Faktorová analýza

4ST512 Vícerozměrná statistika

Lubomír Štěpánek

9. prosince 2017

Obsah

1 Zadání úlohy

Východiskem je soubor `du2_30.sav`, který obsahuje celkem 26 proměnných (sloupců) týkajících se vybraných magazínových témat a 8003 pozorování (řádků) tvořených odpovědmi respondentů na dotazníkové šetření zkoumající jejich tématické preference. Buňky datasetu tedy tvoří typicky odpověď jednoho respondenta na dané téma ve smyslu *ano* (respondent preferuje dané téma), nebo *ne* (respondent dané téma nepreferuje); jde tedy o binární data. Popisky sloupců tvoří vždy první čtyři písmena tématu bez diakritiky (*self-labeling* princip); metadata souboru obsahují původní nezkrácené názvy magazínových témat.

Hlavním cílem je prozkoumat data a pokusit se nalézt menší počet skrytých faktorů, které stále stejně dobře vysvětlují rozložení zájmu o jednotlivá magazínová témata.

(i) Posuďme, zda je použití faktorové analýzy jako nástroje volby vhodné ke splnění hlavního cíle.

(ii) Odhadněme vhodnou dimenzionalitu dat.

(iii) Najděme pomocí faktorové analýzy co nejlepší řešení (konečný počet faktorů se může lišit od předpokladu z předchozího bodu).

(iv) Pokusme se interpretovat získané faktory.

(v) Posuďme úspěšnost analýzy – zejména smysluplnost a využitelnost výsledných faktorů a limitaci použité metodiky.

Vysázeno pomocí R-ka, YAML-u, TEX-u a BibTEX-u

2 Řešení úlohy

2.1 Metodologie a analýza dat

Celá úloha byla řešena v prostředí R, které je určeno pro statistické výpočty a následné grafické náhledy [1]. Datový soubor `du2_30.sav` byl nahrán do prostředí R pomocí balíčku `foreign`.

Jednotlivé výpočty v rámci faktorové analýzy a další přidružené metody byly provedeny pomocí R-kových balíčků `stats` a `psych`.

Vzhledem k binárnímu charakteru dat bylo pro výpočet korelací, které jsou nadále v metodologii používány, možné využít jak Spearmanův, tak Pearsonův korelační koeficient, nebo i Kendallovo τ. Všechny tři míry korelace vrátí shodné hodnoty.

2.1.1 Vhodnost použití faktorové analýzy

Zda jsou vstupní data vhodná k faktorové analýze, lze posoudit pomocí Kaiser-Meyer-Olkinovy míry. Ta je definována jako

$$KMO = \frac{\sum\sum_{j\neq k} r_{jk}^2}{\sum\sum_{j\neq k} r_{jk}^2 + \sum\sum_{j\neq k} p_{jk}^2},$$

kde r_{jk}, resp. p_{jk} je korelační, resp. parciální korelační koeficient mezi j-tou a k-tou proměnnou, [2]. Míra KMO tak zřejmě roste s klesající velikostí parciálních (od vlivu ostatních proměnných očištěných – ty jsou fixovány) korelací; intuitivně tedy rostoucí KMO svědčí pro větší míru vzájemné závislosti proměnných, a tedy vyšší smysluplnost hledaní vysvětlujících skrytých faktorů (pomocí faktorové analýzy), kterých bude méně než původních proměnných, ale komplexitu dat stále ještě dostatečně vystihnou. Empiricky, je-li $KMO \geq 0{,}5$, použití faktorové analýzy je obhajitelné.

Podobně lze použít i Bartlettův test sféricity, který testuje pomocí vhodné testové statistiky, sledující χ^2 rozdělení, nulovou hypotézu H_0 o tom, že korelační matice proměnných odpovídá jednotkové matici, [3]. Zamítnutí nulové hypotézy H_0 naznačuje nenulovost korelačních koeficientů mezi proměnnými, a tedy všeobecně multikolinearitu, kterou je možné analyzovat právě pomocí hledání skrytých faktorů.

2.1.2 Faktorová analýza

Smyslem faktorové analýzy je vysvětlit především *kovarianci* proměnných (narozdíl od analýzy hlavních komponent, která se snaží maximálně vysvětlit variabilitu dat). Předpokladem je, že pozorované proměnné jsou lineární kombinací menšího počtu skrytých, nepozorovaných faktorů, více v [4].

Buď $\boldsymbol{X}_{n\times p}$ datová matice o p proměnných a n pozorováních, kterou jsme apriorně centralizovali, tj. odečetli od všech hodnot vždy střední hodnotu dané proměnné (sloupcový průměr). Předpokládejme,

že existuje matice faktorů $\boldsymbol{F}_{m\times n}$ tak, že její sloupce jsou jednotlivé faktory a $m < p$; dále $\boldsymbol{L}_{p\times m}$ je matice faktorových zátěží a $\boldsymbol{\varepsilon}_{n\times p}$ je matice chyb. Modelem faktorové analýzy je

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}. \tag{1}$$

Nyní odvozujme; za předpokladu, že faktory jsou nezávislé na chybách, tedy $\mathrm{cov}(\boldsymbol{F}, \boldsymbol{\varepsilon}) = \boldsymbol{0}$, a že faktory jsou navzájem nezávislé, tedy $\mathrm{cov}(\boldsymbol{F}) = \boldsymbol{I}$, kde $\boldsymbol{I}$ je jednotková matice příslušného rozměru, platí pro kovarianční matici vstupních centralizovaných dat $\boldsymbol{\Sigma_X} = \mathrm{cov}(\boldsymbol{X})$ po dosazení z (1)

$$\begin{aligned}
\boldsymbol{\Sigma_X} = \mathrm{cov}(\boldsymbol{X}) &= \mathrm{cov}(\boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}) = \\
&= \mathrm{cov}(\boldsymbol{L}\boldsymbol{F}) + \mathrm{cov}(\boldsymbol{\varepsilon}) + 2 \cdot \mathrm{cov}(\boldsymbol{L}\boldsymbol{F}, \boldsymbol{\varepsilon}) = \\
&= \boldsymbol{L} \cdot \mathrm{cov}(\boldsymbol{F}) \cdot \boldsymbol{L}^T + \boldsymbol{\Phi} + \boldsymbol{0} = \\
&= \boldsymbol{L} \cdot \boldsymbol{I} \cdot \boldsymbol{L}^T + \boldsymbol{\Psi} = \\
&= \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}.
\end{aligned}$$

Na faktorovou analýzu je tedy možné nahlížet také jako na dekompozici kovarianční matice centralizovaných vstupních dat $\boldsymbol{\Sigma_X}$ na dvě části – na část vysvětlenou pomocí faktorů, *komunalitu*, $\boldsymbol{L}\boldsymbol{L}^T$, a na část nevysvětlitelnou pomocí faktorů, *jedinečnost*, $\boldsymbol{\Psi}$.

Řešení modelu (1) není jednoznačné, pro odhad parametrů modelu se používá přístup založený na analýze hlavních komponent. Počet faktorů m lze odhadnout expertně, pomocí *sutinového*, tedy *scree diagramu*, dle procenta vysvětlené variability daným počtem komponent či Kaiser-Guttmanovým kritériem, [5].

Všimněme si ještě, že pokud je $\boldsymbol{Q}$ ortogonální matice, tedy $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$, pak se model (1) nezmění, pokud nahradíme $\boldsymbol{L}$ a $\boldsymbol{F}$ za $\boldsymbol{L'} = \boldsymbol{L}\boldsymbol{Q}$ a $\boldsymbol{F'} = \boldsymbol{Q}^T\boldsymbol{F}$. Pokud je tedy model (1) řešen nějakými faktory $\boldsymbol{F}$ a faktorovými zátěžemi $\boldsymbol{L}$, pak je pro každou ortogonální matici $\boldsymbol{Q}$ model (1) řešen i faktory $\boldsymbol{Q}^T\boldsymbol{F}$ a faktorovými zátěžemi $\boldsymbol{L}\boldsymbol{Q}$. Toho se využívá při tzv. *(ortogonálním) rotování* řešení, kdy je matice $\boldsymbol{Q}$ volena vhodně tak, aby struktura faktorových zátěží $\boldsymbol{L}$ byla interpretačně dobře uchopitelná, např. aby některé faktorové zátěže byly maximalizovány a zbylé naopak minimalizovány. Nejčastěji používanou ortogonální *rotací* je tzv. *varimax*, který maximalizuje součet čtverců faktorů, [6].

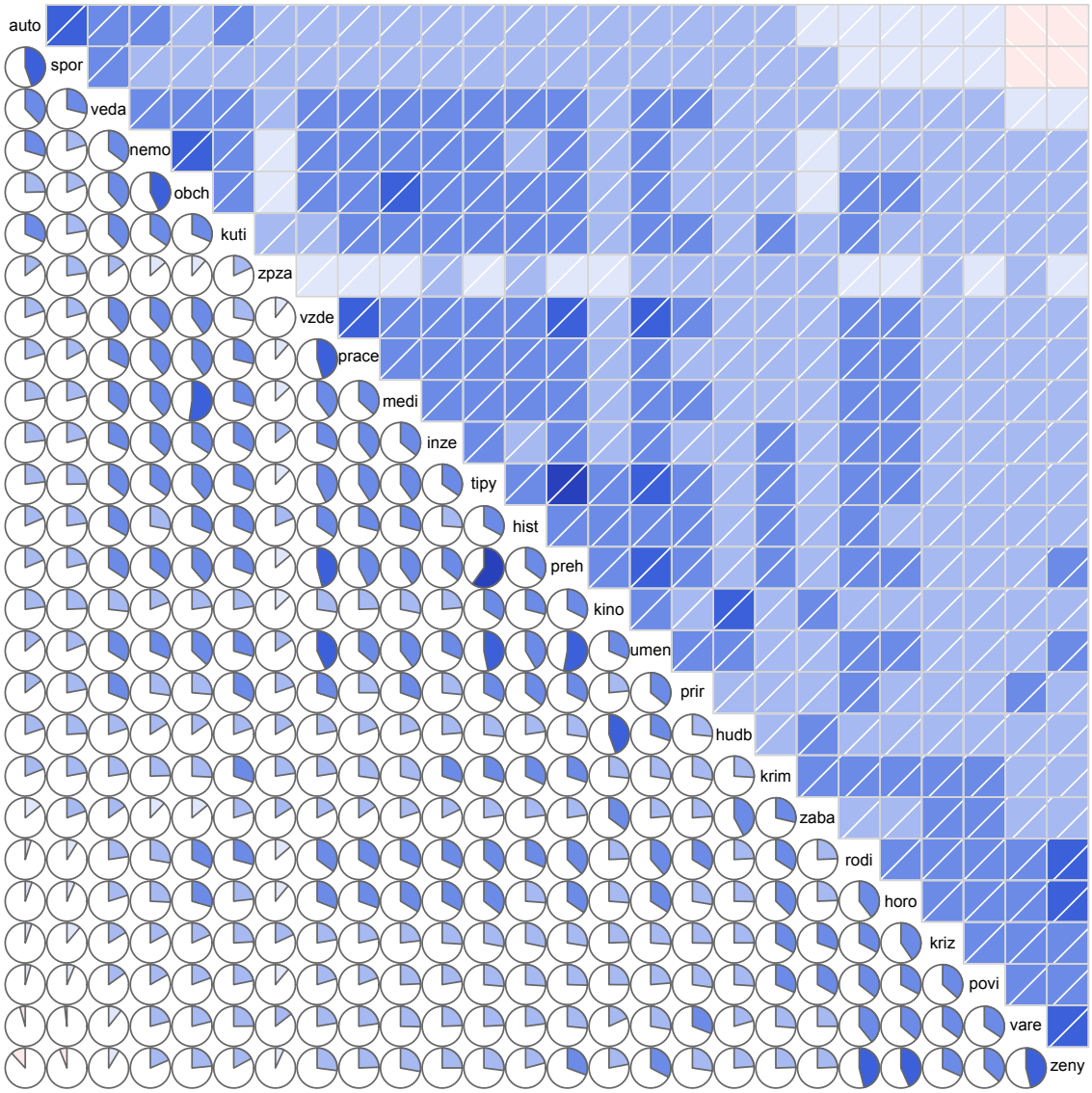
2.2 Výsledky

Explorativní analýza dat. Dataset `du2_30.sav` neobsahuje žádné chybějící hodnoty. Korelační matice byla sestavena pomocí Spearmanova koeficientu. Shodné hodnoty by vracelo ovšem i Kendallovo τ, které se pro kategorická data obzvlášť hodí; avšak časová složitost výpočtu Kendallova τ je $\Theta(n^2)$, kde n je počet pozorování datasetu, oproti rychlejšímu $\Theta(n \log(n))$ u Spearmanova koeficientu. Pro $n = 8003$ je rozdíl délky výpočtu již uživatelsky znatelný.

Náhled na korelační matici je na obrázku 1, z kterého lze podle podobných odstínů a těsné blízkosti silněji korelovaných proměnných (vytvářejících „pyramidová pole“) orientačně již vyčíst i možnou dimenzionalitu dat.

Můžeme nahlédnout, že první malou skupinu podobných si proměnných tvoří motorismus a sport, méně zřetelně i věda a technika či nemovitosti. Cílovou skupinou jsou pravděpodobně čtenáři-muži.

Další velkou skupinu podobných si proměnných tvoří témata typu zpravodajství, obchod, média, inzerce, práce, tipy apod. Již v rámci explorace lze vytušit, že se jedná o skupinu proměnných ve smyslu *aktuálního dění*, resp. *socio-ekonomických témat*.



Obrázek 1: Korelogram sestavený nad maticí Spearmanových korelačních koeficientů mezi proměnnými vstupních dat. Modré odstíny značí pozitivní korelaci, červené odstíny negativní korelaci. Sytost je přímo úměrná síle korelace. Koláčové diagramy vyjadřují velikost korelačního koeficientu jako na ciferníku hodin (tj. např. 3:00 odpovídá korelačnímu koeficientu 0,25).

Poslední nápadnou skupinou podobných si proměnných tvoří témata typu rodina, křížovky, vaření, ženská témata, horoskopy apod. Jde pravděpodobně o *čtení pro ženy*, resp. *„čtení ke kávě“*.

Stranou pak stojí některé izolované proměnné – zejména zpravodajství ze zahraničí, které pravděpodobně atrahují úzkou a specifickou skupinu čtenářů.

Jako zajímavost uveďme, že proměnné typu *automobilismus* a *sport* korelují záporně s proměnnými *vaření* a *ženská témata*, což odpovídá očekávání – je totiž velmi pravděpodobné, že respondenta přitahuje nanejvýš jedna z těchto dvou skupin témat.

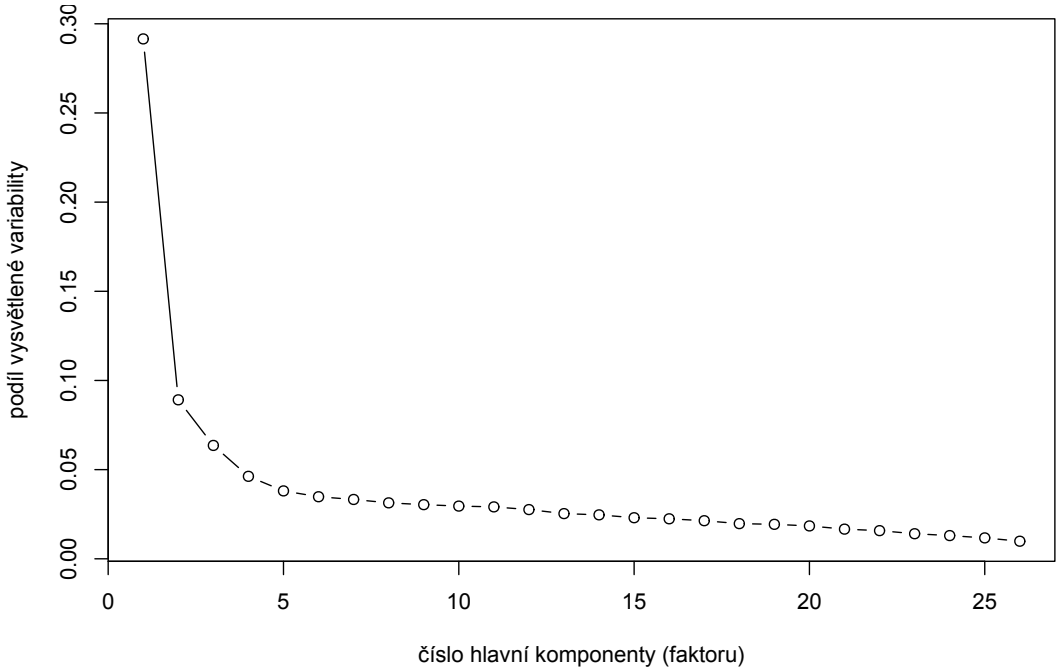
Vhodnost použití faktorové analýzy. Byla spočítána Kaiser-Meyer-Olkinova míra KMO adekvátnosti vstupních dat pro faktorovou analýzu

$$KMO = \frac{\sum\sum_{j\neq k} r_{jk}^2}{\sum\sum_{j\neq k} r_{jk}^2 + \sum\sum_{j\neq k} p_{jk}^2} \doteq \frac{51{,}870}{51{,}870 + 3{,}016} \doteq 0{,}945 \gg 0{,}500,$$

svědčící pro velkou míru vzájemné závislosti proměnných.

Rovněž Bartlettův test sféricity svou testovou statistikou $\chi^2(df = 325) \doteq 64534{,}35$ na hladině významnosti $p \ll 0{,}0001$ zamítl nulovou hypotézu H_0 o podobnosti korelační matice matici jednotkové. Oba výsledky svědčí pro vhodnost užití faktorové analýzy při hledání menšího počtu skrytých faktorů vysvětlujících kovarianční strukturu proměnných.

Dimenzionalita úlohy. Vyjděme z proporce celkové variability, kterou vysvětlí prvních několik faktorů.



Obrázek 2: Modifikovaný *scree diagram*. Diagram popisuje podíl vysvětlené variability prostřednictvím dané hlavní komponenty (faktoru).

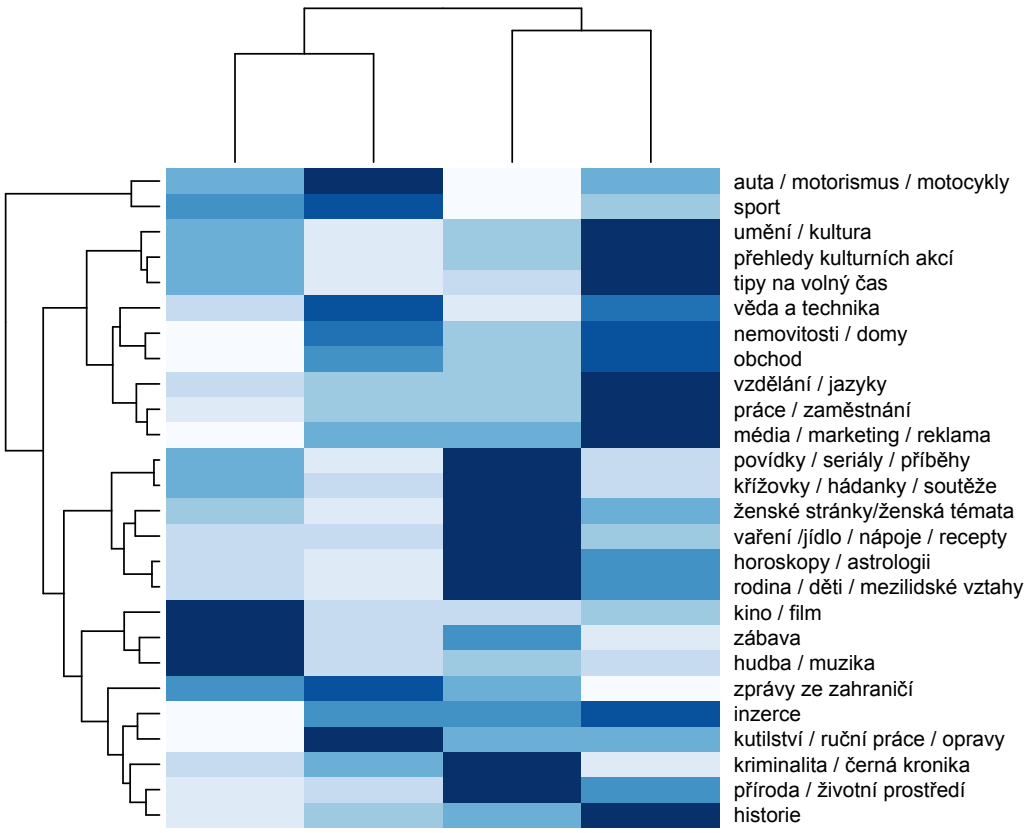
Vhodným grafickým náhledem je *scree diagram*, zobrazující křivku, která ukazuje hodnoty vlastních čísel v závislosti na pořadí jejich komponent; v tomto pořadí komponenty sestupně vysvětlí méně a méně celkové variability. Na křivce scree diagramu je možné hledat tzv. *elbow fenomén*, zlom, ve kterém je (daným počtem komponent) jednak obvykle vysvětleno již významné množství celkové variability, jednak zavzetí dalších komponent již tolik variability navíc nevysvětlí.

Na obrázku 2 vidíme modifikovaný *scree diagram*, který místo hodnot vlastních čísel přímo zobrazuje podíly vysvětlené celkové variability pro každou komponentu (faktor). Bod zlomu, *elbow fenomén* je pravděpodobně u čtvrtého faktoru. To odpovídá i závěru explorativní analýzy dat. Přesto však nemusí jít nutně o nejlepší model (se čtyřmi faktory), z obrázku *scree diagramu* vyplývá, že první čtyři faktory vysvětlí jen cca 50 % celkové variability.

Faktorová analýza. Pro čtyři faktory, jak jsme odhadli v předchozí části, byl vypočítán model faktorové analýzy; model je založen na rovnici (1). V tabulce 1 jsou uvedeny faktorové zátěže $\boldsymbol{L}$ pro jednotlivé proměnné. Díky rotaci *varimax* je možné pomocí velikostí zátěží v tabulce „přiřadit“ jednotlivé proměnné k daným faktorům, které je z velké části vždy vysvětlují – čím větší faktorová zátěž pro daný faktor, tím více se v lineární kombinaci do dané proměnné daný faktor promítá.

	faktor 1	faktor 2	faktor 3	faktor 4
auta / motorismus / motocykly	0,184	-0,166	0,628	0,202
historie	0,350	0,260	0,247	0,215
horoskopy / astrologii	0,309	0,527	0,068	0,137
hudba / muzika	0,125	0,241	0,160	0,559
inzerce	0,378	0,292	0,308	0,074
kino / film	0,243	0,189	0,181	0,492
kriminalita / černá kronika	0,187	0,393	0,287	0,228
křížovky / hádanky / soutěže	0,139	0,496	0,112	0,238
kutilství / ruční práce / opravy	0,282	0,257	0,431	0,088
média / marketing / reklama	0,500	0,272	0,283	0,045
nemovitosti / domy	0,460	0,193	0,387	-0,059
obchod	0,536	0,220	0,322	-0,057
povídky / seriály / příběhy	0,119	0,515	0,082	0,233
práce / zaměstnání	0,534	0,220	0,218	0,040
přehledy kulturních akcí	0,688	0,187	0,046	0,280
příroda / životní prostředí	0,292	0,336	0,228	0,197
rodina / děti / mezilidské vztahy	0,356	0,535	0,072	0,110
sport	0,167	-0,110	0,481	0,333
tipy na volný čas	0,631	0,177	0,115	0,281
umění / kultura	0,576	0,253	0,070	0,262
vaření /jídlo / nápoje / recepty	0,175	0,612	0,007	0,069
věda a technika	0,431	0,054	0,447	0,113
vzdělání / jazyky	0,587	0,194	0,177	0,106
zábava	0,034	0,336	0,154	0,495
zprávy ze zahraničí	0,048	0,152	0,251	0,177
ženské stránky/ženská témata	0,259	0,638	-0,145	0,090

Tabulka 1: Tabulka faktorových zátěží $\boldsymbol{L}$ pro model faktorové analýzy o čtyřech faktorech.



Obrázek 3: Heatmapa faktorových zátěží $\boldsymbol{L}$ pro model faktorové analýzy o čtyřech faktorech (pořadí sloupců odpovídá pořadí faktorů).

Mnohem lepší grafický náhled však získáme pomocí *heatmapy* na obrázku 3, která je v podstatě jen škálovaným podbarvením tabulky 1.

Faktory dobře odpovídají závěrům grafické exploratorní analýze dat. První faktor má vysoké zátěže pro témata typu obchod, média, inzerce, práce, kultura, vzdělání apod. Jedná o skupinu proměnných ve smyslu *aktuálního dění*, resp. *socio-ekonomických témat*. Druhý faktor kombinuje proměnné typu rodina, křížovky, vaření, ženská témata, horoskopy, příroda apod. Jde pravděpodobně o *čtení pro ženy*, resp. *„čtení ke kávě“*. Třetí faktor kombinuje motorismus a sport, vědu a technika, kutilství a (možná překvapivě) zprávy ze zahraničí. Cílovou skupinou jsou pravděpodobně čtenáři-muži. Poslední faktor kombinuje proměnné typu zábava, kino a hudbu. Jde jistě o homogenní skupinu zábavních magazínů (a jejich čtenářů).

Řešení se zdá relativně přijatelné, zkusme však ještě navýšit počet faktorů tak, aby (i) interpretace dat podle faktorů byla smysluplná, (ii) vysvětlená variabilita byla co největší, (iii) komunality všech proměnných byly co možná největší (tím bude kovarianční matice původních centralizovaných dat co možná nejvíce popsána faktorovou složkou, nikoliv nefaktorovou (*jedinečností*)) a (iv) aby heatmapa byla co „nejčistší“, tj. diferencovaná, obsahovala jen syté a bledé odstíny (což souvisí s kvalitou řešení, rotace a faktorových zátěží L). Očividně jde o „trade-off“ (body (i) a (iv) jsou určitém v protikladu s body (ii) a (iii)) a jednoznačně optimální řešení neexistuje.

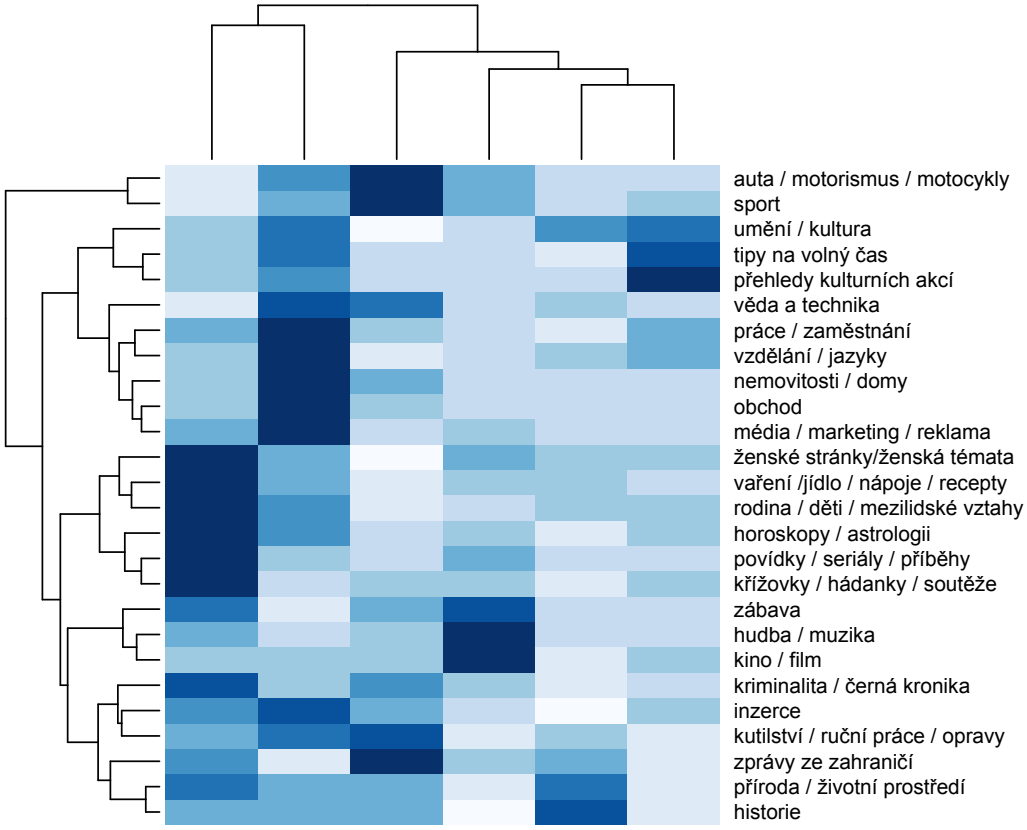
Iterativně dojdeme k počtu faktorů $m = 6$, pro který faktorová analýza relativně dobře splňuje body (i) až (iv). Naopak pro $m = 7$ se již objevil faktor s obecně malými (blízkými nule) faktorovými zátěžemi.

	faktor 1	faktor 2	faktor 3	faktor 4	faktor 5	faktor 6
auta / motorismus / motocykly	0,316	-0,148	0,590	0,176	0,013	-0,080
historie	0,265	0,259	0,268	0,129	0,166	0,348
horoskopy / astrologii	0,285	0,550	0,026	0,136	0,152	-0,030
hudba / muzika	0,090	0,229	0,168	0,626	0,084	0,071
inzerce	0,424	0,297	0,222	0,092	0,123	-0,029
kino / film	0,204	0,182	0,173	0,531	0,150	0,055
kriminalita / černá kronika	0,179	0,440	0,321	0,135	0,106	0,010
křížovky / hádanky / soutěže	0,079	0,546	0,170	0,141	0,126	0,020
kutilství / ruční práce / opravy	0,334	0,256	0,388	0,038	0,050	0,152
média / marketing / reklama	0,593	0,228	0,106	0,148	0,093	0,063
nemovitosti / domy	0,559	0,164	0,229	0,009	0,077	0,047
obchod	0,666	0,164	0,110	0,061	0,066	0,073
povídky / seriály / příběhy	0,106	0,524	0,085	0,201	0,057	0,058
práce / zaměstnání	0,519	0,217	0,119	0,074	0,229	0,023
přehledy kulturních akcí	0,415	0,239	0,106	0,162	0,598	0,117
příroda / životní prostředí	0,235	0,326	0,232	0,132	0,106	0,346
rodina / děti / mezilidské vztahy	0,307	0,524	0,023	0,102	0,147	0,175
sport	0,175	-0,072	0,531	0,233	0,109	0,039
tipy na volný čas	0,415	0,234	0,166	0,177	0,518	0,046
umění / kultura	0,387	0,245	0,069	0,214	0,373	0,336
vaření /jídlo / nápoje / recepty	0,165	0,594	-0,038	0,071	0,034	0,148
věda a technika	0,476	0,031	0,360	0,116	0,107	0,187
vzdělání / jazyky	0,524	0,172	0,084	0,138	0,258	0,181
zábava	0,023	0,342	0,187	0,476	0,042	0,050
zprávy ze zahraničí	0,035	0,176	0,307	0,079	0,029	0,126
ženské stránky/ženská témata	0,224	0,607	-0,231	0,160	0,083	0,118

Tabulka 2: Tabulka faktorových zátěží $\boldsymbol{L}$ pro model faktorové analýzy o šesti faktorech.

Pomocí tabulky 2 a zejména obrázku 4 se můžeme pokusit o interpretaci faktorů. První faktor sdružuje proměnné typu média, marketing, nemovitosti, obchod, zaměstnání, věda a technika, vzdělávání. Jde o vysoce aktuální, *socio-ekonomicko-hospodářská* témata. Druhý faktor kombinuje proměnné typu horoskopy, atrologie, kriminalita, křížovky, povídky, rodina, vaření, jídlo, ženská témata. Jde spíše o oddychová, volnočasová témata ve smyslu *čtení pro ženy* či *čtení ke kávě*. Třetí faktor slučuje proměnné typu motorismus, kutilství, sport a zahraniční zprávy. Audiencí jsou pravděpodobně *mužští čtenáři*. Čtvrtý faktor spojuje hudbu, kino, film a zábavu – jde zřejmě o *zábavní témata*. Pátý faktor slučuje přehledy kulturních akcí a tipy na volný čas, jde tedy o *aktuální témata a dění nepolitického charakteru*. Šestý faktor provažuje *„pomalá“*, *nadčasová* témata typu historie a příroda.

V tabulce 3 jsou komunality pro jednotlivé proměnné. Velká většina z nich je větší než 0,5, tj. faktorová složka vysvětluje více než polovinu variability dané proměnné, což je žádoucí, nebo se kolem 0,5 pohybuje.



Obrázek 4: Heatmapa faktorových zátěží $\boldsymbol{L}$ pro model faktorové analýzy o šesti faktorech (pořadí sloupců odpovídá pořadí faktorů).

	komunalita		komunalita
auta / motorismus / motocykly	0,493	práce / zaměstnání	0,611
historie	0,626	přehledy kulturních akcí	0,362
horoskopy / astrologii	0,573	příroda / životní prostředí	0,636
hudba / muzika	0,508	rodina / děti / mezilidské vztahy	0,569
inzerce	0,658	sport	0,615
kino / film	0,588	tipy na volný čas	0,444
kriminalita / černá kronika	0,642	umění / kultura	0,487
křížovky / hádanky / soutěže	0,630	vaření /jídlo / nápoje / recepty	0,591
kutilství / ruční práce / opravy	0,645	věda a technika	0,583
média / marketing / reklama	0,550	vzdělání / jazyky	0,571
nemovitosti / domy	0,600	zábava	0,617
obchod	0,505	zprávy ze zahraničí	0,851
povídky / seriály / příběhy	0,660	ženské stránky/ženská témata	0,482

Tabulka 3: Tabulka komunalit $\boldsymbol{L}\boldsymbol{L}^T$ pro model faktorové analýzy o šesti faktorech.

2.3 Závěr

Pomocí Kaiser-Meyer-Olkinovy míry a Bartlettova testu sféricity jsme ověřili, že vstupní data jsou vhodná pro následující faktorovou analýzu.

Grafická explorační analýza ukázala opět svou nepostradatelnost v rámci analýzy a pomocí vhodného náhledu na data byla dokonce schopná do jisté míry predikovat výsledek prvního modelu faktorové analýzy.

Pomocí scree diagramu jsme odhadli přibližný počet faktorů. Pro čtyři faktory dával model faktorové analýzy relativně dobrý smysl, avšak pro sedm faktorů bylo možné precizovat interpretaci faktorů.

Praktickou aplikací závěrů faktorové analýzy může být pro vydavatele časopisů například následující

- První faktor, který jsme v rámci interpretace nazvali *socio-ekonomicko-hospodářská*, pokrývá největší podíl témat[1]. S tím by měl vydavatel počítat a věnovat tématům z této skupiny náležitě velký prostor.
- Další faktory pokrývají postupně stále menší skupiny témat, ale přesto je cenným závěrem to, která témata jsou kombinována v rámci jednoho faktoru – toho lze využít v zásadě dvěma způsoby:
  - s výhodou sestavovat jednotlivá vydání časopisů tak, aby byla dostatečně úzce profilovaná a zajistila určitou prodejnost v malé, ale „zaručené“ cílové populaci čtenářů;
  - anebo naopak (a to spíše) sestavovat čísla časopisů vhodně kombinující témata z více skupin faktorů (v rámci faktoru pak stačí vybrat třeba jen jediné), čímž bude naopak zajištěna určitá prodejnost nespecificky napříč populací všech čtenářů.
- Pokud bychom na data nahlíželi nikoliv jako na dotazníkové šetření, ale jako na *nákupní košík* složený z odborných časopisů vždy věnovaných danému jednou tématu, lze v tomto kontextu faktorovou analýzu připodobnit k *analýze nákupního košíku* (analýza *apriori*, *asociační pravidla*) sledující, která témata si daní čtenáři kupují „společně“. Její výsledky mohou využít marketingově např. manažeři velkých prodejen s tiskovinami a vhodně řadit odborné časopisy do regálů – témata „nejčtenějšího“ (prvního) faktoru nejdále od vstupu do prodejny, témata středně rozsáhlých faktorů řadit na více míst v prodejně a co nejdál od sebe apod.

Faktorová analýza tedy potvrdila svou stále významnou roli v kvantitativní analýze dat, kdy s výhodou kombinuje jednoduchost a výpočetní efektivitu analýzy hlavních komponent, ale přidává i relativně dobrou možnost uchopitelné interpretace závěrů.

3 Apendix

Zde je uveden kód v jazyce R, ve kterém byly zpracovávány veškeré výpočty a rovněž generovány diagramy.

[1] Vzhledem k sestrojení faktorových zátěží pomocí analýzy hlavních komponent pokrývá první faktor i největší část celkové variability vstupních dat.

```
################################################################################
################################################################################
################################################################################

## instaluji a loaduji balíčky ------------------------------------------------

invisible(
    lapply(
        c(
            "xtable",
            "openxlsx",
            "foreign",
            "psych",
            "corrgram",
            "ppcor",
            "RColorBrewer"
        ),
        function(my_package){

            if(!(my_package %in% rownames(installed.packages()))){

                install.packages(
                    my_package,
                    dependencies = TRUE,
                    repos = "http://cran.us.r-project.org"
                )

            }

            library(my_package, character.only = TRUE)

        }
    )
)


## ----------------------------------------------------------------------------

################################################################################

## nastavuji handling se zipováním v R ----------------------------------------

Sys.setenv(R_ZIPCMD = "C:/Rtools/bin/zip")


## ----------------------------------------------------------------------------

################################################################################
```

```
## nastavuji pracovní složku ------------------------------------------------

while(!"__domaci_ukol_2__.R" %in% dir()){
    setwd(choose.dir())
}

mother_working_directory <- getwd()


## ------------------------------------------------------------------------

################################################################################

## vytvářím posložky pracovní složky ----------------------------------------

setwd(mother_working_directory)

for(my_subdirectory in c("vstupy", "vystupy")){

    if(!file.exists(my_subdirectory)){

        dir.create(file.path(

            mother_working_directory, my_subdirectory

        ))

    }

}


## ------------------------------------------------------------------------

################################################################################

## loaduji data --------------------------------------------------------------

setwd(
    paste(mother_working_directory, "vstupy", sep = "/")
)

my_data <- read.spss(

    file = "du2_30.sav",
    to.data.frame = TRUE

)
```

```r
setwd(mother_working_directory)


## ------------------------------------------------------------------------

################################################################################

## (pre)processing dat ---------------------------------------------------------

#### přetypovávám všechny proměnné datasetu na textové ------------------------

for(i in 1:dim(my_data)[2]){

    my_data[, i] <- as.character(my_data[, i])

}


#### nyní měním hodnotu "Ano" na numerickou 1 a hodnotu "Ne" na numerickou 0 --

for(i in 1:dim(my_data)[2]){

    my_data[, i] <- ifelse(

        my_data[, i] == "Ano",
        1,
        0

    )

}


## ------------------------------------------------------------------------

################################################################################

## explorativní analýza dat ----------------------------------------------------

#### počítám korelační matici -------------------------------------------------

my_correlations <- cor(

    x = my_data,
    method = "spearman"

)
```

```
#### ukládám korelogram ----------------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = "korelogram.eps",
    width = 8,
    height = 8,
    pointsize = 14
)

par(mar = c(0.1, 0.1, 0.1, 0.1))

corrgram(
    x = my_correlations,
    order = TRUE,
    lower.panel = panel.pie
)

dev.off()

setwd(mother_working_directory)


## ------------------------------------------------------------------------

##############################################################################

## vhodnost použití faktorové analýzy ----------------------------------------

#### počítám Kaiser-Meyer-Olkinovu míru -------------------------------------

KMO(my_data)

# Kaiser-Meyer-Olkin factor adequacy
# Call: KMO(r = my_data)
# Overall MSA =  0.95
# MSA for each item =
#  auto  hist  horo  hudb  inze  kino  krim  kriz  kuti  medi  nemo  obch  povi
#  0.85  0.96  0.95  0.91  0.97  0.94  0.96  0.95  0.96  0.95  0.96  0.94  0.95
# prace  preh  prir  rodi  spor  tipy  umen  vare  veda  vzde  zaba  zpza  zeny
#  0.96  0.94  0.96  0.96  0.88  0.95  0.96  0.93  0.95  0.96  0.92  0.92  0.91


KMO(my_data)$MSA   # 0.9450468


#### anebo též
```

```r
my_partial_correlations <- pcor(my_data, method = "spearman")$estimate

sum(
    my_correlations[upper.tri(my_correlations, diag = FALSE)] ^ 2,
    my_correlations[lower.tri(my_correlations, diag = FALSE)] ^ 2
) / sum(
    my_correlations[upper.tri(my_correlations, diag = FALSE)] ^ 2,
    my_correlations[lower.tri(my_correlations, diag = FALSE)] ^ 2,
    my_partial_correlations[
        upper.tri(my_partial_correlations, diag = FALSE)
    ] ^ 2,
    my_partial_correlations[
        lower.tri(my_partial_correlations, diag = FALSE)
    ] ^ 2
)


#### Bartlettův test sféricity ------------------------------------------------

cortest.bartlett(
    R = my_correlations,
    n = dim(my_data)[1],
    diag = FALSE
)

# $chisq
# [1] 64534.35

# $p.value
# [1] 0

# $df
# [1] 325


## ----------------------------------------------------------------------------

################################################################################

## dimenzionalita úlohy --------------------------------------------------------

#### elbow fenomén nalézáme u 4. faktoru --------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = "scree_plot.eps",
    width = 8,
    height = 5,
```

```
    pointsize = 12
)

par(mar = c(4.1, 4.1, 0.5, 0.3))

plot(
    summary(prcomp(my_data))[["importance"]]["Proportion of Variance", ],
    type = "b",
    xlab = "číslo hlavní komponenty (faktoru)",
    ylab = "podíl vysvětlené variability"
)

dev.off()

setwd(mother_working_directory)


## ------------------------------------------------------------------------

################################################################################

## faktorová analýza pro čtyři faktory ---------------------------------------

first_factor_analysis <- factanal(
    x = my_data,
    factors = 4,
    rotation = "varimax"
)

my_table <- unclass(first_factor_analysis[["loadings"]])

rownames(my_table) <- attr(my_data, "variable.labels")
colnames(my_table) <- paste("faktor", 1:dim(my_table)[2], sep = " ")

print(
    xtable(
        my_table,
        align = rep("", ncol(my_table) + 1),
        digits = 3
    ),
    floating = FALSE,
    tabular.environment = "tabular",
    hline.after = NULL,
    include.rownames = TRUE,
    include.colnames = TRUE,
    format.args = list(decimal.mark = ","))
)
```

```
#### ukládám heatmapu ------------------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = "heatmap_first_factanal.eps",
    width = 8,
    height = 5,
    pointsize = 12
)

par(mar = c(4.1, 4.1, 0.1, 0.1))

heatmap(
    my_table,
    margins = c(0.2, 4),
    #Rowv = NA,
    #Colv = NA,
    col = brewer.pal(9, "Blues"),
    revC = TRUE
)

dev.off()

setwd(mother_working_directory)


## ------------------------------------------------------------------------

##############################################################################

## faktorová analýza pro sedm faktorů ----------------------------------------

second_factor_analysis <- factanal(
    x = my_data,
    factors = 6,
    rotation = "varimax"
)

my_table <- unclass(second_factor_analysis[["loadings"]])

rownames(my_table) <- attr(my_data, "variable.labels")
colnames(my_table) <- paste("faktor", 1:dim(my_table)[2], sep = " ")

print(
    xtable(
        my_table,
        align = rep("", ncol(my_table) + 1),
        digits = 3
```

```
    ),
    floating = FALSE,
    tabular.environment = "tabular",
    hline.after = NULL,
    include.rownames = TRUE,
    include.colnames = TRUE,
    format.args = list(decimal.mark = ",")
)


#### ukládám heatmapu ------------------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = "heatmap_second_factanal.eps",
    width = 8,
    height = 5,
    pointsize = 12
)

par(mar = c(4.1, 4.1, 0.1, 0.1))

heatmap(
    my_table,
    margins = c(0.2, 4),
    #Rowv = NA,
    #Colv = NA,
    col = brewer.pal(9, "Blues"),
    revC = TRUE
)

dev.off()

setwd(mother_working_directory)


#### komunality ------------------------------------------------------------

my_communalities <- data.frame(second_factor_analysis$uniqueness)

rownames(my_communalities) <- attr(my_data, "variable.labels")
colnames(my_communalities) <- "komunalita"

my_table <- data.frame(cbind(
    rownames(my_communalities)[
        1:(dim(my_communalities)[1] / 2)
    ],
    my_communalities[
```

```
        1:(dim(my_communalities)[1] / 2),
    ],
    rownames(my_communalities)[
        (dim(my_communalities)[1] / 2 + 1):dim(my_communalities)[1]
    ],
    my_communalities[
        (dim(my_communalities)[1] / 2 + 1):dim(my_communalities)[1],
    ]
))

for(i in c(2, 4)){

    my_table[, i] <- as.numeric(as.character(my_table[, i]))

}

print(
    xtable(
        my_table,
        align = rep("", ncol(my_table) + 1),
        digits = c(0, 0, 3, 0, 3)
    ),
    floating = FALSE,
    tabular.environment = "tabular",
    hline.after = NULL,
    include.rownames = FALSE,
    include.colnames = FALSE,
    format.args = list(decimal.mark = ",")
)


## ------------------------------------------------------------------------

################################################################################
################################################################################
################################################################################
```

4 Reference

[1] R CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. Vienna, Austria: R Foundation for Statistical Computing, 2016. Dostupné z: https://www.R-project.org/

[2] KAISER, Henry F. An index of factorial simplicity. *Psychometrika* [online]. 1974, **39**(1), 31–36. Dostupné z: doi:10.1007/bf02291575

[3] BARTLETT, M. S. The Effect of Standardization on a χ^2 Approximation in Factor Analysis. *Biometrika* [online]. 1951, **38**(3/4), 337–344. ISSN 00063444. Dostupné z: http://www.jstor.org/

stable/2332580

[4] JÖRESKOG, K.G. *Statistical Estimation in Factor Analysis: A New Technique and Its Foundation* [online]. B.m.: Almqvist & Wiksell, 1963. Selected publications / University of Uppsala, Department of Statistics. Dostupné z: https://books.google.cz/books?id=VoC4AAAAIAAJ

[5] YEOMANS, Keith A. a Paul A. GOLDER. The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors. *Journal of the Royal Statistical Society. Series D (The Statistician)* [online]. 1982, **31**(3), 221–229. ISSN 00390526, 14679884. Dostupné z: http://www.jstor.org/stable/2987988

[6] KAISER, Henry F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* [online]. 1958, **23**(3), 187–200. Dostupné z: doi:10.1007/bf02289233