CA-paperin jäsentely /5.9.2018 (tulostettu 19.9.2018)

SISÄLLYS

capaper /4.9.2018, editoitu 14.9.2018

- kommentit capaper tulosteesta
- bd-capaper – kommentit (?)
- galku- kommentit ja ideat

Kokeillaan jäsentely Wordillä, ensin. Rmarkdown on vähän turhan kömpelö ja raskas ensimmäiseen vaiheeseen. Teen capaper -dokumenttiin vastaavan jäsentelyn otsikkotasolla (3 tasoa maks.). Tähän on kopioitu capaper- bookdown –tekstit ja laajennettu jäsentelyä Galku-dokkarin pohjalta. Tavoite on kirjoittaa jäsentely kokonaan kerran läpi. Bookdown-teksti on tässä myös editointia varten. capaper-bookdown on nyt hetken jäähyllä.

zxy lähteet kursorisesti

1. JOHDANTO

#capaper

zxy Kirjoitetaan disposition pohjalta, keräillään kaikki yleiset ca-luonnehdinnat yhteen paikkaan eli johdantoon.

Mahdollisia lisäyksiä:

1. Lyhyt esitys CA:n historiasta
2. Tavoitteet, sisältö, rajaukset (jota voi myöhemmin täydentää)
3. Muutamat puutteet (erit. ”maapainot” ja MG:n ”reweighting”) CA:n laajennuksiin
4. data: ei huomioida sitä, että otoskoot vaihtelevat aika paljon eli “maapainot” eri suuruisia
5. ei huomioida muitakaan otantaan liittyviä asioita (tämä ainakin mainittava data-osuudessa): ISSP-datassa on mukana painot maakohtaisesti, mutta ei maiden vertailuun. ESSP-datassa on toisin, ja aika monessa muussakin (kai?).
6. kuvaileva menetelmä, mutta mikä on tutkimusongelma? Sellainen pitäisi olla. (HS:Kantola, laajempi kehys naisten työmarkkina-asema josta löytyy vaikka kuinka paljon tutkimusta. Pitäisikö dataan yhdistää työmarkkinadataa ISSP:n maakohtaisista referenssitiedoista?).

zxy Mitä on korrespondenssianalyysi? Muutamalla kappaleella. Yksi kappale historiasta.

TUTKIELMAN TAVOITE

#capaper vai tutkimusongelma?

zxy 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.

zxy -mitä ca on? - dimensioiden vähentäminen ja visualisointi - mihin dataan se soveltuu - määrittele graafinen, deskriptiivinen, eksploratiivinen data-analyysi - yksinkertainen ca, useamman muuttujan ca

ks: miksi on levinnyt hitaasti?

ks: CA ei ole saari tai koulukunta (pelkästään…?), vaan osa laajempaa menetelmien perhettä. Vaihtoehtoisia kokoavia käsitteitä (Geometrinen data-analyysi(LeRoux, deskrptiivisen ja eskploratiiviset data-analyysin menentelmä (Vehkalahti, Mustonen). Eivätkä nämäkään ole saaria.

zxy Miksi eksploratiivinen (määrittele!) ja deskriptiivinen (määrittele!) menetelmä on esitettävä “in vivo”, toiminnassa? Oppikirjoissa (viitteitä) erityisesti MG on havainnolistanut CA:n matemaattista ja geometristä taustaa synteettisillä aineistoilla. Turha kopioida tähän. Menetelmän ydin on yksinkertaisen graafisen esityksen – kartan – avulla tulkita monimutkaisen empiirisen aineiston muuttujien riippuvuuksia. Yhteyksiä ei tiivistetä todennäköisyyspäättelyn kriteereillä tilastolliseen malliin, vaan deskripriivisen 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. Eksploratiivinen data-analyysi on vuoropuhelua aineiston kanssa. Analyysiä tarkennetaan, rajataan ja muokataan, kun aineisto paljastaa jotain kiinnostavaa tai yllättävää. Tästä saa jonkinlaisen aasinsillan matriisiyhtälöiden puolustukseksi. Saksan ja Belgian datan jakaminen on hyvä esimerkki, on “osattava tarttua” menetelmän tulosmatriiseihin.

zxy esitystavan perustelu

- oikea data, kyselytutkimusaineisto, iso ja mutkikas
- esimerkkiaineistoja ja muista on esim. ca-paketissa ja oppikirjoissa; ei toisteta niitä

ks Tämän voi tehdä yksinkertaisen korrespondenssianalyysin avulla. Yksinkertainen kahden luokittelumuuttujan korrespondenssianalyysi antaa graafisen analyysin “…perussäännöt tulkinnalle. Kaikki muut korrespondenssianalyysin muodot ovat saman algoritmin soveltamista toisen tyyppiisiin datamatriiseihin, ja tulkintaa sovelletaan vastaavasti (with the consequent adaptation of the interpretation)”(Greenacre ja Hastie 1987 , s. 437)

zxy 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).

TÄRKEIMMÄT LÄHTET JA OHJELMISTOT

zxy Miksi tämä luku?

LÄHTEET

Michael Greenacre luennoi lyhyen kurssin korrespondenssianalyysistä Helsingin yliopistossa keväällä 2017(M. Greenacre 2017). Luennot ja laskuharjoitukset perehdyttivät minut ensimmäistä kertaa tähän menetelmään, ja kurssin materiaaleihin olen usein palannut. Niihin voi tutustua Helsingin yliopiston [Moodle-palvelussa] (https://moodle.helsinki.fi) (käyttäjätunnus vaaditaan). Greenacren selkeät perusoppikirjat ovat tehneet menetelmää laajasti tunnetuksi.

Ranskalaisen lähestymistan perusoppikirja(Roux ja Rouanet 2004) esittelee menetelmän matemaattiset perusteet. Lyhyt historiallinen katsaus ja menetelmä soveltamisen perusajatusten esittely valaisevat ranskaa taitamattomalle data-analyysin koulukunnan ideoita. Kirjoittajat esittelevät perusteellisesti joitain empiirisiä tutkimuksia, ja lyhyt mutta naseva matriisilaskennan kritiikki on hyvä panna merkille. MOOC-verkkokurssin materiaalit.

OHJELMISTOT JA TEKNINEN YMPÄRISTÖ

KS Käytetyt ohjelmistot, tekninen ympäristö ml. bookdown-asetukset. Tarkat speksit liitteeseen (”tekninen ympäristö?”). Johdantoon yksi kappale (vai tässä omana alalukuna? Liite on ehkä tarpeeton.

zxy R, ca-paketti. löytyy myös muita paketteja. Rmarkdown(Yihui Xie 2018), ja bookdown ((Xie 2016) ja toinen viite (Xie 2018)). Mikäs tuo jälkimmäinen on? PDF-lähdeluettelossa ei ole url-osoitteita.

zxy Helposti toistettavan tutkimukset periaatteet

1. Datasta (löytyy netistä, samoin kattava dokumentaatio) lyhyt matka analyysiin.
2. Koodi selkeää ja dokumentoitua
3. R, LaTeX, pandoc - versiot dokumentoidaan
4. R-koodi ”dokumentoi itse itsensä”, sisältää myös testauksia ja välitulostuksia.

Tarkemmin liittessä.

CA-paketit R:ssä, FActoMine ja manuaalin kuvaus. SAS, SPSS (MG:n CAip-kirjan epilogin kriittiset huomiot, tarkista päteekö vielä!).

KORRESPONDENSSIANALYYSIN HISTORIAA

zxy Tiivis esitys lähteineen. Mieti, mikä on suhde muuhun tekstiin!

”Irlantilainen kaveri” – linkkejä hänen käännöksiinsä Benzecrin yms. teoksista. DeLeeuw on mollannut ranskalaisia ”esoteerisestä tieteenfilosofiasta”. Toisaalta kiinteä yhteys Bourdieun tuontantoon, joka taas on sosiologian valtavirtaa omana aikanaan.

Carme-n.org , ”normalisoituminen”. Ehkä tästä jännite yhteenveto-luvun Gifi – julistukseen?

Beh, LeRoux ja Rouanet, Hull (1974), MG, MG&Blasius – johdannot yms.

Malinvaud ja RStaSoc – debatti, JASA-debatti.

Yleisiä pointteja

1. ”Bourbakilaisuus”
2. Tilastotieteessä usein menetelmät keksitään moneen kertaan (eri tieteenalojen tutkimusongelmien kautta)
3. Teoria – data (evidenssi, empiria tms) – menetelmät. Tässä kolmiossa pyöritään.
4. Olisi houkuttelevaa dramatisoida eroja ja ”data-analyysin filosofiaa”, mutta ehkä liioittelua?

ks: CA Suomessa.

Korrespondenssianalyysi tuli osaksi suomalaista Survo-ohjelmistoa jo vuonna (????), ja menetelmää on esitelty ainakin kahdessa oppikirjassa(Mustonen 1995) ja (Vehkalahti 2008). Kauman väikkärin johdanto, Oksanen Oulusta (kasvibiologi). Bourdieu teki tunnetuksi, mutta ei lyönyt läpi.

2. DATA

zxy Voisi miettiä paremman otsikon. Galku-paperin alusta on lisäilty viitteitä Refworksiin, mutta hieman hankalaa. www.gesis.org - sivusto on aika sekava. Virallinen (heidän määrittelemä) sitaatti löytyy, ja linkkejä. Tässä voisi ehkä käyttää alaviitettä, jossa tarjoaisi linkit? Tai ihan oma lyhyt kappale? Alla virallinen viite, ja tässä kaksi

ISSP_data1.docx – muistiinpanoja ISSP-dokumentaatiosta.

viitteet pois- ehkä tekstiin linkkeinä? (siis issp-datalähteisiin).

ks ISSP (International social survey) on tehnyt laajoja kansainvälisiä kyselytutkimuksia eri teemoista. Yksi teemoista on perhe ja muuttuvat (sosiaalisesti määräytyvät) sukupuoliroolit (Jorat ym. 2016).

ks

zxy Miksi data sovelutuu korrespondenssianalyysin esittelyyn? Iso ja monimutkainen (kansainvälinen, datan laatu? kts. Blasius-viite alempana), sisällölliset muuttuja nominaaliasteikolla (kysymyspatterit, Likert), laadukas hyvin dokumentoitu aineisto.

zxy Onko itse asia kiinnostava? (Kantolan kolumni, HS).

zxy Miksi data on kiinnostava sisällöllisesti? Viite Kantola (HS). Lisäksi laadukas, usealta vuodelta, tarkasti dokumentoitu.

ks Kokoava kappale, ja sen perään tarkentavat

ks1

ks2

ks-n

zxy Aineiston ongelmat ja puutteet (tavanomaisten surveyaineistojen ongelmien lisäksi, erityisesti vastauskadon). Kato erikseen, oikeastaan hyvä juttu koska CA soveltuu sen analyysiin. Tavallaan ”ongelma”, survey-aineistot ovat mutkikkaita ja nämä asiat ovat aina läsnä.

zxy Aineisto kuvattava **sisällön** (mitä asiaa, ilmiötä, tällä datalla halutaan valaista), **para- ja metadatan** näkökulmasta (tai ainakin kerrottava mitä on saatavilla). Kolmanneksi aineiston “tilastotieteellinen olemus”: otanta-asetelmat, kansalliset versioinnit, harmonisoinnit (esim. puoluekenttä vertailukelpoiseksi).

1. Kysymyksissä maakohtaisia eroja. Osa perusteltuja, on haluttu tarkentaa tai muuten hifistellä. Osa kummallisa, erityisesti neutraalin vaihtoehdon puuttuminen (Espanja). Nämä maat pitää sivuuttaa.
2. Datassa painot “maatasolle”, otanta sun muu kuvattu tarkasti dokumentaatiossa. Jos tutkimusongelma on maiden erojen analyysi, mitään vertailupainoja ei ole käytössä. Otoskoko on paino. Paha juttu, MG oikaisee ja ja oikaisee myös sukupuolien osuudet. Tarkka viite CAiP-kirjan luku ”massat”.

TIETOSISÄLTÖ

zxy ISSP:n historiaa

zxy ISSP Muuttuvat sukupuoliroolit ja perhearvot – aineiston keruun taustaa ja historiaa

zxy tietosisältö

zxy Galkusta tänne!

Perusasiat: kohdeperusjoukko, mitä halutaan selvittää

Substanssimuuttujat (mitä kysytään)

Taustamuuttujat

Meta- ja paradata: milloin kerätytty, miten kerätty, erot joissain kysymyksissä, otantamenetelmät

Laatutiedot: on raportteja

Maakohtaisten tietojen muuntaminen vertailukelpoisiksi: äänestystiedot, ammatti(oma ja puolison), koulutustaso, asuinpakka (rural or not), muita?

ks: aineistoa voi analysoida kahdesta peruslähtökohdasta: maakohtaisesti maa kerrallaan tai usean maan dataa yhtä aikaa. Jälkimmäisessä vaihtoehdossa maiden väliset erot ovat keskeisiä, ja asetelmaa voi muuttaa monipuoliseksi ottamalla myös muita muuttujia tavalla tai toisella mukaan tarkasteluihin. Tässä valitaan top down – reitti, mutta se ei välttämättä ole paras. Voiko CA:lla saada jotain tolkkua mutkikkaasta aineistosta?

AINEISTON RAJAAMINEN

zxy Espanja veke heti alussa, Unkari V6-analyysien jälkeen (+)

”Pelin henki” on se, että dataan palataan, tehdään valintoja. Tämä ei on ”konfimatorista” analyysiä!

zxy miksi rajataan? Menelmän esittely yksinkertaisella esimerkillä (vain kuusi maata), myöhemmin laajennetaan isompaan joukkoon.

zxy kuusi maata valittu tarkoituksella niin, että kahdessa on maan sisäinen ositus. Suomi mukana, eri kokoisia maita ja jonkinlainen valikoima eurooppalaisia valtioita.

zxy Isompi joukko: perustelu ("akvaarioargumentti"), tulkinnan helpottaminen "karkeasti samankaltaisilla mailla". Ei sen syvällisempää pohdintaa.

zxy puuttuvat tiedot pois: rankka rajaus, mutta yksinkertaistaa. Puuttuvan tiedon käsittelyn yksityiskohdat R-koodissa. ISSP-dokumentointi. CA ja MCA kelpo menetelmiä puuttuvien havaintojen ja "en osaa sanoa" - vastausten analyysiin. Ovat vain havaintoja nominaaliasteikolla.

zxy alussa vähäistä, kun vain muutamia muuttujia (substanssi- ja taustamuuttujia). Kun aineistoa laajennettaan (enemmän maita, enemmän muuttujia), puuttuneisuuden vaikutus kasvaa.

TODO: taulukko "ison aineiston" puuttuvista tiedoista. EI pureuduta tarkempaan puuttuneisuuden luokitteluun.

ks Eksploratiivinen data-analyysi on vuoropuhelua aineiston kanssa. Tavoite on saada jonkilainen ote tai ymmärrys aineiston riippuvuuksista ja rakenteista. Luodaan tulkinnan kehystä, ja aineistoa voidaan (ja pitääkin) rajata eri tavoin, tarvittaessa. Valmista ei synny kerralla, vaan pala kerrallaan.

AINEISTON KUVAILU

zxy ehkäpä taulukoiden lisäksi Likert-kuva? **Vai tarvitaanko tässä, mitä tässä pitäisi kuvailla?**

viimeinen kappale

Miten aineistoa on käytetty? "ISSP - saitilla" löytyy bibliografia, ja hakupalveluillakin voi haravoida. Michael Greenacre on käyttänyt aineistoa eri vuosilta luentomateriaaleissa (Helsinki 2017 MCA, viite Moodleen?) ja kahdessa oppikirjassa ((Greenacre 2010), (M. J. Greenacre 2017)).ISSP - aineisto vuodelta 1989 on käytetty myös neljän "singuaariarvohajoitelmaan perustuvan menetelmän" vertailuun(Greenacre 2003).

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

Vihreä kirja: MG ja joku muu, subset analysis.

Blasius ja Thiessen ((Blasius ja Thiessen 2006)) arvioivat aineiston laatua ja ja maiden vertailtavuutta vuoden 1994 aineistolla.

“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.”

zxy www.gesis.org - sivustolta löytyy myös julkaisuluettelo, voiko linkin laittaa alaviitteeksi tai suoraan leipätekstiin? Viitteissä yksi tutkimus, jossa naisten osuutta työvoimasta vertaillaan uskonto-tietojen avulla. Olkoon esimerkki substantiaalisesta käytöstä, MG ja Blasius käyttävät aineistoa menetelmän esittelyyn.

3. YKSINKERTAINEN KORRESPONDENSSIANALYYSI - ESIMERKKI

zxy Tässä yksi kysymys, kuusi maata, peruskäsitteet lopussa

zxy Luvun tärkeimmät asiat; miten tarina kulkee

zxy tästä tarkempaa jäsennystä Galkussa!

ÄITI TÖISSÄ

zxy Edellisessä luvussa on esitelyt aineisto, ja kerrottu rajaukset. Voidaan siis mennä suoraan asiaan. Luvun alussa kerrotaan, mikä juoni luvussa on.

KAHDEN MUUTTUJAN FREKVENSSITAULUKON ANALYYSI

zxy graafiset tulokset ja niiden tulkinnan perusteet

zxy ensimmäinen symmetrinen kartta – tulkinnan alkeet (ca:n esittelyn ensimmäinen sykli – mitä kuva kertoo, tai miten sitä tulkitaan). Esittely johdattaa eri etäisyysmittoihin, geometriaan jne.

KAHDEN MUUTTUJAN TAULUKKO JA CA-TERMINOLOGIA PERUSTEET

zxy profiilit ja reunajakaumat

zxy massat

zxy profiilien etäissyydet ja khii2-etäisyys

zxy khii2 – etäisyys, riippumattomuushypoteesi, inertia

ENSIMMÄINEN SYMMETRINEN KARTTA

akselit (kartan koordinaatisto)

pisteiden sijannit ? etäisyydet ? (kartta-metafora)

rivi- ja sarakepisteiden yhteys (”korrespondenssi”), kartta-metafora

akseleiden prosentti-merkintä

zxy Benzecrin ohje: mitä on oikealla, mitä on vasemmalla (akseleiden tulkinta. Alkaa sillä akselilla, jossa kuvattu suurin osa inertiasta eli hajonnasta. Sitten toinen akseli – mitä on ylhäällä ja alhaalla. ”Ääripäät” – mitkä rivi- ja sarakepisteet ovat suhteellisesti kauimpana toisistaan

KORRESPONDENSSIANALYYSIN PERUSKÄSITTEET

CA:n esittelyn toinen sykli: ei formaaleja perusteluja, vaan graafisen analyysin perusteita. Ei kaavoja (ehkä?).

zxy laajennetaan esimerkkiä asymmetrisellä kartalla ja kontribuutiokartalla

ASYMMETRINEN KARTTA JA IDEAALIPISTEET

KONTRIBUUTIOT KARTALLA

YHTEENVETO

CA:ssa kaikki on suhteellista (MG kalvot)

dimensioiden vähentäminen, graafinen 2- tai ehkä kolmiulotteinen aliavaruus johon pistejoukot projisoidaan. Ei eksakti

ca luokitteluasteikon muuttujien PCA:na – projektiossa säilytetään maksimaalinen määrä alkuperäisen pistejoukon hajonnasta.

4. KORRESPONDENSSIANALYYSIN TEORIAA JA TULKINNAN PERUSTEITA

zxy Galku!

ks CA:n numeeriset tulokset ja niiden tulkinta. Herkkyysanalyysi**.** Nämä selitetään tässä, mutta loppupuolella.

Dimensioiden vähentäminen, pistepilvi euklidisessa avaruudessa: PCA – esimerkki

ihmettelyn aihe 1: khii2 –etäisyys (toteuttaa distributional eqvivalence – periaatteen. Bencecrin mukaan tärkein periaate). vrt compositional data ja ”subcompositional coherence – kriteeri).

ihmettelyn aihe 2: rivi- ja sarakeratkaisun symmetria

kaksoiskuvien yleisempi näkökulma (?). Auttaako oikeasti ymmärtämään?

vaikea asia 1: symmetrinen ja asymmetrinen kartta

vaikea asia 2: havaintojen pilvi ja muuttujien pilvi – rivi-ja sarakeongelmien symmetria

vaikea asia 3: ”kaikki on suhteellista” – rivipisteiden suhteelliset etäisyydet, sarakepisteiden suhteelliset etäisyydet

vaikea asia 4: akseleiden tulkinta – ydinkysymys

ca painotettuna MDS:na

ca painotettuna pns-ratkaisuna

Akseleiden (mittayksikön) skaalaus ominaisarvoilla

Inertian maksimimäärä, täydellinen separointi

Guttman : onko MG yleisemmin oikeassa, ja LeRoux spesifisti?

- LeRoux: kahden järjestysasteikon muuttujan korrelaatio
- MG: ainoa reitti simpleksin yhdestä kulmasta toiseen

5. YKSINKERTAISEN KORRESPONDENSSIANALYYSIN LAAJENNUKSIA

zxy Mihin BE/DE – splitti? Luvun 3 loppuun, vai tämän alkuun? Tämän alkuun, niin päästään tutkimaan ca:n numeerisia tuloksia, kontribuutioita, kvaliteettia, inertian dekomponointia.

#zxy Saksan ja Belgian jako liittyy tietysti khii2 – etäisyyteen ja distr.equivalence – periaatteeseen tai pikemminkin ominaisuuteen. Splitataan, ja asetelma ml. kuvan tulkinta ei dramaattisesti muutu.

#zxy S-B-split on aika hyvä esimerkkitapaus, jossa voidaan katsoa miten kartta muuttuu, miten numeerisia tuloksia tulkitaan. Konkreettista!

#ks Maiden otoskoot määräävät niiden massat, ja se ei ole järkevää. Aineisto pitäisi painottaa uudelleen (MG tekee niin) siten, että maiden massat ovat yhtä suuret. MG korjaa myös sukupuolijaon painotuksen. Tässä ohitetaan. Analyysit voi toistaa helposti uudelleenpainotuksella (reweighting). CAiP – kirjassa esitetty lyhyesti. Voidaan ohittaa, koska pääasia menetelmän esittely, ei ”oikea tutkimus”.

zxy siirtymä 1: useampi maa mukaan, lyhyet kuvailut (lyhyet!); vai mennäänkö aluksi kuudella maalla?

ks maat ja otoskoot, painotuksen pulma. ”Yli siitä missä aita on matalin” – säilytetään BE/De – jako?

zxy 1 Kysymys V6, ei puuttuvia tietoja, peruskuvat

> **zxy** miten laveasti diagnostiikkaa? Ei ehkä liikaa… Liitteeksi?
>
> **zxy** mikä muuttuu, muuttuuko tulkinta!

zxy siirtymä 2: tutkimuskysymykset (ei voi vain pyöritellä dataa…)

zxy Galku!

#zxy siirtymä: täydentävät pisteet yleisemmin

TÄYDENTÄVÄT HAVAINNOT (SUPPLEMENTARY POINTS)

#zxy Edellä esimerkkinä täydäntävät pisteet, DE ja BE. Ovat osiensa keskiarvopisteitä. Tässä luvussa esitellään lisäpisteiden/täydentävien pisteiden idea yleisemmin.

#zxy Suppoints.docx /CAip s.89 alkaa…

Aktiiviset pisteet: voi ajatella, että jokaisella aktiivisella pisteellä on oma vetovoima pääakseleihin (princip axes. Voiman suuruus riippuu sijannista ja massasta. Kaukana keskiarvosta sijatsevilla pisteillä on enemmän “vipuvoimaa” in orienting the map towards them (?). Kääntää kooridinaatistoa suuntaansa? Suuri massa taas vetää karttaa puoleensa, keskiarvoa. “have a greater pull on the map”.

Supplementary point (joskus passive point): on sijainti mutta massa nolla, joten kontribuutio inertiaan on nolla eikä niillä ole vaikutusta pääakseleihin.

Hyödyllisiä kolmessa tavallisessa tapuksessa:

1. uusi sarake Y, erityisrahoitus nuorille tutkijoille, juuri tullut mukaan rahoitussysteemiin

2. uusi rivi “museot” (vastakohtana yliopistoille, muut rivit)
3. uusi rivi math&stats, summa kahdesta rivistä

Case1: piste joka on luonnostaan/luonteeltaan/ sisällöllisesti (inherently) erilainen kuin muut. Tutkimusongelman määräämä asia.

- museotutkijat ovat eri porukkaa kuin yo-jengi, mutta voidaan silti visualisoida yo-tutkijoiden “avaruudessa”.
- akseleiden contribuutiot pisteelle (relative contributions or squared cosines or squared correlations) → paljonko kartta selittää pisteen inertiasta.

Monenlaista lisätietoa voidaan ottaa mukaan

- edellisten tutkimusten frekvenssitaulukko, tai uudempien
- tavoiteprofiilit; kuinka kaukana ollaan

Case2: outlier (täysin erilainen profiili), pieni massa . Jos mukana, voi vaikuttaa paljon. Pieni massa taas on sisällöllinen syy jättää pois, vähän väkeä tässä ryhmässä.

Ole tarkkana etäällä keskiarvosta olevien pisteiden kanssa, joilla on pieni massa ja ne (silti) vaikuttavat paljon ratkaisun inertiaan. Kun passiviisiksi, niiden sijanti voidaan esittää kartalla , eikä ratkaisuavaruus (eli kartta) vääristy.

Toinen ratkaisu: yhdistä tulkinnan kannalta järkevästi johonkin olemassaolevaan riviin tai sarakkeeseen.

s. 93 eivät ole kovin vakava ongelma CA:ssa, sillä pieni massa pienentää vaikutusta (etäisyys x massa). Todellinen ongelma on se, että ne ovat niin kaukana muista pisteistä- → Ch 13 alternative scalings.

Case 3: luokkien ryhmiä tai luokan alaryhmiä esittämässä

- esim math&stats – aggregaatti (sentroidi, painoina massat) (Saksa! Belgia!)

Edellä ylimääräisiä profiilipisteitä, jotka projisoidaan valmiiseen karttaan. Vaihtoehto on sijoittaa / laskea koordinaatit “relative to the set of vertex points in an asymmetric map”.

Kun riviprofiilien pääakselit on ratkaistu, sarakkeiden standardikoordinaatit ovat “column vertex point” -pisteiden projektioita pääakseleille (onko sama kuin ideaalipiste? Ei, on se projisoitu piste!). Täydentävä rivipiste saadaan uuden profiliin elementeillä painottamalla (verteksejä ?) s. 94.

Pisteen representaation tulkintaa: vaikka täytentävillä pisteillä ei ole massaa (ja siis ei kontribuutiota to the principal inertias), niiden suhteelliset kontribuutiot voidaan silti tulkita ja arvioida kuinka hyvin ne on esitetty. Suht.kontrib. liittyvä profiilien ja akseleiden välisiin kulmiin, massat eivät mukana. Hyvä läpiluenta suhteellisen kontribuution tulkinnasta : kulmien kosinit akseleittain, ja niiden neliöt kertovat kuinka suuri osuus sijainnista on saatu tasolla kuvattua (contained in the plain), paljonko jää muihin dimensioihin.

Vertex points (käännös!) ovat täydentäviä pisteitä, tulkinnan avuksi projisioidaan karttaan mutta ei käytetä sen määrittämisessä. Siispä voidaan lisätä riveinä “ideaalirivit” (jos sarakkeet a, b,… , niin rivi a = 1,0,0…, b= 0,1,0) jne. Sarakkeiden identiteettimatriisi datan jatkoksi! Koordinaatit ovat sarakkeiden standardikoordinaatit, eli ovat column vertices).

Huom! tätä ei pidä sekoittaa dummy-koodaukseen, joka liittyy MCA:han.

Y on lukumäärämuuttuja, siinä vain sattuu olemaan arvoja 0 ja 1, ei indikaattorimuuttuja.

Jatkuvat supp.muuttujat: vaatii pohdintaa s. 95

#zxy Suppoints.docx loppuu …

#zxy Muuttujalle tai profiililille – täydentävä piste – voidaan laskea koordinaatit, vaikka se ei ole mukana vaikuttamassa akselien määrittelyyn. Pisteiden massa on nolla, mutta ne voidaan esittää muun datan määrittelemässä koordinaatistossa

#zxy Täydentävät muuttujat – idea lavenee MCA-esimerkeissä, joissa juuri tulkinnan tueksi karttaan tulostetaan luokkakeskiarvoja ja luottamusellipsejä.

#vaikea juttu: miten noiden osajoukkojen tunnuslukujen laskennassa voisi maatasolla huomioida otanta-asetelman?

VUOROVAIKUTUSMUUTTUJAT (”INTERACTIVE CODING”)

zxy: ikä ja sukupuoli esimerkkeinä, aineisto (ehkä) vielä vain kuusi maata ja yksi kysymys? Vai laajennetaanko, yksi kysymys ja isompi joukko maita?

TAULUKOIDEN YHDISTÄMINEN (STACKED AND CONCATENATED TABLES)

#zxy CA taipuu moneen, sillä lähtökohta on yleinen (tämä pitää selostaa teoria-luvussa). Vaikeaa on tulkinta, mitä yhteyksiä (”korrelaatioita”) kartta kuvaa, ja mitä se erityisesti **ei** kuvaa.

#zxy Lisämuuttujat; luokittelumuuttujia ja niiden yhteisvaikutusmuuttujia (engl. kankeasti “interactively coded variables”), esimerkkinä ikäluokka ja sukupuoli. Tämä esitellään ensin, ja samalla käy ilmi että kovin pitkälle ei päästä, interaktiomuuttujien määrä kasvaa nopeasti. Joustava työkalu kuitenkin.

#KS Taulukoiden yhdistäminen (pinoaminen, liittäminen sarakkeiden suuntaan) on **vaihtoehto** vuorovaikutusmuuttujien koodaukselle. Tulkinnasta CAiP-kirjassa tarkka selostus, hieman hankala.

#KS between or within sets (MG kalvot – viikko 1) Mitä datalta kysytään, mikä on tutkimusasetelma? Yksinkertaisin tapaus on kahden luokittelumuuttujan yhteyksien analyysi (rivit ja sarakkeet). Jos muuttujia on enemmän, on kaksi mahdolllisuutta (ainakin?). Muuttujat voidaan jakaa kahteen ryhmään, “selittäviin” tai taustamuuttujiin (demografia, koulutus yms.) ja “selitettäviin” muuttujiin (esim. mielipiteet, vastaukset kysymyksiin). Tutkimusongelma on muuttujaryhmien välinen suhde. Muuttujat voivat myös olla tavalla tai toisellla samankaltaisia (homogeenisia), ja tutkimusongelma on muuttujien yhteydet tämän joukon sisällä. Yleensä muuttujat ovat tässä tapauksessa “vastemuuttujia” faktorianalyysin tapaan, ja mitta-asteikko on sama (esim. vastaukset kysymyksiin).

#Tärkeitä käsitteitä ja erotteluja (MG2017hy – kalvot)

zxy Tähän tarvitaan lisää viitteitä, ja asian esittely omin sanoin. Tärkeä asia! Varianssianalyysin tapainen tutkimusasetelma. Viikon 2 kalvoissa myös erottelu “between and within-set inertia w.r.t samples” ja “between and within-set analysis w.r.t variables (tämä yllä). Näissä ei ole mitään uppouutta, mutta tuo hyvin esiin tutkimusasetelmien erot, ja ne taas juontuvat (tai pitäisi) tutkimusongelmista (datan ehdoilla).

between – within sets

wrt samples

wrt variables

Tämä johdataa MCA-kysymyksiin, asettaa rajat CA:lle. Oleellinen on myös käytännöllinen rajoitus: vuorovaikutusmuuttujien määrä kasvaa älyttömyyksiin.

ABBA JA "MATRIISIPARIT" (MATCHED MATRIXES)

zxy ABBAn perusidea ja inertian dekomponointi, inertia on "alimatriisien" inertioiden summa (tarkista!).

zxy Miten tehdään ca-paketilla? Selitetty 2017 laskareissa ja CAiP-kirjassa, juurta jaksain. ABBA-artikkelissa varmaan tarkemipi perustelu.

Viitteet: MG-kalvot, CAiP-kirja, ABBA-artikkeli

zxy Täsmennä: square tables, matched matrixes jne. Yleistettävissä useammille "yhteensopiville matriiseille", sama lohkodiagonaalinen rakenne.

6. USEAMMAN MUUTTUJAN KORRESPONDENSSIANALYYSI

zxy vain lyhyesti, johdatteluna. Tämä luku paisuttaa sivumäärää, koska nyt graafinen analyysi pääsee valloilleen

zxy yleistys yksinkertaisesta – painopiste graafisessa analyysissä

zxy tulkinnan hankaluudet – vastuu on käyttäjällä

USEAN MUUTTUJAN KORRESPONDENSSIANALYYSI

zxy Perusideat:

- on joko indikaattorimatriisin tai Burtin matriisin CA
- havaintojen / yksilöiden pilvi → rivit, muuttujien pilvi → sarakkeet
- inertian selitysosuuden "pessimistinen" vääristyminen, käsite ei ihan sovellu indikaattori- tai Burtin matriisille suoraan
- ratkaisut tähän, JCA (joint ca)

ISSP -DATA – ESIMERKKI

- tutkimusasetelma "naiset työelämässä", muuttujina n vastausta, muutamia taustamuuttujia
- kuvia muutama
  (tämä on vain maistiainen)

AINEISTON RAJAAMINEN GRAAFISESSA ANALYYSISSÄ

zxy hieman lisää, perustelu alla. Luultavasti rajaan varsinaisen subset-analyysin pois, mutta tuleepa mainittua. CA/MCA – kehikon joustavuus esitellään esimerkein.

zxy subset analysis – MG:n artikkeli "vihreässä kirjassa"

zxyMCA:n tulokset ovat usein itsestään selviä. Tärkeää tarkentaa ja rajata, joko graafisesti tai myös analyyttisemmin. Ensimmäinen vaihtoehto vaatii näppärää koodaamista, esimerkit löytyvät MG2017Hy-kurssin ja harjoitustehtävien materiaaleista. Nämä ovat muuten ensimmäiset CA-karttani! Jälkimmäinen vaihtoehto on analyyttisempi (vai onko?), käytetään koordinaatiston ja akselien määrittämiseen koko dataa, mutta esitetään koordinaatistossa vain kiinnostuksen kohteena olevat osajoukot.

zxy Voi esitellä vain yksinkertaisen MCA-esimerkin avulla graafisia tekniikoita, joilla “tukkoisia” kuvia terävöitetään (havaintopisteiden sijaan sentroidit tai luottamusellipsit, zoomaaminen osaan kuva-alueesta, onko muita?). subset –analyysin voi mainita hienompanan vaihtoehtona.

#KS s. 197 : yhteiskuntatieteissä MCA:n tärkein sovellusalue on survey-tutkimusten vastauskategorioiden yhteyksien visualisointi. MG, “vihreä kirja” (Ch8 MCA of subsets of Response Categories) #V (tähän tarvitaan oma viite?).

zxy MG, “vihreä kirja” s. 198:

“The maps obtained by MCA are frequently overpopulated with points…” Ongelmaan on esitetty ratkaisuja, esim. “not plotting points that contribute weakly to the principal axes of the map, but this would be undesirable when we are truly intereted in each category across all the questions. Furthermore, it is commonly found that the principal dimensions of MCA tell an obvious and unsurprising story about the dat at hand while the more interesting patterns are hidden in higher dimensions. Exploring further dimenisons is not a simpel task because all the category points appear on and contribute to every dimension, to a greater or lesser extent. The basic problem is that the MCA map is trying to show many different types of relationships simultaneously, and these relationships are not isolated to particular dimensions. While the technique does its best to visualize all the response categories, the maps may notbe easily conducive to visualizing those relationships of particualr interest to the researcher.” Data ISSP 1994 artikkelin esimerkeissä.

7. YHTEENVETO – MENETELMIEN VUOROPUHELU

zxy provokatoorinen aloitus: Gifin kirja bookdown.org – sivustolla (Multivariate analysis with optimal scaling), kirja on osin keskeneräinen. Alfred Gifi on (tietenkin) nom de plume (https://en.wikipedia.org/wiki/Jan_de_Leeuw).

”The type of multivariate analysis (MVA) we discuss in this book is sometimes called descriptive or exploratory, as opposed to inferential or confirmatory. It is located somewhere on the line between computational linear algebra and statistics, and it is probably close to data analysis, Big Data, machine learning, knowledge discovery, data mining, business analytics, or whatever other ill-defined label is used for the mode du jour”

https://bookdown.org/jandeleeuw6/gifi/

Aika lavea näkemys, datatiede tulee ja jyrää, jättimäiset tekniset korkeakoulut nitistävät pikkuruiset tilastotieteen laitokset! Kiista on jo ohi!

”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.”

Neither do men put new wine into old bottles: else the bottles break, and the wine runneth out, and the bottles perish: but they put new wine into new bottles, and both are preserved. (Matthew 9:17)

Mitä jäi pois?

Kanoninen CA

Vaihtoehtoisen koodaustavat, havaintojen ”tuplaus”, sumea (fuzzy) koodaus.

8. LIITTEET

TAULUKOITA JA TUNNUSLUKUJA?

zxy paisuttavat ehkä muuten liikaa varsinaista sisältöä, mutta niihin pitää voida viitata.

R-KOODI

voiko r-koodissa selostaa myös r- aiheiset jutut ja niksit?

OHJELMISTOT JA VERSIOT

R, Pandoc, MikTeX

työasema?

BOOKDOWN JA GITHUB

ehkä edellinen kohta voisi olla myös tässä?