Korrespondenssianalyysi - graafinen ja geometrinen data-analyysin menetelmä

Jussi Hirvonen

Versio 0.04, tulostettu 2018-08-10

Alkutoimia

Ladataan r-paketit, ei tulosteta dokumenttiin. Pelkkä YAML- 'front matter', lisäkonfiguroinnit tiedostoissa _bookdown.yml ja _output.yml.

Dokumettiin kuuluvat Rmd-tiedostot luetellaan eksplisiittisesti (ei vielä).

Ideoita

1. Ehkä automaattista R-kirjastojen dokumentointia voisi harkita?
2. Saako gitbook-tulosteessa päälle asetuksen code_folding: hide? Vaatii teeman (theme), jos tarpeen voi lisätä _output.yml - tiedostoon esim. html_book - formaatin.
3. Lähdeviitteiden doi-linkit eivät toimi? Ja näköjään lisätyt ISSP-viitteet eivät toimi PDF-tulostuksessa. ! Package natbib Error: Bibliography not compatible with author-year citations. Error: Failed to compile JH_ca.tex. See JH_ca.log for more info. Execution halted Exited with status 1. Koitin korjailla RefWorksiin, ei auttanut (9.8.18), lokitiedostossa listattu ref-id:t joissa ongelmia. Poistin ehkä tarpeettomat http-linkki-viitteet dokumentteihin, ne voi ehkä esittää tekstissä.

Luku 1

Johdanto

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

Mahdollisia lisäyksiä

1. Lyhyt esitys CA:n historiasta (vai omaksi luvuksi, luku 2)?
2. Käytetyt ohjelmistot, tekninen ympäristö ml. bookdown-asetukset. Ehkä paremmin omaksi liitteeksi?
3. Tavoitteet, sisältö, rajaukset (jota voi myöhemmin täydentää)
4. Muutamat puutteet, onko kerrottava tässä?

- data: ei huomioida sitä, että otoskoot vaihtelevat aika paljon eli "maapainot" eri suuruisia
- ei huomioida muitakaan otantaan liittyviä asioita (tämä ainakin mainittava data-osuudessa)
- kuvaileva menetelmä, mutta mikä on tutkimusongelma? Sellainen pitäisi olla.

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

Tutkielman tavoite (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 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 Miksi eksporatiivinen (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

- 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).

Käytetyt ohjelmistot

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

Tarkemmin liittessä.

Korrespondenssianalyysin historiaa

zxy Tiivis esitys lähteineen. Ehkä asiaan palataan kun itse menetelmä on esitelty?

Luku 2

Data

zxy Voisi miettiä paremman otsikon. Galku-paperin alusta on lisäilty viitteitä Refworksiin, mutta hieman hanklaa. 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 muuta ([RefWorks:doc:5b6c7f6ce4b0e4e15164ab1a] ja [RefWorks:doc:5b6c7debe4b0e4e15164ab00]). Löytyy myös seurantaraportti([RefWorks:doc:5b155e0ce4b044dfd738458f]). **viitteet pois- ehkä tekstiin linkkeinä?**

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

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

ks

zxy Mksi data sovelutuu korrespondenssianalyysin esittelyyn?

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.

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.

Aineiston kuvailu (tietosisältö)

Jäsennys:

1.
2.

3.

Aineiston valinta tutkimukseen

zxy Viitteitä myös r-ratkaisuihin, jotka selostetaan koodissa. Erityisesti (a) puuttuvat tiedot ja (b) likert-asteikko faktorina (ilman järjestystä).

zxy Aluksi kuusi maata

zxy Sitten monta maata

Aineiston kuvailu (tunnuslukuja)

zxy ehkäpä taulukoiden lisäksi Likert-kuva?

Luku 3

Yksinkertainen korrespondenssianalyysi

zxy Tässä yksi kysymys, kuusi maata, peruskäsitteet lopussa

zxy Luvun tärkeimmät asiat; mitä on luvassa?

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

Luku 4

Yksinkertaisen korrespondenssianalyysi - tulkinnan syventäminen

xyz Tarkasti läpi keskeiset tulokset ja niiden tulkinta, kaavat, ja ytimenä eri kuvat eli kartat.

Luku 5

Yksinkertaisen korrespondenssianalyysin laajennuksia

xyz Yksinkertainen korrespondenssianalyysi on menetelmän tulkinnan perusta. Perusasetelmaa kahden luokittelumuuttujan ristiintaulukoinnista voidaan laajentaa monipuolisempiin tutkimusasetelmiin. Varsinainen useamman muuttujan korrespondenssianalyysi (MCA - multiple correspondence analysis) esitellään seuraavassa luvussa.

Greenacre, Michael, ja Trevor Hastie. 1987. "The Geometric Interpretation of Correspondence Analysis". *Journal of the American Statistical Association* 82 (398): 437–47. https://doi.org/10.1080/01621459.1987.10478446.

Jorat, Jorge R., Ann Evans, Franz H ollinger, Lilia Dimova, Canada Carleton University Survey Centre Ottawa, Lulu Li, Carolina Segovia, ym. 2016. "International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012". https://doi.org/10.4232/1.12661".

Xie, Yihui. 2016. *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman; Hall/CRC. https://bookdown.org/yihui/bookdown/.

———. 2018. *bookdown: Authoring Books and Technical Documents with R Markdown*. https://CRAN.R-project.org/package=bookdown.

Yihui Xie, Garrett Grolemund, J. J. Allaire. 2018. *R Markdown: The Definitive Guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown/.