G Luku 1 Yksinkertainen korrespondenssianalyysi

Jussi Hirvonen

9.4.2018

Contents

0.1 Data

Ladataan käytettävät paketit:

```
# pitääkö laittaa järjestykseen, vanhemmat ensin?
library(rgl)
library(ca)
library(haven)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## v ggplot2 2.2.1     v readr   1.1.1
## v tibble  1.4.2     v purrr   0.2.4
## v tidyr   0.7.2     v stringr 1.2.0
## v ggplot2 2.2.1     v forcats 0.2.0
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#library(forcats) latautuu haven-paketissa
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(rmarkdown)
library(ggplot2)
library(furniture)
```

```
## Warning: package 'furniture' was built under R version 3.4.4
```

```
## furniture 1.7.3: learn more at tysonbarrett.com
```

```
# ehkä viimeiseksi sessionInfo() ja tämä koodi piiloon
```

Yksinkertainen korrespondenssianalyysi on kahden luokitteluasteikon muuttujan riippuvuuksien geometrista analyysiä. Lähtökohta on kahden muuttujan ristiintaulukointi, alkuperäinen data voi olla muillakin asteikoilla mitattua. Menetelmän ydin on tarkastella molempien muuttujien – taulukon rivien ja sarakkeiden – riippuvuuksia kaksiulotteisena kuvana. Kuvaa kutsutaan myös kartaksi, ja tulkinnan ensimmäinen askel on kartan "koordinaatiston" tulkinta. Kaikki etäisyydet kuvassa ovat suhteellisia, mutta koordinaatiston tulkinnassa voidaan lähteä siitä, että suhteellisesti kaukana toisistaan olevat pisteet ovat kaukana myös alkuperäisessä datassa.

0.1.0.1

(1) Data - tässä tiiviimmin, aineiston kuvailu tarkemmin liitteeseen. Pitää perustella rajaukset ja kertoa miten ne tehdään.

(2) Ensimmäinen taulukko: profiilit, massat, keskiarvoprofiilit, khii2 - riippumattomuustesti ja etäisyysmitta

(3) Hyvin tiivis esitys CA:n perusideasta, mutta ilman aivan simppeleitä kolmiulotteisia kuvia (niitä on jo)

(@)Ensimmäinen symmetrinen kartta, perustulkinta (mitä kuvasta voidaan sanoa, mitä ei) (@) Lyhyt viittaus graafisen esityksen tulkintapulmiin, jotka eivät ole kovin pahoja. Niihin palataan kaksoiskuva-jaksossa.

(4) Tulkinnan syventäminen

0.2 Data

ISSP Research Group (2016): International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012. GESIS Data Archive, Cologne. ZA5900 Data file Version 4.0.0, doi:10.4232/1.12661

https://search.gesis.org/research_data/ZA5900 2012

Muuttujakuvaukset ja muut tiedot (linkit?) http://zacat.gesis.org/webview/index.jsp?object=http://zacat.gesis.org/obj/fStudy/ZA5900

Suomenkielinen lomake (ZA5900_q_fi-fi.pdf), https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e&doi=10.4232/1.12661

Käyttöehdot: https://www.gesis.org/en/services/data-analysis/more-data-to-analyze/data-archive-service/

Tiedonkeruumenetelmä ja otoskoko: https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e&doi=10.4232/1.12661 Viimeisin Portugali 29.06.2014 - 31.01.2015, ensimmäinen Bulgaria 16.08.2011 - 20.09.2011. Suurin osa muista 2012-13, kuten Suomi (21.09.2012 - 07.12.2012).

edit: aineiston kuvailua voi ja kannattaakin jatokssa tarkentaa, ja laittaa se liitteeksi. Dokumentointi on hyvin tarkka, tiedot löytyvät haastattelumenetelmista (parerilomake, tietokoneavusteinen haastattelu, jne),

maakohtaisten taustamuuttujien harmonisoinnista maittain, otantamenetelmistä jne. Esittelen vain aineiston tärkeimmät rajaukset.

0.3 Aineiston rajaaminen

Ainestossa (jatkossa ISSP2012) on kyselytutkimusen tulokset 41 maasta. Lisäksi aineistossa on runsaasti demografisia ja muita taustatietoja. R-koodista selviää käytetty versio (SPSS-tiedoston nimi) ja rajauksessa käytetyt muuttujat.

0.3.1 Rajaukset

1. Eurooppa ja samankaltaiset maat (28)

Bulgaria, Czech Republic, Denmark, Finnland,France, Germany, Great Britain, Ireland, Latvia, Lithuania, Norway, Poland, Sweden, Slovakia Slovenia, Spain, Switzerland, Australia, Austria, Canada, Croatia, Iceland, Russia, United States, Belgium, Hungary, Netherlands, Portugal

Pois jätettiin 13: Argentiina,Turkki, Venezuela, Etelä-Afrikka, Korea, Intia, Kiina, Taiwan, Filippiinit, Meksiko, Israel, Japani, Chile.

2. Maat joissa varsinaisissa tutkimuskysymyksissä on käytetty poikkeavia luokitituksia tms. Esimerkiksi Espanjan datassa on jätetty pois neutraali “en samaa enkä eri mieltä” - vaihtoehto, Unkarin datassa on omia versioita kysymyksistä jne. Nämä maat jätetään pois ainakin aluksi, vertailukelpoisuuden vuoksi.

3. kaikki havainnot, joissa on puuttuvia tietoja. Tämä rajaus on kyselytutkimuksessa ankara, tai oikeastaan kelvoton. Oikea menettely olisi imputoida jollain menetelmällä puuttuvat tiedot, mutta rajaan otantatutkimuksen menetelmät tutkielman ulkopuolelle (aiheesta löytyy artikkeleita. . .).Yksittäisten vastausten puuttuminen eli erävastauskato ohitetaan aluksi, mutta siihen palataan. Korrespondenssianalyysiin on helppo ottaa mukaan myös puuttuvat tiedot, sillä data on luokitteluasteikon dataa. Yksikkövastauskato eli otokseen poimitut joita ei ole tavoitettu ollenkaan on kansallisen tason ongelma, joka on ratkaistu vaihtelevin tavoin. Tiedot löytyvät aineiston dokumentaatiosta. Aineistossa on myös mukana painomuuttujat, mutta ne soveltuvat vain jokaisen maan omaan aineistoon.

```
# kolme maa-muuttujaa datassa. V3 erottelee joidenkin maiden alueita, V4 on koko maan
#two country code variables based on the ISO Code 3166. One identifies
#countries as a whole, the other one possible subsamples, such as East and West Germany. The cross
#tabulations shown in this Variable Report are based on a third, alphanumerical country code variable,
#which also identifies subsamples."
#V3 - Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)
# V3 erot valituissa maissa
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East
# 62001 PT-Portugal 2012: first fieldwork round (main sample)
# 62002 PT-Portugal 2012: second fieldwork round (complementary sample)
# Myös tämä on erikoinen, näyttää olevan vakio kun V4 = 826:
# 82601 GB-GBN-Great Britain
# Portugalissa ainestoa täydennettiin, koska siinä oli puutteita. Jako ei siis ole oleellinen,
# mutta muuut ovat.
# Maat:
# 36 AU-Australia
# 40 AT-Austria
# 56 BE-Belgium
```

```
# 100 BG-Bulgaria
# 124 CA-Canada
# 191 HR-Croatia
# 203 CZ-Czech Republic
# 208 DK-Denmark
# 246 FI-Finland
# 250 FR-France
# 276 DE-Germany
# 348 HU-Hungary
# 352 IS-Iceland
# 372 IE-Ireland
# 428 LV-Latvia
# 440 LT-Lithuania
# 528 NL-Netherlands
# 578 NO-Norway
# 616 PL-Poland
# 620 PT-Portugal
# 643 RU-Russia
# 703 SK-Slovakia
# 705 SI-Slovenia
# 724 ES-Spain
# 752 SE-Sweden
# 756 CH-Switzerland
# 826 GB-Great Britain and/or United Kingdom
# 840 US-United States

#valittavien maiden kolminumeroinen ISO 3166 - koodi vektoriin
incl_countries <- c(36, 40, 56,100, 124, 191, 203, 208, 246, 250, 276, 348, 352, 372, 428, 440,
                    528, 578, 616, 620, 643, 703, 705, 724, 752, 756, 826, 840)
ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav")
#str(ISSP2012.data) #61754 obs. of  420 variables
ISSP2012jh1.data <- filter(ISSP2012.data, V4 %in% incl_countries)
#length((ISSP2012jh1.data))
#names(ISSP2012jh1.data)
#str(ISSP2012jh1.data) #37816 obs. of  420 variables
#V5 - V67 kysymyksiä, joillain mailla omat vastaukset joihinkin omina muuttujina, esim. ES_V5
#ja muutkin Espanjan kysymykset (kaikki?)
#$ V5      :Class 'labelled'  atomic [1:37816] 5 1 2 2 1 NA 2 4 2 2 ...
#  .. ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
#  .. ..- attr(*, "format.spss")= chr "F1.0"
#  .. ..- attr(*, "labels")= Named num [1:8] 0 1 2 3 4 5 8 9
#  .. .. ..- attr(*, "names")= chr [1:8] "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree
# $ ES_V5   :Class 'labelled'  atomic [1:37816] NA NA NA NA NA NA NA NA NA NA ...
#  .. ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
#  .. ..- attr(*, "format.spss")= chr "F1.0"
#  .. ..- attr(*, "display_width")= int 4
#  .. ..- attr(*, "labels")= Named num [1:7] 0 1 2 3 4 8 9
#  .. .. ..- attr(*, "names")= chr [1:7] "NAP: other countries" "Strongly agree" "Agree" "Disagree" ...
#HU_V18
#V18$label
#attr(ISSP2012jh1.data$V5,'labels')
#attr(ISSP2012jh1.data$ES_V5,'labels')
```

Yllä esimerkiksi muuttujan V5 metatiedot. Perusvaihtoehdot ovat 1 - 5, ja joillain mailla on vaihtoehtona

ollut myös “Can’t choose”, muilla taas on vain puuttuva tieto (No answer, 9).

Espanjan aineiston metatiedot muuttujulla ES_V5 taas ovat

```
attr(ISSP2012jh1.data$ES_V5,'labels')
```

```
## NAP: other countries       Strongly agree                Agree
##                    0                    1                    2
##             Disagree    Strongly disagree         Can't choose
##                    3                    4                    8
##            No answer
##                    9
```

```
temp1 <- ISSP2012jh1.data %>% filter(V4 == 724) %>% select(ES_V5, C_ALPHAN)
#str(temp1)
temp1$ES_V5 <- factor(temp1$ES_V5 )
summary(temp1)
```

```
##   ES_V5        C_ALPHAN
##  1   : 572   Length:2595
##  2   :1157   Class :character
##  3   : 633   Mode  :character
##  4   : 199
##  NA's:  34
```

```
#typeof(ISSP2012jh1.data)
#class(ISSP2012jh1.data)
#storage.mode(ISSP2012jh1.data)
#attributes(ISSP2012jh1.data)
```

0.4 Puuttuvat tiedot (erävastauskato)

Datassa ei ole eroteltu vastausvaihtoehtoa “Can’t choose” (8) ja “No answer” (9), ne on (luultavasti) yhdistetty ja koodattu puuttuviksi havainnoiksi. Dokumentaatiosta selviää (s.13), että vaihtoehdon 8 on valinnut 30 ja loput neljä “puuttuvaa tietoa” ovat erävastauskatoa (tai kieltäytymistä tms.). Jokaisen kysymyksen vastauksista löytyy aineiston dokumentaatiossa taulukko, joissa puuttuva tieto on eritelty tarkemmin.

Muiden kuin Espanjan vastaukset kysymykseen V5 jakautuvat näin:

```
temp2 <- ISSP2012jh1.data %>% filter(!(V4 == 724)) %>% select(V5, C_ALPHAN)

#str(temp1)
temp2$V5 <- factor(temp2$V5 )
temp2$maa <-  factor(temp2$C_ALPHAN)
summary(temp2)
```

```
##     V5          C_ALPHAN              maa
##  1   :11763   Length:35221       FR     : 2409
##  2   :13449   Class :character   BE     : 2202
##  3   : 3786   Mode  :character   CZ     : 1804
##  4   : 4266                      DE     : 1766
##  5   : 1072                      AU     : 1612
##  NA's:  885                      RU     : 1525
##                                  (Other):23903
```

```
temp2 %>% tableX(V5,maa,type = "count")
```

```
##          maa
## V5        AT   AU   BE   BG   CA  CH   CZ   DE   DK   FI   FR   GB-GBN
##   1       431  358  730  140  278 375  597  1041 849  457  1238 267
##   2       409  715  789  425  400 591  502  481  372  420  696  448
##   3       111  167  247  157  91  95   316  45   48   98   160  101
##   4       150  270  225  206  136 152  216  141  81   122  196  99
##   5       47   60   63   36   57  19   110  37   44   25   74   18
##   Missing 34   42   148  39   10  5    63   21   9    49   45   17
##   Total   1182 1612 2202 1003 972 1237 1804 1766 1403 1171 2409 950
##          maa
## V5        HR   HU   IE   IS   LT   LV   NL   NO   PL   PT   RU   SE   SI
##   1       295  297  357  492  100  317  178  341  198  244  412  387  428
##   2       413  323  500  523  528  345  597  680  491  508  571  420  436
##   3       82   194  109  72   256  111  193  138  103  73   233  122  71
##   4       153  124  189  74   232  167  216  207  253  149  215  80   63
##   5       51   56   34   9    25   55   63   26   51   20   31   22   9
##   Missing 6    18   26   2    46   5    68   52   19   7    63   29   27
##   Total   1000 1012 1215 1172 1187 1000 1315 1444 1115 1001 1525 1060 1034
##          maa
## V5        SK   US   Total
##   1       614  342  11763
##   2       273  593  13449
##   3       102  291  3786
##   4       84   66   4266
##   5       30   0    1072
##   Missing 25   10   885
##   Total   1128 1302 35221
```

Esimerkiksi Ruotsin puuttuviksi tiedoiksi koodatuista 29 havainnosta 19 valitsi “can’t choose”(8) ja 10 kieltäyti vastaamasta (9) tms. Dokumentti, s.12.

Tarkastellaan aineston puuttuvia havaintoja hieman tarkemmin. Puuttuvat tiedot on koodattu aineistoon näin: 0: Not applicapble (NAP), Not available (NAV) 7: (97,997, 9997,. . .): Refused 8: (98, 998, 9998,. . .): Don’t know 9: (99, 999, 9999,. . .): No answer

NAP ja NAV määritellään

“GESIS adds ‘Not applicable’(NAP) codes for questions that have filters. NAP indicates that only a subsample and not all of respondents were asked. Also in the case of country spesific variables, all the other countries are coded NAP.

GESIS adds ‘Not available’ for variables, which in singe countries may not have been conducted for whatever reason."

```
#puut1 <- filter(ISSP2012jh1.data, na.rm == 1)
```

0.4.1 Poikkeavat kysymyset

Espanjan lisäksi Unkarin osatutkimuksessa kysymyksen V18 V19 V20 vastausvaihtoehdot ovat poikkeavat (em.dok, s. 48)

Islannissa kysymykseen V28 (Consider a couple who both work full-time and now have a new born child. One of them stops working for some time to care for their child. Do you think there should be paid leave available and, if so, for how long?) on tarjolla oma vastausvaihtoehto ((97) “Yes, but don’t know how many months”). Kysymyseen “V29 - Q9 Paid leave: Who should pay ja V30(Paid leave: How to divide between parents) Bulgarian kysely on poikkeava (0 NAP (code 0,98 in V28), s. 91).

Hollannin vastausvaihtoehdoissa kysymykseen V35 (Elderly people: Provider of domestic help) on oma variantti “5 Empoyers”, jonka kuitenkin on valinnut vain 6 vastajaa (0,5 %).

V39, V40, V41, V42, V43, V44, V45, V46, V47, V48, V50, V51, V52, V53, V54: paljon poikkeamia, aika vaikeaselkoisia kysymyksiä. Näitä ehkä pitää tutkailla... V55 (Life in general: How happy on the whole) ok.

V56-57 poikkeamia, V58 (Health status) ok V59 “ketjutettu kysymys”, samoin V60-V64. s. 174 - puolison koulutus...

0.5 Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko

```
# pitääkö laittaa järjestykseen, vanhemmat ensin?
library(rgl)
library(ca)
library(haven)
library(dplyr)
library(knitr)
library(tidyverse)
#library(forcats) latautuu haven-paketissa
library(lubridate)
library(rmarkdown)
library(ggplot2)
library(furniture)
```

Tässä esitellään yksinkertainen esimerkki, yksi kysymys (esim. V6) ja muutamia maita ristiintaulukoituna. Johdatteluna aiheeseen esitellään ca-käsitteet profiili, massa ja reunajakauma. Havainnollistetaan rivi- ja sarakeprofiilien vertailua vastaaviin keskiarvoprofiileihin.

Toiseksi riippumattomuushypoteesi ja χ^2 - riippumattomuustesti (pieni huomautus - on monta tapaa testata taulukon riippuvuuksia). Riippumattomuushypoteesi ehdollisena todennäköisyytenä reunajakauman suhteen.

χ^2 - etäisyys, yhteys hajontaa eli inertiaan ca-terminologiassa.

Dimensioiden vähentämisen idea.

Ensimmäinen symmetrinen kartta, tulkinnat ja yksinkertaisimmat perussäännöt (“mitä on oikealla ja vasemmalla”). Jos pisteet ovat alkuperäisessä “pilvessä” kaukana toisistaan, ne ovat sitä myös projektiossa. Kartta, mutta etäisyyksillä ei suoraa tulkintaa paitsi eteisyyksinllä origoon. Rivipisteiden suhteelliset etäisyydet, samoin sarakepisteiden, mutta ei muut.

0.6 Tulkinnan perusteita

```
# pitääkö laittaa järjestykseen, vanhemmat ensin?
library(rgl)
library(ca)
library(haven)
library(dplyr)
library(knitr)
library(tidyverse)
```

```r
#library(forcats) latautuu haven-paketissa
library(lubridate)
library(rmarkdown)
library(ggplot2)
library(furniture)
```

Luvussa syvennetään esimerkin tulkinnan perusteita. Miksi symmetrinen kartta on yleensä paras vaihtoehto, siksi se oletusarvoisesti esitetäänkin. Milloin voi käyttää vaihtoehtoisia esitystapoja?
