Vzorová úloha 4.3 *Vyčíslení faktorů z korelační matice*

Spearman (1904) analyzoval známky 200 žáků ze tří předmětů. Po vyčíslení korelační matice ***R*** uvažoval jeden faktor a následující faktorový model:

$$x_1 = \lambda_1 f + u_1,$$
$$x_2 = \lambda_2 f + u_2,$$
$$x_3 = \lambda_3 f + u_3.$$

V tomto případě můžeme pojmenovat faktor *f* jako všeobecnou inteligenci žáka a specifické proměnné u_1, u_2, u_3 mají malé rozptyly, když jsou dotyčné proměnné x_i těsně spjaty s faktorem *f*. Z korelační matice plyne, že

$$\lambda_1\lambda_2 = 0.83, \quad \lambda_1\lambda_3 = 0.78, \quad \lambda_2\lambda_3 = 0.67,$$
$$u_1 = 1 - \lambda_1^2, \quad u_2 = 1 - \lambda_2^2, \quad u_3 = 1 - \lambda_3^2$$

a řešením vyjde $\lambda_1 = 0.99$, $\lambda_2 = 0.84$, $\lambda_3 = 0.79$,
$u_1 = 0.83$, $u_2 = 0.78$, $u_3 = 0.67$.

Vzorová úloha 4.4 *Ukázka pojmů a podstaty faktorové analýzy*

Na úloze **B4.02** *Účinky neuroleptik při tlumení rozličných psychóz* si ukážeme pomůcky vícerozměrné analýzy dat. K analýze užijeme také škálovaná data.

Řešení: byl použit program NCSS2000. Výstup metody Factor Analysis programu NCSS2000 pro nestandardizovaná data úlohy B402 obsahuje:

1. Popisné statistiky měr polohy a rozptýlení:

Proměnná	***n***	$\bar{x}$	***s***	*Komunalita H*
B402X1	20	20.05	33.89997	1.004443
B402X2	20	18.6	33.84236	1.005217
B402X3	20	2.95	5.206019	0.883469
B402X4	20	10.35	36.64951	0.846859

Klasické odhady parametrů polohy a rozptýlení pro jednotlivé proměnné informují o faktu, že proměnné byly správně vybrány. Komunalita ukazuje, jak dobře je tato proměnná predikována vybranými faktory.

2. (a) Korelační matice:

Proměnná	*B402X1*	*B402X2*	*B402X3*	*B402X4*
B402X1	1.000000	0.990529	0.835934	0.844519
B402X2	0.990529	1.000000	0.786439	0.851776
B402X3	0.835934	0.786439	1.000000	0.823784
B402X4	0.844519	0.851776	0.823784	1.000000

φ = 0.857883, Ln(Det|*R*|) = -7.336319, Bartlettův test = 123.49, SV = 6, Spočtená hladina významnosti α = 0.000000

Tabulka přináší korelace k posouzení celkové korelační struktury dat. Je zde několik případů vysokého korelačního koeficientu. Jsou-li všechny korelace nízké, menší než 0.3, není žádný důvod k užití faktorové analýzy. **Gleasonova-Staelinova míra redundance φ** = 0.8579 je veliká. Měří sílu vztahu mezi proměnnými. Nulová hodnota φ značí nulovou korelaci mezi proměnnými, zatímco hodnoty blízké jedné indikují silnou korelaci. I když je φ < 0.5, stále ještě *může být* nějaká struktura v datech. K vyčíslení Gleasonovy-Staelinovy míry φ se užívá vzorec

$$N = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} r_{ij}^2 - m}{m(m-1)}}.$$

Ln(Det*R*) značí přirozený logaritmus determinantu korelační matice. Při použití kovariance půjde o přirozený logaritmus determinantu kovariační matice. **Bartlettův test, SV, Spočtená hladina významnosti α**: jde o Bartlettův test sféricity k testování nulové hypotézy, že korelační matice je jednotková matice, všechny mimodiagonální prvky jsou nuly. Je-li velikost spočtené hladiny významnosti α větší než zadaná hodnota 0.05, neměli bychom aplikovat na tato data faktorovou analýzu ani metodu hlavních komponent. Test platí pro velké výběry (n>150) a užívá χ^2 rozdělení s $m(m-1)/2$ stupni volnosti: test lze užít pouze pro korelační, nikoliv však kovarianční matici. Testovací kritérium je vyčísleno vztahem

$$P^2 = \frac{(11 + 2m - 6n)}{6} \ln|\boldsymbol{R}|.$$

(b) Čárový diagram absolutních hodnot korelační matice:

Proměnná	*B402X1*	*B402X2*	*B402X3*	*B402X4*
B402X1		||||||||||||||||||	||||||||||||||||	||||||||||||||||
B402X2	||||||||||||||||||		||||||||||||||||	||||||||||||||||
B402X3	||||||||||||||||	||||||||||||||		||||||||||||||||
B402X4	||||||||||||||||	||||||||||||||||	||||||||||||||||	

Diagram zobrazuje absolutní hodnoty korelací a ukazuje největší a nejmenší korelaci proměnných.

3. Vyšetření indexového grafu úpatí vlastních čísel (Scree Plot):

Index	Vlastní číslo λ_i	Individualní procento	Kumulativní procento	Kumulativní čárový graf úpatí
1	3.507191	93.92	93.92	||||||||||||||||
2	0.187628	5.02	98.94	||
3	0.045168	1.21	100.15		
4	-0.005689	-0.15	100.00		

Jde o vlastní čísla matice $\boldsymbol{LL^T}$. Často se užívají jako rozlišovací kritérium při výběru počtu faktorů. Užívá se těch faktorů, jejichž vlastní čísla jsou větší než 1. Suma vlastních čísel je rovna počtu proměnných. Odtud platí, že první faktor obsahuje informaci obsaženou v 3.507191 původních proměnných. Zatímco všechna vlastní čísla jsou v PCA kladná, vlastní čísla ve FA mohou být i záporná. Obyčejně se tyto faktory vypouští a analýza se potom opakuje. **Individuální procento:** první sloupec přináší procento celkové proměnlivosti v proměnných, vystižené tímto faktorem a druhý sloupec pak **Kumulativní procento**. **Kumulativní čárový graf úpatí** určí hranu, index, rovnající se počtu užitých faktorů.

4. (a) Vlastní vektory pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	-0.525227	-0.361706	0.523702
B402X2	-0.519584	-0.545676	-0.236211
B402X3	-0.473506	0.692071	0.400954
B402X4	-0.479542	0.304046	-0.713566

Jde o vlastní vektory matice $\boldsymbol{L\,L^T}$.

(b) Čárový diagram absolutních hodnot vlastních vektorů pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	||||||||||	||||||||	||||||||||
B402X2	||||||||||	||||||||||	|||||
B402X3	|||||||||	||||||||||||	|||||||||
B402X4	|||||||||	|||||||	||||||||||||||

Diagram absolutních hodnot vlastních vektorů umožňuje rychle posoudit velikost vlastních vektorů, totiž, která původní proměnná x_j silně koreluje s dotyčným faktorem. Tak se znázorní struktura obou faktorů.

5. (a) Faktorové váhy pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	-0.983619	-0.156677	0.111301
B402X2	-0.973051	-0.236365	-0.050201
B402X3	-0.886759	0.299778	0.085213
B402X4	-0.898062	0.131701	-0.151652

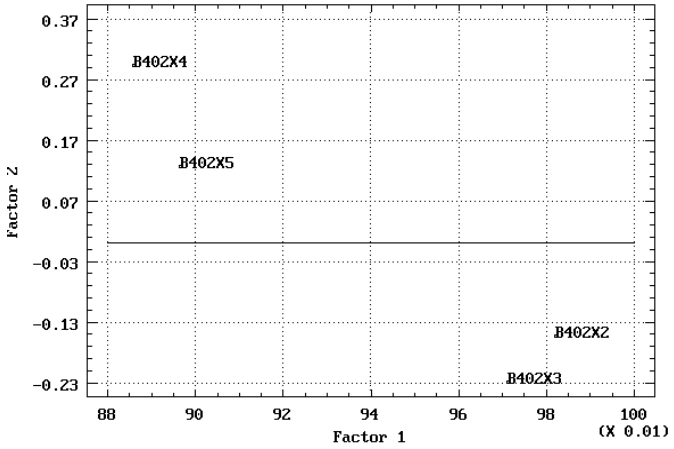
Tabulka numericky znázorňuje korelace mezi proměnnými a faktory.

(b) Čárový diagram absolutních hodnot faktorových vah pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	||||||||||||||||||	||||	|||
B402X2	||||||||||||||||||	|||||	||
B402X3	||||||||||||||||	||||||	||
B402X4	||||||||||||||||	|||	||||

Diagram znázorňuje absolutní hodnotu faktorových zátěží a vyjadřuje korelační strukturu jednotlivých původních proměnných s dotyčnými faktory. Faktor je obyčejně ovlivněn všemi původními proměnnými. Faktor1 je nejvíce ovlivněn *B402X1* a *B402X2*. Faktor2 pak nejvíce *B402X3* a také *B402X2* a nejméně proměnnými *B402X1* a *B402X4* .

6. Graf faktorových vah:



Obr. 4.12 Graf faktorových vah pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* pro data úlohy B402, ***STATGRAPHICS.*** *Proměnné B402X1 a B402X2 leží v diagramu blízko sebe, a proto silně korelují. Proměnné B402X3 a B402X4 jsou poněkud dál od sebe, proto méně korelují. Méně korelují s ostatními dvěma proměnnými B402X1 a B402X2, jsou totiž umístěny daleko od nich.*

7. Příspěvky daného faktoru do komunality:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*	*Kommunalita*
B402X1	0.967507	0.024548	0.012388	1.00
B402X2	0.946828	0.055869	0.002520	1.00
B402X3	0.786341	0.089867	0.007261	0.88
B402X4	0.806515	0.017345	0.022998	0.85

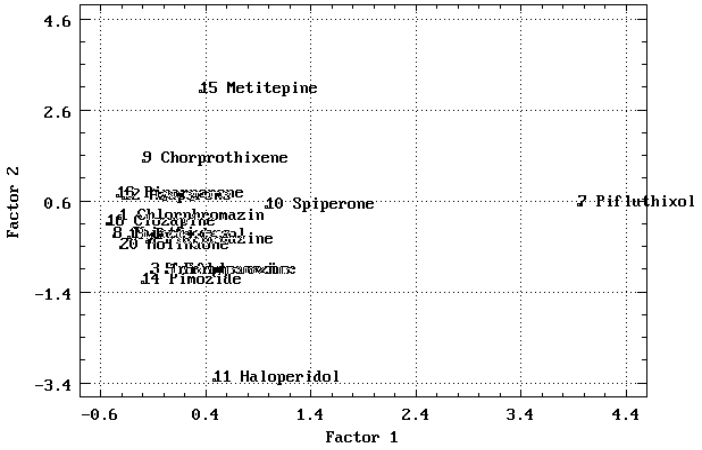
Komunalita představuje podíl proměnlivosti proměnné, vyjádřené dotyčným faktorem. Je podobná hodnotě R^2, kterou dostaneme, když budeme původní proměnné regresovat vybranými faktory. Tabulka obsahuje příspěvek daného faktoru do komunality. Diagram přináší příspěvky vybraných faktorů do komunality.

8. Faktorová skóre jednotlivých faktorů:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	-0.2804579	-0.8350396	2.464167
B402X2	-0.2774445	-1.259754	-1.111442
B402X3	-0.2528401	1.597723	1.886601
B402X4	-0.256063	0.7019241	-3.357533

V tabulce jsou koeficienty, které jsou užity k vytvoření faktorového skóre. Faktorová skóre jsou hodnoty faktorů pro jednotlivé řádky dat. Tyto koeficienty skóre jsou podobné vlastním vektorům. Protože byly předem normovány, přináší skóre jednotkový rozptyl a nikoliv rovný vlastním číslům. To způsobuje, že každý z faktorů má stejný rozptyl. Uživatel může použít tato skóre, jestliže chce vypočítat faktorové skóre pro nové řádky, jež nebyly zatím zařazeny do analýzy.

9. Rozptylový diagram faktorového skóre: diagram ukazuje na závislost faktoru proti faktoru. Prvních *k* faktorů (kde *k* je počet největších vlastních čísel) ukazuje na hlavní strukturu, která byla nalezena v datech. Zbytek faktorů ukazuje odlehlé hodnoty a lineární závislosti.



Obr. 4.13 Rozptylový diagram faktorových skore pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* ze standardizovaných dat úlohy B402, **STATGRAPHICS.** *Kromě tří objektů 7, 11 a 15 leží ostatních 17 objektů v jediném shluku. Objekty 7, 11 a 15 tvoří každý samostatný shluk. Co do podobnosti v těchto vlastnostech, vystižených dvěma hlavními komponentami v rovině, lze hovořit o 4 shlucích: první 15 Metitepin, druhý 7 Pifluthixol, třetí 11 Haloperidol, čtvrtý zbytek.*