Korrespondenssianalyysi - graafinen ja geometrinen data-analyysin menetelmä

Jussi Hirvonen

Versio 0.02, tulostettu 2018-08-08

Sisältö

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.

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.

1.1 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 and 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).

1.2 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.

1.3 Korrespondenssianalyysin historiaa

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

Luku 2

Data

xyz Voisi miettiä paremman otsikon

Luku 3

Yksinkertainen korrespondenssianalyysi

xyz Tässä yksi kysymys, kuusi maata, peruskäsitteet lopussa

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.

Lähteet

Greenacre, M. and Hastie, T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82(398):437–447. doi: 10.1080/01621459.1987.10478446.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.7.

Yihui Xie, J. J. Allaire, G. G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC.