Úkol 2 - Průzkumová faktorová analýza

4ST512 Vícerozměrná statistika

Jiří Filip

Úvod

V této práci se zaměříme na analýzu preferencí čtenářů. V šetření se vydavatel časopisu zeptal 800 respondentů, o která z 26 nadefinovaných témat se zajímají. Pokusíme se odpovědět na otázku, zda zájem o jednotlivá témata lze vysvětlit menším počtem faktorů. Tyto faktory se pokusíme interpretovat a následně identifikovat a shrnout výsledky analýzy, případně její nedostatky.

Posouzení dat

Jako první posoudíme, zda jsou data vhodná pro použití faktorové analýzy. K tomu nám poslouží dva testy - KMO a Bartlettův test. Míra KMO 0.92 je velice dobrý výsledek a znamená, že data jsou vhodná pro faktorovou analýzu.

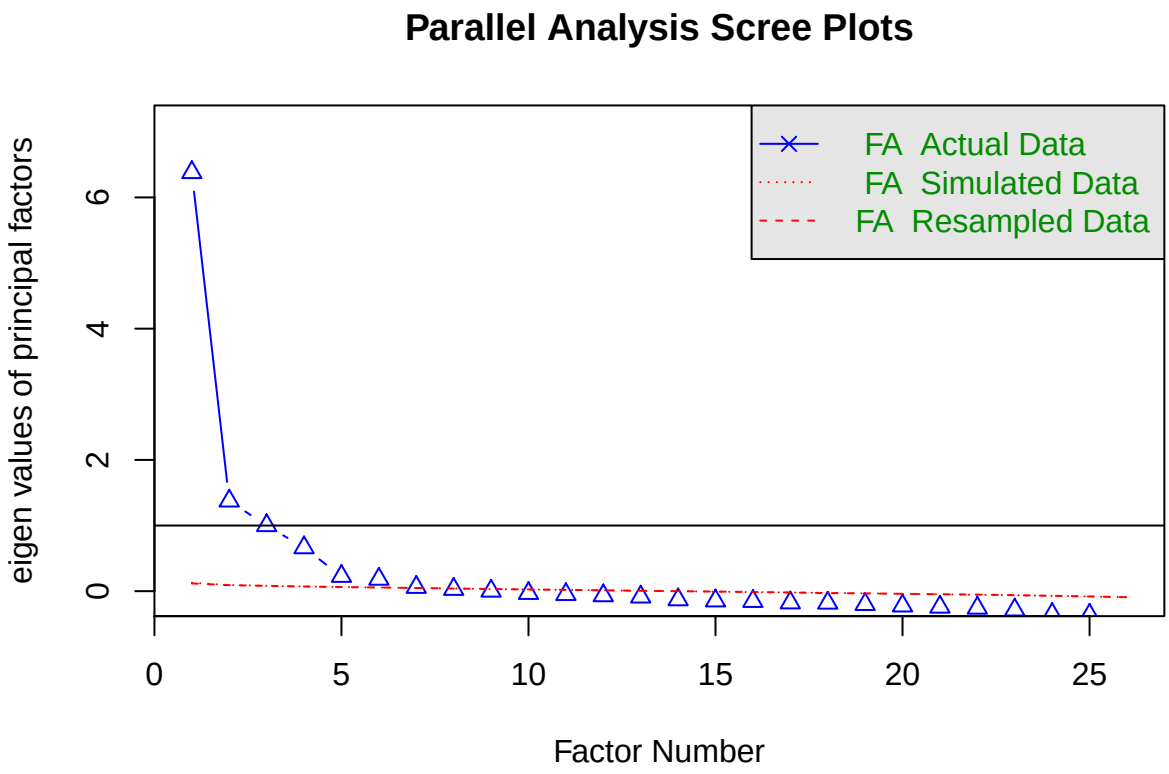
Bartlettův test si klade otázku, zda jsou rozptyly jednotlivých proměnných stejné nebo ne. Výsledek nám říká, že na hladině významnosti 0.05 (kterou jsme si předem zvolili) můžeme zamítnout hypotézu o homogenitě rozptylů jednotlivých proměnných.

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data.raw)
## Overall MSA =  0.92
## MSA for each item =
##  bydl  ekon  horo  hudb  inze  kino  kras  kres  krim  kriz  nemo  obch
##  0.95  0.93  0.94  0.90  0.96  0.93  0.90  0.96  0.95  0.94  0.95  0.93
##  poci  poli prace  prog  preh  prir  regi  erot  spor  zaba  zahr  zpcr
##  0.91  0.92  0.95  0.96  0.97  0.96  0.91  0.96  0.89  0.92  0.93  0.77
##  zpza  zeny
##  0.79  0.87
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data.raw
## Bartlett's K-squared = 7791.6, df = 25, p-value < 2.2e-16
```

Pro zjistění vhodného počtu faktorů použijeme metodu Parallel Analysis. Ta srovnává vlastní hodnoty faktorové analýzy pro naše data a pro náhodná data vygenerovaná pomocí monte carlo simulace. Vidíme, že u pátého faktoru se již blížíme k červené čáře, která udává vlastní hodnoty faktorů náhodných dat. Ideální počet faktorů by tak mohl být 3-4. Doplňující výstup programu (pro přehlednost neuvedeno) sice navrhuje počet faktorů jako 7, ale je vidět, že od pátého faktoru se reálné faktory a faktory na náhodných datech liší minimálně.

I tak faktorovou analýzu se sedmi faktory pro úplnost zkusíme spustit.



Faktorová analýza

Před provedením samotné analýzy ještě uvažujme, co jednotlivé faktory budou znamenat. Jelikož naše jednotlivé proměnné jsou témata, budou jednotlivé faktory představovat jakási "nadtémata" či okruhy témat, např. zpravodajství, koníčky, zábava. Z tohoto důvodu by dávalo smysl, že by se jednotlivá nadtémata mohla z části překrývat. Zkusme však nejprve nalézt vhodné řešení, kde jsou faktory nekorelované (tedy témata se nepřekrývají). Pokud budeme s výsledky nespokojeni, připustíme korelaci jednotlivých faktorů.

Proveďme tedy faktorovou analýzu a zobrazme faktorové zátěže, které jsou větší než 0.3. Vyextrahujeme 3 faktory a použijeme rotaci "varimax", protože chceme, aby jednotlivé faktory byly nekorelované.

Jak lze vidět, faktor č. 3 lze interpretovat jako zájem o zpravodajství. Faktor 2 lze interpretovat jako zájem o jakousi "zábavu", i když tuto interpretaci poněkud kazí fakt, že do něj výrazně přispívá ekonomika, obchod a méně výrazně politika. Faktor č. 1 je potom další "zábavní" mix, hůře intepretovatelný než faktor č. 2.

Vidíme, že faktorová analýza se třemi faktory má řadu nedostatků, zkusme proto zvolit faktory čtyři.

```
##
## Loadings:
##                                MR1    MR3    MR2
## bydlení / domov                 0.458
## ekonomika                              0.615
## horoskopy / astrologie          0.602
## hudba / muzika                  0.413
## inzerce                         0.353  0.453
## kino / film                     0.384
```

```
## krása a vlasy                         0.658
## kresby a karikatury                   0.411  0.484
## kriminalita / černá kronika           0.429
## křížovky / hádanky / soutěže          0.503
## nemovitosti / domy                           0.552
## obchod                                       0.606
## počítače                                     0.549
## politika                                     0.317  0.475
## práce / zaměstnání                    0.311  0.505
## programy televize / přehledy pořadů   0.417
## přehledy kulturních akcí              0.421  0.477
## příroda / životní prostředí           0.384
## regionální zprávy z okolí kde žiji                  0.649
## sex / erotika                                0.436
## sport                                        0.407
## zábava                                0.431
## zahrada / zahrádka / pěstitelství     0.375
## zprávy z České republiky                            0.736
## zprávy ze zahraničí                                 0.796
## ženské stránky / ženská témata        0.700
##
##                  MR1   MR3   MR2
## SS loadings    3.741 3.294 2.281
## Proportion Var 0.144 0.127 0.088
## Cumulative Var 0.144 0.271 0.358
```

Zde nám čtvrtý faktor trochu konkretizuje onen termín "zábava". Zahrnuje zde kino, film, hudbu, sport, zábavu a programy televize a pořadů. Tento faktor tak můžeme intepretovat jako "zájem o pasivní zábavu" (s výjimkou sportu, ale toto téma lze intepretovat např. jako články o fotbalových, hokejových zápasech apod.).

Třetí faktor ukazuje zájem o zpravodajství. Druhý faktor lze kvůli převaze některých témat (krása a vlasy, horoskopy, ženská témata) trochu stereotypně charakterizovat jako "ženská témata" nebo témata, o něž mají zájem převážně ženy. První by pak šlo - opět se značným stereotypem - charakterizovat jako "mužská" témata, především kvůli zahrnutí počítačů, nemovitostí a domů, ekonomiky a sportu. Je jasné, že tento termín plně první faktor nevystihuje, ale pokud bychom ho chtěli nějak kontrastovat vůči druhému faktoru, můžeme ho pojmenovat právě takto.

```
##
## Loadings:
##                                    MR3    MR1    MR2    MR4
## bydlení / domov                           0.455
## ekonomika                          0.622
## horoskopy / astrologie                    0.557
## hudba / muzika                                          0.614
## inzerce                            0.458  0.301
## kino / film                                             0.546
## krása a vlasy                             0.626
## kresby a karikatury                0.492  0.360
## kriminalita / černá kronika               0.332
## křížovky / hádanky / soutěže              0.424
## nemovitosti / domy                 0.579
## obchod                             0.637
## počítače                           0.516
```

```
## politika                              0.317         0.481
## práce / zaměstnání                    0.519
## programy televize / přehledy pořadů           0.308         0.316
## přehledy kulturních akcí              0.477  0.350
## příroda / životní prostředí           0.301  0.321
## regionální zprávy z okolí kde žiji                  0.649
## sex / erotika                         0.415
## sport                                 0.361                0.385
## zábava                                                     0.523
## zahrada / zahrádka / pěstitelství             0.377
## zprávy z České republiky                            0.752
## zprávy ze zahraničí                                 0.792
## ženské stránky / ženská témata                0.728
##
##                  MR3   MR1   MR2   MR4
## SS loadings    3.281 2.855 2.290 1.813
## Proportion Var 0.126 0.110 0.088 0.070
## Cumulative Var 0.126 0.236 0.324 0.394
```

Analýza s pěti faktory byla rovněž vyzkoušena, ale nijak se jí nepodařilo "rozmělnit" první a druhý faktor na více detailů. Rovněž bylo vyzkoušeno i 6 a 7 faktorů, které zde však neuvádím kvůli přehlednosti. Ani jedna varianta nebyla uspokojivější ani lépe interpretovatelná než výše uvedené.

Analýza se čtyřmi faktory pomocí rotace varimax byla celkově uspokojující, ale pro zlepšení výsledků ještě zkusme rotaci oblimin, která povoluje korelaci jednotlivých faktorů.

Zde faktory dva až čtyři představují zpravodajství, ženská témata a pasivní zábavu, přičemž první faktor by se taky dal považovat za jakousi zábavu, nyní však s mixem politiky, ekonomiky apod.

Tato varianta poskytuje ze všech nejslibnější výsledky. Právě toto řešení bych volil jako výsledné a prezentoval bych ho vydavateli časopisu. Je zde vhodný počet faktorů a každý z nich (s výjimkou trochu více obecnějšího prvního faktoru) nese ucelenou informaci o zájmu čtenáře. Jak již bylo uvedeno, přidáním dalších faktorů se nedaří "rozmělnit" první faktor "zábavy" na něco více konkrétního, a tak zůstáváme u faktorů čtyř.

```
##
## Loadings:
##                                    MR1    MR2    MR3    MR4
## bydlení / domov                                  0.363
## ekonomika                           0.663
## horoskopy / astrologie                           0.470
## hudba / muzika                                          0.663
## inzerce                             0.449
## kino / film                                             0.576
## krása a vlasy                                    0.552
## kresby a karikatury                 0.486
## kriminalita / černá kronika
## křížovky / hádanky / soutěže                     0.341
## nemovitosti / domy                  0.630
## obchod                              0.695
## počítače                            0.500
## politika                            0.306  0.449
## práce / zaměstnání                  0.538
## programy televize / přehledy pořadů
## přehledy kulturních akcí            0.452
## příroda / životní prostředí
```

```
## regionální zprávy z okolí kde žiji          0.656
## sex / erotika                         0.380
## sport                                              -0.361  0.375
## zábava                                                     0.562
## zahrada / zahrádka / pěstitelství
## zprávy z České republiky                    0.776
## zprávy ze zahraničí                         0.804
## ženské stránky / ženská témata                      0.709
##
##                   MR1   MR2   MR3   MR4
## SS loadings     3.149 2.156 1.851 1.691
## Proportion Var  0.121 0.083 0.071 0.065
## Cumulative Var  0.121 0.204 0.275 0.340
```

Shrnutí

V předchozím odstavci již byla uvedena pozitiva výsledného řešení, a sice charakteristika jednotlivých "nadtémat" a rozpoznání preference čtenářů.

Nedostatky analýzy vidím v příliš všeobecném prvním faktoru. I tak ale výsledná analýza poskytuje zadavateli přehled o rozdělení zájmu jednotlivých čtenářů časopisu. Rozděluje témata na pasivní zábavu, zpravodajství, ženská témata a zbytkovou "všehochuť". Vydavatel může informace z této analýzy využít k přizpůsobení obsahu čtenářům či výsledky může využít jako odrazový můstek k provedení podrobnější analýzy, např. s jinými nástroji.