Domácí úkol 2

Faktorová analýza

Marie Melínová

Druhý úkol pojednává o vydavateli časopisů, který by si rád udělal představu o tom, která témata oslovují podobné čtenáře. Podle zadání má datový soubor 26 proměnných a 800 pozorování.

Po předběžném prozkoumání mají data skutečně 26 proměnných, avšak v datech je 8003 pozorování. Nejsem si jistá, čemu bych tuto nesrovnalost měla přikládat.

```
library(foreign)
data = read.spss("du2_9.sav",to.data.frame=TRUE)
dim(data)
```

```
## [1] 8003   26
```

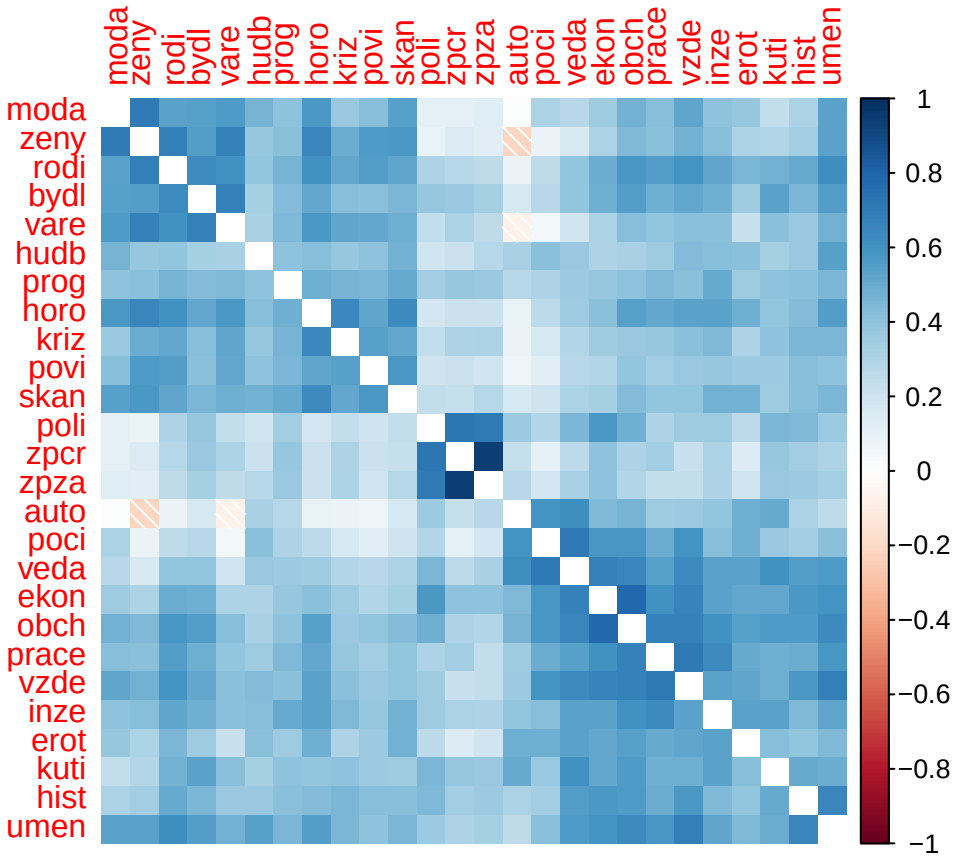
Posouzení vhodnosti dat pro faktorovou analýzu

Hlavní problém, které mohou data vykazovat je to, že jednotlivé hodnoty jsou binární. Standartní faktorová analýza počířá s korelační maticí, která předpokládá, že její hodnoty jsou spojité.

Musíme tedy najít jiný vhodný způsob pro výpočet matice asociací jednotlivých proměnných. To nám umožňuje například funkce `hetcor()`. Funkce `hetcor()` se podívá na každý pár proměnných a vypočítá vhodnou heterogenní korelaci.

```
library("corrplot")
library("polycor")

korMatice <- hetcor(data)$correlations
corrplot(korMatice, order = "hclust", method = "shade", diag = F)
```



Odhad vhodné dimenzionality úlohy

Na základě předchozího grafu bych odhadla, že v datech můžou figurovat 3 faktory. S touto informací tedy budeme pracovat a vhodný počet faktorů po případě následně upravíme.

Model faktorové analýzy

Nyní se již pustíme do samotné faktorové analýzy. Tu provedeme pomocí funkce `factanal()`, pro kterou nemusíme doinstalovávat žádný balíček, jelikož se již nachází v základních funkcích.

```
fa <- factanal(covmat = korMatice, factors = 3, rotation = "varimax")
fa
```

```
##
## Call:
## factanal(factors = 3, covmat = korMatice, rotation = "varimax")
##
## Uniquenesses:
##  auto  bydl  ekon  hist  horo  hudb  inze  kriz  kuti  moda  obch  poci  poli
## 0.366 0.460 0.331 0.526 0.378 0.682 0.473 0.572 0.518 0.431 0.283 0.353 0.378
##  povi prace  prog  rodi  erot  skan  umen  vare  veda  vzde  zpcr  zpza  zeny
## 0.565 0.429 0.605 0.343 0.532 0.524 0.379 0.375 0.246 0.321 0.005 0.105 0.176
##
## Loadings:
##       Factor1 Factor2 Factor3
## auto  -0.172   0.754   0.190
## bydl   0.630   0.293   0.237
```

```
## ekon   0.332   0.690   0.287
## hist   0.419   0.497   0.227
## horo   0.741   0.262
## hudb   0.425   0.353   0.113
## inze   0.478   0.513   0.188
## kriz   0.601   0.176   0.189
## kuti   0.338   0.539   0.277
## moda   0.727   0.197
## obch   0.472   0.685   0.158
## poci           0.800
## poli   0.119   0.384   0.679
## povi   0.637   0.147
## prace  0.473   0.555   0.200
## prog   0.482   0.298   0.272
## rodi   0.747   0.286   0.131
## erot   0.351   0.587
## skan   0.644   0.223   0.106
## umen   0.588   0.500   0.155
## vare   0.771           0.174
## veda   0.179   0.836   0.152
## vzde   0.513   0.640
## zpcr   0.177           0.977
## zpza   0.142   0.163   0.921
## zeny   0.908
##
##                Factor1 Factor2 Factor3
## SS loadings      6.986   5.743   2.913
## Proportion Var   0.269   0.221   0.112
## Cumulative Var   0.269   0.490   0.602
##
## The degrees of freedom for the model is 250 and the fit was 3.4
```

Při použití tří faktorů si můžeme všimnout, že tyto faktory dohromady vysvětlují 60.2 % variability. Mohla bych zvýšit počet faktorů tak, abychom se dostali na vyšší procento vysvětlené variability, avšak při vyšším počtu faktorů už dávají tyto faktory nelogické interpretace.

Celkový počet faktorů bych musela zvednout až na 8, abychom se dostali alespoň na 70 % vysvětlené variability. Z těchto důvodů jsem se rozhodla ponechat původní tři.

Interpretace jednotlivých faktorů

```
faktor1 <- as.data.frame(t(names(which(fa$loadings[,1] >= 0.5))))
faktor2 <- as.data.frame(t(names(which(fa$loadings[,2] >= 0.5))))
faktor3 <- as.data.frame(t(names(which(fa$loadings[,3] >= 0.5))))
```

```
faktor1
```

```
##     V1   V2   V3   V4   V5   V6   V7   V8   V9  V10  V11
## 1 bydl horo kriz moda povi rodi skan umen vare vzde zeny
```

Faktor 1 představuje zájemce o typicky ženskou tématiku (móda, rodina, skandály, vaření, ženy,...). Můžeme si zároveň všimnout, že tento faktor jako jediný s jako jedinou proměnnou `zeny` má negativní korelaci - řekla bych, že to souhlasí s obecným názorem žen na toto téma.

```
faktor2
```

```
##     V1   V2   V3   V4   V5   V6    V7   V8   V9  V10  V11
## 1 auto ekon inze kuti obch poci prace erot umen veda vzde
```

O faktoru 2 bych naopak řekla, že spojuje zájemce o typicky mužskou tématiku (auto, počítače, věda,...).

```
faktor3
```

```
##     V1   V2   V3
## 1 poli zpcr zpza
```

Pro poslední faktor jsem měla trochu problém nakonec najít interpretaci. Nakonec bych řekla, že lidi, kteří se velmi zajímají po politiku a dění ve světě i u nás.

Poznámka

Myslím si, že faktorů opravdu mohlo být více, bohužel jsem nikdy nenašla vhodnou interpretaci jednotlivých faktorů tak, aby odpovídali i případným záporným korelacím atd.