Házi feladatok megoldása 11.

Klasztermegoldás validálása a MORI együtthatóval

Smahajcsik-Szabó Tamás, M9IJYM

1. A 8. óra 1. feladatában elmentett klaszterváltozókkal végezz MORI-elemzéseket a random permutáció módszerével, rep = 25 ismétlésszámmal! Melyik klaszterszám megoldása tűnik a legjobbnak?

A random permutáció módszerével, 25 független ismétléssel végzett MORI-elemzés eredményéről az alábbi táblázat tájékoztat.

Scenario	EESS%	Pontbisz	XBmod	SilCoef	HCatlag	CLdelta	GDI24
KClus6	0.24	-0.10	0.04	0.05	0.23	-0.14	-0.29
KClus7	0.23	-0.08	0.18	0.09	0.23	-0.14	-0.11
KClus8	0.23	-0.06	0.12	0.06	0.23	-0.14	-0.06
KClus9	0.22	-0.04	0.10	0.06	0.22	-0.14	-0.12
KClus10	0.21	-0.05	0.13	0.06	0.21	-0.15	-0.07

Az egyes MORI-mutatók alapján a k=7 megoldás tűnik megfelelőbbnek, mert PB és a módosított Xie-Beni mutató MORI-indexe nyomán random permutáció mellett jobb szeparációt jelent a többi struktúránál.

2. A 8. óra 1. feladatában elmentett klaszterváltozókkal végezz MORI-elemzéseket a korreláló random normális kontroll módszerével, rep = 25 ismétlésszámmal! Melyik klaszterszám megoldása tűnik a legjobbnak?

A kontrolláló random normális kontrollhoz varimax rotációval készült faktorsúly-mátrixot használtam. A súlyok értékeit az alábbi táblázat összegzi.

	PC1	PC2	PC3
PTelj	-0.3116836	-0.1410578	0.9396574
PBoldog	-0.8924786	-0.2799002	0.3537482
Pmagany	0.2318846	0.9636181	-0.1329273

A faktorképzés, a rotáció menetét, és az outputok ROPSTat számára értelmezhető "floadingbetolt.txt" fájlba rendezését az alábbi script segítségével végeztem el.

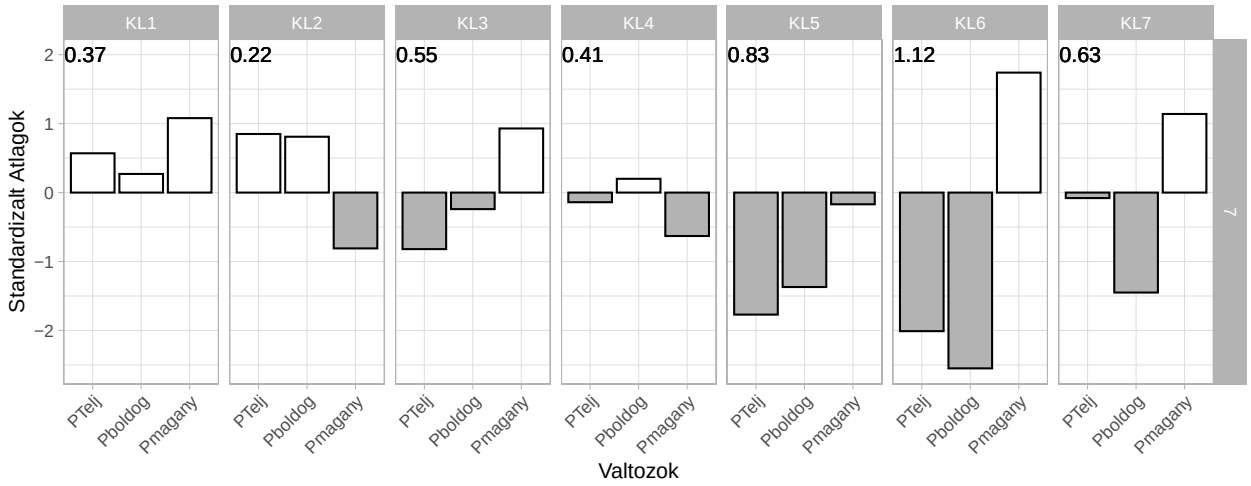
```
library(vegan)
library(pracma)
library(tidyverse)

dataset <- read_csv("../../data/data.csv")
dataset <- dataset[c("PTelj", "PBoldog", "PMagány")]

pca <- rda(dataset, scale = TRUE)
loading <- scores(pca, choices = c(1, 2, 3))$species
rotated_loading <- varimax(loading)$loadings
iloading <- t(pinv(rotated_loading))
scores <- scale(dataset) %*% iloading


weight_matrix <- cor(scale(dataset), scores)


write.table(weight_matrix, "output/pca_weights.txt", row.names = FALSE, col.names = FALSE)
weights <- read.table("output/pca_weights.txt")

weight_content <- ""
for (r in 1:3) {
  actual_row <- paste0(unname(unlist(weights[r, ])), collapse = "\t")
  weight_content <- paste0(weight_content, actual_row, "\n")
}

line1 <- "3 (Number of variables)\n"
line2 <- "PTelj\n"
line3 <- "PBoldog\n"
line4 <- "PMagány\n"

file_content <- paste0(line1, line2, line3, line4, weight_content)
writeLines(file_content, con = "output/floadingbetolt.txt")
```

A MORI-értékeket az alábbi táblázatban foglaltam össze.

Scenario	EESS%	Pontbisz	XBmod	SilCoef	HCatlag	CLdelta	GDI24
KClus6	0.39	0.05	-0.15	0.24	0.39	0.00	-0.64
KClus7	0.40	0.05	0.03	0.24	0.40	-0.03	-0.52
KClus8	0.41	0.03	-0.13	0.19	0.40	-0.06	-0.55
KClus9	0.41	0.03	-0.11	0.19	0.40	-0.08	-0.58
KClus10	0.41	0.02	-0.07	0.19	0.41	-0.10	-0.56

Az egyes MORI-mutatók alapján a k=7 megoldás tűnik megfelelőbbnek, különösen a szeparációs mutatók, így XBmode alapján.

A k=7 struktúra bemutatására az alábbi ábrát készítettem, a standardizált klaszterátlagok és a klaszterenkénti homogenitási együtthatók segítségével.



KL6 kivételével viszonylag jó homogenitással jellemezhető struktúrát kapunk, melynél adott egy boldog, jól teljesítő, de magányos csoport (KL1), egy hasonlóképp boldog, jobban teljesítő és kevésbé magányos klaszter (KL2), egy magányos, boldogtalan rosszul teljesítő (KL3), melynek hasonló mintázatot mutató, de extrémebb rokona a KL6. A KL4 olyanokat sűrít, akik enyhén boldogok, kevésbé magányosak, de nem is teljesítenek jól.

A KL5 a KL1 ellentételezése (minden mutatóban alacsony értéket mutat); a KL7 pedig a KL3 szélsőségesebb értékmintázatot mutató testvére.

3. A 10. óra 6. feladatában elmentett klaszterváltozóval végezz MORI-elemzést a random permutáció módszerével, rep = 25 ismétlésszámmal! Vesd össze a kapott eredményt az 1. feladatban kapottal!

A 25-szörös permutációval számolt MORI-értékeket az alábbi táblázat közli.

Scenario	EESS%	Pontbisz	XBmod	SilCoef	HCatlag	CLdelta	GDI24
BIC3	0.12	-0.14	-0.08	-0.08	0.12	-0.16	-0.43

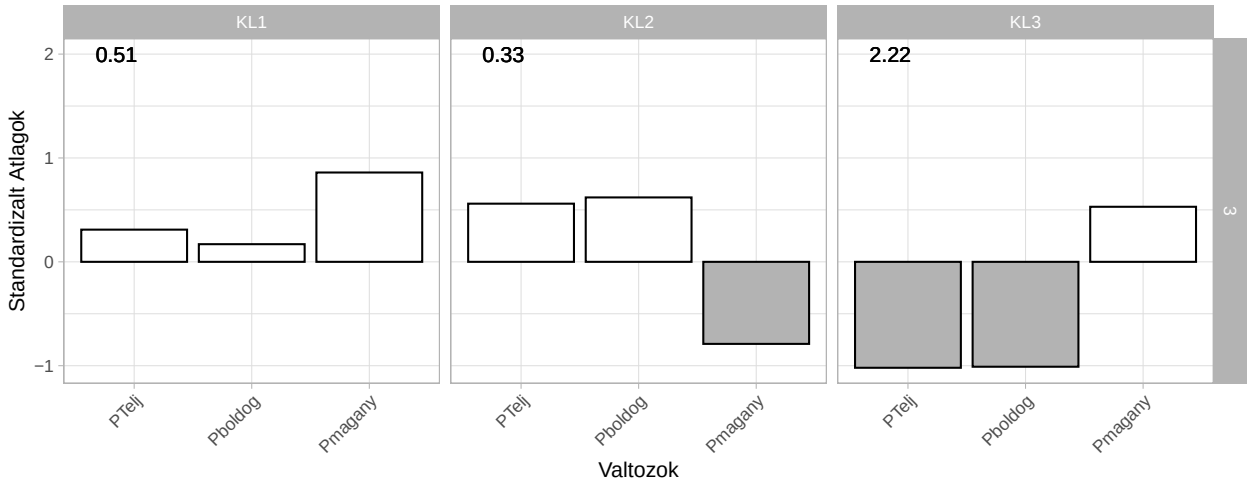
A modell-alapú klaszterezés különböző struktúrái (G=3, 7, 8, 10) közül ezen G=3 struktúra szolgáltatta a legjobb belső validitási mutatókat, melyeket akkor úgy interpretáltam, megközelíti egy HKA eredményét.

A három klaszteres struktúra ugyanakkor permutációval becsült MORI-jai alapján kedvezőtlen szeparációval bír (PB, XBmod, Silhouette együtthatók), viszonylag jó HCÁtlaggal bír (0.21).

4. A 10. óra 6. feladatában elmentett klaszterváltozóval végezz MORI-elemzést a korreláló random normális kontroll módszerével, rep = 25 ismétlésszámmal! Vesd össze a kapott eredményt a 2. feladatban kapottal!

Random, kontrolláló normális kontrollal végzett külső validitás becslés alapján azt látjuk, elfogadható EESS%, XBmod, Solhouette, HCÁtlag, és alacsony CLdelta MORI-értékeket kapunk.

Scenario	EESS%	Pontbisz	XBmod	SilCoef	HCatlag	CLdelta	GDI24
BIC3	0.21	0.04	0.14	0.18	0.21	0.07	-0.46



A tényleges klaszterstruktúra tekintetében elmondható, hogy értlemezhetőbb, kevésbé redundáns (minden fent közölt, értelmezhető mintázat megjelenik e háromban), mint a fenti k=7, ahol összesen két olyan klaszterpár is volt, mely teljesen azonos értékmintázatot mutatott a standard átlagokban, mindössze az értékek maximális és minimális értékeiben mutattak eltérést.