UNIVERZITA KOMENSKÉHO V BRATISLAVE

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

FAKTOROVÁ ANALÝZA A JEJ PRAKTICKÉ POUŽITIE

BAKALÁRSKA PRÁCA

2014 **Michaela FLORIÁNOVÁ**

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

FAKTOROVÁ ANALÝZA A JEJ PRAKTICKÉ POUŽITIE

BAKALÁRSKA PRÁCA

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2014

Michaela FLORIÁNOVÁ

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Michaela Floriánová
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov: Faktorová analýza a jej praktické použitie / *Factor analysis and its practical applications*

Cieľ: Zoznámiť sa s faktorovou analýzou. Naprogramovať aspoň niektoré v nej používané metódy. Predviesť jej použitie a intrepretáciu výsledkov na konkrétnych dátach.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 18.10.2013

Dátum schválenia: 14.11.2013 doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie

Veľká vďaka patrí môjmu vedúcemu práce, Mgr. Jánovi Somorčíkovi za jeho cenný čas, ochotu pomôcť, ako aj za schopnosť motivácie k pravidelnej a priebežnej práci. Tiež by som sa rada poďakovala mojim blízkym a priateľom, ktorí pri mne stoja a podporujú ma.

Abstrakt

FLORIÁNOVÁ, Michaela: *Faktorová analýza a jej praktické použitie* [Bakalárska práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2014, 45s.

Predmetom tejto bakalárskej práce je predstaviť štatistickú metódu faktorovej analýzy, pričom sme si zvolili dva primárne ciele. Prvým je vniknutie do numerických schém nachádzajúcich sa za rozličnými metódami faktorovej analýzy, ktorá sa snaží odhaliť skryté faktory popisujúce vzťahy medzi jednotlivými premennými. Druhým cieľom je praktické prevedenie týchto metód na získané dáta ankety Fakulty matematiky, fyziky a informatiky Univerzity Komenského. Pomocou výsledkov získaných jednotlivými metódami sa snažíme nájsť faktory, ktoré podávajú informáciu o predmetoch vyučovaných na našej fakulte.

Kľúčové slová: faktorová analýza, metóda hlavných komponentov, metóda maximálnej vierohodnosti, VARIMAX

Abstract

FLORIÁNOVÁ, Michaela**:** *Factor analysis and its practical applications* [Bachelor Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2014, 45p.

The subject of this Bachelor's thesis is to present a statistical method of factor analysis, whereby we have chosen two primary objectives. The first is to take a deep look into numerical schemes behind various methods of factor analysis, which seeks to discover the hidden factors describing the relationships between the variables. The second objective is the practical execution of these methods to survey data collected from inquiry of Faculty of Mathematics, Physics and Informatics, Comenius University. Results obtained using different methods try to find the factors that give information on the subjects taught at our faculty.

Keywords: factor analysis, method of principal components, maximum likelihood estimation, VARIMAX

Obsah

Úvod

Faktorová analýza patrí medzi viacrozmerné štatistické metódy a pokúša sa popísať vlastnosti množiny premenných pomocou menšieho počtu nových skrytých premenných, nazývaných faktory. Pomocou faktorov sa potom snaží vyvodzovať závery o podstate vzájomných závislostí pôvodných premenných s tým, že popíše podstatnú časť informácie. Počiatky faktorovej analýzy siahajú do začiatkov 20. storočia, kedy začala byť obľúbená v spoločensko–vedných výskumoch, hlavne v oblasti psychológie. O prvé praktické uvedenie sa zaslúžili matematik Karl Pearson a známy psychológ Charles Spearman pri ich slávnom meraní inteligencie [8]. Tým sa faktorová analýza začala predovšetkým využívať v oblasti psychometrie, no neskôr sa jej aplikácie rozšírili na mnohé iné odbory, napr. i do olympijských hier – známy príklad desaťboja [7], ako aj do ekonómie, sociológie, či iných oblastí.

Cieľom tejto práce je uviesť faktorovú analýzu a po podrobnom vysvetlení a vniknutí do numerických schém ukrytých za rozličnými metódami aplikovať túto štatistickú metódu na školskú anketu vytvorenú pre potreby Fakulty matematiky, fyziky a informatiky Univerzity Komenského. Našou snahou bude analyzovať odpovede na otázky kladené študentom v ankete rozličnými metódami a nájsť isté skryté faktory, ktoré nám popíšu názory študentov na jednotlivé predmety, ktoré navštevujú.

V prvej kapitole sa zameriame na uvedenie do problematiky faktorovej analýzy. Dôležitým nástrojom bude korelačná matica, z ktorej budeme vychádzať pri odhadovaní modelu. Predstavíme dve metódy odhadu, tzv. *metódu hlavných komponentov* a *metódu maximálnej vierohodnosti*. Následnými vhodnými transformáciami vieme zaručiť jednoduchšiu interpretáciu faktorov. Predstavíme teóriu skrytú za týmito transformáciami a tiež popíšeme, akým spôsobom sa budeme snažiť zjednodušiť štruktúry, aby boli faktory ľahšie interpretovateľné. Po vysvetlení teórie sa v druhej kapitole zameriame na numerické pozadie zabudovaných metód, ktoré si podrobne vysvetlíme a tiež naprogramujeme v štatistickom softvéri R [4]. V poslednej kapitole prevedieme aplikáciu na získané dáta, ktoré sa týkajú názorov študentov na jednotlivé predmety, ako aj ich vyučujúcich. Každý

študent môže ohodnotením navštevovaných predmetov prispieť ku kvalitnej výučbe, prípadne inšpirovať ku zmenám, ktoré by mohli zlepšiť vyučovanie. V tejto práci sa budeme na základe získaných dát snažiť odhaliť a pomenovať faktory, ktoré by nám popísali vzťahy medzi odpoveďami na jednotlivé otázky zo školskej ankety. Výsledky budeme analyzovať pomocou viacerých metód, pričom budeme hľadať čo najlepšie interpretácie samotných faktorov.

1. Faktorová analýza

Faktorová analýza je viacrozmerná štatistická metóda, zameraná na vytvorenie nových nepozorovateľných premenných , tzv. *faktorov*, pomocou ktorých sa zredukuje a zjednoduší pôvodný počet dát pri zachovaní podstatnej časti informácie. Lineárna kombinácia faktorov aproximuje pôvodné pozorovanie, pričom zachytáva skryté vzťahy medzi pôvodnými premennými.

Primárna otázka vo faktorovej analýze znie, či sú dáta konzistentné s predpísanou štruktúrou.

Úvodnú teóriu k modelu faktorovej analýzy spracujeme na základe [6].

1.1 Základná charakteristika modelu

Máme daný *p*- rozmerný, náhodný pozorovateľný vektor $X = (X_1, X_2, \dots, X_p)'$, so strednou hodnotou $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$ a kovariančnou maticou Σ. Model predpokladá, že *X* je lineárne závislý od niekoľkých náhodných premenných $F_1, F_2, \dots, F_m$ nazývaných spoločné faktory a p dodatočných zdrojov variability $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$, zvaných špecifické faktory, alebo jednoduchšie – chyby modelu.

Model faktorovej analýzy vyzerá teda nasledovne:

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1$$

$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2$$

$$\vdots \qquad\qquad \vdots$$

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p,$$

čo môžeme v maticovom zápise napísať ako:

$$(X - \mu)_{p\times 1} = L_{p\times m}F_{m\times 1} + \varepsilon_{p\times 1} \qquad (1)$$

Koeficienty l_{ij} voláme náklady *i*-tej premennej na *j*-ty faktor, preto matica *L* je matica faktorových nákladov. Každý špecifický faktor ε_i závisí iba od *i*-teho merania X_i. V modeli faktorovej analýzy je p odchýlok $X_1 - \mu_1, X_2 - \mu_2, \dots, X_p - \mu_p$ vyjadrených pomocou $p + m$ náhodných premenných $F_1, F_2, \dots, F_m, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$, ktoré sú nepozorovateľné, čo odlišuje tento model od viacrozmerného regresného modelu, v ktorom sú nezávislé premenné (ktorých pozíciu teraz zastávajú *F*) pozorovateľné.

Vzhľadom na množstvo nepozorovateľných veličín by model nebolo možné priamo riešiť. Preto pridávame niekoľko predpokladov pre náhodné vektory *F* a ε, ktoré implikujú isté kovariančné vzťahy v modeli.

Predpokladáme, že : $\mathrm{E}(F) = 0_{\mathrm{m}\times 1}, \qquad \mathrm{Cov}(F) = \mathrm{E}(FF') = I_{m\times m}$

$$\mathrm{E}(\varepsilon) = 0_{\mathrm{p}\times 1}, \qquad \mathrm{Cov}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon') = \Psi_{\mathrm{m}\times\mathrm{m}} = \begin{pmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{pmatrix} \quad (2)$$

$$\mathrm{Cov}(\varepsilon, F) = \mathrm{E}(\varepsilon F') = 0_{\mathrm{p}\times\mathrm{m}}\ , \quad \text{teda } F \text{ a } \varepsilon \text{ sú nezávislé}$$

Spomínaný model je ortogonálny a implikuje aj štruktúru kovariančnej matice vektora X. Z (2) vieme vyjadriť:

$$(X-\mu)(X-\mu)' = (LF+\varepsilon)(LF+\varepsilon)' = (LF+\varepsilon)((LF)'+\varepsilon')$$

$$= \quad LF(LF)' + \varepsilon(LF)' + LF\varepsilon' + \varepsilon\varepsilon',$$

čo využijeme pri vyjadrení kovariančnej matice:

$$\Sigma = \mathrm{Cov}\,(X) = \mathrm{E}(X-\mu)(X-\mu)' = L\mathrm{E}(FF')L' + \mathrm{E}(\varepsilon F')L' + L\mathrm{E}(F\varepsilon') + \mathrm{E}(\varepsilon\varepsilon')$$
$$= LL' + \psi$$

a $\mathrm{Cov}\,(X,F) = \mathrm{E}(X-\mu)F' = L\mathrm{E}(FF') + \mathrm{E}(\varepsilon F') = L$,

čiže

$$\mathrm{Var}(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i$$

$$\mathrm{Cov}(X_i, X_k) = l_{i1}l_{k1} + \cdots + l_{im}l_{km}$$

$$\mathrm{Cov}(X_i, F_j) = l_{ij}\ .$$

V našom modeli je veľmi dôležitý predpoklad linearity, bez ktorého by nebolo možné vyjadriť kovariančnú maticu ako $\Sigma = LL' + \Psi$, čo je východiskový vzťah pre faktorovú analýzu.

Množstvo variancie *i*-tej premennej $Var(X_i) = \sigma_{ii}$ zapríčinené *m*- spoločnými faktormi nazývame *i*-ta komunalita a zvyšok pridaný vplyvom špecifických faktorov nazývame špecifická variancia.

Keď *i*-tu komunalitu označíme ako h_i^2 a *i*-tu špecifickú varianciu ako ψ_i, dostávame:

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2$$

$$\sigma_{ii} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i = h_i^2 + \psi_i \qquad i = 1,2,\dots,p$$

Faktorový model usudzuje, že $p + \frac{p(p-1)}{2} = \frac{p(p+1)}{2}$ variancií a kovariancií matice *X* môže byť vytvorených z pm faktorových nákladov l_{ij} a p špecifických variancií ψ_i. V prípade, že *m*=*p*, každá kovariančná matica Σ sa dá vyjadriť v tvare *LL'* , čiže ψ je nulová matica. V takom prípade však faktorová analýza nemá celkom zmysel, pretože práve keď *m* je relatívne malé voči *p* je faktorová analýza užitočná, keďže ponúka "jednoduché" vysvetlenie kovariancií v *X* pomocou menej parametrov ako pôvodných *p*.

1.2 Metódy odhadu

Faktorová analýza si pre dané vektory pozorovaní $x_1, x_2, \dots, x_n$ pre p väčšinou korelovaných premenných kladie otázku, či faktorový model (1) s malým počtom faktorov adekvátne reprezentuje dáta. Tento problém sa potom rieši snahou overiť platnosti kovariančných vzťahov v (2).

Výberová kovariančná matica S z dát $x_1, x_2, \dots, x_n \in R^p$ je odhadom neznámej kovariančnej matice Σ. Ak sú prvky výberovej kovariančnej matice S mimo diagonály malé, (alebo môžeme vziať korelačnú maticu, v tom prípade, ak sú prvky výberovej korelačnej matice R mimo diagonály takmer nulové), vtedy premenné nie sú korelované, teda nie je medzi nimi podstatný súvis, a teda metóda faktorovej analýzy nebude užitočná. Za týchto okolností hrajú hlavnú rolu špecifické faktory, kým cieľom faktorovej analýzy je určiť niekoľko spoločných faktorov.

Ak sa však Σ signifikantne líši od diagonálnej matice, potom môžeme previesť faktorovú analýzu na model (1), kde základnou úlohou je odhadnúť faktorové náklady l_{ij} ako aj špecifické variancie ψ_i. Uvedieme dve z najznámejších metód odhadu parametrov: *metódu hlavných komponentov*, skr. PCA (z angl. principal component analysis), a *metódu maximálnej vierohodnosti*, skr. MLE (z angl. maximal likelihood estimation). Potom predstavíme rotáciu faktorov, ktorá zjednoduší interpretáciu faktorov.

1.2.1 Metóda hlavných komponentov

Základ tejto metódy, ktorá je odvodená od analýzy hlavných komponentov, spočíva v úvodnom spektrálnom rozklade kovariančnej, resp. korelačnej matice Σ. Nech má matica Σ p vlastných hodnôt $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, a ku každej prislúcha vlastný vektor e_i.

Potom:

$$\Sigma = \lambda_1 e_1 e_1{}' + \lambda_2 e_2 e_2{}' + \cdots + \lambda_p e_p e_p{}' = \left[\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \ldots \sqrt{\lambda_p} e_p\right] \begin{bmatrix} \sqrt{\lambda_1} e_1{}' \\ \sqrt{\lambda_2} e_2{}' \\ \vdots \\ \sqrt{\lambda_p} e_p{}' \end{bmatrix} \qquad (3)$$

spĺňa predpísanú kovariančnú štruktúru pre model faktorovej analýzy, ktorý má práve toľko faktorov, koľko je premenných ($m = p$) a špecifické variancie $\psi_i = 0 \quad \forall i,$ preto $\Sigma = LL'$.

Táto reprezentácia kovariančnej matice je síce presná, avšak nie je pre nás užitočná, pretože my požadujeme taký model, ktorý vysvetlí kovariančnú štruktúru iba pomocou niekoľkých spoločných faktorov ($m < p$). V druhej kapitole uvedieme preto pravidlá, ktoré určujú koľko faktorov dobre vysvetlí vzťahy medzi premennými. Pre metódu hlavných komponentov sa za pravidlo považuje zväčša zanedbať posledných *p-m* vlastných čísel, ktoré sú malé, a preto $\lambda_{m+1} e_{m+1} e_{m+1}{}' + \cdots + \lambda_p e_p e_p{}'$ má malý prínos do Σ.

Tak dostávame aproximáciu Σ :

$$\Sigma \doteq \lambda_1 e_1 e_1{}' + \lambda_2 e_2 e_2{}' + \cdots + \lambda_m e_m e_m{}' = \left[\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \ldots \sqrt{\lambda_m} e_m\right] \begin{bmatrix} \sqrt{\lambda_1} e_1{}' \\ \sqrt{\lambda_2} e_2{}' \\ \vdots \\ \sqrt{\lambda_m} e_m{}' \end{bmatrix} =$$

$$\tilde{L}_{p\times m} \tilde{L}{}'_{m\times p},$$

kde špecifické faktory ε nie sú veľmi podstatné, a preto môžu byť tiež ignorované pri faktorizovaní Σ. Ak zahrnieme do modelu aj špecifické faktory, ich variancie vezmeme ako diagonálne prvky matice $\Sigma - \tilde{L}_{p\times m} \tilde{L}{}'_{m\times p}$.

Keď vezmeme do úvahy špecifické faktory, aproximácia bude $\Sigma \doteq \tilde{L}\tilde{L}' + \Psi$

$$\Sigma \doteq \left[\sqrt{\lambda_1}e_1\ \sqrt{\lambda_2}e_2\ \ldots\ \sqrt{\lambda_m}e_m\right]\begin{bmatrix}\sqrt{\lambda_1}e_1{}'\\ \sqrt{\lambda_2}e_2{}'\\ \vdots\\ \sqrt{\lambda_m}e_m{}'\end{bmatrix}+\begin{bmatrix}\psi_1 & 0 & \cdots & 0\\ 0 & \psi_2 & 0 & \vdots\\ \vdots & 0 & \ddots & 0\\ 0 & \cdots & 0 & \psi_p\end{bmatrix},$$

$$\text{kde } \psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2 \qquad i = 1 \ldots, p\,.$$

Pre riešenie metódy hlavných komponentov môžeme poznamenať, že odhadnuté náklady pre dané faktory sa nemenia s pridaním ďalších faktorov.

1.2.2 Metóda maximálnej vierohodnosti

V prípade, že o spoločných faktoroch *F* a špecifických faktoroch ε môžeme predpokladať, že sú normálne rozdelené, potom pre ne môžeme vytvárať odhady metódou maximálnej vierohodnosti.

Keď F_j a ε_j sú združene normálne, potom sú aj pozorovania $X_j - \mu = LF_j + \varepsilon_j$ normálne a funkciou vierohodnosti bude:

$$L(\mu, \Sigma) = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}}e^{-\left(\frac{1}{2}\right)\mathrm{tr}\left[\Sigma^{-1}\left(\sum_{j-1}^{n}(x_j-\bar{x})(x_j-\bar{x})' + n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]} =$$

$$= (2\pi)^{-\frac{(n-1)p}{2}}|\Sigma|^{-\frac{(n-1)}{2}}e^{-\left(\frac{1}{2}\right)\mathrm{tr}\left[\Sigma^{-1}\left(\sum_{j-1}^{n}(x_j-\bar{x})(x_j-\bar{x})'\right)\right]} \times (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\left(\frac{n}{2}\right)(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)}, \quad (4)$$

kde $\Sigma = LL' + \Psi$. Aby bol tento model jednoznačne definovaný, musíme na neho položiť dodatočnú podmienku, ktorá ovplyvní výber matice *L*. V ďalšej časti uvedieme dôvod nejednoznačnosti tohto výberu.

V metóde maximálnej vierohodnosti tak uvažujeme, že máme n náhodných meraní $X_1, X_2, \dots X_n$ pochádzajúcich z normálneho rozdelenia $N_p(\mu, \Sigma)$, kde $\Sigma = LL' + \psi$. Odhady získané touto metódou, $\hat{L}, \hat{\psi}$, a $\hat{\mu} = \bar{x}$ maximalizujú (4) s podmienkou:

$$\hat{L}'\hat{\Psi}^{-1}\hat{L} = D\,, \qquad \text{kde } D \text{ je diagonálna matica.}$$

Odhady komunalít získané touto metódou sú:

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2 \qquad i = 1,2, \dots, p\,.$$

1.2.3 Porovnanie metód

Štúdie zaoberajúce sa problematikou faktorovej analýzy odporúčajú pre dobrú analýzu dát vykonať viacero druhov odhadu a potom ich výsledky navzájom porovnať. Skúsime si priblížiť výhody a nevýhody spomenutých dvoch metód.

Jedným z kritérií pri rozhodovaní o adekvátnosti prevedenia faktorovej analýzy je určenie kumulatívneho pomeru vysvetlenej variancie v modeli k celkovej variancii, čo v samostatnej podstate vyjadruje množstvo zachytenej informácie.

Ideálne by mal byť príspevok prvých pár faktorov do variancie premenných veľký, aby sme mohli povedať, že daný počet faktorov je dostačujúci na vysvetlenie modelu. Počet faktorov obsiahnutých v modeli teda zvyšujeme, kým nevysvetlíme požadované množstvo celkovej variancie.

Pri metóde PCA príspevok prvého spoločného faktoru do variancie s_{ii} je l_{i1}^2, preto príspevok prvého faktora do celkovej variancie $\sum_{\mathrm{i=1}}^{\mathrm{p}} s_{ii} = \mathrm{tr}(S)$ potom bude

$$l_{11}^2 + l_{21}^2 + \cdots + l_{p1}^2 = \left(\sqrt{\widehat{\lambda_1}}\widehat{e_1}\right)' \left(\sqrt{\widehat{\lambda_1}}\widehat{e_1}\right) = \widehat{\lambda_1}$$

Vo všeobecnosti teda môžeme napísať, že množstvo variancie zachytenej i-tym faktorom bude:

- v prípade kovariančnej matice S: $\frac{\hat{\lambda}_i}{\mathrm{tr}(S)}$
- v prípade korelačnej matice R: $\frac{\hat{\lambda}_i}{p}$, pretože $\mathrm{tr}(R) = p$.

V metóde MLE bude množstvo variancie zachytenej i-tym faktorom:

$$\frac{\hat{l}_{1j}^2 + \hat{l}_{2j}^2 + \cdots + \hat{l}_{pj}^2}{\mathrm{tr}(S)}.$$

Toto kritérium zvýhodňuje PCA, pretože týmto prístupom k riešeniu sme schopní vysvetliť viac percent z celkovej variancie. Dôvodom môže byť fakt, že faktorové náklady získané metódou PCA sa viažu na hlavné komponenty, ktoré majú vlastnosti optimalizujúce obsiahnutú varianciu.

Ďalším kritériom je reziduálna matica $S - (LL' + \Psi)$, čiže rozdiel medzi skutočnou a odhadnutou kovariančnou maticou (prípadne môžeme využiť aj korelačnú maticu R). Táto matica má nulové prvky na diagonále a ak aj ostatné prvky sú relatívne malé, môžeme považovať zvolený počet m za vhodný.

V tomto kritériu má zvyčajne výhodu práve metóda MLE.

1.3 Rotácia faktorov

Pre $m \geq 2$ sa s faktorovým modelom spája vždy istá nejasnosť, pretože matica L má rovnaké štatistické vlastnosti v danom modeli aj keď ju prenásobíme ľubovoľnou ortogonálnou maticou $T_{\mathrm{m\times m}}$, pre ktorú platí $T'T = TT' = I$, a teda pre $L^* = LT$ a $F^* = T'F$ vieme náš model prepísať na

$$X - \mu = LF + \varepsilon = LTT'F + \varepsilon = L^*F^* + \varepsilon$$

A keďže

$$E(F^*) = T'E(F) = 0 \text{ a } \quad Cov(F^*) = T'Cov(F)T = T'T = I \quad ,$$

tak hoci náklady L a L^* sú vo všeobecnosti odlišné matice, obe tvoria rovnakú kovariančnú maticu Σ:

$$\Sigma = LL' + \Psi = LTT'L' + \Psi = L^*L^{*\prime} + \Psi,$$

a teda udávajú rovnakú reprezentáciu dát. Rotáciou však môžeme dosiahnuť jednoduchšiu interpretovateľnosť faktorov, ktorú požadujeme.

Keď aplikujeme rotáciu matice odhadnutých faktorových nákladov rotačnou maticou *T*, zistíme tak , že aj reziduálna matica $S - \hat{L}\hat{L}' - \widehat{\Psi} = S - \widehat{L^*}\widehat{L^*}' - \widehat{\Psi}$ zostane taktiež nezmenená.

"Jednoduchá štruktúra" faktorových nákladov, ktorú by sme radi dosiahli, spočíva najmä v tom, že každá premenná má vysoké náklady na jeden z faktorov a na ostatné faktory nízke. Tiež v každom stĺpci matice *L* chceme nájsť skupinu vysokých faktorových nákladov, a skupinu zanedbateľných nákladov. V prípade použitia korelačnej matice tak hľadáme v celej matici *L* buď čísla blízke 1 v absolútnej hodnote, čo značí vysoký vplyv daného faktora na danú premennú alebo náklady blízke 0, v tom prípade má faktor na premennú nízky vplyv.

Podľa toho, či sú faktory navzájom korelované, alebo nie, rozlišujeme dva rôzne spôsoby rotácie. V prípade veľmi nízkej korelovanosti až nekorelovanosti používame ortogonálne metódy, ako napríklad *varimax, promax, quartimax,* a iné. Tieto zabezpečujú kolmosť osí po rotácií. Ak korelácia presiahne *0,32,* tak podľa článku [3] sa odporúča rotácia nepriamymi metódami, medzi ktoré patria *oblimin*, či *promax*.

Jednotlivými metódami rotácie sa v tejto práci nebudeme zaoberať, s výnimkou metódy VARIMAX, ktorú podrobne rozoberieme v druhej kapitole.

2 Numerické pozadie faktorovej analýzy

Okrem aplikovania metódy faktorovej analýzy na reálne dáta, ktorú uvedieme v poslednej kapitole, sa v tejto práci pokúsime vniknúť aj do numeriky, ktorá sa ukrýva za jednotlivými metódami, ktoré sú automaticky prevedené v štatistickom softvéri R funkciami "*factanal*" a "*varimax*". Pre lepšie pochopenie súvislostí v jednotlivých metódach sa ich pokúsime všetky sami naprogramovať. Okrem metódy PCA, ktorá v softvéri R vôbec nie je naprogramovaná pre účely faktorovej analýzy, sa zameriame na metódu MLE, ako aj rotáciu faktorov metódou VARIMAX.

Najprv však uvedieme, aké jednotlivé pravidlá pre výber počtu faktorov je možné zohľadniť, vzhľadom na to, že v praktickom časti budeme musieť sami stanoviť počet faktorov, ktoré použijeme.

2.1 Pravidlá pre výber počtu faktorov

Podľa [2] väčšina zo zaužívaných pravidiel sú skôr "zastavujúce pravidlá", čo znamená, že pri výskume postupne pridávame faktory, až kým nám ukazovateľ daného pravidla nepovie, že pridaním ďalšieho faktora viac stratíme, ako získame.

Pri porovnávaní metód PCA a MLE sme spomenuli kritérium, ktoré nám pomáha stanoviť počet zvolených faktorov a to *počet percent z kumulatívnej variancie*. Požadované minimálny počet percent nie je celkom jasne určený, ale minimum vysvetlenej variancie považujeme za 70%, pričom čím viac percent, tým lepšie, pretože tým viac informácie zachytíme.

Ďalším pravidlom je tzv. *Kaiserovo zastavujúce pravidlo,* ktoré tvrdí, že iba toľko faktorov treba zvažovať v analýze problému, koľko je vlastných čísel korelačnej matice vytvorenej z X, ktoré sú väčšie ako 1. Ukazuje sa, že súčet druhých mocnín vlastných hodnôt menších ako jedna, ktoré zanedbáme (čo v metóde PCA predstavuje práve spomenutých $p - m$

vylúčených vlastných hodnôt) dáva v súčte malé číslo, a teda implikuje to aj malú hodnotu súčtu druhých mocnín chýb aproximácie.

Iný spôsob ako otestovať počet faktorov pomocou vlastných hodnôt je tzv. *Scree test.* Táto stratégia zahŕňa vytvorenie grafickej vizualizácie znázorňujúcej vzťah medzi relatívnou veľkosťou vlastných hodnôt a počtom faktorov. Po vytvorení grafu v softvéri R pomocou "*screeplotu*" zistíme vhodný počet faktorov tak, že nájdeme bod zlomu, v ktorom priamka prestane stúpať strmo a začne sa vyrovnávať. Konkrétny príklad uvedieme v praktickej časti.

Výhodným kritériom, ktoré však často nie je možné aplikovať je tzv. *A priori kritérium.* Toto je možné využiť v prípade že už existuje výskum, na ktorý nadväzujeme, a ktorý už špecifikoval adekvátny počet faktorov, alebo pri prácou s konkrétnym prieskumom, ktorý má stanovený počet požadovaných faktorov. Toto kritérium vzhľadom na problém, ktorý sme si sami navrhli, pre túto prácu nebude možné využiť.

2.2 Rotácia metódou VARIMAX

Jedna z najznámejších rotácií, metóda VARIMAX je ortogonálna metóda, ktorá sa snaží o otočenie osí tvorených faktormi F o taký uhol, aby zrotované náklady matice L mali požadovanú štruktúru – ideálne, aby každá premenná mala vysoké náklady pre práve jeden faktor a aby suma štvorcov faktorových nákladov v jednotlivých stĺpcoch bola čo najväčšia.

Takto dosiahneme, že daný faktor, ktorý predstavuje stĺpec matice L ma vysoký vplyv iba na niektoré premenné, vďaka ktorým sme schopní odhadovať význam tohto skrytého faktora.

Metóda VARIMAX, je pevne zabudovaná v softvéri R vo funkcii "*factanal*", ktorý pre vstupné dáta prevedie faktorovú analýzu metódou MLE a ako výstup dostávame zrotovanú maticu nákladov L. Pre prevedenie rotácie VARIMAX v metóde PCA používame príkaz "*varimax*", ktorý ako vstup požaduje maticu nákladov L a ako výstup dostávame opäť zrotovanú maticu L ako aj maticu rotácie.

Pokúsime sa vniknúť do tejto metódy aj numerickou formou, čo znamená, že sa budeme snažiť nájsť uhly rotácie, vďaka ktorej vieme jednoduchšie interpretovať dáta. Naše výsledky potom porovnáme so zabudovanými funkciami. Spravíme tak pre *m=2* a *m=3* , pretože pre vyšší počet faktorov by bolo náročné získať štruktúru rotačnej matice *T*.

Pri zostavovaní algoritmu, sa budeme riadiť podľa [6] pánom Kaiserom, ktorý navrhol predeliť zrotované náklady odmocninami z komunalít (teda $l_{ij}^{\sim} = l_{ij}^{*}/h_i$) , čo spôsobí, že premenné, ktorých komunality sú malé, budú mať väčšiu váhu pri určovaní zjednodušenej štruktúry.

Metóda VARIMAX potom vyberie takú rotačnú maticu *T*, že hodnota:

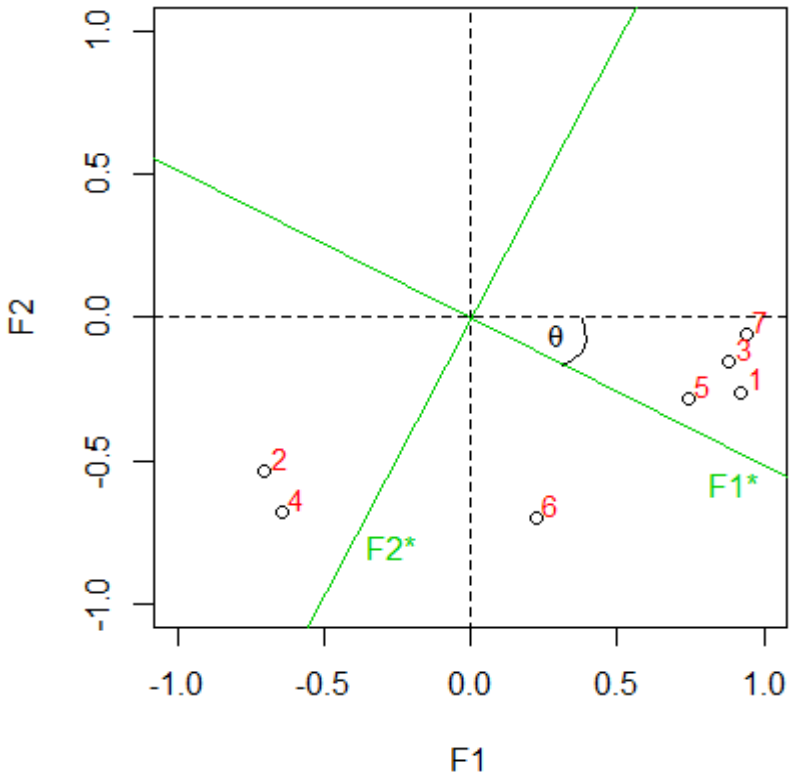
$$V = \sum_{j=1}^{m} (\text{variancia štvorcov preškálovaných nákladov pre j-ty faktor})$$

$$= \frac{1}{p}\sum_{j=1}^{m}\left(\sum_{i=1}^{p} {l_{ij}^{\sim}}^4 - \frac{1}{p}\left(\sum_{i=1}^{p} {l_{ij}^{\sim}}^2\right)^2\right)$$

bude čo najväčšia. Maximalizáciou *V* dosiahneme rozdelenie štvorcov nákladov na každý faktor najviac ako sa dá. Keďže rotačnou maticou rotujeme osi, na ktorých ležia faktory, znamená to, že otočíme osi o taký uhol (uhly), aby body, určené nákladmi na faktory ležali čo najbližšie k týmto osiam. Podľa ich polohy budeme vedieť povedať, ktorý faktor na danú premennú vykazuje vysoké náklady a ktorý malé.

Pre *m=2* volíme štruktúru rotačnej matice $T = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$, ktorá otočí maticu *L* v smere hodinových ručičiek o uhol θ. Po skonštruovaní funkcie *V* sme ju optimalizovali pomocou funkcie "*optimize*", čím sme našli hľadaný uhol a dostali identický výsledok so zabudovanou funkciou "*varimax*". Funkcia "*optimize*" hľadá minimum, resp. maximum zadanej funkcie na určenom intervale , ktorým je v našom prípade (0,2π), vzhľadom na prvý zadaný parameter, u nás uhol θ. Správnosť riešenia sme si overili aj vykreslením samotnej funkcie *V* na intervale (0,2π). V tomto prípade sme v dvojdimenzionálnom

priestore, môžeme tak graficky znázorniť rotáciu faktorov o nami získaný optimálny uhol $\theta = 27{,}6°$.



Obr.1 Rotácia faktorov metódou VARIMAX , p=7, m=2

Na *Obr. 1* si môžeme všimnúť rozmiestnenie siedmych premenných vzhľadom na osi F1 a F2 predstavujúce faktory. Otočenie osí o uhol θ vytvorilo nové, zrotované osi F1* a F2*, pre ktoré platí, že sú najlepšie možné, tak aby platilo, že každá premenná znázornená na obrázku guličkou bola silno ovplyvnená jedným faktorom (teda blízko k jednej osi) a druhým faktorom ovplyvňovaná minimálne (teda vzdialenosť k druhej osi veľká).

Vidíme, že F1 ovplyvňuje skupinu premenných 1,3,5,7 , kým F2 skupinu tvorenú premennými 2,4 . Premenná 6 je približne rovnako ovplyvnená oboma faktormi, je teda možné, že *m=2* nie je ideálny počet faktorov. Tieto výsledky podrobnejšie vysvetlíme v poslednej kapitole.

V prípade *m=3,* je výber matice *T* zložitejší, pretože máme niekoľko možností. Hoci by bolo možné maticu vyrobiť v závislosti od dvoch uhlov, my sme zvolili rotáciu okolo

základných osí *F1, F2, F3*, čím sme obišli problém hľadania roviny, podľa ktorej by sme rotovali v prípade 2 uhlov. Rotáciu prevedieme v poradí: okolo osi *x* o uhol ϕ, potom okolo *y* o uhol θ a nakoniec okolo osi *z* o uhol φ, čím vznikne matica:

$$T = \begin{pmatrix} \cos\theta\ \cos\varphi & \sin\phi\sin\theta\cos\varphi - \cos\phi\sin\theta & \cos\phi\sin\theta\cos\varphi + \sin\phi\sin\varphi \\ \cos\theta\sin\varphi & \sin\phi\sin\theta\sin\varphi + \cos\phi\cos\theta & \cos\phi\sin\theta\sin\varphi - \sin\phi\cos\varphi \\ -\sin\theta & \sin\phi\cos\theta & \cos\phi\cos\theta \end{pmatrix}$$

V tomto prípade musíme funkciu *V* optimalizovať v závislosti od troch parametrov, vybrali sme tak funkciu *"optim"*. Pri optimalizácií sme narazili na problém počiatočného vstupu, od ktorého závisel výsledok. Pri prevedení metódy na nami získané dáta sme sa tak nedostali k identickému výsledku so zabudovanou metódou, no pre vstupný vektor uhlov $(\Phi, \theta, \varphi) = (\pi, \pi, \pi)$ sú si výsledné rotačné matice ako aj spätne dopočítané uhly veľmi podobné :

$$T_{zabudovaná} = \begin{pmatrix} 0.8335431 & 0.5344137 & 0.1400282 \\ -0.3597509 & 0.7174264 & -0.5965556 \\ -0.4192674 & 0.4468796 & 0.7902617 \end{pmatrix}, \Phi \doteq 212°\ \theta \doteq 152°\ \varphi \doteq 166°$$

$$T_{optimálna} = \begin{pmatrix} 0.8335613 & 0.5544868 & 0.1400737 \\ -0.3596783 & 0.7084556 & -0.5965855 \\ -0.4192936 & 0.4469091 & 0.7902310 \end{pmatrix}, \Phi \doteq 210°\ \theta \doteq 155°\ \varphi \doteq 157°$$

Optimalizáciou funkciou "*optim*" sme sa tak veľmi blízko priblížili výsledku. Zistili sme že hodnota funkcie *V*, ktorú sme minimalizovali, je pri zabudovanej metóde $V_{zab} = 0{,}4359$, kým pri nami vyrátaných uhloch je $V_{opt} = 0{,}4501$, z čoho vidíme, že zabudovaná metóda poskytuje presnejšie riešenie, ako tá naša. Je možné, že pre iné vstupné hodnoty by sme sa ešte viac priblížili k riešeniu získanému zo zabudovanej metódy.

2.3 Metóda maximálnej vierohodnosti

Ďalšia zo zabudovaných metód, do ktorých sa snažíme numericky vniknúť je základná metóda odhadu MLE vykonávaná automaticky funkciou "*factanal*". Ponorenie sa do tejto metódy je značne zložitejšie. Budeme postupovať podľa odporúčanej metódy v [5]. Keďže kovariančná matica S je odhadom pre Σ, problém, ktorý treba vyriešiť je odhad matice nákladov L a špecifickej variancie Ψ . Tak musíme pomocou $\frac{p(p+1)}{2}$ známych prvkov matice odhadnúť $pm + p$ neznámych prvkov. Dôležitý je predpoklad, že p premenných pochádza z viacrozmerného normálneho rozdelenia. Lawley navrhol previesť si log-likelihood funkcie (4) na funkciu

$$h(\psi, L) = ln|\Sigma| - ln|S| + tr(\Sigma^{-1}S) - p \,, \tag{5}$$

kde $\Sigma = LL' + \Psi$. Funkcia h sa minimalizuje vzhľadom na ψ, pri pevnom L, za podmienky uvedenej v prvej kapitole:

$$\hat{L}'\widehat{\Psi}^{-1}\hat{L} = D \,, \qquad \text{kde } D \text{ je diagonálna matica.}$$

Po získaní optimálneho Ψ vyrátame nové L, ktoré závisí od Ψ podľa nižšie uvedených vzťahov. Tento postup opakujeme, až kým neskonvergujeme , teda kým

$$\sigma_{ii} - l_{i1}^2 - l_{i2}^2 - \cdots - l_{im}^2 - \psi_i \approx 0$$

Pôvodnou myšlienkou podľa pána Lawleya bolo hľadanie riešení rovníc $\frac{dh}{d\hat{L}} = 0$ a $\frac{dh}{d\widehat{\Psi}} = 0$, avšak väčšina iteračných metód konvergovala veľmi pomaly, preto Jöreskog a Lawley navrhli iný spôsob.

Navrhnutá iteračná schéma vyzerá nasledovne:

1.) Určíme počiatočný odhad diagonálnej matice Ψ, vytvorenej z vektora $\hat{\psi}$:

$\hat{\psi}_i = (1 - \frac{m}{2p})(\frac{1}{s_{ii}})$, kde s_{ii} sú prvky na diagonále inverznej kovariančnej matice S^{-1}.

2.) Po vytvorení matice Ψ budeme počítať vlastné hodnoty preškálovanej kovariančnej matice $S^* = \hat{\Psi}^{-1/2} S_n \hat{\Psi}^{-1/2}$, kde $S_n = \frac{n-1}{n} S$. Získame diagonálnu maticu $\hat{\Lambda}$, utvorenú z vektora prvých *m* zostupne zoradených vlastných hodnôt (väčších ako 1), a maticu im prináležiacich normalizovaných vlastných vektorov $\hat{E}$. Pomocou nich dostávame odhad

$$\hat{L} = \hat{\Psi}^{1/2} \hat{E} (\hat{\Lambda} - I)^{1/2} \tag{6}$$

3.) Odhadnuté $\hat{L}$ dosadíme do (5) a minimalizujeme vzľadom na $\hat{\psi}_1, \hat{\psi}_2, \dots, \hat{\psi}_p$ pomocou kvázinewtonovskej metódy navrhnutej pánmi Fletcherom a Powellom. Získané hodnoty $\hat{\psi}_i$ následne dosadíme do vzťahu (6) pre vytvorenie nového odhadu $\hat{L}$. Postup opakujeme, až kým neskonvergujeme k optimálnym hodnotám ψ a *L*.

Optimalizáciu v softvéri R sme sa snažili previesť funkciou *"constr.optim"*. Tento výber bol skomplikovaný vstupnou hodnotou vektora ψ, ktorá sa nenachádzala v požadovanom obore. Problém prípustnosti sme skúšali riešiť tak, že pomocou projekčných matíc sme vytvorili náhodný počiatočný vektor, ktorý spĺňal počiatočné podmienky. Tento spôsob však vyžadoval aspoň jednu prevedenú iteráciu, a tak sme sa dostali do zacykleného kruhu.

Použili sme preto funkciu *"auglag"* z knižnice *"alabama",* ktorá nepožaduje prístupnosť vstupného vektora a zároveň optimalizuje s reštrikciami. Za gradient sme dosadili podľa návrhu v [5]:

$$\frac{dh}{d\hat{\Psi}} = diag(\hat{\Psi}^{-1}(\Sigma - S)\hat{\Psi}^{-1})$$

Podmienku diagonálnosti $\hat{L}'\hat{\Psi}^{-1}\hat{L} = D$ si môžeme prepísať ako reštrikciu v tvare rovnosti:

$$\left(l_i * l_j\right)\left(\frac{1}{\psi}\right) = 0 \text{ pre } i \neq j, i, j \in \{1, \dots, m\}$$

Ďalšiu reštrikciu v tvare nerovnosti tvorí podmienka kladnosti:

$$\psi_i > 0 \ \ \forall\, i \in \{1, .., m\},$$

čím sa vyhneme neželaným prípadom zápornosti vektora ψ, tzv. *Heywood cases*.

Ďalší problém vytvárala kovariančná matica, pretože determinant inverznej kovariančnej matice bol záporný, prípadne veľmi malý, čo spôsobovalo chybné hlásenia v algoritme pri rátaní (5). Z uvedeného zdroja bolo povolené použiť namiesto kovariančnej matice *S* korelačnú maticu *R*. Táto zmena zabezpečila vhodný determinant.

Opakovaním uvedeného postupu sme odskúšali nami naprogramovaný algoritmus na zozbieraných dátach pre *p*=7 premenných. Po 8 iteráciách sme skonvergovali k riešeniu, ktoré sme následne porovnali so zabudovanou metódou MLE vo funkcii *"factanal"*.

V *Tabuľke 1* si môžeme pre porovnanie pozrieť optimálne hodnoty ψ_i:

	naše psi	psi z MLE
psi[1]	0.049	0.005
psi[2]	0.219	0.005
psi[3]	0.185	0.219
psi[4]	0.158	0.005
psi[5]	0.327	0.420
psi[6]	0.842	0.810
psi[7]	0.074	0.101

Tabuľka 1: Porovnanie optimálnych hodnôt vektora ψ

Musíme skonštatovať, že hoci sa naše výsledky približujú, nepodarilo sa nám skonvergovať k riešeniu zabudovanej metódy MLE. Je zaujímavé, že hodnota funkcie *h* pre naše ψ je menšia od tej pre ψ získané zabudovanou metódou: *h(nasepsi)= 0,1534555*, *h(psiMLE)=0,8279718.*

Uvedieme však aj získané matice faktorových nákladov:

	získaný F1	získaný F2	získaný F3	zabudovaný F1	zabudovaný F2	zabudovaný F3
[1]	0.957	-0.185	-0.021	0.968	-0.193	0.143
[2]	-0.577	-0.634	0.157	-0.271	0.957	0.075
[3]	0.892	-0.075	0.090	0.828	-0.287	0.113
[4]	-0.544	-0.736	-0.031	-0.351	0.680	0.640
[5]	0.714	-0.219	-0.400	0.697	-0.180	0.248
[6]	0.221	-0.350	0.070	0.243	0.049	0.358
[7]	0.955	0.026	0.089	0.899	-0.293	-0.071

Tabuľka 2: Porovnanie matíc faktorových nákladov

Opäť vidíme, že medzi jednotlivými faktorovými nákladmi sú rozdiely, avšak pomocou obidvoch matíc L by sme boli schopní interpretovať rovnaké faktory, ako uvedieme v nasledujúcej kapitole.

Dôvod, prečo sme nedokonvergovali k presnému riešeniu nevieme celkom určiť. Môže byť spôsobený nami – chybou v algoritme, avšak tiež nemôžeme vylúčiť možnosť, že zabudovaná metóda pre daný algoritmus použila úpravy pre vylepšenie výsledku, o ktorých my nemusíme vedieť.

3 Analýza školskej ankety

Vo tejto práci sme sa rozhodli použiť praktickú aplikáciu faktorovej analýzy na výsledkoch zo školskej ankety. Školská anketa je prostriedok pre študentov, ako aj profesorov a učiteľov Fakulty matematiky, fyziky a informatiky, aby pomocou ohodnotenia kvality predmetov a samotných vyučujúcich bolo možné štúdium zlepšiť, alebo prinajmenšom získať potrebnú spätnú väzbu zo strany študentov. Systém ankety je nastavený tak, že každému študentovi sa po prihlásení zobrazia predmety, ktoré v daný semester navštevoval,

a tie následne ohodnotí zaškrtnutím jednej z možností. Potom prejde na ďalšiu sekciu, a to je hodnotenie prednášajúceho a cvičiacich, prináležiacich k danému predmetu.

3.1 Návrh experimentu

Ku hodnoteniu predmetov sa viaže vždy 7 kritérií, a to:

1.	celkové hodnotenie kvality predmetu	(kvalita)
2.	náročnosť preberanej látky	(náročnosť)
3.	zaujímavosť obsahu	(zaujímavosť)
4.	množstvo práce vzhľadom na počet kreditov	(množstvo práce)
5.	jasnosť stanovených požiadaviek a spôsobu hodnotenia	(jasnosť hodnotenia)
6.	percento navštívených hodín	(návštevnosť)
7.	odporúčanie ďalším študentom	(odporúčanosť)

Pomocou faktorovej analýzy sa pokúsime zistiť, či je možné rozdeliť uvedené kritéria do menšieho počtu skupín, tak, že každej skupine bude možné nájsť faktor, ktorý ju popisuje. Našim cieľom je tak zredukovanie počtu otázok v ankete, pričom identifikujeme väčšinu informácií zachytených v odpovediach na pôvodné otázky.

V ankete sa nachádza veľké množstvo predmetov a každý hodnotil rozličný počet študentov, spomedzi tých, ktorí mali daný predmet zapísaný. Keďže všetky tieto údaje sú voľne dostupné v ankete, v tejto práci sme si mohli určiť pravidlo, ktoré vymedzilo hodnotenie iba tých predmetov, ktoré ohodnotil istý počet ľudí, resp. percento z ľudí, ktorí si zapísali daný predmet. Prvý nápad bol stanoviť si minimálny počet hodnotiacich ľudí na 20, bez ohľadu na počet ľudí, ktorí mali daný predmet zapísaný. Väčšinou odpovedalo okolo 50% všetkých respondentov. Takto sme získali informácie pre 30 rozličných predmetov. Neskôr sme zvážili toto pravidlo, ktoré sa ukázalo ako trochu prísne, keďže automaticky vyčleňovalo hodnotenia predmetov, ktoré mal zapísaný menší počet študentov.

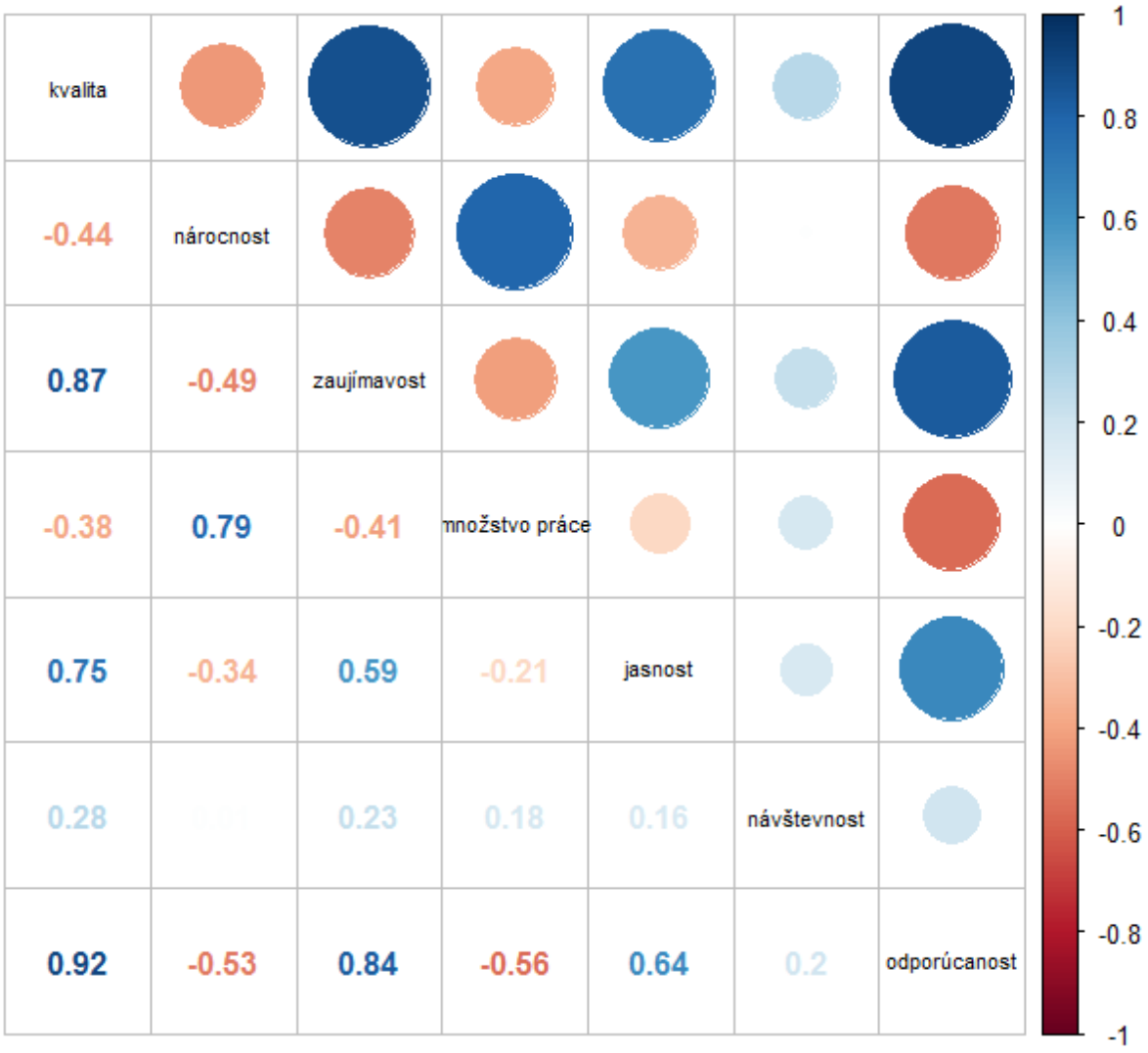
Rozhodli sme sa zaradiť do zoznamu aj predmety, ktoré sme považovali za adekvátne ohodnotené. Celkovo boli naše kritéria také, aby predmet ohodnotilo minimálne 10 ľudí a zároveň aspoň 50% všetkých respondentov. Niektorým predmetom sme však dali výnimku; buď ich ohodnotilo veľa ľudí, no stále to bolo malé percento z celkového počtu ľudí, ktorí si daný predmet zapísali (napr. Telesná výchova), alebo naopak, respondentov bolo síce len medzi 7-10, no tvorili veľkú časť všetkých zapísaných. Takto sme získali

údaje až pre 70 predmetov absolvovaných študentmi matematických, informatických či fyzikálnych študijných odborov.

3.2 Analýza výsledkov

Pri každom z predmetov treba odpovedať na 7 otázok, pričom možnosti ako odpovedať sa líšia. Niektoré kritériá sa hodnotia jednou až piatimi hviezdičkami, teda od 1-5. (Návštevnosť) sa hodnotí počtom percent navštívených hodín. Iné kritéria treba hodnotiť výberom jednej z ponúkaných slovných odpovedí, ktoré potom tvorcovia ankety prenásobia zvoleným koeficientom a dostanú tak číselné ohodnotenie. Pre daný predmet sa potom ohodnotenia sčítajú a predelia počtom všetkých respondentov, čím sa dosiahne priemerné skóre, ktoré dané kritérium na predmete získalo.

Vyrátané priemery pre každú zo siedmych odpovedí si pre každý predmet načítame do matice dát X, ktorá obsahuje 7 stĺpcov, ktoré predstavujú jednotlivých p=7 kritérií predmetu, a n=70 riadkov znázorňujúcich merania pre 70 rôznych predmetov. Na maticu X aplikujeme v softvéri R faktorovú analýzu. Najprv vyrátame korelačnú maticu, ktorú znázorníme aj graficky. Tak preskúmame vzťahy medzi odpoveďami na jednotlivé otázky, teda či sú silne korelované (navzájom silne súvisia), alebo či sú slabo korelované (a teda medzi odpoveďami na dané otázky nie je významný súvis). Tiež si budeme všímať znamienka – záporné znamienko naznačuje, že čím jedna hodnota viac stúpa, opačná klesá, čiže sú navzájom komplementárnymi.



Obr. 2: Korelačná matica , p=7

Na *Obr. 2* vidíme korelačnú maticu *R*, získanú z matice dát *X*. Keďže *R* je symetrická, rozdelili sme si ju podľa diagonály na dve časti. V spodnej časti matice sú vypísané výberové korelačné koeficienty, ktoré sú v hornej časti zakreslené guličkou. Čím väčšia a tmavšia je gulička, tým viac dané premenné korelujú. Červená farba znázorňuje zápornú korelovanosť, kým modrá kladnú. Premenné, ktoré sú silne korelované by mali pravdepodobne vytvárať skupiny, ktoré budú formované jednotlivými faktormi. Vidíme, že jednou silnou skupinou sú (kvalita), (zaujímavosť), (jasnosť hodnotenia) a (odporúčanosť), a ďalšiu tvoria (náročnosť) a (množstvo práce). Premenná (návštevnosť) nemá so žiadnou inou premennou silnú koreláciu.

Po úvodnej analýze našich dát sa zameriame na prevedenie samotnej faktorovej analýzy, pomocou ktorej sa budeme snažiť zredukovať počet premenných *p* na menší počet faktorov *m* a to cez dve metódy, ktoré sú predstavené v prvej kapitole.

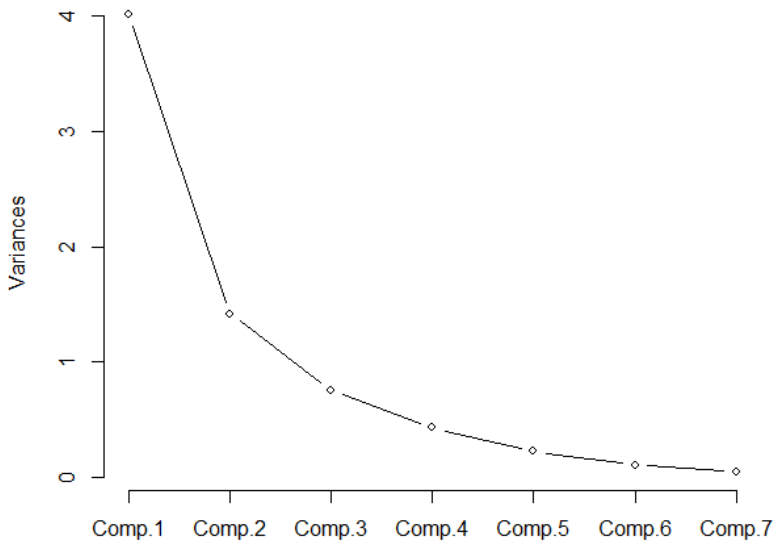
Ako prvú aplikujeme metódu PCA. Výsledky vidíme v *Tabuľke 3* a *4.*

	1	*2*	*3*	*4*	*5*	*6*	*7*
lambda	4.02	1.42	0.76	0.43	0.22	0.11	0.05
kumulatívna variancia	0.57	0.78	0.89	0.95	0.98	0.99	1

Tabuľka 3: Veľkosť vlastných hodnôt matice R a kumulatívna variancia

Lamda zobrazuje vlastné hodnoty korelačnej matice matice *R* (7x7) získanej z matice dát *X*. Pre metódu PCA odporúča *Kaiserovo zastavujúce pravidlo* zvoliť počet faktorov rovný počtu vlastných čísel väčších ako jedna. Z tohto dôvodu usudzujeme, že pre našu analýzu by mohol byť vhodné hľadať dva, prípadne tri faktory, vzhľadom na to, že 0,76 nie je až tak veľmi vzdialená od 1, a preto tretí faktor by mohol byť relevantný. *Kumulatívna variancia* znázorňuje podiel celkovej variancie vysvetlenej pomocou daných faktorov. Môžeme si všimnúť, že pridaním tretieho faktora zvýšime popísanú varianciu až o 11%, čím získame až 89% popísanej variancie pri *m*=3.

Pomôžeme si aj spomínaným *scree plotom*:



Obr. 3: Screeplot pre 70 predmetov

Z *Obr. 3* vidíme, že zlom v grafe sa nachádza niekde medzi Comp.2 a Comp. 3, ktoré predstavujú počet vlastných hodnôt. Z hľadiska kritéria vysvetlenej *kumulatívnej variancie* sme sa rozhodli, že za vhodný počet vysvetľujúcich faktorov budeme považovať *m*=3.

	Faktor1	**Faktor2**	**Faktor3**
kvalita	0.922	0.262	0.139
nárocnost	-0.702	0.536	0.314
zaujímavost	0.882	0.155	0.052
množstvo práce	-0.642	0.679	0.218
jasnost	0.742	0.286	0.366
návštevnost	0.221	0.703	-0.671
odporúcanost	0.942	0.06	0.061

Tabuľka 4: Výsledná matica L z metódy PCA bez rotácie

V *Tabuľke 4* vidíme maticu faktorových nákladov *L* pre *m*=3. Je pozitívne, že vieme pre niektoré faktory zakrúžkovať hodnoty, ktoré sú vysoké (blízke 1) a ktoré naopak značne nízke (blízke 0), a teda daný faktor má vysoký, resp. nízky vplyv na dané kritérium. Skúsme si pomenovať jednotlivé faktory. Faktor 1, ktorý predstavuje prvý stĺpec matice *L*, vysoko ovplyvňuje kritéria (odporúčanosť), (kvalita), (zaujímavosť), a tiež (jasnosť hodnotenia), (náročnosť) a (množstvo práce), pričom si treba všimnúť, že kritéria (náročnosť) ako aj (množstvo práce) tvoria faktorové náklady s opačným znamienkom, a teda čím je predmet menej náročnejší a množstvo práce adekvátnejšie počtu kreditov, tým je do istej miery zaujímavejší , kvalitnejší a hodnejší odporúčania ďalším študentom. Prvý faktor má tak vplyv hlavne na odporúčanosť, celkovú kvalitu a zaujímavosť predmetu, čo sú kritériá, ktoré vytvárajú istý **celkový imidž predmetu**, ako nazveme faktor 1. Faktor 2 má vysoké faktorové náklady pre (návštevnosť),(množstvo práce) a tiež (náročnosť). Tento krát sa všetky vyskytujú s rovnakým znamienkom, a teda čím viac považujem predmet za náročný a potrebný zapracovania pre získanie kreditov, tým viac prednášok absolvujem. Tento faktor sme nazvali **strašidelnosť predmetu**. Posledný, tretí faktor nie je tak jednoduché identifikovať, opäť sa tu vyskytuje návštevnosť, v tomto prípade s opačným znamienkom. Môžeme nechať tento faktor nepomenovaný, pretože pomocou rotácie faktorov budeme môcť získať jednoduchšiu štruktúru faktorových nákladov.

Skúsime teraz zrotovať maticu nákladov *L* metódou VARIMAX a overiť, či sme naozaj schopní zjednodušiť jej štruktúru a interpretovať i tretí faktor.

	Faktor1	Faktor2	Faktor3
kvalita	0.921	-0.243	-0.176
nárocnost	-0.261	0.9	0.027
zaujímavost	0.813	-0.337	-0.175
množstvo práce	-0.199	0.928	-0.143
jasnost	0.875	-0.028	0.015
návštevnost	0.156	0.086	-0.981
odporúcanost	0.832	-0.433	-0.119

Tabuľka 5: Výsledná matica nákladov L z metódy PCA po rotácii VARIMAX

Po ortogonálnej rotácii sa nám matica naozaj zjednodušila – pre každý faktor nachádzame buď vysoké náklady jednotlivých premenných, alebo naopak nízke, ako aj pre každú premennú vieme nájsť práve jeden faktor, ktorý ju silno ovplyvňuje. Faktor 1 zostáva **celkovým imidžom predmetu,** faktor 2 už má vysoké náklady iba pre (množstvo práce) a (náročnosť), čo by sme mohli interpretovať ako vynaloženú **námahu**. Posledný faktor teraz silne ovplyvňuje iba návštevnosť, ale so záporným znamienkom. Môžeme tento faktor nazvať **neúčasť na vyučovaní**.

Ďalej sa zameriame na metódu MLE, pre ktorú je dôležitý predpoklad normality. Po prevedení testu viacrozmernej normality na naše dáta pomocou funkcie " mshapiro.test" bola normalita zamietnutá, keďže $p - value = 6{,}731 \times 10^{-6}$. Napriek tomu, pre porovnanie s metódou PCA, predstavíme v *Tabuľke 6* výsledok získaný metódou MLE bez rotácie faktorov. Neskôr to porovnáme aj s prípadmi, keď v MLE faktory rotujeme metódami VARIMAX a *promax*. Musíme mať však na pamäti, že výsledky kvôli zamietnutej normalite nemusia byť celkom hodnoverné.

	Factor1	Factor2	Factor3
kvalita	0.691	0.719	-0.026
nárocnost	-0.909	0.254	-0.323
zaujímavost	0.686	0.557	0.028
množstvo práce	-0.888	0.335	0.306
jasnost	0.492	0.568	0.124
návštevnost	0.018	0.375	0.220
odporúcanost	0.779	0.521	-0.142

Tabuľka 6: Výsledná matica L z metódy MLE bez rotácie

V tejto metóde je kumulatívna variancia pre tri faktory vypočítaná v softvéri R zabudovanou metódou o čosi menšia, iba 78%, čo je typické pre metódu MLE. Zvolené tri faktory sú podobne interpretovateľné ako pri PCA. Všimnime si, že faktory 1 a 2 sú vo vymenenom poradí, čo však nie je problém, pretože matica L je určená až na poradie jednotlivých stĺpcov, čiže nákladov pre samotné faktory. Pri faktore 1 sú opäť dôležité (náročnosť), (množstvo práce), avšak namiesto (návštevnosť) má vysokú hodnotu faktorových nákladov premenná (odporúčanosť). Znamienka sú opačné a vysvetlenie je také, že čím náročnejší a pracnejší je predmet, tým menej je odporúčaný ďalším študentom. Vysoký súvis majú aj ostatné faktory, čiže (celková kvalita), (zaujímavosť obsahu) a (jasnosť požiadaviek), jedine (návštevnosť) celkom nesúvisí s daným faktorom, náklad je veľmi blízky nule. Faktor 1 by som nazvala **jednoduchosť predmetu** a to práve z dôvodu, že študenti doprajú svojim nasledovníkom, aby mali predmety ktoré neobsahujú náročné učivo a na ktorých sa dajú pomerne ľahko získať kredity. Faktor 2 svedčí ako faktor 1 v metóde PCA o **celkovom imidže predmetu**.

Opäť usudzujeme, že bude dobrý nápad zrotovať maticu faktorových nákladov pre m=3 získanú metódou MLE a porovnať, či dostaneme jednoduchšiu štruktúru, vďaka ktorej

budeme schopnejší pomenovať tretí faktor. Rotáciu prevedieme dvomi metódami: ortogonálnou metódou VARIMAX a nepriamou metódou *promax* a porovnáme výsledky.

	Factor1	Factor2	Factor3
kvalita	0.968	-0.193	0.143
narocnost	-0.271	0.957	0.075
zaujimavost	0.828	-0.287	0.113
množstvo práce	-0.351	0.680	0.640
jasnost	0.697	-0.180	0.248
návštevnost	0.243	0.049	0.358
odporúcanost	0.899	-0.293	-0.071

Tabuľka 7: Výsledná matica L z metódy MLE po rotácii VARIMAX

Tabuľka 7 naznačuje, že rotácia faktorov metódou VARIMAX nám naozaj pomohla s interpretáciou faktorov. Prvý faktor je rovnaký ako pri PCA - **celkový imidž predmetu** a druhý faktor je silne ovplyvňovaný (náročnosťou) a (množstvom práce), čo sme nazvali **námaha**. Posledný faktor sa nám trochu viac objasnil, i keď stále nie jednoznačne. Má vplyv na (množstvo práce), ale aj (návštevnosť), preto čím viac práce na počet kreditov sa vyžaduje, tým viac chodí študent do školy. Toto je pozoruhodné zistenie, ktoré sme sa snažili logicky interpretovať. Keďže z kritéria (množstvo práce) nevieme jasne, či sa myslí práca v škole, resp. doma na projektoch, môžeme usúdiť, že množstvo práce je veľké aj v škole, preto sa študentovi oplatí do nej chodiť, byť aktívnym a získavať body, informácie a podobne. Tento faktor by sme mohli pomenovať ako **predmet, na ktorom treba zabrať.**

Výsledky analýzy metódy MLE po rotácii *promax* sa nachádzajú v *Tabuľke 8.*

	Factor1	Factor2	Factor3
kvalita	1.044	0.079	-0.001
nárocnost	0.132	1.027	0.107
zaujímavost	0.834	-0.078	-0.008
množstvo práce	-0.241	0.331	0.755
jasnost	0.695	-0.077	0.162
návštevnost	0.233	-0.043	0.356
odporúcanost	0.961	0.048	-0.225

Tabuľka 8: Výsledná matica L z metódy MLE po rotácii promax

Opäť nám je jednoduchšie interpretovať výsledky, môžeme si však všimnúť, že výsledky po rotácii spomenutými dvomi metódami sú veľmi podobné. Hoci sú niektoré premenné viac a niektoré menej ovplyvňované, po analýze zistíme, že všetky tri faktory by sme mohli pomenovať rovnako ako pri rotácií VARIMAX.

Zistili sme, že získané dáta zo školskej ankety je možné analyzovať faktorovou analýzou, pričom pri odhadoch troch faktorov sme vždy schopní zachytiť okolo 70-90 % informácie. V prípade metódy PCA sme schopní vysvetliť o niečo viac variancie, keby sme však porovnali reziduálne matice, zvíťazila by metóda MLE. Pri obidvoch metódach nachádzame jeden fixný faktor, reprezentujúci **celkový imidž predmetu**, pričom zvyšné dva faktory objasnené rotáciou sú rôzne, predstavujú však podstatné veci, ako **námaha, strašidelnosť predmetu, neúčasť na vyučovaní, jednoduchosť predmetu,** či **predmet, na ktorom treba zabrať**. Vďaka rotácií sme tak schopní aplikácie faktorovej analýzy založenej na jasnej interpretácii faktorov.

3.3 Zmena číselného skórovania odpovedí

Porovnanie viacerých metód a výsledkov sme už aplikovali, no prišla idea vyskúšať preskórovať pôvodné dáta, a tak porovnať výsledky s pôvodnými. Niektoré otázky v ankete je možné ohodnotiť počtom hviezdičiek od 1 do 5, teda je jednoduché odpovede naskórovať a urobiť priemer. Pre niektoré otázky, ako napr. „ Boli jasne stanovené požiadavky a spôsob hodnotenia predmetu? “ boli však odpovede slovné. Rozhodli sme sa, že odpovediam „určite áno“ a „určite nie“ priradíme väčšiu dôležitosť ako odpovediam „skôr áno“ a „skôr nie“ v porovnaní so skórovaním, ktoré im bolo pridelené tvorcami ankety. Pôvodne boli odpovede obsahujúce nepresvedčivé slovo “skôr“ prenásobené +- 1, podľa toho, či odpoveď bola pozitívna, alebo negatívna. Odpovede obsahujúce “určite“ boli prenásobené +- 2, čo sme sa rozhodli zmeniť na +- 3, pretože si myslíme, že výpoveď “určite“, znie veľmi presvedčivo, študent si je odpoveďou istý. Po preskórovaní sme pre každý predmet vyrátali nové priemery hodnotenia študentov, čo sme opäť zvolili ako dáta do základnej matice X a ďalej sme postupovali ako pri prvej analýze. Pre porovnanie, či preskórovanie ovplyvní našu analýzu, uvedieme niektoré z výsledkov.

Najprv uvedieme korelačnú maticu, z ktorej sa budeme opäť snažiť vytvoriť skupiny premenných, ktoré sú vysoko korelované, a tak potencionálne riadené jedným z faktorov.

Obr. 4 : Korelačná matica pre naskórované odpovede, p=7

Po preškálovaní pôvodných dát vidíme opäť silnú skupinu (kvalita),(zaujímavosť),(jasnosť hodnotenia) a (odporúčanosť), kým ďalšia skupina je už oveľa menej výrazná a okrem (náročnosti) a (množstva práce) môžeme pridať aj (odporúčanosť). Najväčší rozdiel zaznamenávame v korelácií medzi premennými (náročnosť) a (množstvo práce), ktorá klesla z 0.79 na 0.45. Premenná (návštevnosť) opäť nekoreluje významne so žiadnou inou premennou, tvorí akúsi vlastnú skupinu.

Pomocou korelačnej matice prevedieme faktorovú analýzu a zameriame sa na porovnanie výsledkov, a síce, či nápad preskórovania slovných odpovedí získaných v školskej ankete bude mať efekt na interpretáciu faktorov.

	1	*2*	*3*	*4*	*5*	*6*	*7*
lambda	4.02	1.42	0.76	0.43	0.22	0.11	0.05
kumulatívna variancia	0.57	0.78	0.89	0.95	0.98	0.99	1

Tabuľka 9: Veľkosť vl. hodnôt matice R preskórovaných dát a kumulatívna variancia

Z *Tabuľky 9* vidíme, že veľkosť lámbd, ako aj kumulatívna variancia sú takmer rovnaké ako pri pôvodných odpovediach, preto zvolíme opäť *m*=3.

Začneme metódou PCA, ktorá už bez rotácie umožní interpretovať všetky tri faktory.

	Faktor1	**Faktor2**	**Faktor3**
kvalita	0.947	0.17	0.144
nárocnost	-0.572	0.373	0.645
zaujimavost	0.888	0.113	0.204
množstvo práce	-0.552	0.668	0.085
jasnost	0.769	0.136	0.144
návštevnost	0.237	0.803	-0.458
odporúcanost	0.947	0.011	0.102

Tabuľka 10: Výsledná matica L z metódy PCA bez rotácie, naskórované dáta

Faktory v tomto prípade sú podobné ako pred prečíslovaním, až na tretí faktor, ktorý je už bez rotácie interpretovateľný. Prvá dva faktory sme nazvali ako **celkový imidž predmetu**, a **predmet, na ktorom treba zabrať**. Tretí faktor ovplyvňuje (náročnosť) a (návštevnosť), lenže s opačnými znamienkami. Výsledok je celkom pozoruhodný, študenti chodia menej keď je látka ktorá sa preberá náročná a naopak, keď je ľahká, zúčastnia sa podľa výsledku

analýzy školskej ankety na viacerých prednáškach. Mohli by sme ho nazvať **nezrozumiteľnosť predmetu**.

Keďže v tomto prípade sme boli schopní interpretovať jednotlivé faktory už bez rotácie, uvedieme už iba výsledky získané metódou MLE.

Metóda MLE bez rotácie prinesie výsledok:

	Factor1	Factor2	Factor3
kvalita	0.859	0.486	0.031
nárocnost	-0.215	-0.480	0.176
zaujímavost	0.727	0.500	0.153
množstvo práce	0.136	-0.988	0.003
jasnost	0.718	0.307	-0.512
návštevnost	0.399	-0.153	0.176
odporúcanost	0.717	0.633	0.114

Tabuľka 11: Výsledná matica L z metódy MLE bez rotácie, naskórované dáta

Faktor 1 je opäť **celkový imidž predmetu**. Môžeme si však všimnúť, že po zmenení skórovania nám tento faktor silnejšie ovplyvňuje (navštevovanosť), ktorá ním predtým nebola skoro vôbec ovplyvňovaná. Faktor 2 môžeme nazvať **adekvátny počet kreditov**, pretože má na (množstvo práce) veľmi silný vplyv, kým na (odporúčanosť), (zaujímavosť), (kvalitu) a (náročnosť) má tiež vplyv ale okrem (náročnosti) opačný, teda čím menej sú zaslúžené kredity, t.j. menej bolo treba pracovať, tým viac sa predmet berie ako kvalitný a odporúčateľný. Faktor 3 súvisí s (jasnosťou hodnotenia), avšak nevieme ho konkrétne pomenovať.

Skúsme teda previesť rotáciu metódou VARIMAX:

	Factor1	Factor2	Factor3
kvalita	0.799	-0.239	0.528
nárocnost	-0.168	0.398	-0.347
zaujímavost	0.767	-0.293	0.359
množstvo práce	-0.088	0.988	-0.107
jasnost	0.342	-0.086	0.865
návštevnost	0.388	0.247	0.045
odporúcanost	0.763	-0.422	0.409

Tabuľka 12: Výsledná matica L z metódy MLE po rotácii VARIMAX, naskórované dáta

Faktory 1 a 2 môžu byť pomenované ako **celkový imidž predmetu** a **jednoduchosť predmetu** , hoci faktor 1 už veľmi neovplyvňuje (jasnosť hodnotenia) a viac ovplyvňuje (návštevnosť), kým faktor 2 má menšie faktorové náklady na (odporúčanosť). Faktor 3 je podobný faktorom, ktoré vyjadrujú celkový imidž predmetu, tentokrát má však vysoký podiel (jasnosť hodnotenia), a teda navrhujeme faktor nazvať ako **organizovanosť predmetu.**

Nápad prečíslovania hodnôt získaných z odpovedí na otázky školskej ankety, ktoré analyzujeme, chcel vylepšiť počiatočné dáta a overiť, či takéto zmeny môžu mať vplyv na výslednú interpretáciu faktorov získaných pomocou jednotlivých metód.

Číselné skórovania boli pozmenené pre premenné : (náročnosť), (jasnosť hodnotenia), (množstvo práce) a (odporúčanosť). Zvyšné kritéria zostali nezmenené, pretože boli ohodnotené buď podľa počtu hviezdičiek (celková kvalita) a (zaujímavosť), alebo podľa percenta navštevovaných hodín (návštevnosť).

Pomocou metódy MLE sme bez rotácie pomenovali nový faktor nazvaný **adekvátny počet kreditov** a po rotácii sme našli faktor, ktorý sme nazvali **organizovanosť predmetu.**

Môžeme si všimnúť, že tieto faktory ovplyvňujú hlavne premenné, ktorých číselné skórovanie bolo pozmenené. Zaujímavé je pozorovanie premennej (návštevnosť), ktorú sme boli schopní popísať o čosi lepšie.

3.4 Analýza obohatená o ohodnotenie vyučujúcich

Počet premenných p=7, ktorý sa snažíme zredukovať na menší počet faktorov m, nie je celkom ideálny. Preto sme sa rozhodli výsledky rozšíriť aj o ohodnotenia profesorov, ktorí prednášali, prípadne cvičili daný predmet. Aj tieto dáta sú voľne prístupné v školskej ankete. Vzhľadom na predmety ako Telesná výchova či Anglický jazyk, ktorí nemajú jasných vyučujúcich, sa počet našich meraní znížil na n=65, avšak počet premenných sa zvýšil na p=10, pretože každého vyučujúceho študenti hodnotia odpoveďami na tri otázky, ktorými sú:

8.	celkové hodnotenie vyučujúceho na danom predmete	(vyučujúci celkovo)
9.	schopnosť a ochota vyučujúceho komunikovať so študentmi a odpovedať na ich otázky	(komunikácia so študentmi)
10.	ústny prejav a prezentačné zručnosti vyučujúceho	(ústny prejav)

Pre novú maticu $X_{65\times10}$ sme previedli faktorovú analýzu. Opäť ukážeme korelačnú maticu a potom začneme metódou PCA, ktorá nám napovie, koľko faktorov je vhodné hľadať.

Obr. 5 : Korelačná matica pre p=10

Skupiny, ktoré načrtávajú, ktoré premenné budú ovplyvňovať jednotlivé faktory, sú rovnaké ako doteraz, akurát do prvej skupiny sa pridali všetky tri nové premenné súvisiace s ohodnotením prednášajúceho, teda (vyučujúci celkovo), (komunikácia so študentmi) a (ústny prejav). Premenné (náročnosť) a (množstvo práce) opäť negatívne korelujú so všetkými premennými okrem seba navzájom a ešte premennej (návštevnosť), s ktorou však korelujú minimálne.

Výsledky získané metódou PCA vidíme pre *p*=10 v *Tabuľke 13*.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
lambda	6.31	1.57	0.77	0.47	0.35	0.21	0.16	0.09	0.05	0.03
kumulatívna variancia	0.63	0.79	0.86	0.91	0.95	0.97	0.98	0.99	1	1

Tabuľka 13: Veľkosť vlastných hodnôt matice R a kumulatívna variancia, p=10

	Faktor1	**Faktor2**	**Faktor3**
kvalita	0.955	0.125	0.044
nárocnost	-0.603	0.626	0.333
zaujímavost	0.879	0.013	-0.053
množstvo práce	-0.528	0.764	0.219
jasnost	0.779	0.113	0.237
návštevnost	0.201	0.677	-0.693
odporúcanost	0.929	-0.088	-0.061
vyucujúci	0.929	0.163	0.14
komunikácia	0.879	0.163	0.211
ústny prejav	0.904	0.238	0.018

Tabuľka 14: Výsledná matica L z metódy PCA bez rotácie, p=10

Podobne ako predtým, z vlastných čísel korelačnej matice, teda komponentov vektora *lamda* vidíme, že opäť počet dva až tri faktory by mal byť vhodné pre náš model. Zvolili sme ako predtým *m*=3. Prvý faktor je ako doteraz celkový imidž predmetu, avšak má veľký vplyv aj na prednášajúceho, teda jeho celkové ohodnotenie (vyučujúci celkovo), prezentačné schopnosti (ústny prejav), ako aj schopnosť odpovedať študentom na otázky a komunikovať s nimi (komunikácia so študentmi). Tento faktor by teda mohol byť nazvaný **kvalita prednášajúceho a predmetu.** Druhý faktor je, podobne ako pri PCA pre

výsledky ankety bez hodnotení učiteľov, **strašidelnosť predmetu**. Tretí faktor je opäť bez rotácie faktorov ťažko pomenovateľný, má však súvis s (návštevnosťou) a (náročnosťou), čo sme raz nazvali **nezrozumiteľnosť predmetu**, pretože znamienka na jednotlivých faktorových nákladoch sú opačné.

Po rotácii VARIMAXom získavame jednoznačne faktory **kvalita prednášajúceho a predmetu, jednoduchosť predmetu** a **neúčasť na vyučovaní**, ako vidieť v *Tabuľke 15*.

	Faktor1	Faktor2	Faktor3
kvalita	0.914	-0.27	-0.145
nárocnost	-0.265	0.892	0.002
zaujímavost	0.783	-0.372	-0.154
množstvo práce	-0.194	0.917	-0.177
jasnost	0.812	-0.118	0.042
návštevnost	0.154	0.128	-0.969
odporúcanost	0.798	-0.475	-0.111
vyucujúci	0.932	-0.185	-0.083
komunikácia	0.909	-0.131	-0.018
ústny prejav	0.892	-0.172	-0.222

Tabuľka 15: Výsledná matica L z metódy PCA po rotácii VARIMAX, p=10

Ešte si preveríme aj výsledky získané cez metódu MLE, najprv bez rotácie, znázornené v *Tabuľke 16.*

	Factor1	Factor2	Factor3
kvalita	0.935	0.263	0.223
nárocnost	-0.617	0.494	-0.147
zaujimavost	0.821	0.144	0.294
množstvo práce	-0.629	0.774	0.011
jasnost	0.688	0.266	0.133
návštevnost	0.116	0.399	0.111
odporúcanost	0.919	0.014	0.227
vyuèujúci	0.938	0.294	-0.172
komunikácia	0.862	0.273	-0.175
ústny prejav	0.880	0.320	-0.131

Tabuľka 16: Výsledná matica L z metódy MLE bez rotácie, p=10

V prípade MLE sme schopní popísať 75% informácie pomocou dvoch faktorov a 78% pomocou troch faktorov. Skúsme interpretovať výsledky. Faktor 1 nazveme opäť **kvalita prednášajúceho a predmetu**, pričom si môžeme všimnúť, že (kvalita) a (vyučujúci celkovo) majú najväčšie náklady na daný faktor a práve tieto kritéria hodnotia celkovú kvalitu predmetu ako aj učiteľa. Teda je celkom viditeľné, že kvalitný profesor výrazne súvisí s celkovým dojmom kvality predmetu. Faktor 2 predstavuje znova istú **strašidelnosť predmetu**, kým faktor 3 je ťažko pomenovateľný a keďže kumulatívne nám neprispieva významne ku zachyteniu celkovej informácie, je možné, že aj až zanedbateľný.

V *Tabuľke 17* uvidíme, či sa niečo zmení keď maticu *L* získanú pomocou MLE zrotujeme metódou VARIMAX.

	Factor1	Factor2	Factor3
kvalita	0.773	-0.157	0.610
nárocnost	-0.249	0.707	-0.289
zaujímavost	0.603	-0.215	0.610
množstvo práce	-0.216	0.968	-0.104
jasnost	0.610	-0.050	0.433
návštevnost	0.204	0.313	0.213
odporúcanost	0.660	-0.375	0.565
vyuèujúci	0.954	-0.136	0.258
komunikácia	0.885	-0.123	0.222
ústny prejav	0.899	-0.087	0.277

Tabuľka 17: Výsledná matica L z metódy MLE po rotácii VARIMAX, p=10

Vďaka rotácií nám súvislosť schopného prednášajúceho s celkovým imidžom predmetu zostáva zachovaná a teda faktor 1 zostáva ako **kvalita prednášajúceho a predmetu**, kým faktor 2 sa zmení a faktor 3 budeme schopní pomenovať. Faktor 3 ovplyvňuje (zaujímavosť), (kvalitu), (odporúčanosť) ale aj (jasnosť hodnotenia), čo je ako pri pôvodných analýzach faktor nazvaný **celkový imidž predmetu.** Faktor 2 je faktorom **námahy**. Celkové percento popísanej informácie (78%) je pomerne dobré, a tak môžeme interpretáciu považovať za relevantnú.

Analýzou dát obohatených o ohodnotenia vyučujúcich sme dospeli k veľmi podobným záverom ako pri analýze pre p=7 kritérií, ktoré udávali ohodnotenia iba pre samotné predmety. Zásadným rozdielom je faktor 1, ktorý pôvodne hovoril o celkovom imidže predmetu, pre p=10 však reprezentuje aj kvalitu vyučujúceho, preto keď je dobrý prednášajúci, tak je dobrý aj celkový imidž predmetu.

Záver

Je mnoho oblastí, v ktorých je možné využiť faktorovú analýzu na hľadanie súvislostí, ktoré sú určené novými, ale skrytými premennými – faktormi. Obľúbenosť tejto metódy nás podnietila k jej preštudovaniu a aplikovaniu na informácie získané vo fakultnej ankete. Zaoberali sme sa len dvomi najpoužívanejšími metódami, ktorými vieme popísať korelačnú maticu, aby sme boli schopní vniknúť aj do detailov týchto metód. Prvá z nich, metóda hlavných komponentov - PCA, odvodená od analýzy hlavných komponentov, je celkom praktická, pretože nekladie žiadne dodatočné podmienky a pri vhodne zvolenom probléme dokáže zachytiť vysoké percento variancie. Druhá popísaná metóda maximálnej vierohodnosti - MLE vyžaduje predpoklad normality dát, ktorý bol žiaľ pre naše dáta zamietnutý. Táto metóda je priamo zabudovaná v softvéri R a používa zložitejší algoritmus, ktorý sa snaží numericky nájsť riešenie. My sme sa pokúsili naprogramovať tento algoritmus, čím sme sa dobre, no nie presne priblížili k výsledku zabudovanej metódy v Rku. Ďalšou dôležitou súčasťou faktorovej analýzy je rotácia jednotlivých faktorov. Rotáciu, ktorú chápeme ako určitú transformáciu, môžeme previesť viacerými spôsobmi. My sme sa zamerali na najzaužívanejšiu transformáciu VARIMAX. Tá sa snaží o vhodné otočenie faktorov, aby každá premenná bola silne ovplyvňovaná práve jedným faktorom a ostatnými takmer vôbec. Po prevedení rotácie sme schopní vďaka zjednodušenej štruktúre lepšie interpretovať jednotlivé faktory. Aj túto metódu sme si sami naprogramovali v softvéri a dostali očakávané výsledky.

Po dôkladnom pochopení a implementovaní základných metód sme sa pomocou nich snažili analyzovať odpovede zaznamenané v školskej ankete. Zozbierali sme odpovede pre 70 predmetov, pričom pri každom predmete študent odpovedal na 7 otázok týkajúcich sa kvality a priebehu predmetu. Uvážili sme, že je vhodné hľadať tri faktory, ktoré by nám vedeli povedať informácie o danom predmete. Pomocou metód PCA a MLE a následných rotáciách sme dokázali interpretovať faktory, ako napr. **celkový imidž predmetu, strašidelnosť predmetu, námaha, jednoduchosť predmetu,** alebo **neúčasť na vyučovaní.**

Ďalej sme sa rozhodli analyzovať výsledky ankety po jemnom preskórovaní jednotlivých odpovedí, pretože niektorým sme prikladali väčšiu váhu ako bola navrhnutá v ankete.

Chceli sme si tak aj overiť, či malá zmena v skórovaní odpovedí môže viesť k odlišnej interpretácii faktorov.

Okrem už interpretovaných faktorov sme identifikovali ďalšie, ako **adekvátny počet kreditov, organizovanosť predmetu,** či **nezrozumiteľnosť predmetu**. Zistili sme, že aj jemné preskórovanie odpovedí môže viesť k iným výsledkom a presvedčili sme sa, že nové faktory ovplyvňovali práve premenné so zmeneným skórovaním.

Ďalšou ideou bolo zahrnutie ohodnotenia prednášajúceho, ktoré tak zvýšilo počet premenných zo 7 na 10. Opäť sme previedli jednotlivé metódy a dopracovali sa k faktorom, ktoré boli podobné ako predtým. Dôležitým objavom bol faktor nazvaný **kvalita prednášajúceho a predmetu**, ktorý úzko súvisel s už nájdeným faktorom **celkového imidžu predmetu**. Tento jednoznačný výsledok tak naznačuje veľkú potrebu kvalitných pedagógov.

Posledným nápadom, ktorý sme chceli previesť v tejto práci, bol návrh na overenie, či nájdené faktory naozaj určujú vzťahy v dátach školskej ankety. Chceli sme to previesť takým spôsobom, že by sme si naše získané odpovede náhodne rozdelili na dve polovice a overovali, či budeme pre obidve polovice schopní nájsť rovnaké faktory. Tento návrh sme však nepreviedli, pretože počet dát pre obe podskupiny by bol iba 35, čo by nedávalo veľmi hodnoverné výsledky.

Aplikovaním faktorovej analýzy sme prišli k zaujímavým výsledkom, keďže sme dokázali objaviť faktory, ktoré hovorili o tom, že študenti navštevujú prednášky v prípade, že predmet vyžaduje veľa práce a je náročný, alebo tiež, že nechcú mať predmety, ktoré vyžadujú veľa námahy a tak ich ani neodporúčajú ďalším študentom. Pozoruhodný faktor nazvaný **nezrozumiteľnosť predmetu** priniesol paradoxný výsledok – študent ktorý považuje učivo za príliš náročné, chodí menej do školy, pretože nerozumie. Faktor **celkového imidžu predmetu** sme dostali po prevedení akejkoľvek metódy. Naznačuje, že zaujímavý predmet s jasnými požiadavkami a vysokou celkovou kvalitou, je odporúčaný aj ďalším ročníkom, pretože práve taký predmet študenti považujú za dobrý. Po pridaní ohodnotení učiteľov sme potvrdili, že čím lepší učiteľ, tým lepšiu povesť má samotný predmet.

Literatúra

[1] Anketa FMFI , dostupné na internete (19.3.2014) : https://anketa.fmph.uniba.sk/vysledky/2013-2014-zima

[2] Brown, J.D.: *Choosing the Right Number of Components or Factors in PCA and EFA*, Shiken: JALT Testing & Evaluation SIG Newsletter. Vol.13 No.2 (2009), 19– 23, dostupné na internete (14.4.2014): http://jalt.org/test/bro_30.htm

[3] Brown, J.D.: *Choosing the Right Type of Rotation in PCA and EFA*, Shiken: JALT Testing & Evaluation SIG Newsletter. Vol.13 No.3 (2009), 20 – 25, dostupné na internete (10.4.2014): http://jalt.org/test/PDF/Brown31.pdf

[4] Development Core Team (2010), *R: A language and environment for statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, dostupné na internete (20.2.2014): http://www.R-project.org

[5] Chen, R.: *An SAS/IML procedure for maximum likelihood factor analysis*, Behavior Research Methods, Instruments, & Computers Vol. 35 (2003), 310-317, dostupné na internete (3.5.2014): http://www.ncbi.nlm.nih.gov/pubmed/12834089

[6] Johnson, R.A, Wichern, D.W.: Applied Multivariate Statistical Analysis, Pearson Prentice Hall, New Jersey, 2007

[7] Linden, M. : *Factor Analytical Study of Olympic Decathlon Data*, Research Quarterly – American Alliance for Health, Physical Education and Recreation (1977) Volume 48, Issue 3, 562-568, dostupné na internete (20.2.2014): http://www.tandfonline.com/doi/abs/10.1080/10671315.1977.10615462#preview

[8] What are the Different Theories of Multiple Intelligence?, dostupné na internete (15.2.2014):http://general-psychology.weebly.com/what-are-the-different-theories-of-multiple- intelligence.html