Korrespondenssianalyysi - graafinen ja geometrinen data-analyysin menetelmä

Jussi Hirvonen

Versio 0.06, tulostettu 2018-08-12

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 _bookdown.yml-tiedostossa.

RefWorksistä eksportattu bib-tiedosto kannattaa avata ensin (Atomilla), ja korjailla skandit jos niissä on vikaa.

Ideoita

1. Ehkä automaattista R-kirjastojen dokumentointia voisi harkita?
2. Gitbook-tulosteessa ei saa koodia "piilotettua", asetus "code_folding: hide" vaatii teeman (theme). _output.yml - tiedostoon lisätty html_book - formaatti, siinä voi tarvittaessa käyttää piilotusta.
3. Versiointi: 0.0n aloittelua, 0.n jäsentely koko paperille, 1.n.n valmiimpaa tekstiä.

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 Tärkeimmät lähteet ja ohjelmistot

zxy Tarvitaanko tämä, perustelu? Muutamat lähteet aivan keskeisiä, ja MG:n kurssi pitää mainita.

1.2.1 Lähteet

Michael Greenacre luennoi lyhyen kurssin korrespondenssianalyysistä Helsingin yliopistossa keväällä 2017(Greenacre, 2017a). 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 kärsivällisesti kirjoitetut perusoppikirjat ovat tehneet menetelmää laajasti tunnetuksi englantia lukeville.

Ranskalaisen lähestymistan perusoppikirja(Roux and 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.

Korrespondenssianalyysi tuli osaksi suomalaista Survo-ohjelmistoa jo vuonna (**????**), ja menetelmää on esitelty ainakin kahdessa oppikirjassa(Mustonen, 1995) ja (Vehkalahti, 2008).

1.2.2 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

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 et al., 2016).

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

ks

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

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

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.

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.

2.1 Aineiston kuvailu (tietosisältö)

Jäsennys:

1.
2.
3.

2.2 Aineiston rajaaminen

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

puuttuvat vastaukset aluksi pois (rankka rajaus - vain yksinkertaistus, menetelmän esittelyn vuoksi)

zxy Aluksi kuusi maata

zxy Sitten monta maata

2.3 Aineiston kuvailu (tunnuslukuja)

zxy ehkäpä taulukoiden lisäksi Likert-kuva?

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), (Greenacre, 2017b)).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."

Blasius ja Thiessen ((Blasius and 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?

Luku 3

Yksinkertainen korrespondenssianalyysi

zxy Tässä yksi kysymys, kuusi maata, peruskäsitteet lopussa

zxy Luvun tärkeimmät asiat; mitä on luvassa?

3.1 Äiti töissä

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

3.2 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.

Lähteet

Blasius, J. and Thiessen, V. (2006). Assessing data quality and construct comparability in cross-national surveys. *European Sociological Review*, 22(3):229–242.

Greenacre, M. (2003). Singular value decomposition of matched matrices. *Journal of Applied Statistics*, 30(10):1101–1113.

Greenacre, M. (2017a). Multiple correspondence analysis (mca): Theory and practice, spring 2017 (university of helsinki). Course material in moodle.helsinki.fi requires authentication.

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.

Greenacre, M. J. (2010). *Biplots in Practice*. Fundacion BBVA, Bilbao, Spain. Onko tämä kirja, vai monografia?

Greenacre, M. J. (2017b). *Correspondence analysis in practice*. CRC Press, Boca Raton, Florida, third edition edition.

Jorat, J. R., Evans, A., ollinger, F. H., Dimova, L., Carleton University Survey Centre, Ottawa, C., Li, L., Segovia, C., Hamplová, D., Jerolimov, D. M., Clement, S. L., Larsen, C. A., Andersen, J., Andersen, J. G., Melin, H., Fridberg, T., Blom, R., Forsé, M., Lemel, Y., Wolf, C., Park, A., Clery, L., Ágústa E. Bj ornsdóttir, Guðmundsdóttir, H., Cleary, A., Lewin-Epstein, N., Kobayashi, T., Sang-Wook, K., Murata, H., Tabuns, A., Krupavičius, A., Morones, C., Ceballos, V., Palacios, F., Moran, M., Kolsrud, K., Skjåk, K. K., Guerrero, L., Zielinski, M. W., Khakhulina, L., Bahna, M., Malnar, B., Hafner-Fink, M., Tos, N., Struwig, J., Méndez, M., García-Pardo, N., Edlund, J., Joye, D., Sapin, M., hwa Chang, Y., Çarkoglu, A., Kalaycıoğlu, E., Smith, T. W., Marsden, P. V., Hout, M., León, R. B., Ávila, O., Camardiel, A., Deshmukh, Y., Kolosi, T., Carton, A., Vanderkelen, F., Ganzeboom, H. B. G., Vala, J., and Ramos, A. (2016). International social survey programme: Family and changing gender roles iv - issp 2012.

Mustonen, S. (1995). *Tilastolliset monimuuttujamenetelmät*. Survo Systems, Helsinki.

Roux, B. L. and Rouanet, H. (2004). *Geometric data analysis: from correspondence analysis to structured data analysis*. Kluwer Academic Publishers, Dordrecht.

Vehkalahti, K. (2008). *Kyselytutkimuksen mittarit ja menetelmät*. Tammi, Helsinki.

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.