Průzkumová analýza vícerozměrných dat

Vícerozměrná data: vyskytují se v situacích, kdy u každého z n objektů zjišťujeme hodnoty p znaků $X_1, ..., X_p$.
p-rozměrný datový soubor: matice n x p:

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Řádky charakterizují objekty, sloupce znaky.

Např. máme n sportovců, u každého sledujeme tyto znaky: pohlaví (0 – žena, 1 – muž), tělesná výška (v cm), tělesná váha (v kg), nejlepší výkon ve skoku do dálky (v cm), nejlepší výkon ve skoku do výšky (v cm), nejlepší výkon v běhu na 100 m (v s).

Úkoly průzkumové analýzy vícerozměrných dat:
- odhalit vektory pozorování nebo jejich složky, které se jeví jako vybočující
- postihnout závislosti mezi sloupci datového souboru
- identifikovat shluky v datech, které svědčí o nehomogenitě daného výběru
- posoudit vícerozměrnou normalitu dat.

Omezíme se na dva problémy, a to na vizualizaci dat pomocí hlavních komponent a na shlukovou analýzu dat.

Vizualizace vícerozměrných dat

Je-li p = 2 nebo p = 3, můžeme hodnoty znaků chápat jako souřadnice v dvou či třírozměrném prostoru a získáme tak dvourozměrný či třírozměrný tečkový diagram. Ze vzhledu těchto tečkových diagramů lze poznat, zda se v datech vyskytují odlehlá pozorování, zda mezi znaky existuje nějaká závislost nebo zda se objekty sdružují do skupin.
Je-li p > 3, použijeme k vizualizaci dat metodu hlavních komponent, která umožňuje vyjádřit informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků získaných jako lineární kombinace znaků původních. Tyto nové znaky, kterým se říká hlavní komponenty, jsou nekorelované a jsou uspořádané podle svého klesajícího rozptylu. Většina informace o variabilitě původních dat je tedy soustředěna v první hlavní komponentě a nejméně informace je obsaženo v poslední hlavní komponentě. Ukazuje se, že pouze několik prvních hlavních komponent má dostatečně velký rozptyl. Ostatní pak mů-

žeme zanedbat, čímž docílíme snížení dimenze dat. V datovém souboru však musí existovat mezi znaky dostatečně silná korelace, aby bylo možno tuto redukci provést.
Analýza hlavních komponent může být chápána jako transformace z původního do nového souřadnicového systému, jehož osy jsou tvořeny hlavními komponentami. Osy procházejí směry maximálního rozptylu, protože podmínka nezávislosti komponent vede ke kolmosti os.
Data pak znázorníme v prostoru prvních dvou či tří hlavních komponent.
Metodu hlavních komponent (Principal Component Analysis – PCA) popsal v r. 1901 Karl Pearson a ve 30. letech 20. století ji dále rozvinul Harold Hotelling.

Harold Hotelling (1895 – 1973), americký matematik a statistik

Podstata metody hlavních komponent

Uvažme datový soubor, který vznikl tak, že 6 žáků absolvovalo 4 testy, které měří následující veličiny:
X_1 – přírodovědné znalosti,
X_2 – literární vědomosti,
X_3 – schopnost koncentrace,
X_4 – logické myšlení.
Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek)

	1 X1	2 X2	3 X3	4 X4
1	7	9	10	8
2	9	8	8	10
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4

Označení
$\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ – vektor pozorování i-tého objektu, i = 1, 2, ..., n
Např. pro i = 3 máme $\mathbf{x}_3 = (4\ 3\ 1\ 2)^T$

$m_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ - průměr j-tého znaku, j = 1, 2, ..., p

Např. pro j = 1 máme $m_1 = \frac{1}{6}(7+9+4+2+3+1) = 4,\overline{3}$

$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - m_j)^2$ - rozptyl j-tého znaku, j = 1, 2, ..., p

Např. pro j = 1 máme $s_j^2 = \frac{1}{5}\left[(7-4,\overline{3})^2 + \ldots + (1-4,\overline{3})^2\right] = 9,4\overline{6}$

Datový soubor s průměry, směrodatnými odchylkami a rozptyly

	1 X1	2 X2	3 X3	4 X4
1	7	9	10	8
2	9	8	8	10
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4
průměry	4,33	4,17	4,00	5,00
s.o.	3,08	3,49	3,95	3,29
rozptyly	9,47	12,17	15,60	10,80

$z_{ij} = \frac{x_{ij} - m_j}{s_j}$ - (i,j)-tá standardizovaná hodnota, i = 1, 2, ..., n, j = 1, 2, ..., p

Např. pro i = 1, j = 1 máme $z_{11} = \frac{7-4,\overline{3}}{\sqrt{9,4\overline{6}}} = 0,8667$

Datový soubor standardizovaných hodnot

	1 X1	2 X2	3 X3	4 X4
1	0,866703	1,385674	1,519109	0,912871
2	1,51673	1,098983	1,012739	1,521452
3	-0,10834	-0,33447	-0,75955	-0,91287
4	-0,75836	-0,33447	-0,50637	-0,91287
5	-0,43335	-0,90786	-0,50637	-0,30429
6	-1,08338	-0,90786	-0,75955	-0,30429

$\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^T$ – vektor standardizovaných pozorování i-tého objektu, i = 1, 2, ..., n

$\mathbf{m} = (m_1, \ldots, m_p)^T$ – vektor průměrů

$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$ - výběrová varianční matice

V našem případě:

Proměnná	Kovariance (pca)			
	X1	X2	X3	X4
X1	9,46667	9,73333	10,60000	8,80000
X2	9,73333	12,16667	13,20000	9,40000
X3	10,60000	13,20000	15,60000	11,60000
X4	8,80000	9,40000	11,60000	10,80000

$\mathbf{R} = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T$ - výběrová korelační matice

V našem případě:

Proměnná	Korelace (pca)			
	X1	X2	X3	X4
X1	1,000000	0,906937	0,872258	0,870307
X2	0,906937	1,000000	0,958133	0,820031
X3	0,872258	0,958133	1,000000	0,893684
X4	0,870307	0,820031	0,893684	1,000000

(**S** a **R** jsou čtvercové symetrické matice řádu p.)

Základní pojmy

A - čtvercová matice řádu p.

Vlastní číslo matice **A** – takové číslo λ, které pro libovolný nenulový vektor **v** typu p x 1 splňuje rovnici $\mathbf{Av} = \lambda\mathbf{v}$.

Vlastní vektor matice **A** – vektor **v.**

Charakteristický polynom matice **A** - determinant $|\mathbf{A} - \lambda\mathbf{I}|$.

Stopa matice **A** - součet jejích diagonálních prvků (značí se Tr(**A**)).

Výpočet vlastních čísel matice A

Rovnici $\mathbf{Av} = \lambda\mathbf{v}$ upravíme na tvar $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{o}$. Tato soustava p rovnic má netriviální řešení, právě když charakteristický polynom matice **A** je roven 0. Dostaneme rovnici p-tého stupně. Jejím řešením jsou vlastní čísla $\lambda_1, ..., \lambda_p$. Jejich součet je roven stopě matice **A**.

Získání hlavních komponent

Nechť výběrová varianční matice **S** má vlastní čísla $l_1, ..., l_p$ a vlastní vektory $\mathbf{v}_1, ..., \mathbf{v}_p$, přičemž $\mathbf{v}_j^T\mathbf{v}_j = 1$, j = 1, ..., p a $\mathbf{v}_j^T\mathbf{v}_k = 0$ pro $j \neq k$. Znamená to, že vektory $\mathbf{v}_1, ..., \mathbf{v}_p$ jsou ortonormální. Bez újmy na obecnosti předpokládáme, že $l_1 > l_2 > ... > l_p$.

1. hlavní komponenta vznikne jako lineární kombinace znaků $X_1, ..., X_p$, kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru $\mathbf{v}_1$, tedy

$$Y_1 = v_{11}X_1 + ... + v_{1p}X_p.$$

Její rozptyl je l_1.

Dosadíme-li za $X_1, ..., X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_1 = (y_{11}, ..., y_{1n})^T$, kde $y_{1i} = \mathbf{v}_1^T\mathbf{x}_i$.

2. hlavní komponenta vznikne jako lineární kombinace znaků $X_1, ..., X_p$, kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru $\mathbf{v}_2$, tedy

$$Y_2 = v_{21}X_1 + ... + v_{2p}X_p.$$

Její rozptyl je l_2.
Přitom $\mathbf{v}_1^T\mathbf{v}_2 = 0$, tj. 1. a 2. hlavní komponenta jsou lineárně nezávislé.
Dosadíme-li za $X_1, ..., X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_2 = (y_{21}, ..., y_{2n})^T$, kde $y_{2i} = \mathbf{v}_2^T\mathbf{x}_i$.

.....

j-tá hlavní komponenta vznikne jako lineární kombinace znaků $X_1, ..., X_p$, kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru $\mathbf{v}_j$, tedy

$$Y_j = v_{j1}X_1 + ... + v_{jp}X_p.$$

Její rozptyl je l_j. Přitom $\mathbf{v}_j^T\mathbf{v}_k = 0$, j = 1, ..., k-1, tj. j-tá hlavní komponenta je lineárně nezávislá se všemi ostatními hlavními komponentami.
Dosadíme-li za $X_1, ..., X_p$ vektory pozorování $\mathbf{x}_i$, i = 1, ..., n, dostaneme vektor souřadnic $\mathbf{y}_j = (y_{j1}, ..., y_{jn})^T$, kde $y_{ji} = \mathbf{v}_j^T\mathbf{x}_i$.

Lze dokázat, že celková variabilita obsažená v datech je rovna stopě matice **S**, tj. součtu vlastních čísel $l_1 + ... + l_p$.

1. hlavní komponenta tedy vyčerpává $\frac{l_1}{l_1 + \ldots + l_p}100\%$ celkové variability.

Pokud je číslo $\frac{l_1}{l_1 + \ldots + l_p}$ dostatečně blízké 1, znamená to, že 1. hlavní komponenta dobře nahrazuje celý datový soubor. Je-li toto číslo podstatně menší než 1, musíme vzít tolik hlavních komponent, aby jejich součet dělený stopou matice **S** byl dostatečně blízký 1.
(V mnoha aplikacích se stává, že i při velkém počtu znaků stačí poměrně malý počet hlavních komponent.)
Znázorníme-li rozmístění objektů na ploše prvních dvou hlavních komponent, můžeme poznat, které objekty se řadí do skupin neboli shluků.
(Před provedením metody hlavních komponent je třeba se rozhodnout, zda budeme pracovat s původními hodnotami znaků nebo standardizovanými hodnotami.)
Důležité upozornění: Proměnné $X_1, \ldots, X_p$ musí být mezi sebou dostatečně korelované, jinak metoda hlavních komponent nedá dobré výsledky.

Koeficient korelace i-tého znaku X_j s k-tou hlavní komponentou Y_k lze vyjádřit jako $R(X_i, Y_k) = \frac{v_{ki}\sqrt{l_k}}{s_i}$.

Reprodukce výchozí kovarianční matice: V teorii matic se dokazuje vzorec $S = \sum_{i=1}^{p} l_i v_i v_i^T$ (tzv. spektrální rozklad matice **S**) Rozhodneme-li se uvažovat prá-

vě m hlavních komponent ($m \leq p$), pak pomocí tohoto vztahu můžeme posoudit, jak těchto m hlavních komponent reprodukuje rozptyly a kovariance původních proměnných. Lze posoudit i reziduální matici, tj. matici, kterou získáme jako rozdíl výchozí kovarianční matice a reprodukované kovarianční matice.

Doporučený postup při analýze hlavních komponent

a) Provedeme tabulkové a grafické zpracování datového souboru, abychom se blíže seznámili s daty.

b) Sestavíme korelační matici a prověříme, zda jsou korelace natolik silné, aby mělo smysl provádět analýzu hlavních komponent.

c) Rozhodneme, kolika hlavními komponentami lze popsat datový soubor bez podstatné ztáty informace. Označme tento vhodný počet jako m. Při stanovení m můžeme použít tato pomocná kritéria:

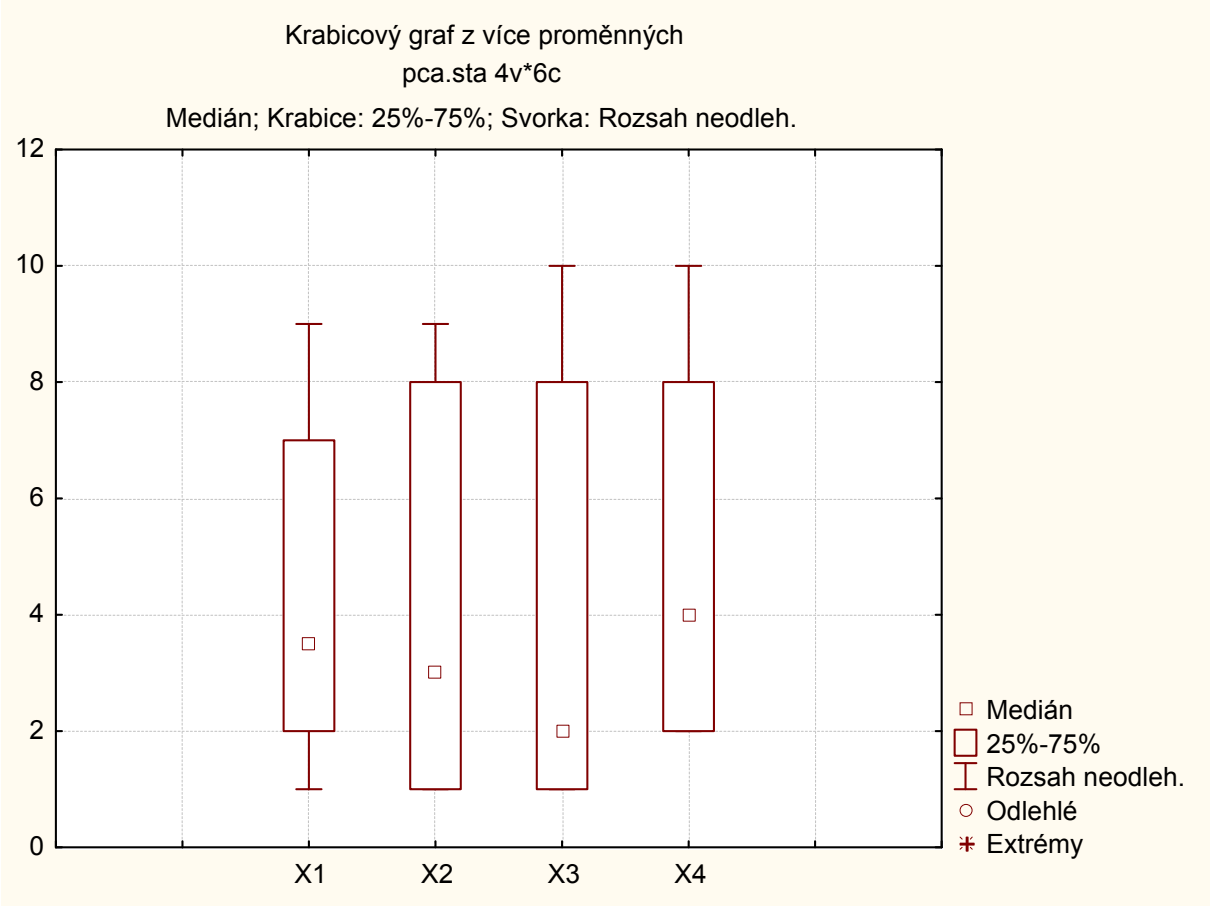
- Kaiserovo kritérium - za m volíme počet těch vlastních čísel matice **R**, která jsou větší než 1.
- Sutinový test (scree test) – grafická metoda, která spočívá v subjektivním posouzení vzhledu sutinového grafu (scree plot), tj. grafu znázorňujícího velikosti sestupně uspořádaných vlastních čísel matice **R**. Objeví-li se v grafu určité zploštění, pak za m vezmeme to pořadové číslo, kde se zploštění projevilo.
- Kritérium založené na kumulativním procentu vysvětleného rozptylu. Požadujeme, aby vybrané hlavní komponenty vysvětlily aspoň 70% celkového rozptylu.
- Kritérium založené na reziduální korelační či kovarianční matici. Požadujeme, aby prvky reziduální matice byly co možná nejmenší.

d) Pokusíme se o interpretaci prvních m hlavních komponent. Zkoumáme přitom, jak jsou jednotlivé vybrané hlavní komponenty utvořeny z původních znaků a jak s nimi korelují.

e) Vypočítáme vektory souřadnic a následně sestrojíme dvourozměrné tečkové diagramy.

Pro náš datový soubor nejprve znázorníme data pomocí krabicových diagramů: Grafy – 2D Grafy – Krabicvé grafy – zvolíme Vícenásobný – Proměnné - Závis-

le proměnné X1-X4 – OK – OK



Nyní vypočte korelační matici: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné X1 až X4, OK – OK – Popisné statistiky – Korelační matice

	Korelace (pca.sta)			
Proměnná	X1	X2	X3	X4
X1	1,000000	0,906937	0,872258	0,870307
X2	0,906937	1,000000	0,958133	0,820031
X3	0,872258	0,958133	1,000000	0,893684
X4	0,870307	0,820031	0,893684	1,000000

Dále vypočteme vlastní čísla a procento vysvětleného rozptylu: na záložce Základní výsledky vybereme Vlastní čísla.

	Vlastní čísla korelační matice a související statistiky (pca Pouze aktiv. proměnné			
Pořadí vl.č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	3,661431	91,53577	3,661431	91,5358
2	0,188636	4,71589	3,850066	96,2517
3	0,134072	3,35181	3,984139	99,6035
4	0,015861	0,39653	4,000000	100,0000

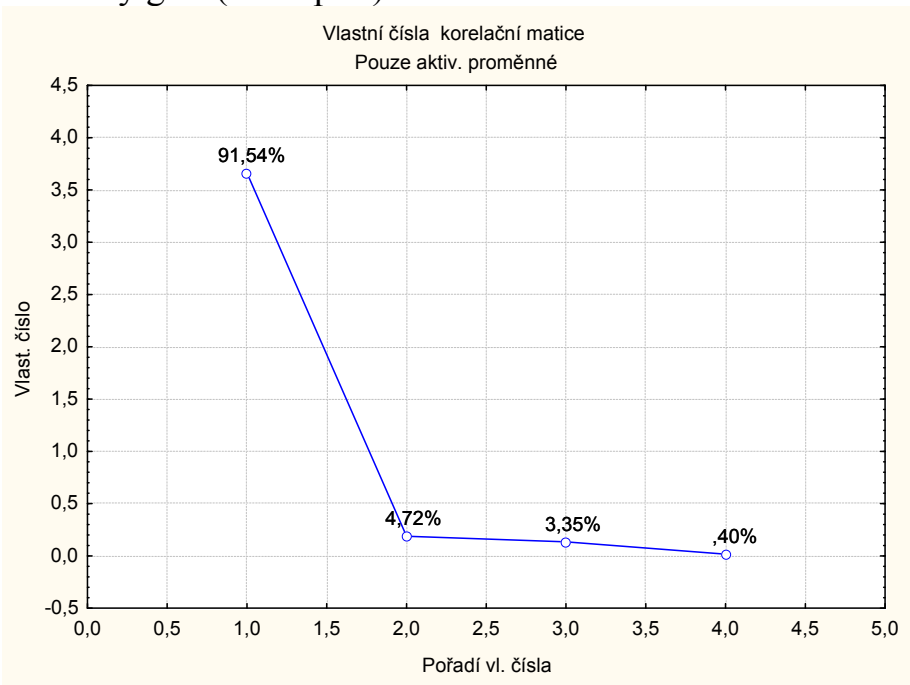
Vidíme, že 1. vlastní číslo $l_1 = 3,66$, tedy 1. hlavní komponenta vyčerpává 91,5% variability dat, 2. vlastní číslo $l_2 = 0,19$, 2. hlavní komponenta vyčerpává 4,7% variability dat atd.
Podle Kaiserova kritéria by stačilo uvažovat pouze 1. hlavní komponentu, protože pouze první vlastní číslo je větší než 1. Kvůli znázornění objektů však budeme uvažovat první dvě hlavní komponenty.

Dále vypočítáme vlastní vektory: na záložce Proměnné vybereme Vlastní vektory

	Vlastní vektory korelační matice (pca) Pouze aktiv. proměnné			
Proměnná	Faktor 1	Faktor 2	Faktor 3	Faktor 4
X1	-0,498301	-0,000518	0,817131	-0,289816
X2	-0,503657	0,582217	-0,082290	0,632916
X3	-0,508833	0,185043	-0,539021	-0,645217
X4	-0,488994	-0,791696	-0,187036	0,314832

1. hlavní komponenta:
$Y_1 = -0,49X_1 - 0,5X_2 - 0,5X_3 - 0,49X_4$,
2. hlavní komponenta:
$Y_2 = -0,0005X_1 + 0,58X_2 + 0,19X_3 - 0,79X_4$ atd.

Sutinový graf (scree plot):



V sutinovém grafu nastává výrazné zploštění po 1. vlastním čísle.

Výpočet koeficientů korelace 1. a 2. hlavní komponenty a původních čtyř proměnných: na záložce Proměnné vybereme Korelace faktorů & proměnných

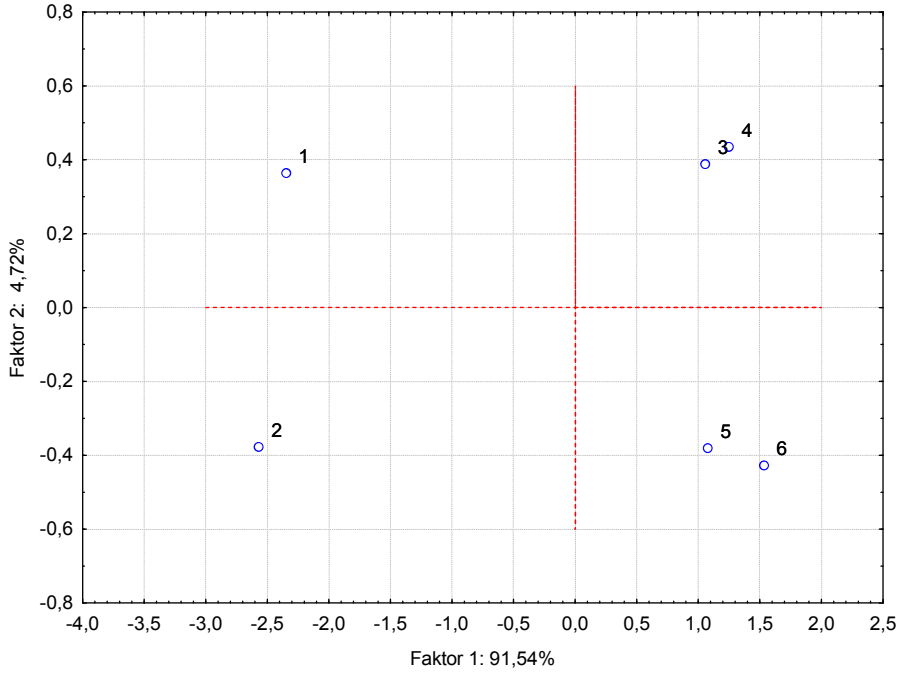
Proměnná	Faktor 1	Faktor 2
X1	-0,953492	-0,000225
X2	-0,963740	0,252869
X3	-0,973645	0,080368
X4	-0,935684	-0,343851

Vidíme, že 1. hlavní komponenta vysoce záporně koreluje se všemi proměnnými. 2. hlavní komponenta slabě kladně koreluje s druhou proměnnou a středně silně záporně koreluje s třetí proměnnou.

Podívejme se rovněž na vektory souřadnic (v systému STATISTICA se jim říká faktorové souřadnice případů): na záložce Případy vybereme Faktorové souřadnice případů.

Případ	Faktor 1	Faktor 2
1	-2,34914	0,364696
2	-2,56859	-0,378068
3	1,05532	0,387487
4	1,25040	0,434674
5	1,07964	-0,381138
6	1,53238	-0,427651

Znázornění objektů (žáků) na ploše prvních dvou hlavních komponent:



Shluková analýza

Cíl shlukové analýzy

Cílem shlukové analýzy je roztřídění n objektů, z nichž každý je popsán p znaky, do několika pokud možno stejnorodých (homogenních) skupin (shluků, clusterů). Požadujeme, aby objekty uvnitř shluků si byly podobné co nejvíce, zatímco objekty z různých shluků co nejméně. Přesný počet shluků většinou není přesně znám.

Shluková analýza nachází uplatnění v celé řadě oborů, např. v biologii. U n populací změříme p biometrických charakteristik a zjišťujeme, zda určité skupiny populací tvoří shluky.

Shluková analýza je ovšem průzkumovou metodou a měla by sloužit jako určité vodítko při dalším zpracování dat.

Podobnost objektů

Podobnost (či rozdílnost) objektů posuzujeme pomocí různých měr vzdálenosti. Pro znaky intervalového či poměrového typu nejčastěji používáme euklidovskou vzdálenost. Nechť k-tý objekt je popsán vektorem pozorování $\mathbf{x}_k = (x_{k1}, ..., x_{kp})^T$ a l-tý objekt vektorem $\mathbf{x}_l = (x_{l1}, ..., x_{lp})^T$.

Euklidovská vzdálenost k-tého a l-tého objektu:

$$d_{kl} = \sqrt{\sum_{j=1}^{p} \left(x_{kj} - x_{lj}\right)^2} \ .$$

Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do matice vzdáleností. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

Matice euklidovských vzdáleností pro datový soubor s údaji o 6 žácích:
Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné X1 – X4 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky) – OK – na záložce Detaily vybereme Matice vzdáleností.

	Euklid. vzdálenosti (pca)					
Případ	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0,0	3,6	12,7	12,7	12,6	14,0
P_2	3,6	0,0	12,8	13,2	12,5	14,1
P_3	12,7	12,8	0,0	2,2	3,2	4,1
P_4	12,7	13,2	2,2	0,0	3,0	3,2
P_5	12,6	12,5	3,2	3,0	0,0	2,2
P_6	14,0	14,1	4,1	3,2	2,2	0,0

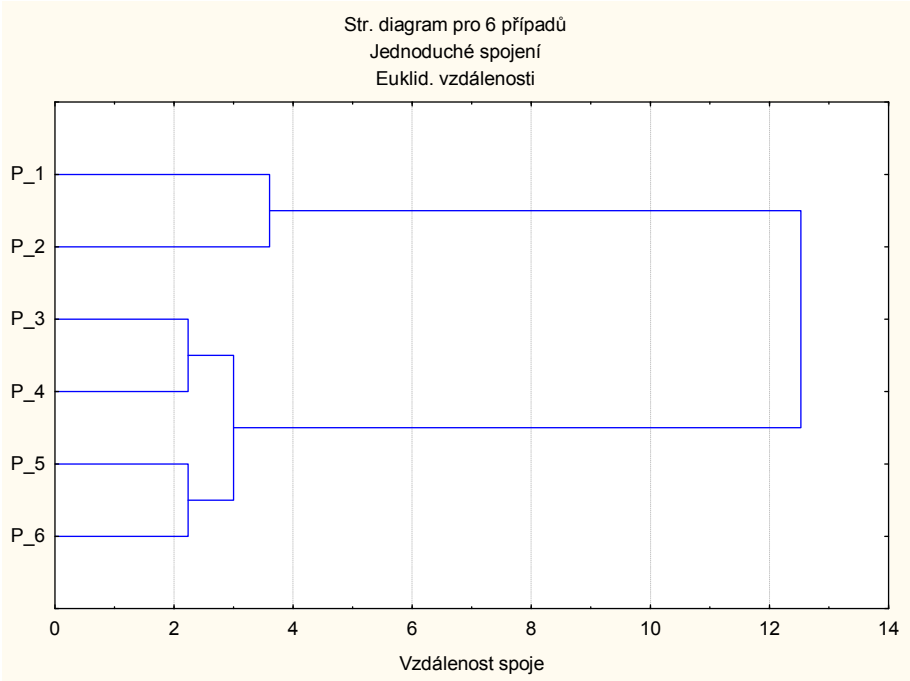
Hierarchické shlukování

Při aplikacích shlukové analýzy se nejčastěji používá aglomerativní hierarchická procedura. Její princip spočívá v postupném slučování objektů, a to nejprve nejbližších a v dalších krocích pak stále vzdálenějších.

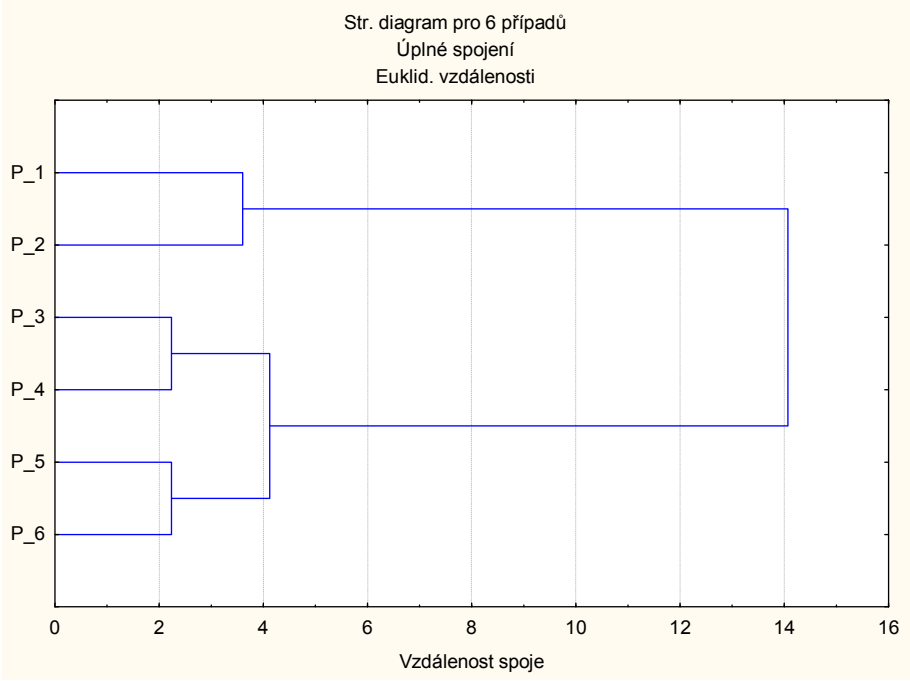
Algoritmus:
1. krok: Každý objekt považujeme za samostatný shluk.
2. krok: Najdeme dva shluky, jejichž vzdálenost je minimální.
3. krok: Tyto dva shluky spojíme v nový, větší shluk a přepočítáme matici vzdáleností. Její řád se sníží o 1. Vrátíme se na 2. krok.
Funkce algoritmu končí, až jsou všechny objekty spojeny do jediného shluku.

Vzdálenost mezi shluky se počítá různými způsoby. Uvedeme tři z nich.
a) Metoda nejbližšího souseda: Vzdálenost mezi dvěma shluky je minimem ze všech vzdáleností mezi jejich objekty.
b) Metoda nejvzdálenějšího souseda: Vzdálenost mezi dvěma shluky je maximem ze všech vzdáleností mezi jejich objekty.
c) Metoda průměrné vazby: Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.
Výsledky aglomerativní hierarchické procedury se zpravidla znázorňují pomocí dendrogramu. Je to graficky znázorněná posloupnost dvojic $\{(\nu_1, S^{(1)}), \ldots, (\nu_n, S^{(n)})\}$, kde $\{\nu_i\}_{i=1}^{n}$ je neklesající posloupnost úrovní spojování a $S^{(i)}$ je roztřídění objektů odpovídající úrovni ν_i, i = 1, ..., n.

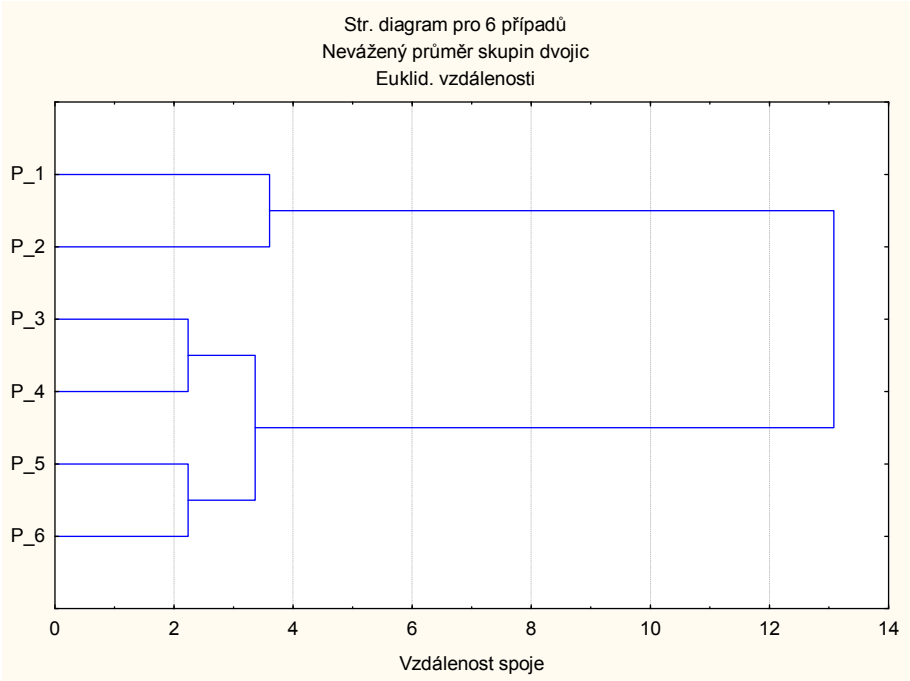
Dendrogram pro metodu nejbližšího souseda:
Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné X1 – X4 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky), pravidlo slučování ponecháme Jednoduché spojení, míru vzdálenosti ponecháme Euklidovské vzd. – OK – Horizontální graf hierarch. stromu



Dendrogram pro metodu nejvzdálenějšího souseda:
Na záložce Detaily vybereme pravidlo slučování Úplné spojení



Dendrogram pro metodu úplné vazby: Na záložce Detaily vybereme pravidlo slučování Nevážený průměr skupin dvojic.



Vidíme, že výsledky všech tří metod jsou velmi podobné a odpovídají rozmístění objektů (žáků) na ploše prvních dvou hlavních komponent.