja ”kutistetun” pisteen etäisyys origosta (”vektori”) kertoo sen kontribuution pääakseleille. Näiden pisteiden välisillä etäisyyksillä ei ole suoraa tulkintaa.

Jako standardi- ja pääkoordinaatteihin on suora seuraus korrespondenssianalyysin matemaattisesta ratkaisusta. Greenacre esittelee kaksoiskuvia käsittelevässä kirjassaan (?) selkeästi koordinaattien yhteyden ratkaisualgoritmiin, singulaariarvohajoitelmaan.

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

$$p\text{koordinaatit} = \text{standardikoordinaatit} \times \sqrt{p\text{akselien inertiat}} \quad (3.7)$$

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

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 lisätä kuvaan, jos se on järkevästi vertailukelpoinen kartan määrittäneiden 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 profiili, joka on jollain lailla sisällöllisesti erilainen kuin muut. Esimerkkiaineistossa kartalle voisi lisätä joitain Euroopan ulkopuolisia maita. Vaikka nämä riviprofiilit eivät vaikuta kartan akselien määräytymiseen, ne voidaan esittää kuuden maan määrittämässä “avaruudessa”. Projektion laatu (suhteelliset kontribuutiot) voidaan myös esittää.

Toinen käyttötapaus on pienen massan profiili. 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.


Kolmas mahdollisuus on jakaa pistejoukkoja osajoukkoihin ja esittää niiden summaprofiili täydentävänä pisteenä. Summaprofiili on osiensa painotettu (barysentriinen) keskiarvo. Kun se esiteteään passiivisena pisteenä, havaintoja ei oteta ratkaisuun kahta kertaa. Profiilien yhdistämiseen liittyy korrespondenssianalyysin tärkein periaate, jakaumaekvivalenssi (*distributional equivalence*). Profileiltaan samanlaiset rivit voidaan yhdistää, analyysin tulokset eivät muutu. Khii2-etäisyysmitta on ainoa etäisyysmitta joka toutettaa tämän periaatteen. En esittele tätä ydinkäsitettä tämän enempää, se on ollut menetelmän kehittämisessä tärkein tavoite. ? esittelevät menetelmän matemaattiset perusteet laajemmin.

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).

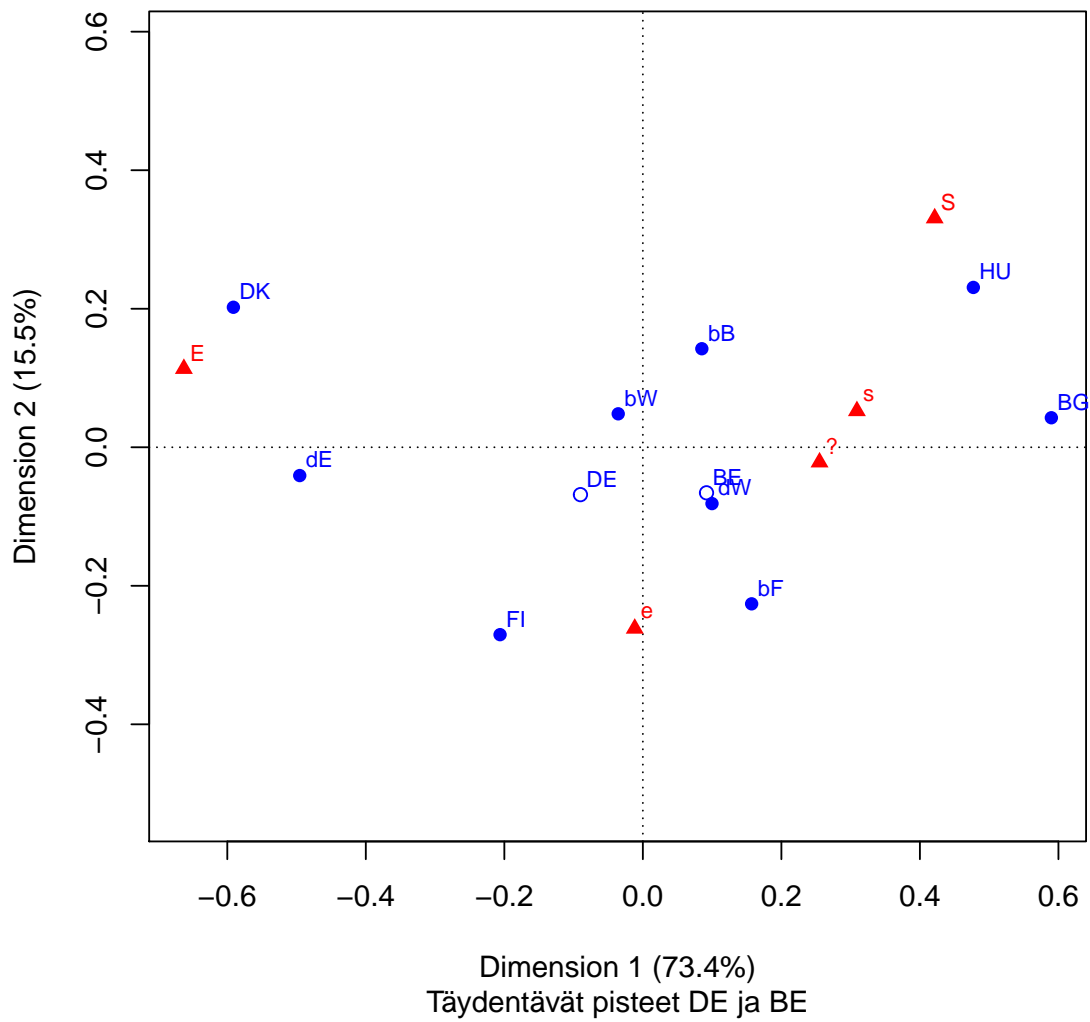
Aineistoon lisätään passiivisina riveinä Saksan ja Belgian maaprofiilit (DE, BE). Maiden massoja ei skaalata yhtä suuriksi, otoskoot vaikuttavat ratkaisuun.

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. Jos karttaa vertaa kuvaan  ei eroja juuri ole. 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 (bW) on liberaalilla

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

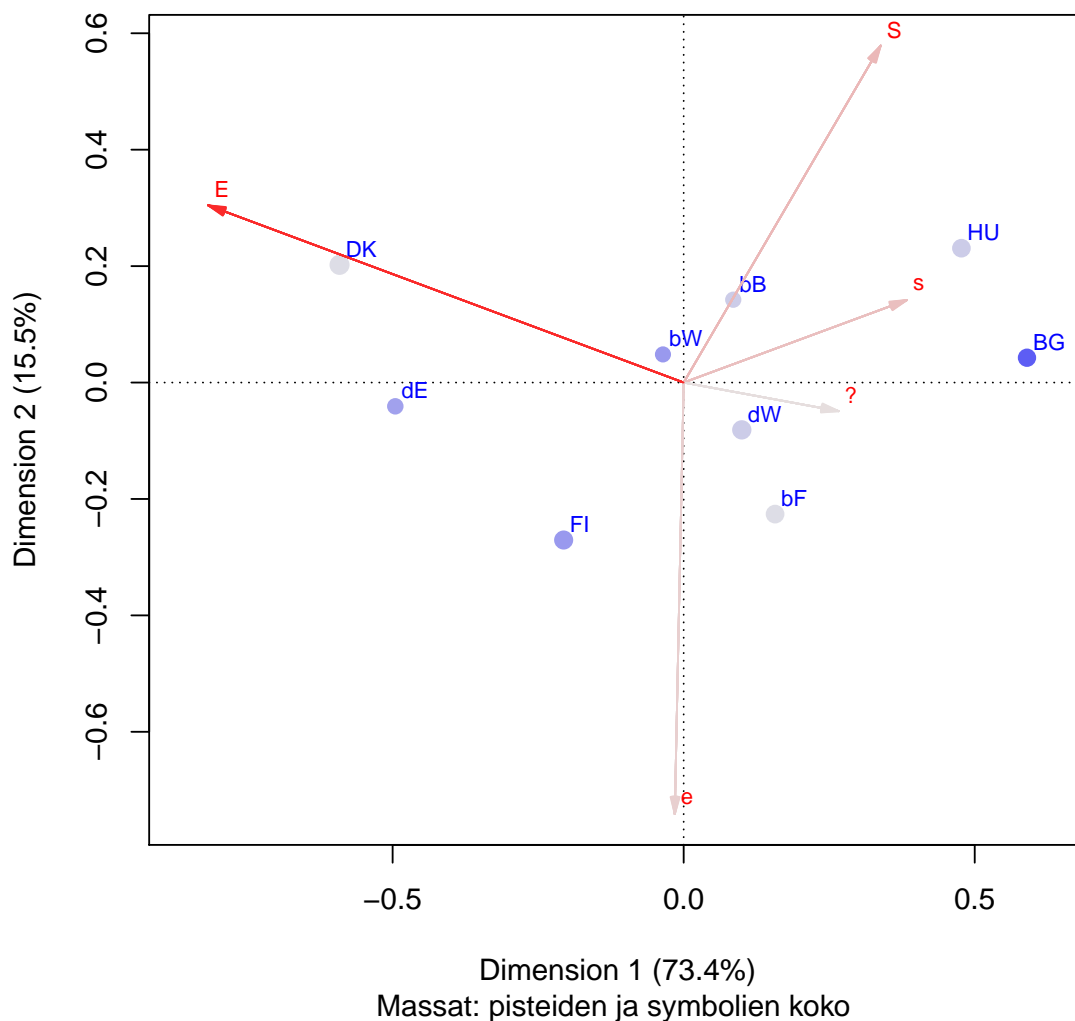


Kuva 4.1: Q1b: Saksan ja Belgian aluejako

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

Bryssel ja Wallonia näyttävä olevan hyvin lievästi U-muotoisen maapisteen 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 kaarieffektin taustalla on järjestysasteikon muuttujan korrelaatio (viite: LeRoux). Kaaren sisäpisteet ovat usein polarisoituneita ensimmäisen dimension “ääripäävastausten” 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-Saksa 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 kuvan ratkaisusta. Nämä tulokset ovat erilaisia kokonaisinertian dekomponentteja. Kokonaisinertia (total inertia) profiilien ja keskiarvoprofiilin khii2-etäisyyksien massoilla painotettu summa ((3.6). Se kuvaa profilipisteiden hajontaa ideaalipisteiden verteksin sisällä. Maksimi-inertia saavutetaan kun profiilit ovat verteksin kärkeissä, jokaisessa profilissa on vain yksi luokittelumuuttujan 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.

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).

```
##
## 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 joukossa kerrotaan pisteen massa, laatu (qlt) ja inertia-kontribuutio.

Inertia-kontribuutio on suhteellinen osuus kokonaisinertiasta. Aktiivisia rivejä on 9, joten tasaisesti jaettu inertia olisi noin 110. Tanska, Bulgaria ja Unkari “selittävät” 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 inertiasta on esitetty kartalla. Kaksiulotteinen kartta kuten tässä on yleisin valinta, laatu kerrotaan valitulle dimensioiden määrälle. 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 kehoitteen esitetty, ja vain Flandersin laatu on kohtuullisen hyvä. Kovin hyvä ei ole täydentävien maapisteidenkään laatu.

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

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

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 the point to the inertia of the axis -> high relative contribution of the axis to the inertia of the point. Ei päde kääntäen. ”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

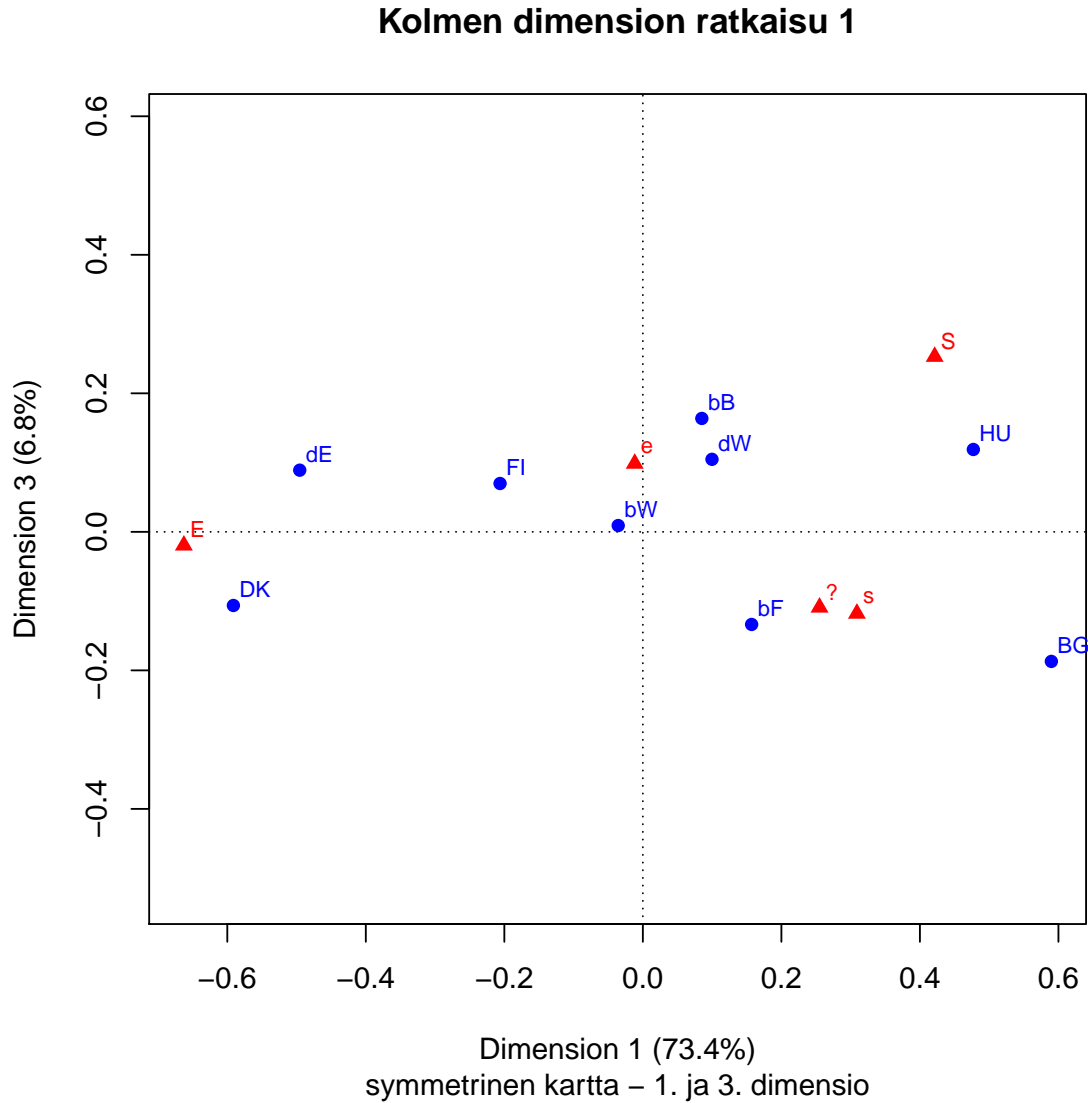
k Ei kovin hyviä kuvia, mutta periaate on tärkeä. Kartta on approksimaatio, pitää päättää milloin se on tarpeeksi hyvä. Tai mille pisteille hyvä, mille huonompi.

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

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

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

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



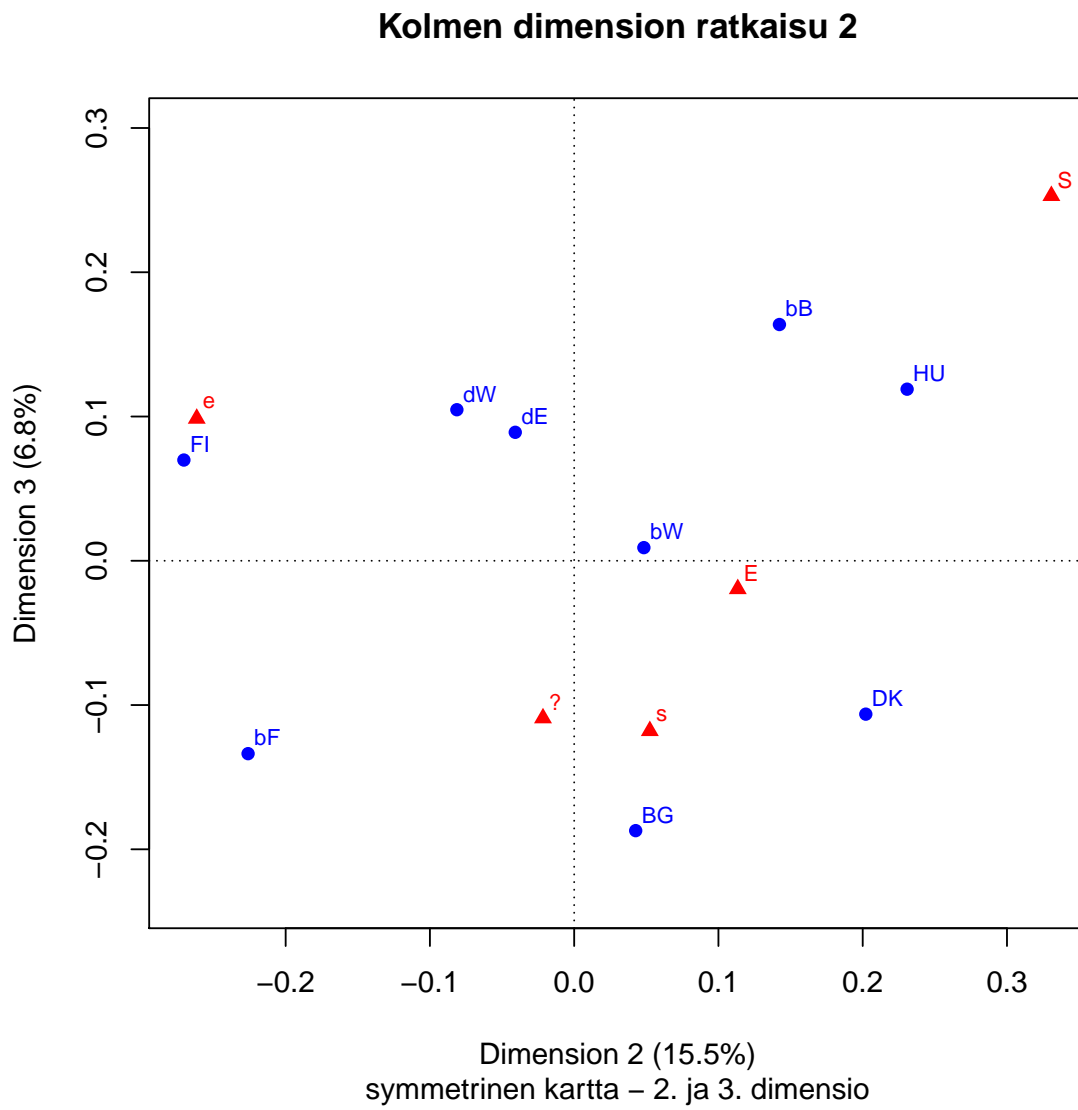
Kuva 4.3: Q1b: Saksan ja Belgian aluejako

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

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

Kahdesta projektiosta 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 kokonaisinertiasta on noin seitsemän prosenttia. Belgian alueiden ero näkyy

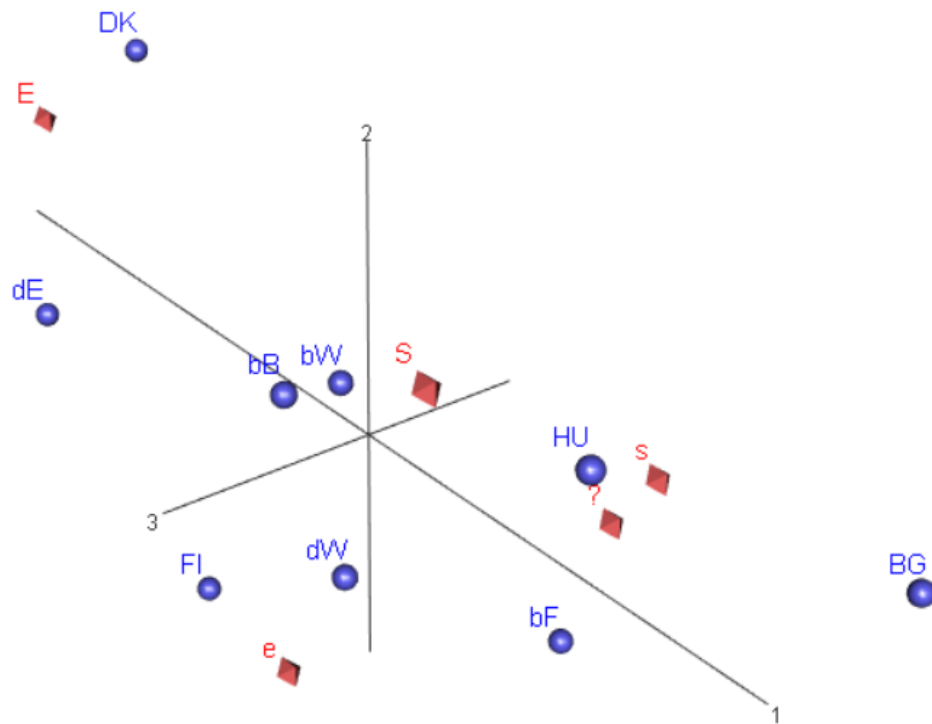


Kuva 4.4: Q1b: Saksan ja Belgian aluejako

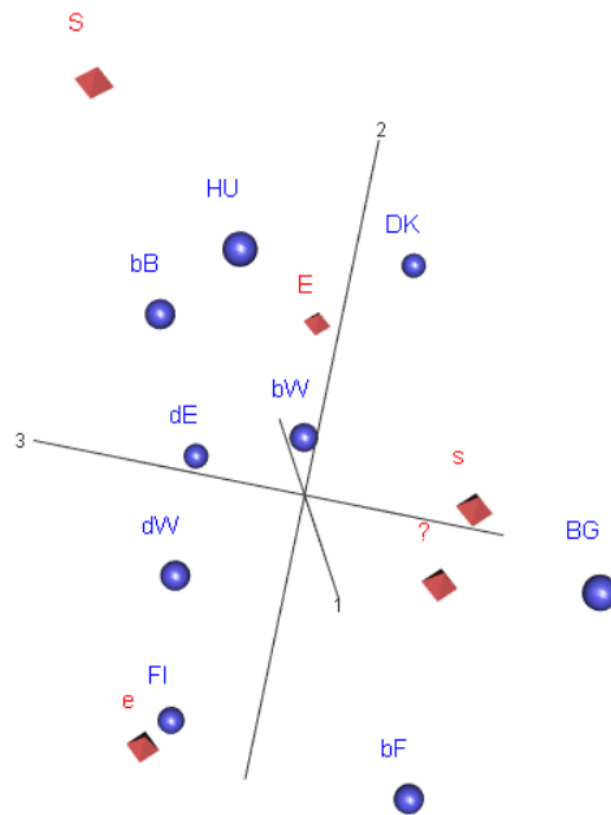
erityisesti kartassa 4.4 ja 4.6, samoin E-sarakkeen kohtalainen muutos.

edit - toinen ehkä riittää?

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



Kuva 4.5: Saksan ja Belgian aluejako - 3d-kuva1



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

Luku 5

Yhteisvaikutusmuuttujat

Yksinkertaisin tapa tutkia taustamuuttujien yhteisvaikutuksia on yhdistää kaksi muuttujaa uudeksi luokitelu-muuttujaksi ("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. Finanssikriisin 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 yhteisvaikutusmuuttajalla, mukaan myös maa. Käytännössä kolmen luokittelumuuttujan yhdistäminen tekee taulukosta jo hieman huteron, joissain soluissa havaintojen määrä pienee. 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.

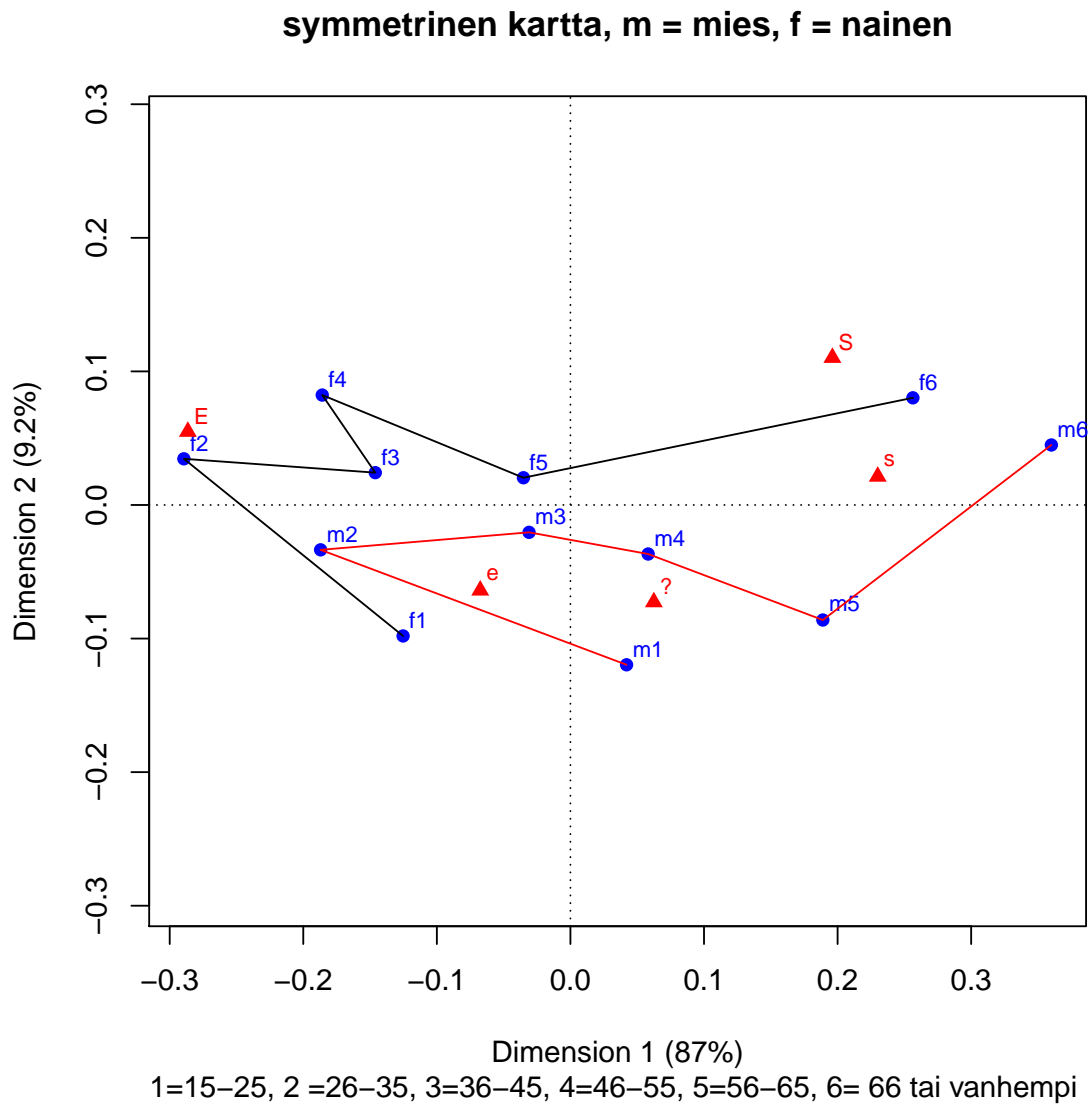
5.1 Ikä ja sukupuoli

Ikäjäkauma painottuu kaikissa maissa jonkinverran 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 pisteet voi yhdistää nuorimmasta vanhimpaan.

Ratkaisu on melko yksiulotteinen, ensimmäinen dimensio kuvaa 87 prosenttia kokonaisinertiasta. Dimenioiden 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 kokonaisinertiasta (sarake inr) on 768. Niiden kontribuutio x-akselille on yhteensä vielä suurempi (849). Muut sarakkeet taas kontribuoivat y-akselin inertiaan, mutta sen osuus kokonaisinertiasta 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 melko saman kokoisia, mutta kolmen ryhmän (f2, f6 ja m6) osuus kokonaisineratiasta on 616 ja niiden kontribuutio ensimmäiselle dimensiolla 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 kokonaisineratiasta 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.

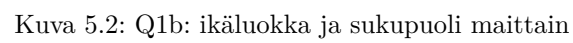
Selvästi kaikissa ikäluokissa miehet ovat konservatiivisempia kuin naiset. Nuorin ikäluokka on hieman vähemmän varma mielipiteistään kuin vanhin. Yksi mahdollinen selitys kartan tulkinnan ongelmille on se, että maiden väliset erot mielipiteissä ovat paljon suurempia kuin sukupuolten väliset maiden sisällä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

5.2 Ikä, sukupuoli ja maa

Kuvasta saa jotenkin selvää, kun sen suurentaa mutta pisteitä on selvästi liikaa. Joitain muuttujien nimiä voisi lyhentää, kuva-alaa voisi rajata joihinkin osiin mutta osajoukon korrespondenssianalyysi tarjoaa pätevimmän vaihtoehdon.

Sarakkeiden järjestys vasemmalta oikealle ja ylhäältä alas on sama kuin edellisissä kartoissa. Dimenisoiden tulkinta on sama, osuus inertiasta pienenee x-akselilla noin 6 prosenttiyksikköä. Pisteiden järjestys liberaalista konservatiiviseen alkaa Tanskan ja Suomen pisteistä, sitten tulee Saksan ja Belgian pisteitä ja konservatiivisimpia ovat oikeassa laidassa Unkarin ja Bulgarian osajoukot. Toisella akselilla maltillisia ja neutraaleja ovat hyvin karkeasti Suomen pisteet ja lähes kaikki Saksan ja Belgian pisteet. Eri maiden osajoukkojen suhteita on hankalampi hahmottaa, erityisesti kartan oikealla laidalla.

Numeeristen tulosten taulukko on pitkä, mutta kartan informaatio pitää tarkistaa. Numeeriset tulokset eivät ole pelkkää diagnostiikkaa ja kartatan esittämien riippuvuuksien varmistamista. Niistä näkee myös tarkemmin mahdolliset kiinnostavat piirteet datassa. Regeressiomallien tulosten raporteissa diagnostiikka on enintään liitteenä, mutta eksploratiivisessa data-analyysissä se ohjaa analyysiä myös eteenpäin.



Kuva 5.2: Q1b: ikäluokka ja sukupuoli maittain

Tässä voi nähdä myös todennäköisyysteoriaan perustuvan tilastollisen mallintamisen vahvan puolen, aineiston rakenne ja muuttujien yhteydet saadaan parhaassa tapauksessa esitettyä paljon tiiviimmin.

```
##
## 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 | FI f1 | 12 980 11 | -417 693 11 | -269 287 21 |
## 50 | FI f2 | 12 927 26 | -730 907 34 | -110 21 4 |
## 51 | FI f3 | 12 984 13 | -423 590 11 | -346 394 36 |
## 52 | FI f4 | 14 991 14 | -398 644 12 | -292 347 32 |
## 53 | FI f5 | 17 952 8 | -240 502 5 | -227 450 23 |
## 54 | FI f6 | 11 835 7 | 151 134 1 | -347 701 35 |
## 55 | FI m1 | 7 787 5 | -115 78 1 | -347 710 22 |
## 56 | FI m2 | 9 977 14 | -598 832 17 | -250 146 14 |
## 57 | FI m3 | 9 998 6 | -345 629 6 | -265 369 16 |
## 58 | FI m4 | 13 837 6 | 19 3 0 | -316 834 33 |
## 59 | FI m5 | 12 734 7 | 220 289 3 | -273 446 23 |
## 60 | FI m6 | 9 911 6 | 336 637 6 | -220 274 12 |
## 61 | HU f1 | 7 723 9 | 499 698 9 | 93 25 1 |
## 62 | HU f2 | 11 689 11 | 438 685 11 | -35 4 0 |
## 63 | HU f3 | 12 808 18 | 484 586 15 | 298 222 27 |
## 64 | HU f4 | 11 768 18 | 491 564 15 | 296 204 25 |
## 65 | HU f5 | 12 850 13 | 474 753 14 | 170 97 9 |
## 66 | HU f6 | 13 671 34 | 637 581 28 | 251 90 21 |
## 67 | HU m1 | 6 935 5 | 426 766 6 | 201 170 6 |
## 68 | HU m2 | 9 381 11 | 344 381 6 | -2 0 0 |
## 69 | HU m3 | 13 957 12 | 441 803 13 | 193 154 12 |
## 70 | HU m4 | 10 999 10 | 468 830 12 | 211 169 11 |
## 71 | HU m5 | 13 942 12 | 472 891 15 | 113 51 4 |
## 72 | HU m6 | 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 kohtalaisen hyvin esitetty, heikoimmin neutraali vaihtoehto (535). Kun sen suhteellinen kontribuutio (cor) on vain 1 toisella dimensiolla jää loppuosa x-akselille. Maltillisuuden dimensiota määrittää e-sarake (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 dimenisolle maltilliseen suuntaan. Tanska taas kontribui vahvasti jyrkempien mielipiteiden suuntaan.

Kaikissa taulukon soluissa on vähintään viisi havaintoa, muutama pienen massan havainto kontribui 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 määrittävät 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.

Korrespondenssianalyysiä ja erityisesti khii2- etäisyysmittaa on arvosteltu siitä, että se on liian herkkä harvinaisille luokittelumuuttujan arvoille. Yhteenvetoartikkelissaan ? tarttuu ”vaikuttavien poikkeavien havaintojen myyttiin”, ja pitää sitä lähes aina perusteettomana.

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

Harvinaisten kategorioiden vaikutus voi olla suuri, joten numeerisista tuloksista on tarkistettava onko hyvin pienen massan pisteillä suuri kontribuutio ratkaisuun. Käytännössä näin voi käydä esimerkiksi silloin, kun jonkun harvinaisen luokittelumuuttujan arvon havainnot ovat keskittyneet muutamaan profiliin joissa niiden osuus on suuri (CAiP, s 298). Luvussa 7 nähdään, miten melko vähäinen määrä puuttuvia vastuuksia kasaantuu samaan vastaajien osajoukkoon ja mitä seurauksia sillä on.

Stabiiliutta voi helposti kokeilla määrittelemällä joitain pisteitä täydentäviksi pisteiksi.

Kartan 5.2 numeerisista tuloksista ei löydy pienen massan pisteitä joilla on merkittävä kontribuutio akseleihin.

Luku 6

Osajoukon korrespondenssianalyysi

Graafisessa data-analyysissä kuvien on oltava selkeitä, mutta korrespondenssianalyysin ja monimuuttujakorrespondenssianalyysin kartat ovat usein liian täynnä pisteitä. Ongelmaa voi lievenätää jättämällä pois ratkaisuun vain vähän vaikuttavia pisteitä, keksimällä mahdollisimman lyhyitä symboleja muuttujille tai rajaamalla kuvaa. Ongelma on kuitenkin syvempi, usein kartta kertoo aika yllätyksettömän ja ilmeisen tarinan. Kiinnostavimmat yhteydet pysyvät piilossa ylemmissä dimensioissa. MCA-kartan perusongelma on se, että siinä yritetään esittää monia erityyppisiä yhteyksiä simultaanisesti ja nämä yhteydet eivät ole “isolated to particular dimensions” (? , 198).

Osajoukon korrespondenssianalyysi (subset CA, subset MCA) on yksi vastaus tähän pulmaan. Teoreettiset perusteet on esitetty Greenacren ja Pardon artikkelissa (?) ja sen laajennetussa versiossa (emt.), esimerkkitaineistona ISSP:n 1994 data. Selkeä oppikirjaesitys on CAiP (ss. 161-).

Eräs selkeä sovelluskohde yhteiskuntatieteellisissä kyselyaineistoissa on puuttuvien vastausten analyysi, johon palataan seuraavassa luvussa.

Osajoukon korrespondenssianalyysin idea on säilyttää koko aineiston massat ja khii2-etäisyyksien painot mutta analysoida vain osaa aineistosta. Koko aineiston sentroidi säilyy kartan keksipisteenä. Osajoukkojen inerttioiden summa on koko aineiston inerttia.

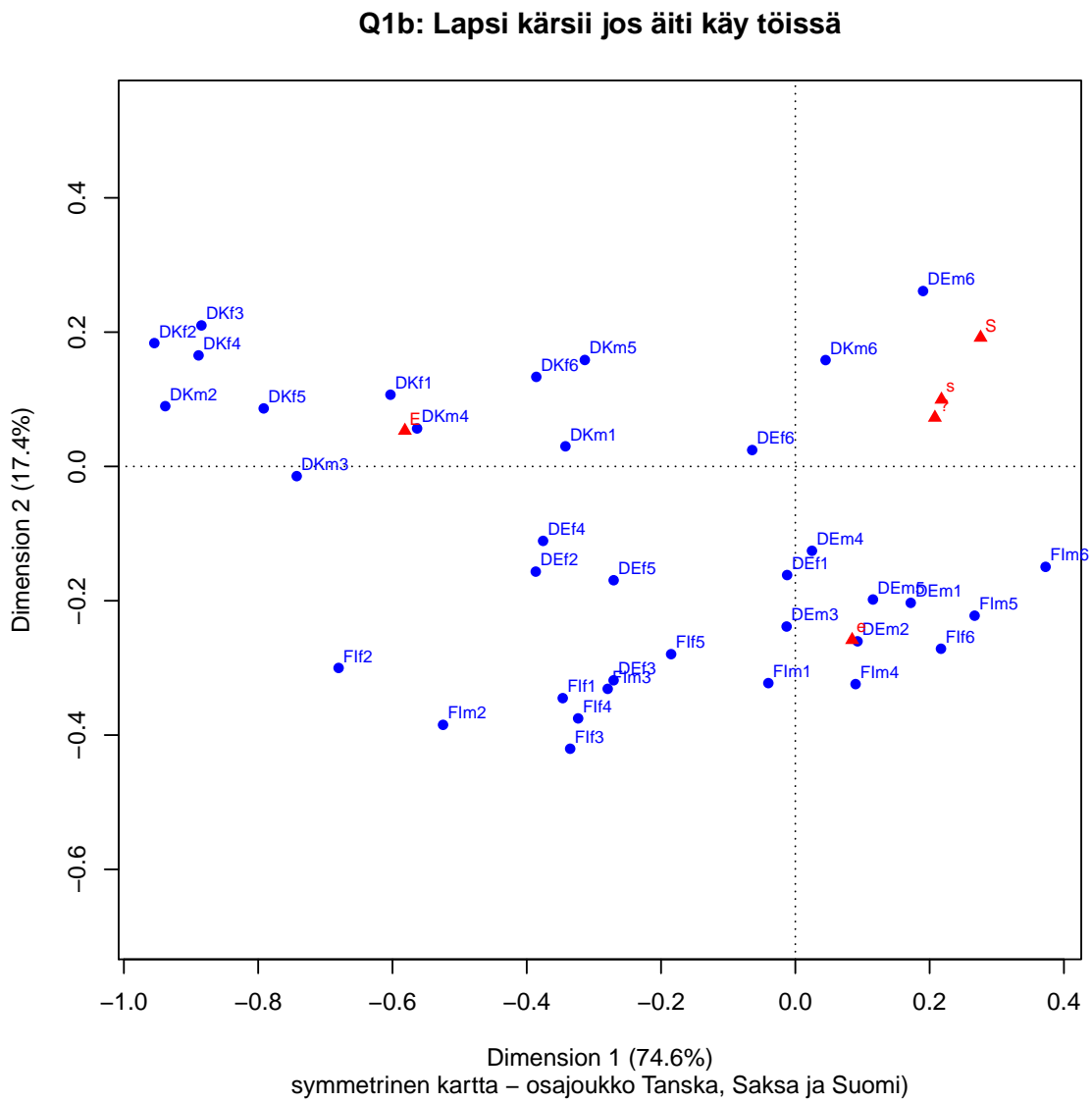
Osajoukon voi valita havaintojen tai muuttujien suhteen. Täydentäviä pisteitä voi helposti lisätä kartalle , jos ne eivät kuulu siihen joukkoon josta osajoukko on valittu. Osajoukon profiilit muuttuvat, niiden summa ei enään ole yksi ja barysentristä periaatetta ei voi suoraan käyttää täydentävän pisteen koordinaattien laskemiseen. Tässä esimerkissä emme voi suoraan ca-paketin avulla sijoittaa esimerkiksi maapisteitä kartoille.

Kartan 5.2 avulla on helppo jakaa aineisto aluksi vain kahteen ryhmään. Suomi, Tanska ja Saksa ovat pääakselin oikealla puolella. Bulgaria ja Unkari yhdessä Belgian kanssa ovat toinen ryhmä.

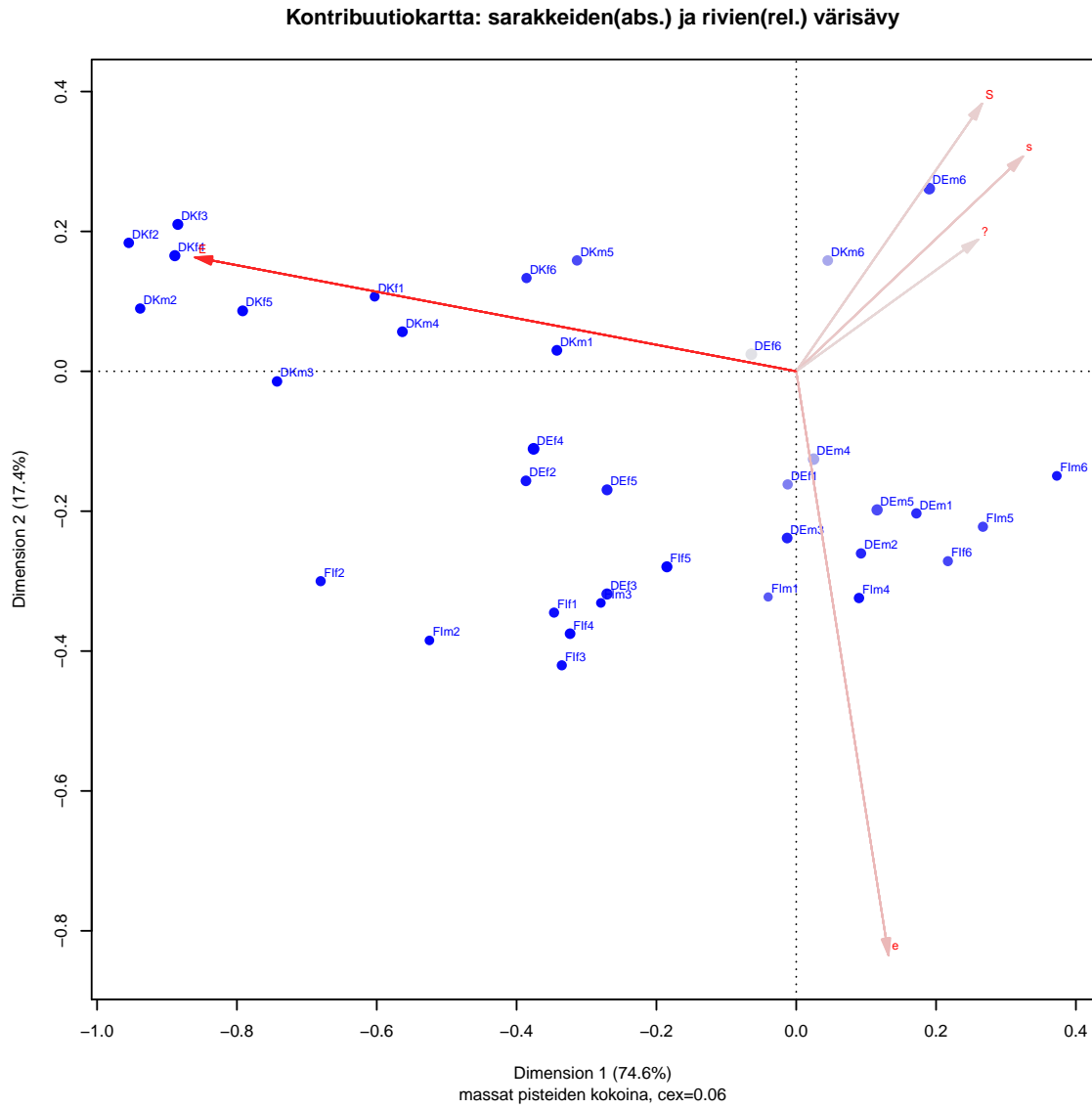
Karttoja 6.1 ja 6.2 joutuu katsomaan aika tarkkaan ennenkuin uskoo, että akseleiden skaalaus on akateeminen pulma vailla käytännön merkitystä (kts.luku 3.3.4). Dataa analysoidaan graafisesti, ja kuvat *näyttävät* erilaisilta. Pääakselien inerttioiden neliöjuuret ovat 0.327 ja 0.158, sarakkeiden etäisyyksiä voisi tulkita myös kontribuutiokuvista.

k Rivipisteiden tulkinta

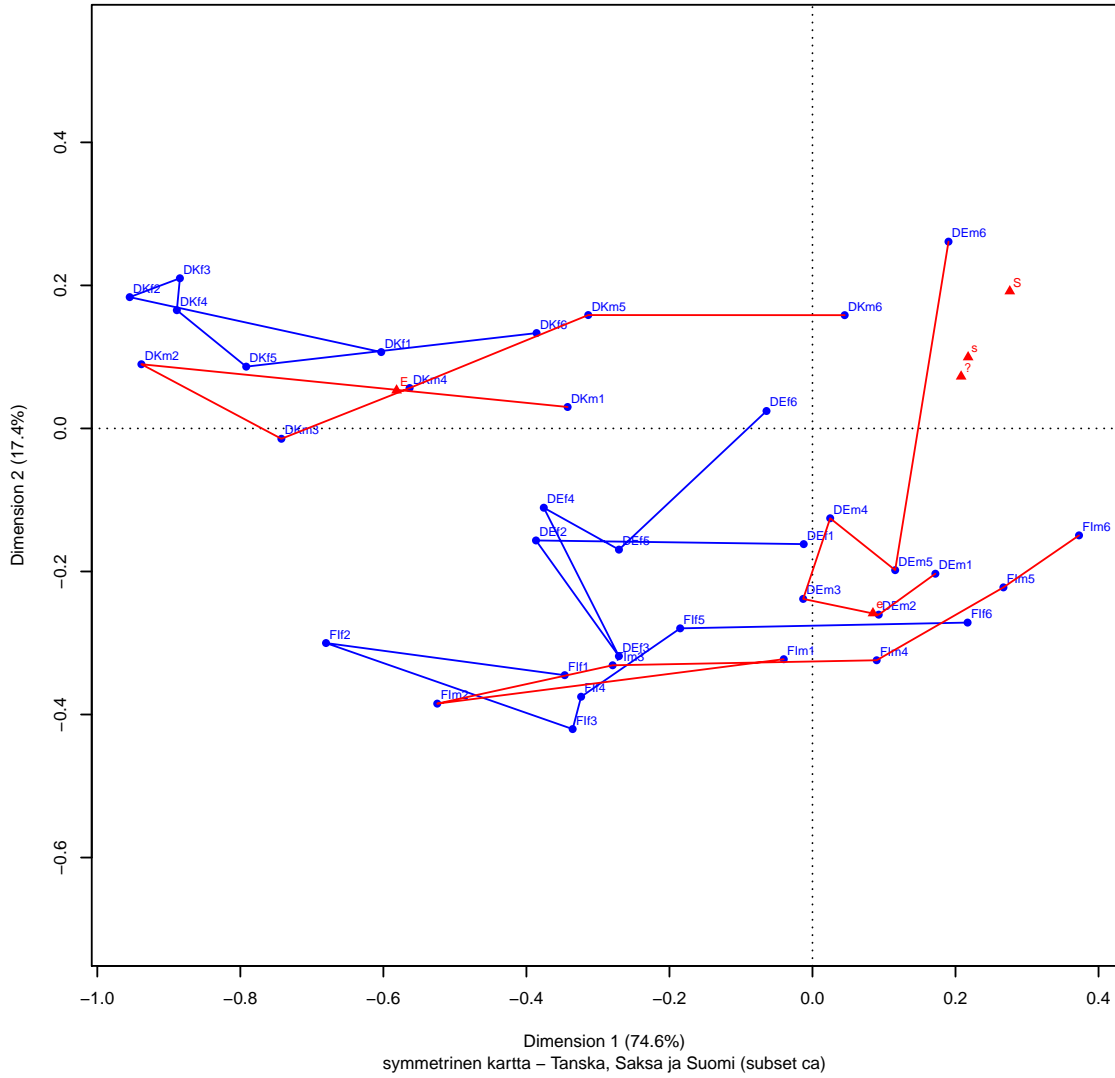
```
##
## Principal inertias (eigenvalues):
##
## dim      value      %   cum%   scree plot
## 1      0.107090  74.6  74.6  *****
## 2      0.024985  17.4  92.0  ****
## 3      0.006594   4.6  96.6  *
## 4      0.004882   3.4 100.0  *
```



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



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



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

```

##          -----
## 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 | FI f1 |    12  951   20 |  -346 478  13 |  -345 474 55 |
## 26 | FI f2 |    12  941   48 |  -680 788  50 |  -300 153 42 |
## 27 | FI f3 |    12  952   24 |  -335 370  12 |  -420 582 82 |
## 28 | FI f4 |    14  999   25 |  -323 426  14 |  -375 573 82 |
## 29 | FI f5 |    17  982   14 |  -185 299  6 |  -280 683 55 |
## 30 | FI f6 |    11  704   13 |   217 274  5 |  -271 430 33 |
## 31 | FI m1 |    7  624    8 |   -40  10  0 |  -323 614 30 |
## 32 | FI m2 |    9  984   26 |  -525 640  22 |  -385 344 52 |
## 33 | FI m3 |    9  990   12 |  -279 412  6 |  -331 578 38 |
## 34 | FI m4 |    13  944   11 |    90  67  1 |  -324 877 54 |
## 35 | FI m5 |    12  722   14 |   267 426  8 |  -222 295 23 |
## 36 | FI m6 |    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 - tiukka konservatiivi) on inertiasta 17 prosenttia, edellä ollut paljon yksiulotteisempia ratkaisuja. Huono kvaliteetti (qlt) on ryhmillä DEf1 (467) ja DEf6 (48), DEm4 (285). Tanskan havainnoista vanhimmat miehet (DKm6,291) ovat kaikkein huonoimmin

Ensimmäisen dimension tulkinta pysyy samana, mutta nyt molemmat erimieliset (E, e) vastauskategoriat ovat selvästi oikealla. Ne ovat lähes x-akselin päällä, kun ensimmäisen osajoukon kartalla e-sarake oli oikealla ja alhaalla kontrastina S- ja s- vastauksille ja myös neutraalille vaihtoehdolle. Kartan toinen dimenisio 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 maltillisemmän ja jyrkemmän konservatiivisuuden välillä pystysuuntaan. Toisen dimension kontrasti on myös hieman yllättäen Bulgarian nuorimpien naisten (BGf1) Unkarin vanhimpiten naisten (HUf6) välillä.

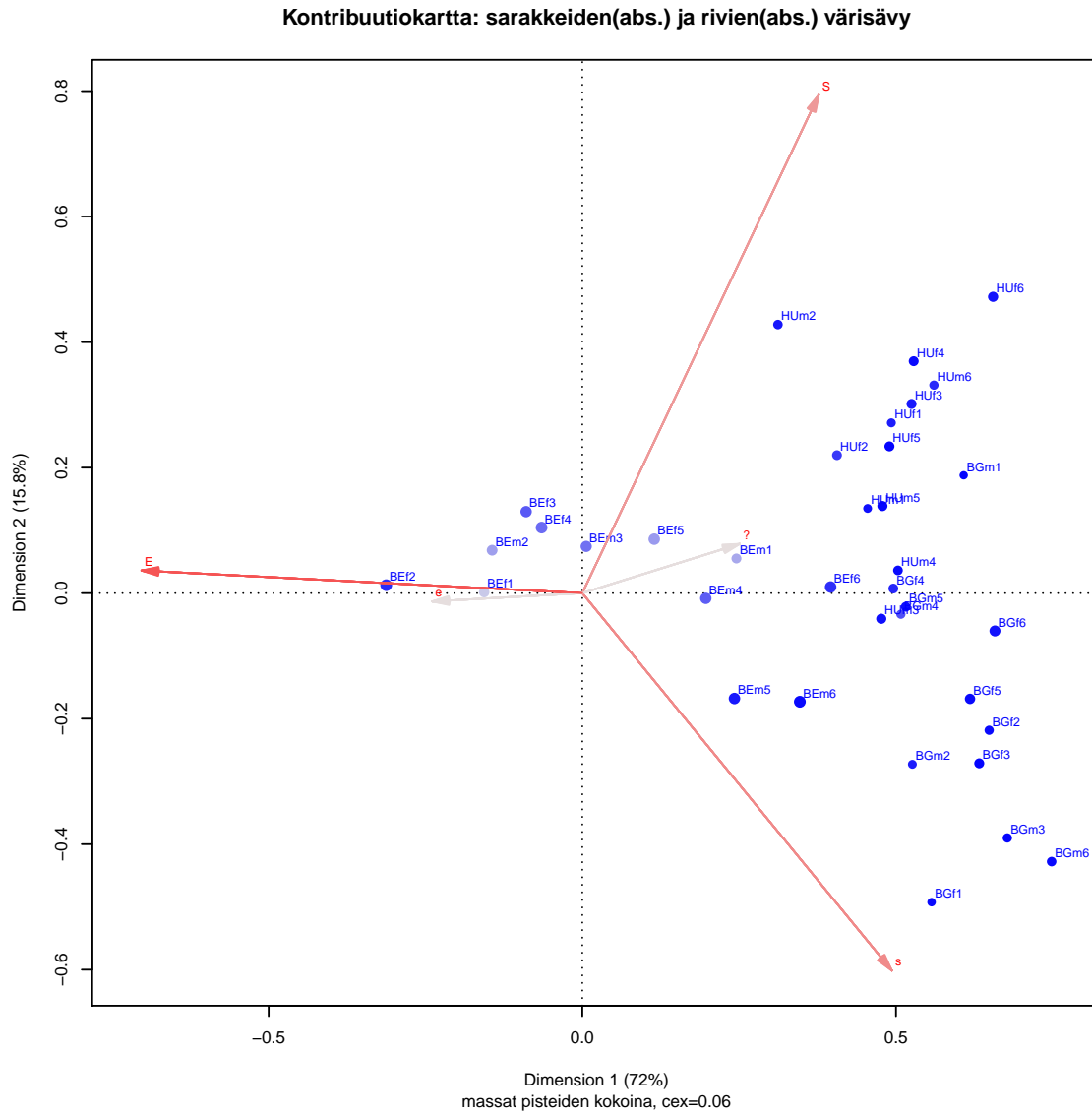
Kuvan 6.3 tapaan ei Bulgarian ja Unkarin ikäluokkia kannata yhdistää. Järjestys toki löytyy, mutta ei ollenkaan niin selkeä. Saksan naisten ikäluokkakuvaa alkaa erkaantua hieman Suomen ja Taskan hyvin samanlaisista kuvioista. Saksan miehillä on jo eroja paljon toisen dimension suuntaan, Unkarin ja Bulgarian 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) tasapainosemmalla 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 yleiskuva. Tässä eksploratiivinen data-analyysi kuitenkin kulkee hieman eri polkuja kuin tavaniomainen tilastollisten mallien analyysi. Seurataan minne data kuljettaa, etsiään 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 | HUf1 | 6 948 12 | 455 871 14 | 135 77 6 |
## 32 | HUf2 | 9 902 24 | 312 313 10 | 428 589 90 |
## 33 | HUf3 | 13 945 26 | 477 938 33 | -41 7 1 |
## 34 | HUf4 | 10 965 22 | 503 960 29 | 36 5 1 |
## 35 | HUf5 | 13 993 26 | 478 916 33 | 139 77 13 |
## 36 | HUf6 | 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 |

```

edit: tässä toistoa

Kahden osajoukon inerttioiden summa on sama kuin koko aineiston ($0.144 + 0.12 = 0.263$), Selitysasteet nousevat hieman, ja aineiston riippuvuuden rakenteesta saadaan esiin selviä eroja. Osajoukkojen analyysi täydentää ja tarkentaa yleiskuvaa (@ref:fig:maagaCA1map1)

Belgian pisteistä osalla on huono kvaliteetti (BEf1, BEf5, BEm1, BEm2). Bulgaria ja Unkari hyvin esitetty. Belgia on pulmallinen tapaus, ehkä taas omissa dimensioissaan. Belgian poikkeavuus (annetulla aluejaolla) on kiinnostava havainto, korrespondenssianalyysin tavoite ei ole pelkästään kohtuullisen luotettava yleiskuva taulukon riippuvuuksista. Poikkeavat havainnot eivät ole ongelma, vaan datan ominaisuus.

Luku 7

Monimuuttuja-korrespondenssianalyysi (MCA)

Usean muuttujan samanaikainen analyysi voidaan korrespondenssianalyysissä jakaa kahteen erialaiseen tutkimusasetelmaan. Ensimmäisessä tutkitaan kahden erilaisen muuttujaryhmän välisiä suhteita, toisessa yhden homogeenisen muuttujaryhmän sisäisiä suhteita.

Esimerkkiaineistossa haastateltavien vastaukset substanssikysymyksiin ovat oma ryhmänsä ja taustamuuttujat toinen ryhmä. Kahden muuttujaryhmän välisiä suhteita voidaan tutkia rakentamalla yhdistetty matriisi useasta kahden muuttujan ristiintaulukoinnista. Tämä pinottujen ja yhdistettyjen (stacked and concatenated) taulukoiden menetelmä (CAiP s. 129) ei kerro muuttujaryhmien sisäisistä yhteyksistä, joita edellä analysoitiin vuorovaikutusmuuttujien avulla.

Toinen asetelma on keskenään homogeenisten muuttujien välisten suhteiden analyysi. Monimuuttujakorrespondenssianalyysi (multiple correspondence analysis, MCA) soveltuu hyvin kyselytutkimusten vastausten analyysiin. MCA - kartoilla voidaan esittää myös havainnot yksittäiset havainnot, mutta usein on järkevämpää käyttää kartalla taustamuuttujien keskiarvopisteitä. Keskiarvopisteisiin voi simuloimalla lisätä luottamusellipsit (CAiP, s. 299).

Esitän molemmista analyyseistä yhden esimerkin. MCA-esimerkissä otan käyttöön kaikki aineiston havainnot ja useita vastausmuuttujia, puuttuvat tiedot ovat mukana yhtenä luokittelumuuttujan arvona.

Yksinkertaisen korrespondenssianalyysin yleistys usean muuttujan samanaikaiseen analyysiin ei ole aivan yksinkertainen asia. Silti MCA toimii käytännössä usein hyvin juuri kyselytutkimusten analyyseissä. Vaikka geometrikan tulkinta ei ole läheskään niin selkeä kuin yksinkertaisessa korrespondenssianalyysissä, skaalaustulkinta on pätevä. MCA-kartat esittävät luokittelumuuttujien arvot optimaalisesti, kartalla esitettyjen havaintopisteiden hajonta on maksimaalinen (? , s. 447, kts. myös Liite 1).

7.1 Pinotut ja yhdistetyt taulukot

Pinottujen taulukoiden idea on esitetty kuvassa 7.1. Taulukossa kaksi “selitettävää” muuttujaa on ristiintaulukoitu kolmen taustamuuttujan kanssa.

Jos reunajakaumat ovat samat eli puuttuvia tietoja ei ole, taulukon kokonaisinertia on osataulukoiden inertioiden keskiarvo. Taulukon analyysi on yhden kysymyksen ja yhden taustamuuttujan parittaisten suhteiden analyysiä, yksi pari kerrallaan.

Pierre Bourdieun tunnettu tutkimus *La Distinction* (1979) sovelsi tätä menetelmää. Ranskan väestö luokitellaan ammattiryhmiin ja taulukoidaan useiden elämäntapaa kuvaavien muuttujien kanssa. Taulukot yhdistetään pinoamalla ne päällekkäin (? , s.21).

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

Kuva 7.1: Pinotut ja yhdistetyt taulut - periaate

Seuraavassa esimerkissä “selitettäviä” luokittelumuuttujia on vain yksi. Tällainen pinotun taulun analyysi on eräänlainen kaikkien siihen kuuluvien taulukoiden “keskiarvokartta” (CAiP, s 136). Kysymyksen Q1b vastauksien ristiintaulkointi ikäluokan ja sukupuolen kanssa liitetään maarivien taulukkoon.

Pisteiden määrä kartoilla kasvaa, ja muuttujanimiä joudutaan tiivistämään. Toinen tekninen detalji on kartan kääntäminen. Kuvat kääntyvät herkästi akselien ympäri, vertailun helpottamiseksi koordinaattien etumerkkejä joutuu joskus muuttamaan (kts. Liite 3: R-koodi).

Tulkinta ei muutu, eikä maapisteiden sijaintikaan.

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. dimensio suuntaan.

Ikäluokkapisteet ovat koko aineiston keskiarvopisteitä, niiden sijaintia voi tulkita pistejoukko kerrallaan kuten maapisteidenkin. Naispiste on tiukassa nipussa ikäluokkien a3 ja a4 kanssa aivan origon vasemmalla puolella. Miesten keskiarvopiste on hieman origosta oikealle, yhdessä ikäluokan a5 kanssa.

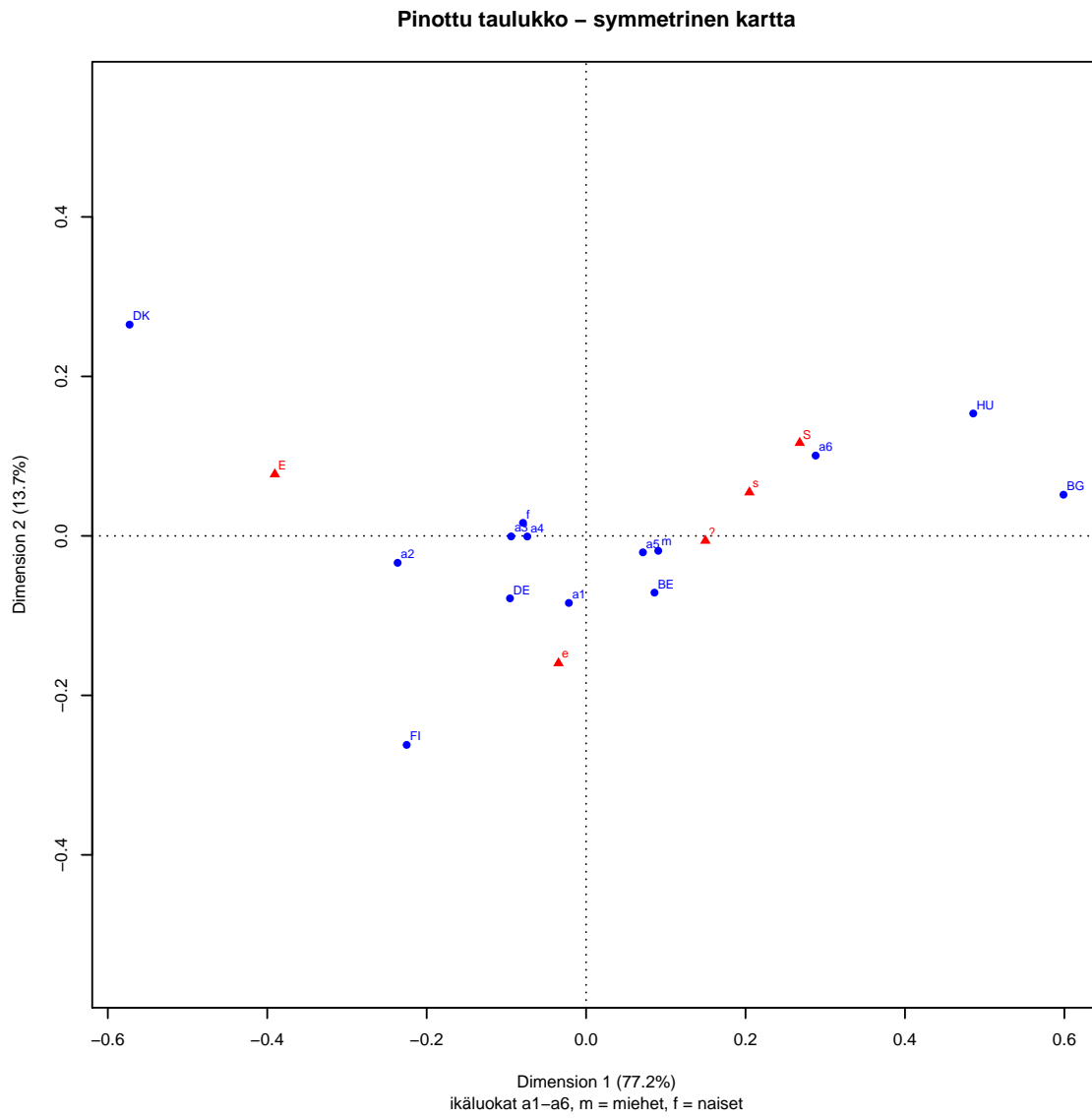
Lisäpisteet on hyvin esitetty, niiden etäisyyksiä voi luotettavasti arvioida kuvasta. Poikkeus on nuorin ikäluokka (a1, q1t = 501). Inertian osuudet (inr) ovat yhtä vaatimattomia kuin Belgian (28) ja Saksan (29), (m = 20, f = 17, a2 = 40, a6 = 83), samoin kontribuutiot akseleiden inertiaan. 1. dimension kontribuutio (ctr) on suuri (>800) kaikilla paitsi nuorimmalla ikäryhmällä (a1) jolla 2. dimension selittää lähes puolet sen inertiaasta (470).

Karttaa vertaa karttaan 5.1 jossa on esitetty iän ja sukupuolen yhteisvaikutusmuuttuja. Pinottu taulu on vaihtoehtoinen tapa, ja kartasta 7.2 voi päätellä samat asiat: miehet ovat konservatiivisempia kuin naiset, iäkkäämmät ovat konservatiivisempia kuin nuoret. Nuorin ikäluokka poikkeaa muista.

tämä poistetaan lopullisesta versiosta

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

##



Kuva 7.2: Q1b: Lapsi kärsii jos äiti käy työssä

```
## 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, inertiaikontribuution keskiarvo
# 1000/14 = 71
```

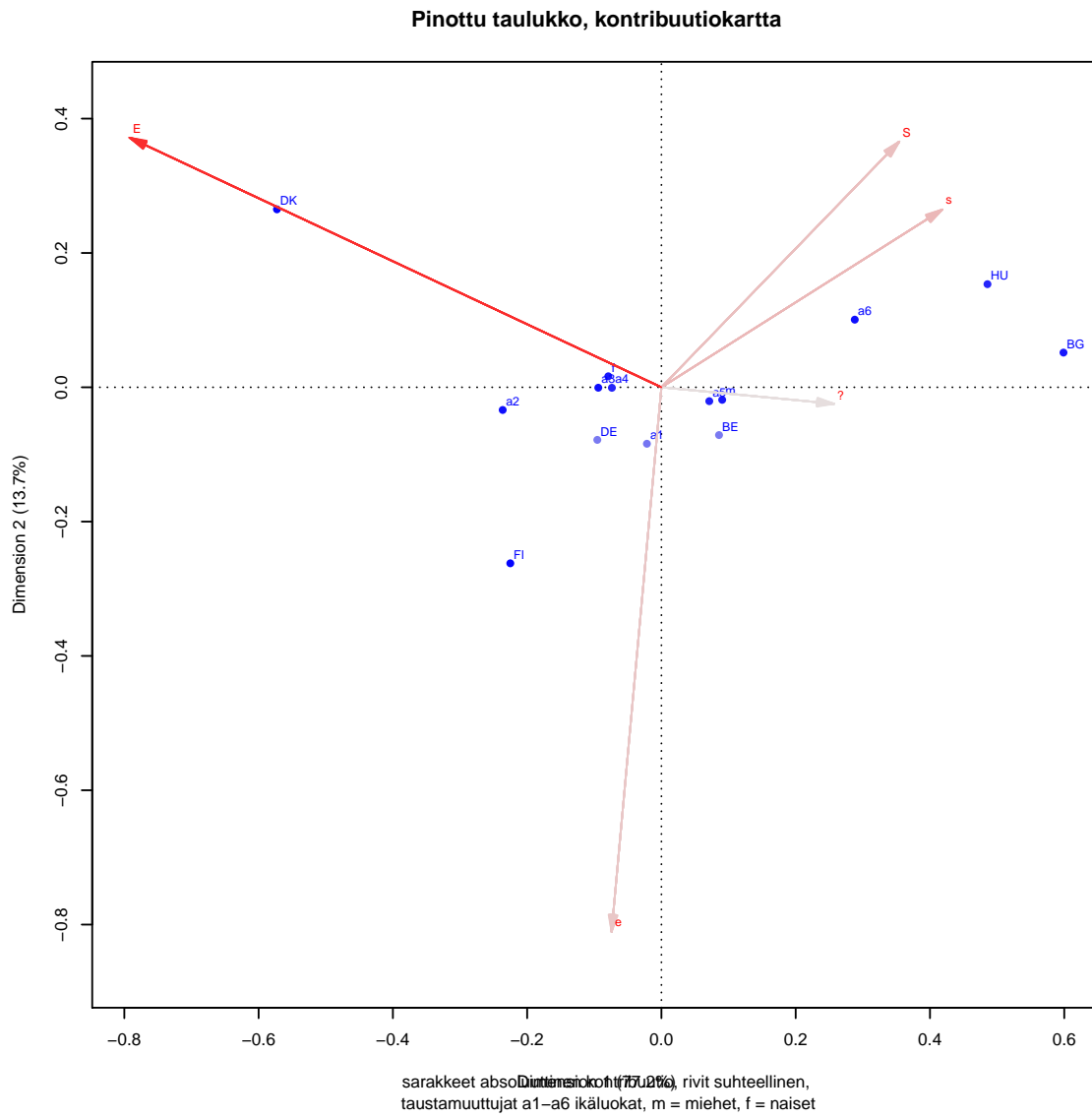
Kontribuutionkartan sarakkeista E-sarake (täysin eri mieltä) määrittää akseleita vahvasti, kontrastina kaksi konservatiivista vastausta (S ja s) ja myös neutraali vaihtoehto (e). Numeeriset tulokset kertovat, että ikäluokat vaikuttavat juuri ensimmäiseen tärkeimpään dimensioon.

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.

Kontribuutiokartalla voi arvioida hieman tarkemmin ikäluokkien ja vastausvaihtoehtojen yhteyttä sarake kerrollaan. Maltillisesti eri mieltä olevien osuus on suhteellisesti suurin nuorimmassa ikäluokassa. Jos a1 - pisteen projisoi konservatiivisen S-vastauksen janalle se ei ikäluokista konservatiivisiin. Ikäluokat a2 ja a6 ovat ääripäitä, muut ikäluokat sijoittuvat lähelle toisiaan mutta liberaalille puolelle.

Esimerkkiaineistossa ei ole puuttuvia tietoja. Ne olisivatkin aika pulmallisia, varianssin dekomponointi ei onnistu jos reunajakaumat ovat alitaulukoissa selvästi erilaisia.

Matriisien yhdistely on monipuolinen laajennus. Eräs kiinnostava malli on ABBA, kahden rakenteeltaan samanlaisen matriisin yhdistäminen lohkoina. Nimi kertoo yhdistetyn matriisin rakenteen (block circulant matrix),



Kuva 7.3: Q1b: Lapsi kärsii jos äiti käy työssä

päällekkäin pinotut A ja B liitetään toiseen pinottuun matriisiin B ja A.

Matriiseilla on samat rivit ja sarakkeet, esimerkiksi miesten ja naisten vastausprofiilit yhteen kysymykseen aittain luokiteltuina. Kokonaisinertia saadaan dekomponoitua ryhmien sisäiseen ja väliseen hajontaan kahdelle kartalle. Toinen kuvaa maiden välisiä eroja ja toinen maiden sisäisiä sukupuolten välisiä eroja (CAiP ss. 177-)

7.2 MCA - monimuuttujakorrespondenssianalyysi

MCA on yhdistettyjen ("pinottujen") taulukoiden erikoistapaus, samantyyppiset muuttujat taulukoidaan keskenään. Tulos riippuu siis vain muuttujien parittaisista yhteyksistä.

Tätä "supertaulukkoa" kutsutaan *Burtin matriisiksi*. *Indikaattorimatriisi* on toinen tapa esittää data. Indikaattorimatriisin sarakkeet ovat luokittelumuuttujan arvoja (kategorioita) ja rivit yksittäisiä vastausprofileja. Profilissa on rivi nollia ja ykkösiä, 1 valitun vaihtoehdon sarakkeessa.

En tässä jaksossa esitä numeerisia tuloksia, niiden tutkimisella voi jatkaa analyysiä.

Tavoite on tutkia seitsemän kysymyksen vastausvaihtoehtojen yhteyksiä, miten ne asettuvat kaksiulotteiselle kartalle. Aikaisemmissa jaksoissa etsittiin yhteyksistä uusia piirteitä ja tarkennettiin analyysiä. Nyt hahmotetaan ison aineiston muuttujien välisiä yhteyksiä ja erityisesti puuttuvien tietojen ongelmaa.

Koko datassa (kts. luku 2) on 32823 havaintoa 25 maasta. Niistä täydellisiä on 71 prosenttia. Jos valitaan kuusi taustamuuttujaa (edu, sosta, urbru, maa, ika, sp) ja seitsemän kysymystä, täydellinen havaintoja on 81 prosenttia.

Pelkissä kysymyksissä (Q1a, Q1b, Q1c, Q1d, Q1e, Q2a, Q2b) puuttuvia tietoja on 14 prosentissa havaintoja (4554). Kaikkien puutteellisten havaintojen poistamien ("listwise delete") on sitä huonompi vaihtoehto mitä enemmän muuttujia on.

Kaikissa kysymyksissä on viisi vastausvaihtoehtoa ja kuudes kategoria puuttuvalle tiedolle. Analyysi kannattaa aloittaa tästä pulmasta, yksi hyvä ratkaisu on osajoukon MCA (subset MCA).

Inertian selitysosuudet ovat paljon pienempiä, ja ratkaisu on selvästi kaksiulotteinen. Puuttuvat vastaukset erotuvat omana ryhmänä, ja varsinaiset vastaukset ovat pakkautuneet y-akselin oikealle puolelle. Niiden erot näkyvät vain toisessa dimensiossa. Ensimmäinen dimensio kuvaa vastaamattomuutta (syystä tai toisesta) vs. kaikkia vastauksia. Kokonaisinertiaa on korjattu pienemmäksi (adjusted inertia, kts. liite 1), selitetyn inertian osuus on realistinen arvio.

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

Karttaa voi parantaa lisäämällä siihen vastaajien pisteet.

Jokainen havainto on sarakevektoreiden keskiarvopiste. Sarakevektoreita ei voi tulkita yhtä selkeästi kuin yksinkertaisessa korrespondenssianalyysissä. Ne eivät edusta kysymystä vaan kysymyksen yhtä vastauskategoriaa.

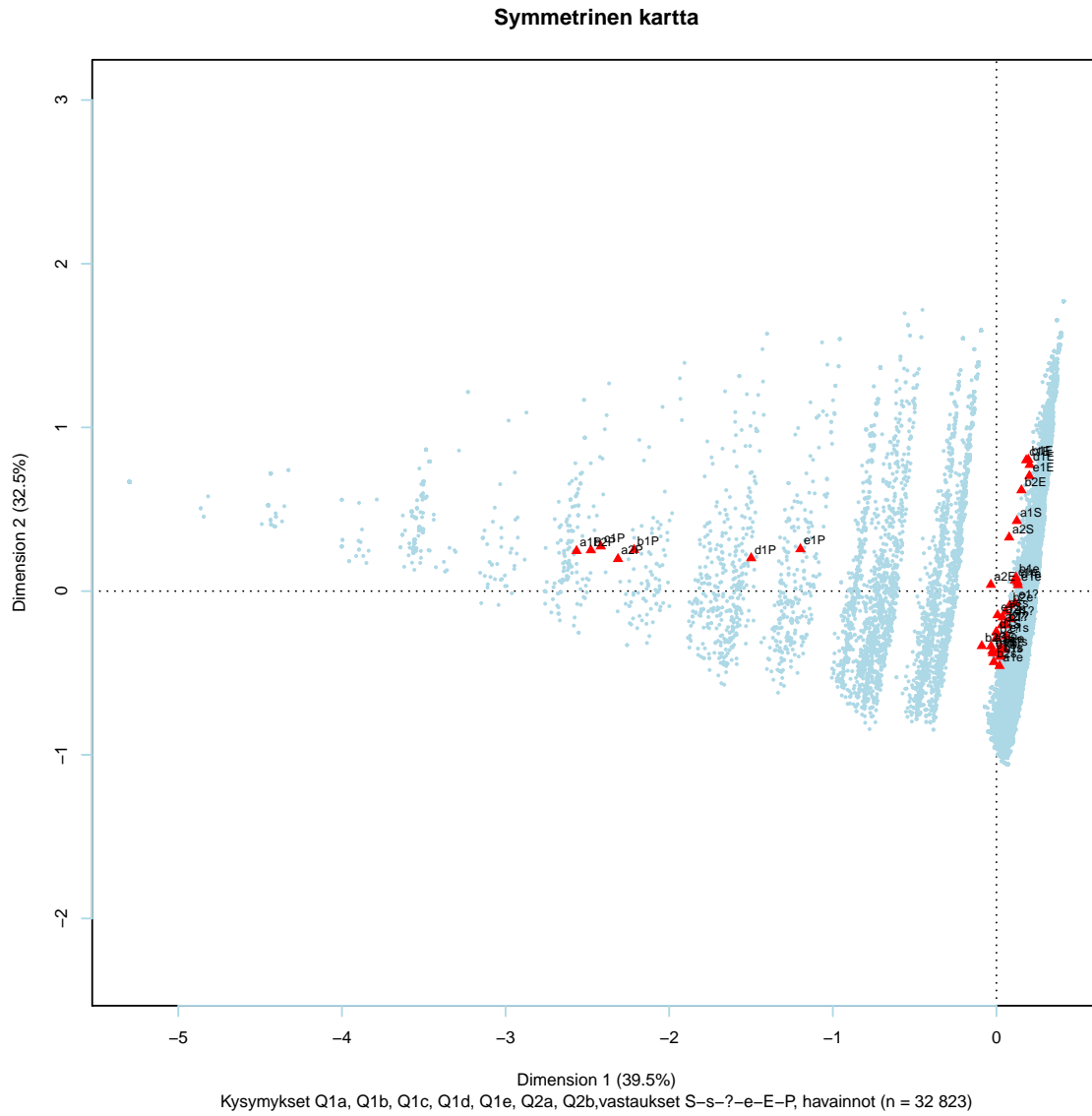
Pistepilven muoto näyttää, miten pienenevä joukko vastaajia lähestyy kiilana puuttuvien tietojen pisteitä. Kaikkiin kysymyksiin vastanneet ovat massana kuvan oikeassa laidassa. Pistepilvet oikealta vasemmalle kuvaavat kuinka moneen kysymykseen on jätetty vastaamatta.

Osajoukon MCA

Osajoukon MCA sopii hyvin sekä puuttuvien tietojen että täydellisten vastausten analyysiin. Asymmetrisessä kartassa sarakkeet skaalautuvat pois orgiosta ja näkyvät paremmin. Kuvaan on piirretty havaintopisteet, joista voi hahmottaa havaintojen sijoittumista kartalla.

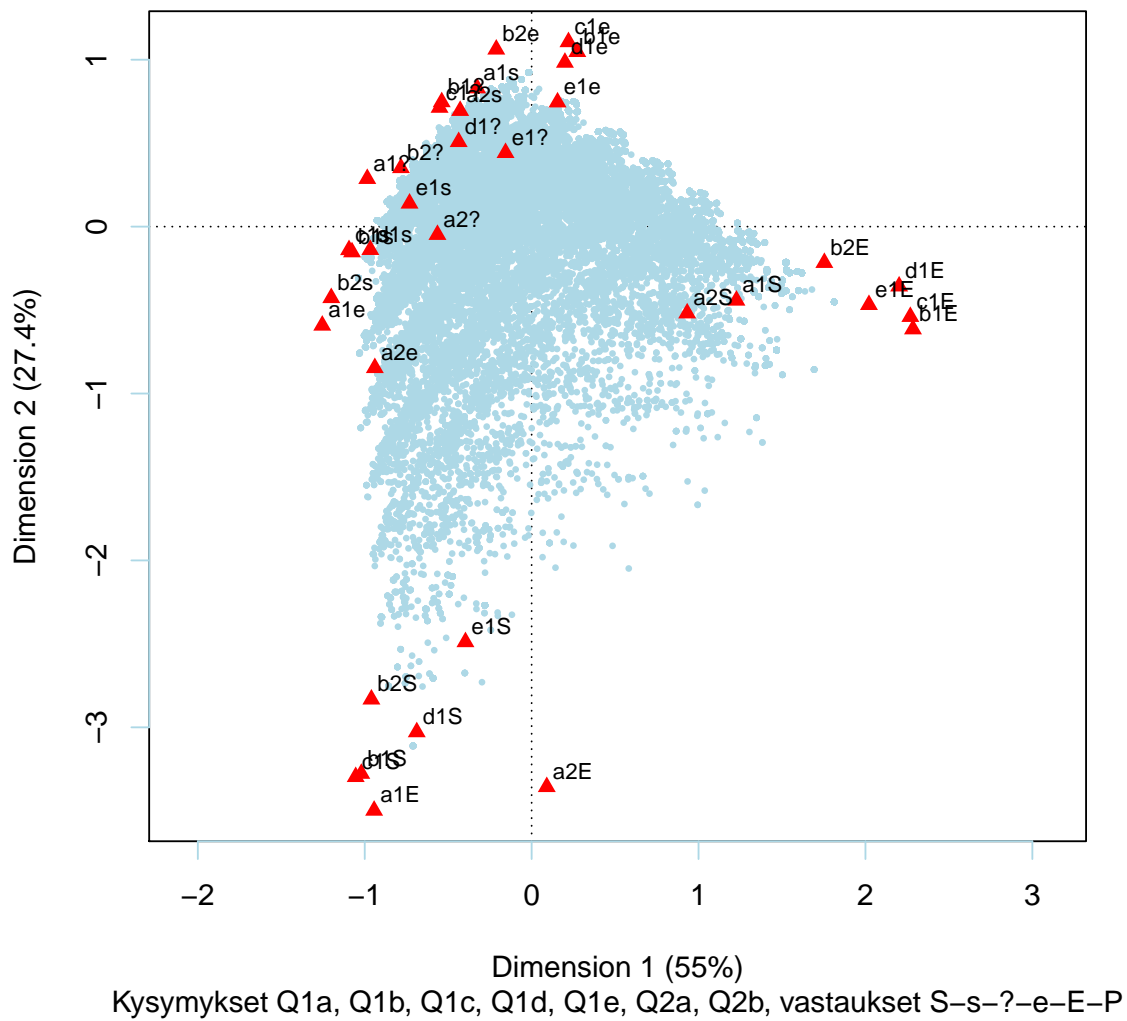
Kontrasti on "ääripäiden" välillä, vahvat mielipiteet (S ja E) hallitsevat vasenta alakulmaa ja oikeaa laitaa x-akselin tuntumassa. Maltilliset vastaukset ja neutraali vaihtoehto ovat ylhäällä vasemmalla. Liberaalit vastaukset ovat oikealla ylhäällä ja jokaiselle löytää vastinparin vasemmalta alhaalta, konservatiivien kulmasta. Konservatiivisten vastausten joukosta lähimpänä liberaalia ryvästä on "e1S" ("Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen" - täysin samaa mieltä). Vastaavasti "a2S" ("Sekä miehen että naisen





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

Asymmetrinen kartta – osajoukko: ei puuttuvia vastauksia



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

tulee osallistua perheen toimeentulon hankkimiseen” - täysin samaa mieltä) on lähimpänä konsertvatiivista kulmaa. Molemmat “ääripäiden maltilliset” pisteet ovat myös omassa ryhmässään lähimpänä maltillisten ja neutraalien vastausten ryhmää.

Karttaan voi hahmotella diagonaalin suuntaisen akselin vahvojen mielipiteiden ryppäiden välille. Muut vastaukset ovat mukaisesti näiden välisellä U-muotoisella linjalla (Guttman-efekti).

Osajoukon korrespondenssianalyysin esitelleissä artikkeleissa ((?) ja samasta teemasta laajentaen kokoomateoksessa (?), ss. 197-217.) Aineistona jälkimmäisessä tutkimuksessa on sama kysymyssarja kuin tässä 1994 datasta. Kysymyksissä on jonkin verran eroja.

? (s. 139-) analysoi lähes samoja kysymyksiä ISSP-datalla 2012. Tulokset ovat hyvin samantapaisia. Karttaa ??fig:subsetMCA1map3) vastaavan kuvan jatkoanalyysi on vasemman yläkulman pisteiden tarkempi analyysi. Greenacre havaitsee, että nämä neutraalit ja maltilliset vastaukset eivät ole hyvin esitettyjä kaksiulotteisella kartalla, vaan ne karkaavat korkeampiin ulottuvuuksiin, “dimenisions of middleness”.

Tässä aineistossa kolmiulotteinen MCA antaa samantapaisia viitteitä, mutta vain osa keskikategorioista on huonosti esitetty myös siinä. Huono kvaliteetti on vain osalla sarakkeita (e1? 475, c1? 71 ja b1? 573).

Luku 8

Yhteenveto

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

k Onko maiden vertailu järkevää? 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 sulkuja) ja suluilla (?)

k Eksploraatiivinen data-analyysi ja todennäköisyysteoreettinen päättely

Gifi-nimimerkillä kirjoittavat Jan De Leeuw jatkavat verkkokirjassaan (?) keskustelua konfirmatorisen ja data-analyttisen eksploraatiivisen lähestymistavan eroista. He ovat tiukasti eksploraatiivisen 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-analyttikko.

****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

Perusyhtälöt on esitetty teoksen “Correspondence Analysis in Practice”(?) liitteen “Theory of Correspondence Analysis” mukaisesti.

Datamatriisilla N on I riviä ja J saraketta ($I \times J$). Alkiot ovat ei-negatiivisia (eli nollat sallittuja) ja samassa mitta-asteikossa. Jos mitta-asteikko on intervalli- tai suhdeasteikko, mittayksiköiden on oltava samoja (esim. euroja, metrejä). Tietyn ehdoin myös negatiiviset luvut ovat sallittuja (?, s. 60).

Taulukon alkioden summa on $\sum_i \sum_j n_{ij} = n$, missä $i = 1, \dots, I$ ja $j = 1, \dots, J$.

Korrespondenssimatriisi P saadaan jakamalla matriisin N alkiot niiden summalla n .

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

Korrespondenssianalyysin ratkaistaan singulaariarvohajoituksen avulla.

Singulaariarvohajoitus (singular value decomposition) tuottaa ratkaisun kun sitä sovelletaan standardoituun 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.

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

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

Toinen esitystapa on hyödyllinen, kun tarkastellaan CA:n yhteyksiä muihin läheisiin menetelmiin. Näitä ovat esimerkiksi “suhteellisten osuuksien datan” analyysi (log ratio analysis of compositiona data), moniulotteinen skaalaus, lineaarinen diskriminanttianalyysi, kanoninen korrelaatioanalyysi, pääkomponenttianalyysi, kaksoiskuvat ja muut SVD-hajoitelmaan perustuvat dimensioiden vähentämisen menetelmät.

Samat kaavat voi esittää myös alkionmuodossa:

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

ja toinen

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

Alkimuodosasa esitetyistä kaavoista näkee 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.

Singulaariarvohajoitelma (singular value decomposition, SVD) matriisille S on

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

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

Matriisit U ja V ovat ortogonaalisia singulaarivektoreiden matriiseja. Singulaariarvohajoituksen merkitys dimensioiden vähentämiselle perustuu Eckart - Young - teoreemaan. Teoreema kertoo että saamme pienimmän neliösumman m - ulotteisen approksimaation matriisille S matriisien U ja V ensimmäisten sarakkeiden ja ensimmäisten singulaariarvojen avulla.

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

Korrespondenssianalyysin ratkaisualgoritmissa tätä tulosta on muokattava niin, että rivien ja sarakkeiden massat huomioidaan pienimmän neliösumman approksimaatiossa painoina.

Näin saadaan standardikoordinaatit ja pääkoordinaatit 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

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

Pääakselien inertiat (principal inertias) λ_k

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

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

Korrespondenssianalyysi ratkaisun akselien inertioita kutsutaan usein ominaisarvoiksi, mutta periaatteessa SVD-ratkaisulla saadaan singulaariarvot. Niiden neliöt ovat akselien inertioita. Ominaisarvojen ja singulaariarvojen yhteys on läheinen ja riippuu diagonalisoitavan matriisin ominaisuuksista.

Korrespondenssimatriisi P voidaan esittää matriisi- ja alkimuodossa 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 riippuvuutta kuvaavat transitioyhtälöt (transition equations).

Pääkoordinaatit standardikoordinaattien funktiona (barysentrisen ominaisuus, barycentric relationships):

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

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

Pääkoordinaatit pääkoordinaattien funktiona:

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

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

Yhtälöt (15) ja (16) esittävät profilipisteet ideaalipisteiden (vertex points) painotettuina keskiarvoina, painoina profiilin elementit. Asymmetriset kartat (rivien tai sarakkeiden suhteen) perustuvat näihin yhtälöihin. Yhtälöiden (17) ja (18) kahdet pääkoordinaatit ovat perusta symmetrisille kartoille. Myös niitä yhdistää barysentrisen painotetun keskiarvon riippuvuus, mutta mukana ovat skaalaustekijät $\frac{1}{\sqrt{\lambda_i}}$.

Pisteet ja projektio aliavaruuteen

Kuva on kurssimateriaaleista(?).

Kuvassa on esitetty korrespondenssianalyysin ratkaisun minimointiongelma. Pisteiden projektio on sitä parempi mitä pienempi kulma on sentroidista pisteeseen piirretyn janan ja pisteen projektion välillä. COR - tunnusluku ca-funktion numeerisissa tuloksissa tämän kulman kosinin neliö. Pisteiden kuvauksen laatu (qlt) ca-tuloksissa on valitun approksimaation akseleiden kvaliteettien (COR) summa.

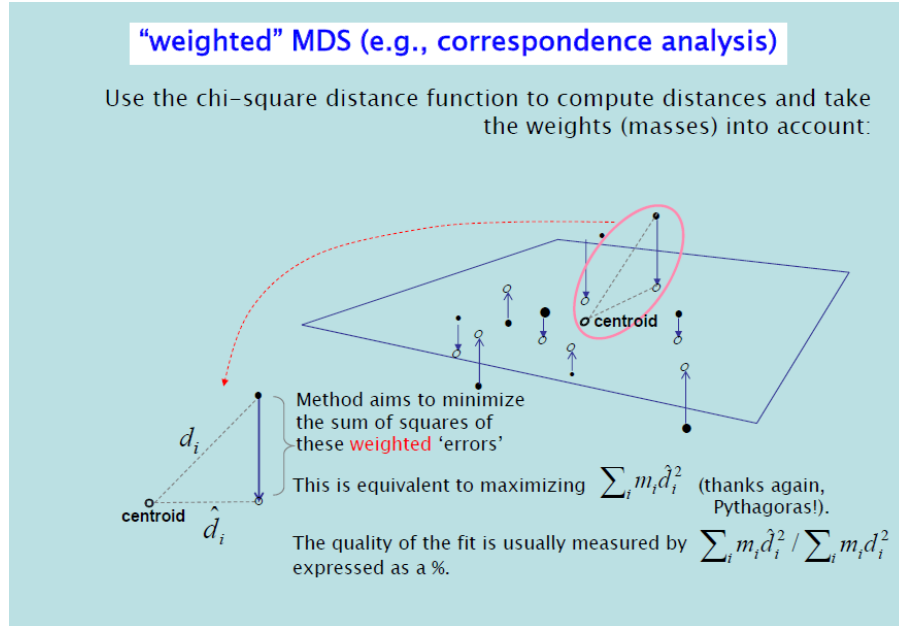
Kuvasta voi myös hahmottaa sen periaatteen, että projektiossa kaukana olevat pisteet ovat kaukana myös alkuperäisessä avaruudessa. Projektiossa lähellä olevat pisteet voivat olla alkuperäisessä avaruudessa kaukana toisistaan, jos niiden projektion laatu on huono.

Matriisit ja niiden havainnollistaminen

edit: kaavaesimerkkejä - poistetaan lopullisesta versiosta

Korrespondenssianalyysin sovelluksissa tutkimusongelman ratkaisu on usein sopivan matriisin rakentaminen.

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



Kuva 1: Pisteen projektio aliavaruuteen

Ehkäpä ABBA onnistuu paremmin tällä notaatiolla?

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

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

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

Yksinkertainen korrespondenssianalyysi on kahden luokittelumuuttujan määrittämien 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. Rivit ovat esimerkiksi maitan summattuja vastauksia, ja niiden alle voidaan lisätä joku toinen luokittelumuuttuja. Havaintojen määrä yhdistetyssä (“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 vääristävät 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 ($f1, \dots, f6$ ja $m1, \dots, m6$) samojen vastausvaihtoehtojen kanssa. Jos tätä taulukkoa merkitään matriisilla $B_{f',J}$, voimme muodostaa yhdistetyn matriisin

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

ABBA on erityistapaus yleisemmästä moniulotteisen taulukon (multiway table) analyysistä, jossa useita kahden muuttujan taulukoita “pinotaan” päällekkäin ja rinnakkain. Voimme ottaa yhden kysymyksen vastausten lisäksi

analyysiin mukaan useamman kysymyksen vastaukset laajentamalla kahden päällekkäisen matriisin taulukkoa oikealle.

Monimuuttuja-korrespondenssianalyysi MCA

Usean muuttujan korrespondenssianalyysissä tutkitaan usean muuttujan välisiä yhteyksiä. Kartan tulkinnan apuna siihen voidaan lisätä havaintojen sijaan niiden keskiarvopisteitä ja niille simuloituja luottamusellipsejä. Kuvien pääongelma on liian suuri määrä 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 pareittain muodostetut taulukot. Diagonaalilla 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 inertia osuudet jäävät melko pieniksi. Ratkaisuna on inertia oikaisu tai korjaus (adjusted inertia), jossa mm. poistetaan kokonaisinertialaskelmista Burtin matriisin diagonaalilla olevat alimatriisit. Näillä korjauksilla ei ole vaikutusta kartan pisteiden sijaintiin. Tämä menetelmä on ca-paketin mja-funktion oletus.

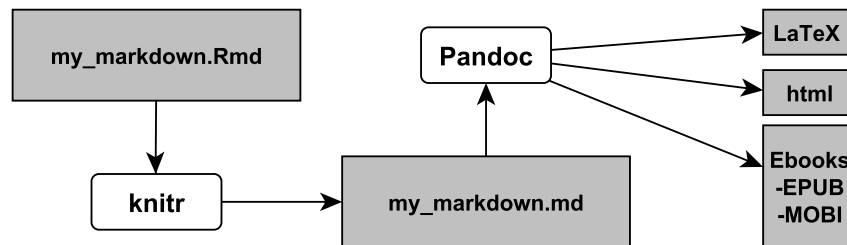
Kolmas vaihtoehto on ns. yhdistetty korrespondenssianalyysi (joint ca).

Greenacre on kirjoittanut useissa yhteyksissä MCA:n geometrisen tulkinnan ongelmista. ? (s. 447) sanovat asian näin: "Finally, although we have motivated simple correspondence analysis from the geometric point of view, the geometry of the indicator matrix in multiple correspondence analysis is admittedly not convincing. Distances between the row profiles of an indicator matrix and projections of artificial column vertices have less intuitive appeal. However, the scaling interpretation remains attractive in this case; the displays are graphical representations of optimal scale values for the categories."

Greenacre jatkaa samasta aiheesta artikkelikokoelmassa (?, , s. 41, s. 61): "Yleistys useammalle kuin kahdelle muuttujalle ei ole ilmeinen eikä hyvin määritelty". Siten MCA onnistuu esittämään hyvin kiinnostavia yhteyksiä muuttujien välillä ("succesfully recovers interesting patterns of association"). Kriittisyys ei kuitenkaan estä häntä soveltamasta MCA-analyysiä. Tulkiten tämän niin, että menetelmää voi aivan hyvin soveltaa, mutta geometrista tulkintaa ei voi suoraan siirtää CA-kartoista MCA-karttoihin. Greenacre esittelee perusteellisesti MCA-sovelluksia kaksoiskuvia käsittelevässä kirjassaan (?). Ehkä korrespondenssianalyysin matemaattinen teoria ei ole vielä täysin valmis?

Liite 2: Tekninen ympäristö ja Bookdown-paketti

Muokataan tiiviimpi pätkä esimerkkipostista bookdown-testi1. Tämä kuva kertoo vain julkaisutekniikan ympäristön.



Kuva 2: Tulostiedoston prosessointi

Käyttöjärjestelmä ja R-ympäristö

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] assertthat_0.2.1  tinytex_0.27      bookdown_0.21     printr_0.1
## [5] reshape2_1.4.4    scales_1.1.1      furniture_1.9.7    rmarkdown_2.5
## [9] lubridate_1.7.9.2 forcats_0.5.0     stringr_1.4.0     purrr_0.3.4
## [13] readr_1.4.0        tidyr_1.1.2       tibble_3.0.4      ggplot2_3.3.2
## [17] tidyverse_1.3.0    knitr_1.30        dplyr_1.0.2       haven_2.3.1
## [21] ca_0.71.1         rgl_0.100.54
##
```

```
## loaded via a namespace (and not attached):
## [1] httr_1.4.2                jsonlite_1.7.1          modelr_0.1.8
## [4] shiny_1.5.0              highr_0.8               cellranger_1.1.0
## [7] yaml_2.2.1               pillar_1.4.6            backports_1.2.0
## [10] glue_1.4.2               digest_0.6.27           manipulateWidget_0.10.1
## [13] promises_1.1.1          rvest_0.3.6             colorspace_2.0-0
## [16] htmltools_0.5.0         httpuv_1.5.4            plyr_1.8.6
## [19] pkgconfig_2.0.3         broom_0.7.2            xtable_1.8-4
## [22] webshot_0.5.2           later_1.1.0.1           generics_0.1.0
## [25] farver_2.0.3            ellipsis_0.3.1         withr_2.3.0
## [28] cli_2.1.0               magrittr_1.5            crayon_1.3.4
## [31] readxl_1.3.1            mime_0.9                evaluate_0.14
## [34] fs_1.5.0                fansi_0.4.1             xml2_1.3.2
## [37] tools_3.6.3             hms_0.5.3              lifecycle_0.2.0
## [40] munsell_0.5.0           reprex_0.3.0           compiler_3.6.3
## [43] rlang_0.4.8             grid_3.6.3             rstudioapi_0.13
## [46] htmlwidgets_1.5.2       crosstalk_1.1.0.1      miniUI_0.1.1.1
## [49] labeling_0.4.2          gtable_0.3.0           DBI_1.1.0
## [52] R6_2.5.0                fastmap_1.0.1          stringi_1.4.6
## [55] Rcpp_1.0.5              vctrs_0.3.4            dbplyr_2.0.0
## [58] tidyselect_1.1.0        xfun_0.19
```