

# Vícerozměrné analýzy a jejich využití

## Obsah

<b><u>1</u></b>	<b><u>VÍCEROZMĚRNÉ ANALÝZY A JEJICH VYUŽITÍ</u></b>	<b>2</b>
<b><u>1.1</u></b>	<b><u>CO JE VÍCEROZMĚRNÁ ANALÝZA?</u></b>	<b>2</b>
<b><u>1.2</u></b>	<b><u>PROČ VÍCEROZMĚRNÁ ANALÝZA?</u></b>	<b>3</b>
<b><u>1.3</u></b>	<b><u>PRINCIPY VÍCEROZMĚRNÝCH ANALÝZ</u></b>	<b>3</b>
<b><u>1.4</u></b>	<b><u>DATOVÉ PODKLADY</u></b>	<b>3</b>
<b><u>2</u></b>	<b><u>ZÁKLADNÍ TYPY VÍCEROZMĚRNÝCH ANALÝZ</u></b>	<b>8</b>
<b><u>3</u></b>	<b><u>KORESPONDENČNÍ ANALÝZA (CORRESPONDENCE ANALYSIS)</u></b>	<b>9</b>
<b><u>4</u></b>	<b><u>ANALÝZA HLAVNÍCH KOMPONENT A FAKTOROVÁ ANALÝZA</u></b>	<b>14</b>
<b><u>4.1</u></b>	<b><u>ANALÝZA HLAVNÍCH KOMPONENT (PRINCIPAL COMPONENT ANALYSIS PCA)</u></b>	<b>14</b>
<b><u>4.2</u></b>	<b><u>FAKTOROVÁ ANALÝZA (FACTOR ANALYSIS)</u></b>	<b>17</b>
<b><u>5</u></b>	<b><u>DISKRIMINAČNÍ ANALÝZA (CANONICAL VARIATE ANALYSIS)</u></b>	<b>20</b>
<b><u>6</u></b>	<b><u>SHLUKOVÁ ANALÝZA</u></b>	<b>23</b>
<b><u>6.1</u></b>	<b><u>HIERARCHICKÉ SHLUKOVÁNÍ</u></b>	<b>23</b>
<b><u>6.2</u></b>	<b><u>NEHIERARCHICKÉ SHLUKOVÁNÍ</u></b>	<b>26</b>

# 1 Vícerozměrné analýzy a jejich využití

Vícerozměrné analýzy představují velice užitečný nástroj pro uchopení, zjednodušení a vizualizaci velmi složitých dat.

Na druhou stranu mohou v případě nesprávného užití vést k zavádějícím výsledkům, jejichž chybnost nemusí být ovšem na první pohled zřejmá, protože je skryta za složitou strukturou dat a komplikovaností výpočtu.

Cílem tohoto výukového dokumentu je představit ve stručné a přehledné formě základy využití různých typů vícerozměrných analýz včetně jejich interpretace a potenciálně slabých míst.

## 1.1 Co je vícerozměrná analýza?

Veškerý svět kolem nás je vícerozměrný, kromě vnímání třírozměrného tvaru můžeme každý objekt popsat celou řadou dalších charakteristik, jako je třeba barva, hmotnost, chuť atd. atd. Přes tuto skutečnost, kterou vnímáme každý den je pro nás ovšem problémem představit si tuto skutečnost popsanou ve formě datové tabulky a nebo ji dokonce nějakým způsobem popsat jinému člověku – nastává zde tedy místo pro speciální typ analýzy, tedy vícerozměrnou analýzu.

**Obrázek** Přímé vnímání vícerozměrné reality vs. naše schopnost jejího záznamu.



ID	Sex	Rok narození	Výška	Hmotnost	Vzdělání
1	muž	1943	183	90	VŠ
2	žena	1953	162	61	Z
3	muž	1955	172	86	Z
4	žena	1956	164	65	SŠ s mat.
5	muž	1927	170	85	VŠ
6	žena	1928	158	71	SŠ s mat.
7	žena	1983	182	68	SŠ s mat.
8	muž	1977	195	85	SŠ s mat.
9	žena	1976	173	73	VŠ
10	muž	1960	180	103	SŠ s mat.
11	muž	1956	189	96	SŠ s mat.
12	žena	1955	179	74	VŠ
13	žena	1976	164	57	SŠ
14	muž	1966	172	63	Z
15	muž	1938	172	96	SŠ
16	muž	1955	178	110	SŠ
17	muž	1979	180	63	VŠ
18	muž	1951	185	126	VŠ
19	muž	1955	172	78	SŠ
20	žena	1974	169	98	SŠ

## 1.2 Proč vícerozměrná analýza?

Ačkoliv klasická statistika zná řadu způsobů popisu jednotlivých měřených nebo pozorovaných parametrů, je pro nás v případě hodnocení velkého množství parametrů velmi obtížné si tyto výstupy složit v mozku do jednoduššího obrazu vedoucího k pochopení podstaty. Právě vícerozměrná analýza dat je nástrojem sloužícím k usnadnění tohoto procesu a její přínos lze shrnout následovně:

- Nalezení smysluplných pohledů na data popsaná velkým množstvím parametrů
- Nalezení a popsání skrytých vazeb mezi parametry a tak zjednodušení jejich struktury
- Jednoduchá vizualizace dat, kdy v jediném grafu se skrývá informace např. z 20 proměnných
- Umožnění a/nebo zjednodušení a interpretace dat na základě jejich zjednodušení a vizualizace vícerozměrnou analýzou

## 1.3 Principy vícerozměrných analýz

Ačkoliv je v případě vícerozměrných analýz používána celá řada matematických postupů, jedno mají všechny tyto analýzy společné – hledají, které naměřené parametry nebo objekty spolu nějakým způsobem souvisí a které je tedy možné jako podobné sloučit a tak snížit složitost naměřených dat.

Extrémním případem by byla např. situace, kdy by výška člověka byla měřena zároveň v centimetrech a v milimetrech (=2 parametry), oba parametry představují ve skutečnosti pouze parametr jediný. V praxi samozřejmě tuto absurdní situaci nenajdeme, nicméně různé vazby mezi parametry existují vždy a umožňují nám tak zjednodušit získaná data.

Pokud by mezi naměřenými parametry žádná vazba neexistovala nebo byla velmi slabá nemá smysl vícerozměrnou analýzu vůbec počítat !!!

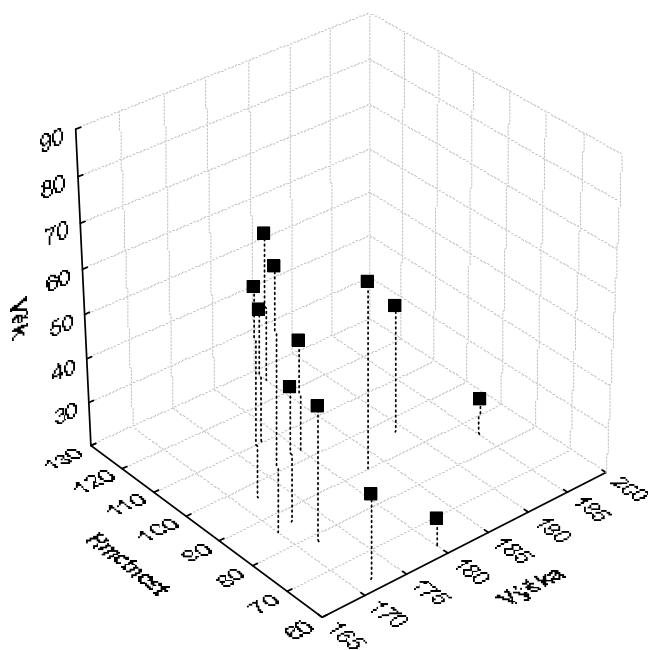
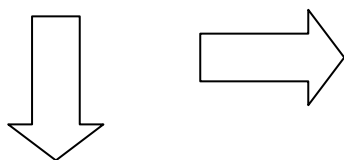
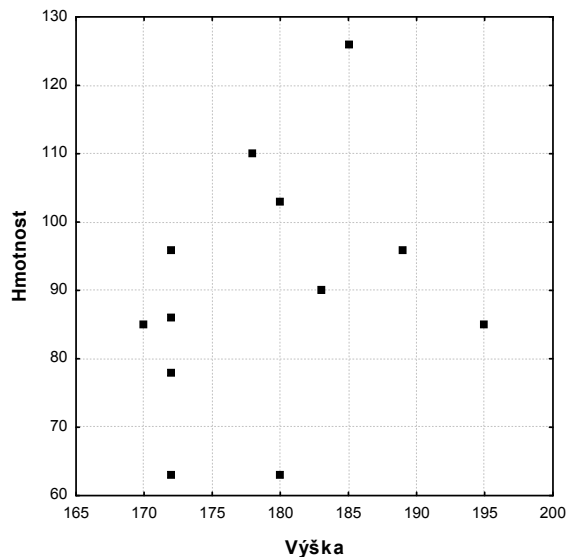
## 1.4 Datové podklady

Podkladem každé vícerozměrné analýzy je vždy tabulka obsahující v řádcích jednotlivé měřené objekty (respondenty) a ve sloupcích parametry měřené na těchto objektech. Každý parametr představuje jeden rozměr (dimenzi) objektu. Ukázka datové tabulky je níže.

Číslo přihlášky	Název programu	Test fyzika	Test chemie	Test biologie	Součet (max. 180)	Body pro přijetí	Přijat
1608025	Všeob. lékařství	52	44	28	124	104	ano
1624822	Všeob. lékařství	48	46	52	146	104	ano
1607697	Všeob. lékařství	34	36	42	112	104	ano
1685031	Všeob. lékařství	32	24	46	102	104	ne
1681248	Všeob. lékařství	44	36	42	122	104	ano
1356491	Všeob. lékařství	18	36	46	100	104	ne
1355689	Všeob. lékařství	40	30	46	116	104	ano
1384362	Všeob. lékařství	38	34	32	104	104	ano
1503013	Všeob. lékařství	34	44	52	130	104	ano
1683334	Všeob. lékařství	22	42	40	104	104	ano
1478234	Všeob. lékařství	48	48	38	134	104	ano
1569543	Všeob. lékařství	28	34	46	108	104	ano
1642344	Všeob. lékařství	44	42	54	140	104	ano
1660003	Všeob. lékařství	44	30	50	124	104	ano
1654911	Všeob. lékařství	30	32	40	102	104	ne

Hodnoty parametrů definují pozici objektů ve vícerozměrném prostoru, v nejjednodušším případě jde dvourozměrný prostor, který lze zobrazit v XY grafu. Problémem samozřejmě je, že pro více než tři dimenze nejsme schopni vytvořit odpovídající zobrazení a potřebujeme vícerozměrná analyzu, která příslušná data zjednoduší a umožní zobrazit.

ID	Sex	Věk	Výška	Hmotnost
1	muž	62	183	90
3	muž	50	172	86
5	muž	78	170	85
8	muž	28	195	85
10	muž	45	180	103
11	muž	49	189	96
14	muž	39	172	63
15	muž	67	172	96
16	muž	50	178	110
17	muž	26	180	63
18	muž	54	185	126
19	muž	50	172	78

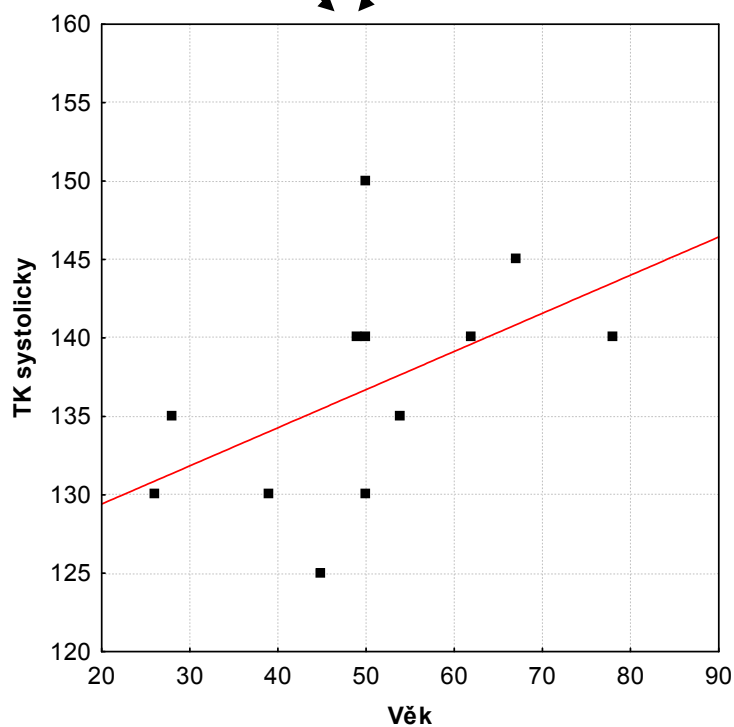


Protože vícerozměrná analýza zjednodušuje naměřená data na základě analýzy jejich vzájemných vazeb je nezbytným krokem v další analýze vytvoření měřítka vazby:

- Naměřených parametrů
- Naměřených objektů
- Skupin objektů k parametrům

Jako měřítko vazby parametrů je nejčastěji využívána korelace a kovariance, vzniklá tzv. asociační matice parametrů, která je podkladem pro faktorovou analýzu a analýzu hlavních komponent.

ID	Věk	Výška	Hmotnost	Krvní tlak systolický	Krvní tlak diastolický
1	62	183	90	140	60
3	50	172	86	130	85
5	78	170	85	140	80
8	28	195	85	135	90
10	45	180	103	125	75
11	49	189	96	140	90
14	39	172	63	130	85
15	67	172	96	145	95
16	50	178	110	140	80
17	26	180	63	130	80
18	54	185	126	135	90
19	50	172	78	150	95

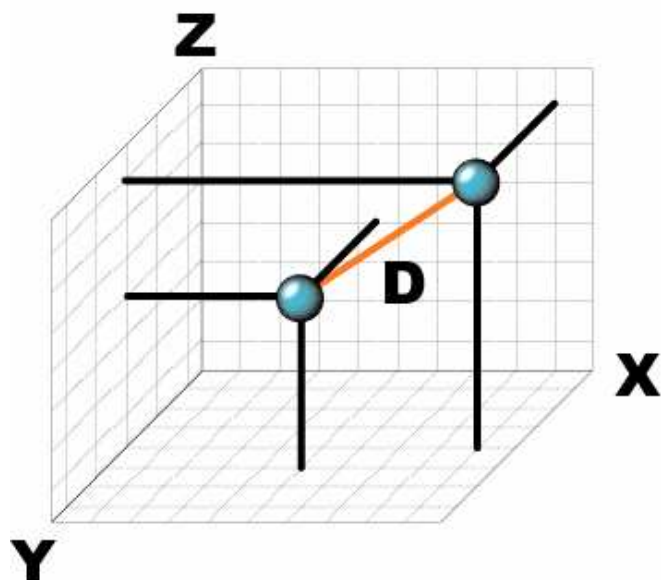


**Tabulka.** Ukázka asociační matice (matice korelací mezi parametry):

	Věk	Výška	Hmotnost	TK systolický	TK diastolický
Věk	1.00	-0.47	0.37	0.50	-0.12
Výška	-0.47	1.00	0.30	-0.16	-0.05
Hmotnost	0.37	0.30	1.00	0.10	0.03
TK systolický	0.50	-0.16	0.10	1.00	0.32
TK diastolický	-0.12	-0.05	0.03	0.32	1.00

Pro zjištění vazby objektů (nejčastěji za účelem nalezení jejich skupin) jsou používány tzv. vícerozměrné vzdálenosti objektů. Jde o vzájemnou vzdálenost objektů ve vícerozměrném prostoru; tabulka všech vzdáleností mezi měřenými objekty vytváří tzv. matici vzdáleností objektů. Matice vzdáleností objektů je nejčastěji podkladem pro hlučkovou analýzu.

**Obrázek.** Vzdálenost objektů v třírozměrném prostoru ( $D$  = Euklidovská vzdálenost).



**Tabulka.** Jednotlivá přestavení a jejich počet v kinech Palace Cinemas ze dne 16.3.2006..

	Slovanský dům (Praha)	Nový Smíchov (Praha)	Park Hostivař (Praha)	Letňany (Praha)	Olympia Centrum (Brno)	Velký Špalíček (Brno)
Capote	0	2	0	0	0	0
Casanova	4	3	3	2	3	3
Faktótum	0	0	0	0	0	1
Zmražená pojistka	5	6	4	3	2	4
Bambi 2	2	2	0	1	2	0
Dobrou noc a hodně štěstí	3	0	0	0	0	0
Erasmus 2	2	0	0	0	0	0
Fimfárum 2	1	3	2	0	0	0
Gejša	3	2	0	2	2	1
Hostel	1	2	1	1	2	0
Jak se krotí krokodýly	0	1	2	1	2	0
Karoolka	3	4	1	1	2	2
Letopisy Narnie	0	2	0	0	0	0
Pýcha a předsudek	1	3	0	0	0	1
Raftáci	9	15	6	6	6	7
Růžový panter	4	6	3	3	3	3
Underworld: Evolution	1	3	1	2	2	1
Univerzální uklízečka	1	0	0	0	0	0
Válka světů	1	0	0	0	0	0
Walk the Line	1	0	0	0	0	0
Zkrocená hora	3	4	2	2	1	3

**Tabulka.** Tabulka vzdáleností kin Palace Cinemas podle hraných představení (ukázka vícerozměrných vzdáleností; použita byla Euklidovská vzdálenost).

	<b>Slovanský dům (Praha)</b>	<b>Nový Smíchov (Praha)</b>	<b>Park Hostivař</b>	<b>Letňany (Praha)</b>	<b>Olympia Centrum (Brno)</b>	<b>Velký Špalíček (Brno)</b>
<b>Slovanský dům (Praha)</b>	0	9.110	7.211	6.708	7.071	5.916
<b>Nový Smíchov (Praha)</b>	9.110	0	11.790	11.916	12.124	10.863
<b>Park Hostivař</b>	7.211	11.790	0	3.606	4.472	3.873
<b>Letňany (Praha)</b>	6.708	11.916	3.606	0	2.646	3.464
<b>Olympia Centrum (Brno)</b>	7.071	12.124	4.472	2.646	0	5.000
<b>Velký Špalíček (Brno)</b>	5.916	10.863	3.873	3.464	5.000	0

Z tabulky vidíme, že podle nabídky filmů jsou si nejbližší kina Letňany (v Praze) a Olympia (v Brně).

Konečně třetí možností je vyhodnocení vazby mezi skupinami objektů a parametry. Pro hodnocení skupiny respondentů je spočítána zástupná statistika parametrů (např. počet hodnot (zde vzniká tzv. kontingenční tabulka), průměr, suma nebo jiná smysluplná statistika). Vzniklá tabulka je nejčastěji vstupem pro korespondenční analýzu.

**Tabulka.** Ukázka kontingenční tabulky. Počet studentů přihlášených na programy lékařské fakulty z různých krajů ČR.

<b>Studijní program Kraj</b>	<b>Ošetrovatelství</b>	<b>Specializace ve zdravotnictví</b>	<b>Všeobecné lékařství</b>	<b>Zubní lékařství</b>	<b>Celkový součet</b>
Jihočeský	30	45	101	31	<b>207</b>
Jihomoravský	47	115	170	53	<b>385</b>
Karlovarský	11	27	49	15	<b>102</b>
Kralovehradecký	23	42	89	22	<b>176</b>
Liberecký	21	35	68	12	<b>136</b>
Moravskoslezský	60	123	215	59	<b>457</b>
Olomoucký	25	54	96	30	<b>205</b>
Pardubický	14	57	85	18	<b>174</b>
Plzeňský	25	46	81	22	<b>174</b>
Praha	53	118	186	49	<b>406</b>
Středočeský	44	104	165	58	<b>371</b>
Ústecký	35	78	127	40	<b>280</b>
Vysočina	21	43	85	20	<b>169</b>
Zlínský	31	48	98	31	<b>208</b>
<b>Celkový součet</b>	<b>440</b>	<b>935</b>	<b>1615</b>	<b>460</b>	<b>3450</b>

## 2 Základní typy vícerozměrných analýz

Jak již bylo naznačeno existuje celá řada vícerozměrných metod s rozdílným cílem analýzy a aplikovatelné na různé typy dat:

- Shluková analýza hledá v datech skupiny obdobných objektů nebo parametrů, lze ji dále rozdělit:
  - Hierarchické aglomerativní metody vytváří strom objektů (parametrů) podle jejich podobnosti, strom vzniká postupným narůstáním
  - Nehierarchické metody rozdělují objekty do zadaného počtu shluků tak aby si objekty v jednotlivých shlucích byly navzájem co nejpodobnější
  - Hierarchické divizivní metody vytváří strom objektů podle jejich podobnosti, strom vzniká postupným rozdělováním dat
- Ordinační metody se snaží data zjednodušit a výsledek analýzy např. 20 parametrů prezentovat ideálně v dvou rozměrech XY grafu, který by zároveň měl nést většinu informace z původních dat.
  - Korespondenční analýza
  - Faktorová analýza a analýza hlavních komponent
- Klasifikace a modelování jsou využívány pro predikci charakteristik respondentů nebo jejich zařazení do známých skupin, z této skupiny je v dokumentu zmíněna detailněji pouze diskriminační analýza
  - Diskriminační analýza
  - Rozhodovací stromy a lesy
  - Neuronové sítě
  - Atd.



### 3 Korespondenční analýza (Correspondence analysis)

Korespondenční analýza je nástrojem pro analýzu vztahů mezi řádky a sloupci kontingenčních tabulek.

Kontingenční tabulky jsou základním nástrojem pro zkoumání vztahů mezi dvěma proměnnými. Kontingenční tabulkou rozumíme frekvenční tabulku (klasicky dvouúrovňovou), která nám zaznamenává kumulativní četnosti dvou nominálních (kategoriálních) proměnných. Každý sloupec a každý řádek tabulky pak reprezentuje jednu kategorii dané proměnné.

**Příklad:** Uvažujme  $n$  druhů společností a  $p$  měst, ve kterých tyto společnosti sídlí. Chceme zjistit, jestli určitý druh společnosti preferuje nějaké město (nebo jinak, jestli existuje nějaký location index, který se vztahuje k danému typu společnosti).

<b>X</b>	4	0	2	←	finance
	0	1	1	←	energie
	1	1	4	←	hi-tech
	↑	Frankfurt			
		↑	Berlín		
			↑	Mnichov	

Hodnota  $x_{ij}$  v tabulce  $X$  o rozměrech  $n \times p$  označuje počet pozorování ve vzorku, které současně náleží do  $i$ -té “řádkové” kategorie a  $j$ -té “sloupcové” kategorie pro  $i = 1, \dots, n$  a  $j = 1, \dots, p$ .

Základní myšlenkou metody korespondenční analýzy je vytvořit či odvodit indexy (pokud možno “jednoduché”), které budou nějakým způsobem označovat (kvantifikovat) vztahy mezi řádkovými a sloupcovými kategoriemi. Z těchto indexů pak budeme schopni odvodit, která sloupcová kategorie má větší či menší váhu v daném řádku a naopak.

Metoda je zaměřena na kategoriální proměnné, proto je třeba v případě spojitých proměnných nejprve provést jejich diskretizaci – rozdělení do kategorií podle nějakých intervalů.

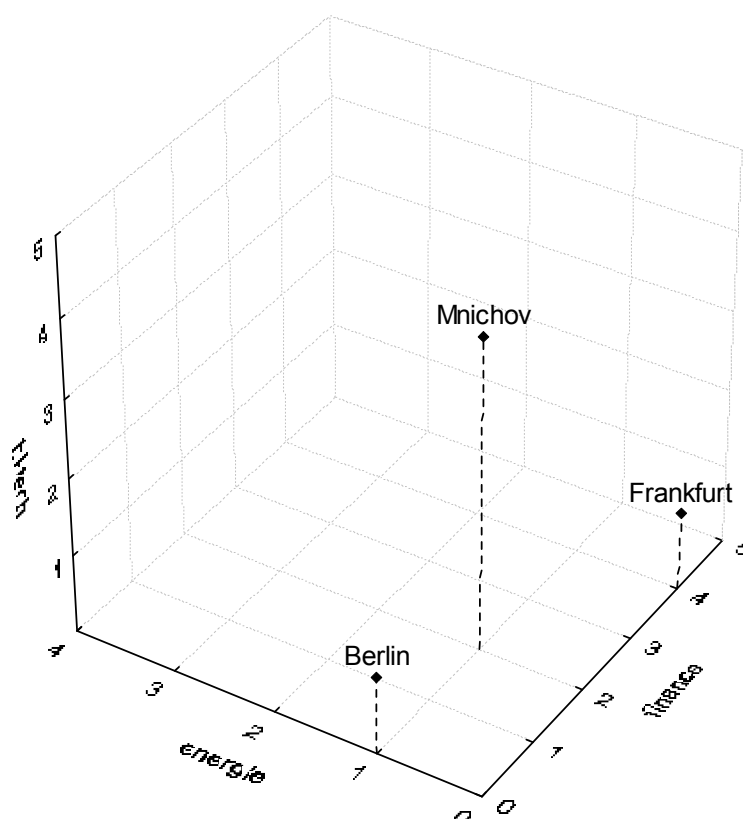
Korespondenční analýza se také vztahuje k otázce snížení dimezionality dat podobně jako např. analýza hlavních komponentů (principal component analysis: PCA) a ke snaze o dekompozici tabulky na faktory.

V případě korespondenční analýzy se snažíme získat indexy v klesajícím stupni důležitosti tak, aby se hlavní informace obsažená v tabulce dala shrnout do prostoru s co možná nejmenší dimenzí. Např. pokud jsou použity pouze dva faktory (indexy), můžeme výsledky zobrazit pomocí dvourozměrného grafu.

Grafické znázornění vztahů, které obdržíme z korespondenční analýzy, je založeno na myšlence reprezentovat všechny sloupce a řádky a interpretovat relativní pozice bodů jako váhy příslušné danému sloupci a řádku. Systém indexů, který si pomocí této metody odvodíme, nám tedy bude poskytovat souřadnice každého sloupce a řádku. Tyto souřadnice zakreslíme do grafu, ze kterého můžeme poznat, které sloupcové kategorie jsou více důležité v řádkových kategoriích a naopak.

Výše uvedený jednoduchý příklad umíme zobrazit v třírozměrném grafu.

**Obrázek.** Znázornění tří měst podle druhu společností v třírozměrném prostoru.



Výsledkem korespondenční analýzy je graf, tzv. ordinační diagram, kde platí, že:

- blízkost dvou řádků (sloupců) značí podobný „profil“ v těchto dvou řádcích (pojmem profil označujeme distribuci podmíněné četnosti)
- pokud jsou od sebe řádky či sloupce vzdáleny, jejich „profil“ je značně odlišný
- blízkost určitého řádku a určitého sloupce znamená, že tento řádek má důležitou váhu v daném sloupci
- pokud jsou od sebe určitý řádek a sloupec daleko, nejsou v daném sloupci téměř žádná pozorování příslušející danému řádku

#### **Příklad:**

Jako příklad použití této metody můžeme uvést analýzu dat, které byly získány při analýze prodeje novin v Belgii.

Obyvatelstvo bylo rozděleno do 10 skupin podle oblasti, ve které daná osoba žila (*Antwerp, Western Flanders, Eastern Flanders, Hainant, Liege, Limbourg, Luxembourg, Flemish-Brabant, Wallon-Brabant, Brussels*). Občané byli dotazováni, který druh novin pravidelně čtou. Existovalo 15 možných druhů odpovědí (novin), které byly rozděleny do tří kategorií (vlámské – předpona *v*, francouzské – předpona *f*, oba jazyky – předpona *b*). V průzkumu byla zkoumána mimo jiné závislost mezi bydlištěm a druhem novin, které lidé z dané oblasti čtou. V průzkumu byl získán následující soubor dat.

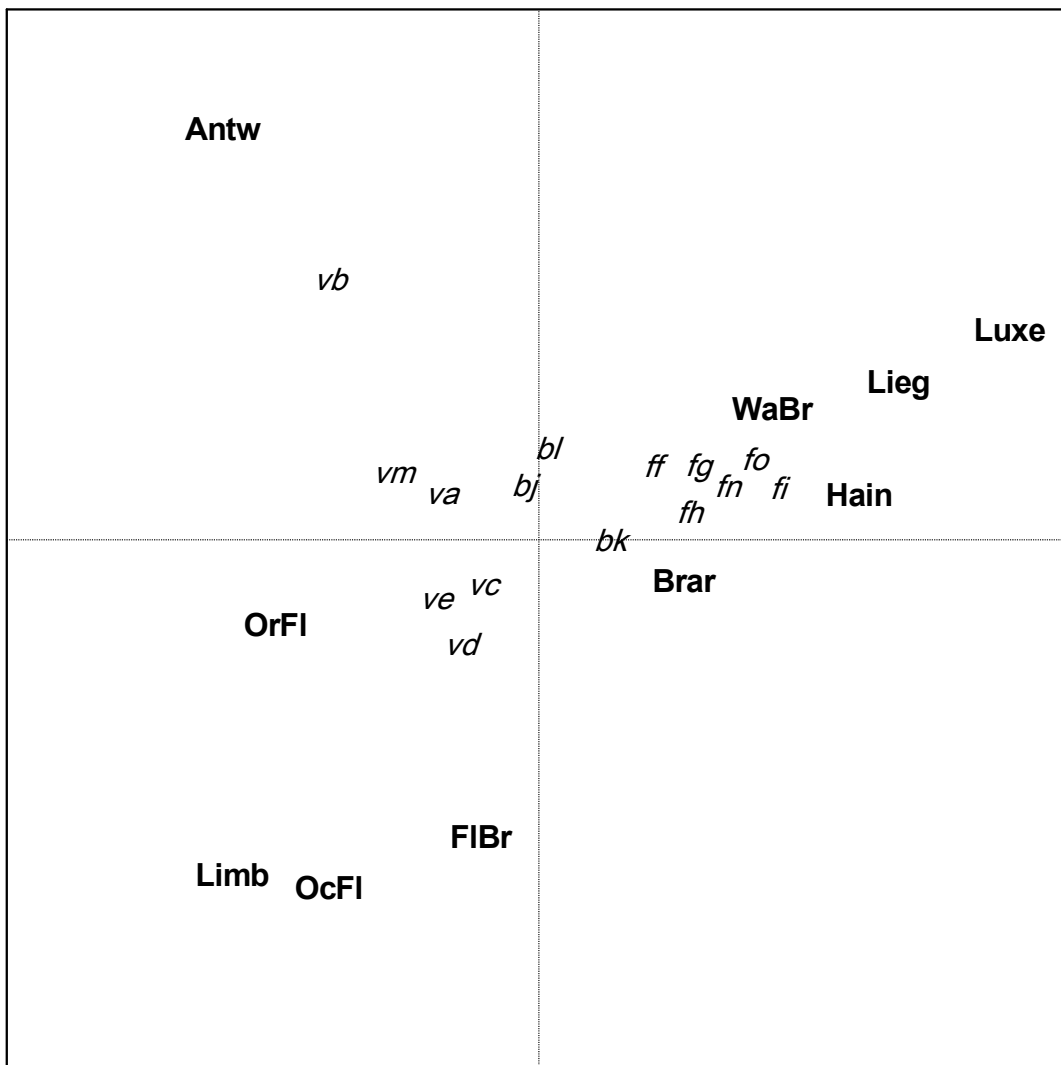
X1: WaBr	Walloon Brabant
X2: Brar	Brussels area
X3: Antw	Antwerp
X4: FIBr	Flemish Brabant
X5: OcFl	Occidental Flanders
X6: OrFl	Oriental Flanders
X7: Hain	Hainaut
X8: Lieg	Liege
X9: Limb	Limburg
X10: Luxe	Luxembourg

	WaBr	Brar	Antw	FIBr	OcFl	OrFl	Hain	Lieg	Limb	Luxe
va	1.8	7.8	9.1	3.0	4.3	3.9	0.1	0.3	3.3	0.0
vb	0.1	3.4	17.8	1.0	0.7	4.1	0.0	0.0	0.2	0.0
vc	0.1	9.4	4.6	7.7	4.4	5.8	1.6	0.1	1.4	0.0
vd	0.5	15.6	6.1	12.0	10.5	10.2	0.7	0.3	5.4	0.0
ve	0.1	5.2	3.3	4.8	1.6	1.4	0.1	0.0	3.5	0.0
ff	5.6	13.7	3.1	2.4	0.5	1.7	1.9	2.3	0.2	0.2
fg	4.1	16.5	1.9	1.0	1.0	0.9	2.4	3.2	0.1	0.3
fh	8.3	29.5	1.8	7.3	0.8	0.4	5.1	3.2	0.2	0.3
fi	0.9	7.8	0.2	2.6	0.1	0.1	5.6	3.8	0.1	0.8
bj	6.1	18.2	10.8	4.1	4.5	5.3	2.0	2.6	3.4	0.2
bk	8.3	35.4	6.2	11.0	5.0	6.1	5.5	3.3	1.5	0.3
bl	4.4	9.9	6.7	3.4	1.1	3.9	2.1	1.5	2.1	0.0
vm	0.3	11.6	14.2	4.7	5.1	7.9	0.3	0.5	3.0	0.0
fn	5.1	21.0	1.3	3.4	0.2	0.2	2.3	4.4	0.0	0.4
fo	2.2	9.8	0.1	0.3	0.0	0.7	2.3	3.0	0.3	1.0

Prvním krokem analýzy je výpočet tzv. vlastních čísel matice (eigenvalues) a určení části rozptylu vysvětleného jednotlivými ordinačními osami.

Spravidla několik málo prvních os vysvětluje většinu celkového rozptylu dat. V našem případě první dvě vlastní čísla přinášejí přibližně 81% celkového rozptylu, což je poměrně dost, a tedy že si vystačíme s dvoudimenzionální reprezentací.

**Obrázek.** Výsledek korespondenční analýzy. V grafu jsou zobrazeny jak řádků (noviny) tak i sloupce (oblasti) původní tabulky. V prezentovaném znázornění je patrná poměrně silná závislost mezi oblastí, ve které dotyčná osoba žije, a novinami, které pravidelně čte. V Belgii je tento fakt daný především jazykovými dispozicemi (Valoni a Vlámové).



**Vstup korespondenční analýzy:**  
Kontingenční tabulka

**Výstup korespondenční analýzy:**  
Ordinační diagram  
Skóre (souřadnice) řádků a sloupců na ordinačních osách

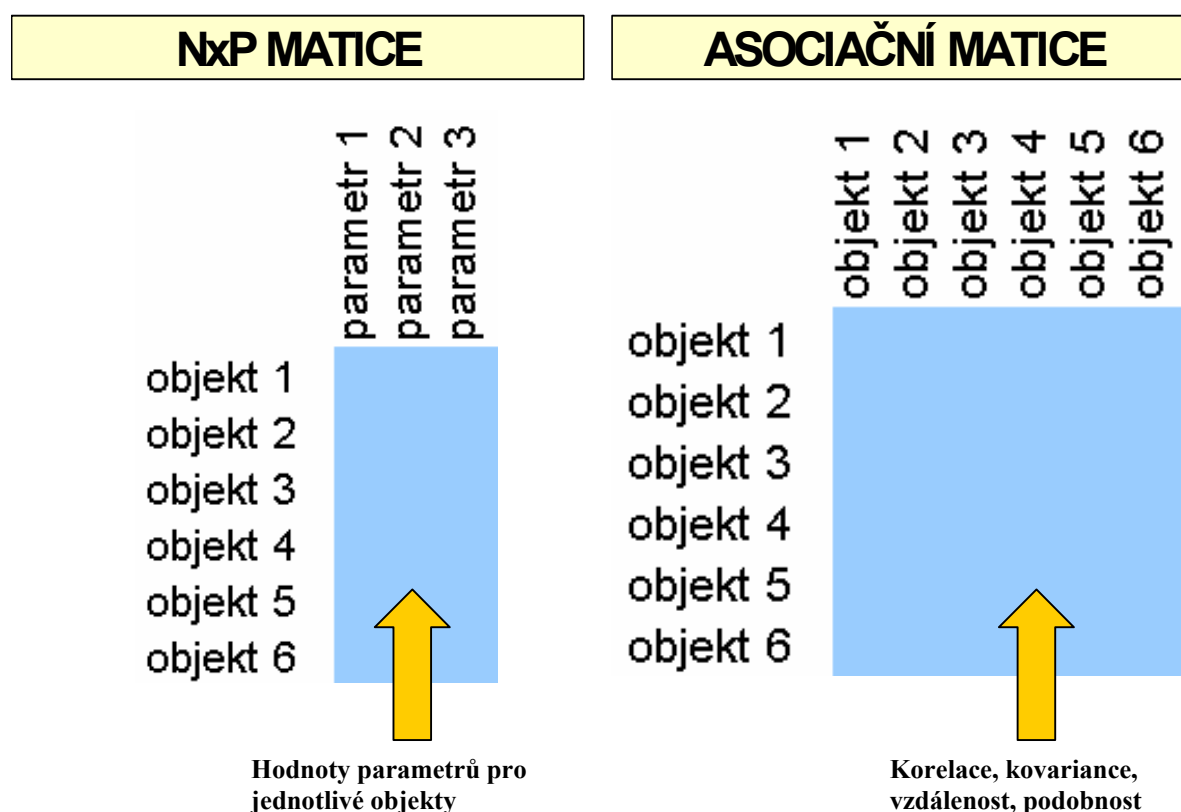
**Při použití korespondenční analýzy je nutno pamatovat na níže uvedená omezení:**

- velký počet malých skupin respondentů může způsobit problematickou interpretaci výsledků a nestabilitu výpočtu

## 4 Analýza hlavních komponent a faktorová analýza

### 4.1 Analýza hlavních komponent (Principal component analysis PCA)

Předpokládejme že naše mnohorozměrná pozorování se skládají z  $p$  proměnných měřených na  $n$  objektech a vytvářejí tak datovou matici o rozměru  $n \times p$ . Z této datové matice získáme kovarianční i korelační matici rozměru  $p \times p$ . Analýza hlavních komponent je definována pro kovarianční a korelační matici. Korelace jsou kovariance standardizovaných proměnných. Hlavní komponenty, které získáme z korelační matice neodpovídají komponentům získaným z kovarianční matice. Vzdálenosti mezi objekty v těchto dvou případech nejsou stejné.



Jsou-li jednotlivé proměnné vyjádřeny ve zcela rozdílných jednotkách měření, nemají lineární kombinace původních hodnot – kterými hlavní komponenty jsou – velký význam, a je proto vhodnější založit PCA na normovaných proměnných, tj. na korelační matici. Jsou-li jednotlivé proměnné vyjádřeny v příbuzných jednotkách, je ze statistického hlediska vhodnější použít kovarianční matici.

PCA vytvoří hlavní komponenty, které jsou lineární kombinací původních proměnných. Zároveň platí, že všechny komponenty jsou ortogonální (tj. nezávislý, kolmý na sebe). Tento proces je postupný; nejdříve se vytvoří první hlavní komponenta, tj. taková proměnná, která vysvětluje největší část variability původních dat. Po nalezení první hlavní komponenty je nalezena druhá hlavní komponenta – lineární kombinace, která vysvětluje největší část zbytkové variability původních dat a zároveň je nezávislá (ortogonální) na první hlavní komponente. Proces pokračuje hledáním dalších komponent, až kým dosáhneme stavu, kdy nalezené hlavní komponenty vysvětlují více než určité procento variability původní datové matice. Po sobě jdoucí komponenty vyčerpávají maximum zbývajících rozptylu u souboru

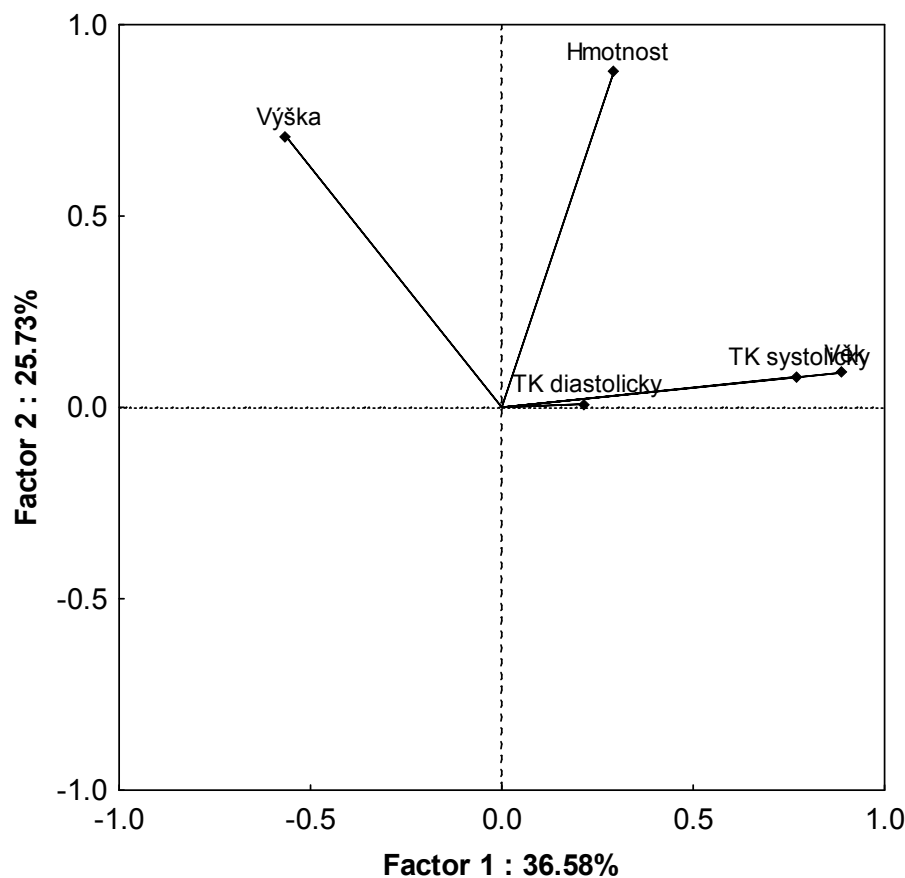
proměnných a jsou vzájemně nezávislé. Algebraicky PCA hledá vlastní hodnoty a vlastní vektory matice.

Ve většině případů pracujeme s korelační matricí, kde rozptyl (variabilita) všech proměnných je rovna 1.0. Pak je celková variabilita datové matice rovná počtu proměnných. Například, když máme 10 proměnných, každou s rozptylem 1, pak je celková variabilita, která může být vysvětlená, rovná 10. Jedním z výsledků analýzy hlavních komponent (PCA) je také tzv. vlastní hodnota (eigenvalue) hlavní komponenty, která vyjadřuje variabilitu vysvětlenou hlavní komponentou. Jedním z nejčastěji používaných kritérií k volbě počtu hlavních komponent, které zachováme je tzv. Kaiserovo kritérium, které navrhuje ponechat pouze komponenty s vlastní hodnotou větší než 1. V takovém případě totiž hlavní komponenta vysvětluje alespoň takovou část variability, jako jedna původní proměnná.

### Geometrický význam hlavních komponent

Výběr rozsahu  $n$  z  $p$ -rozměrného rozdělení si můžeme geometricky představit jako „shluk“  $n$  bodů v  $p$ -rozměrném euklidovském prostoru, jehož osy odpovídají jednotlivým proměnným  $X_1, X_2, \dots, X_p$ . Bez snížení obecnosti můžeme předpokládat, že za střed souřadnicového systému byl vzat bod se souřadnicemi danými výběrovými průměry proměnných. Nalezení lineárních kombinací proměnných (PCA) geometricky odpovídá rotaci původní souřadnicové soustavy provedené tak, že nové osy procházejí směry maximálního rozptylu shluku bodů. Tato transformace souřadnicového systému tedy umožňuje zachytit na několika prvních osách maximum informace o prostorové struktuře souboru vícerozměrných pozorování.

**Obrázek.** Výsledek analýzy hlavních komponent (PCA korelací 5 parametrů; data z tabulky na straně 5).



Z grafu je zřejmá silná kladná korelace věku a systolického tlaku krve. To znamená, že s nárůstem hodnot jednoho parametru rostou i hodnoty druhého parametru (s věkem se zvyšuje krevní tlak). Výška respondentů s jejich hmotností koreluje kladně, ovšem ne velmi výrazně.

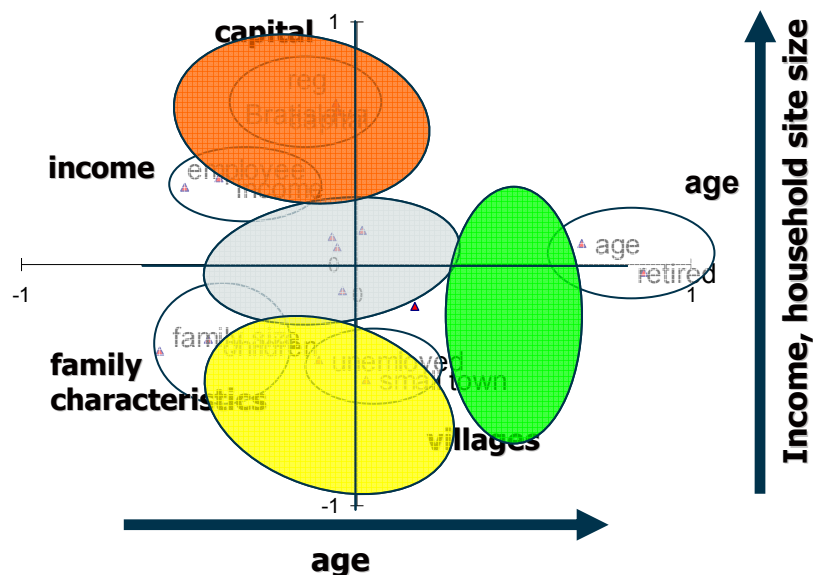
### Interpretace hlavních komponent

PCA transformuje původní proměnné na ortogonální, popř. ortonormální veličiny sumarizující rozptyly původních proměnných. Zda jsou tyto nové veličiny umělými charakteristikami, či zda skutečně odrážejí určité reálné faktory, tj. mají určitý předmětný obsah, je otázkou interpretace, kterou je třeba provádět na základě věcných znalostí zkoumaných proměnných.

Při interpretaci hlavních komponent (podobně jako při interpretaci faktorů ve faktorové analýze) vyhledáváme proměnné, které s jednotlivými komponentami korelují, a snažíme se tyto komponenty interpretovat jako vzájemně nezávislé, zobecněné, v pozadí stojící vlivy, vyvolávající variabilitu a ovlivňující strukturu závislosti proměnných. Při pokusu o interpretaci hlavních komponent je vhodné se omezit především na první komponenty s vysokými vlastními hodnotami. Interpretace komponent s vyššími pořadovými čísly bývá nezdědka obtížná a problematická.

Pro každý objekt lze určit jeho pozici v prostoru definovaném kombinací parametrů charakterizujících tento objekt. V ideálním případě se podél hlavních komponentů vyčlení skupiny objektů, jak je tomu v případě uživatelů mobilních telefonů (obrázek níže). Pomocí hlavních komponent (souřadnic objektů na prvních dvou komponentech) lze definovat několik smysluplných skupin, které lze dále analyzovat jinými metodami.

**Obrázek.** Využití pozice respondentů v prostoru PCA k definici hodnotových skupin na telekomunikačním trhu





## 4.2 Faktorová analýza (Factor analysis)

Faktorová analýza je vícerozměrná statistická metoda, jejíž podstatou je rozbor struktury vzájemných závislostí proměnných na základě předpokladu, že tyto závislosti jsou důsledkem působení určitého menšího počtu v pozadí stojících nezměřitelných veličin. Tyto veličiny jsou označovány jako *společné faktory*. Cílem faktorové analýzy je na základě závislostí pozorovaných proměnných odpovědět na otázku, jaká je struktura společných faktorů stojících za vzájemně korelovanými proměnnými. Faktorová analýza se přitom snaží odvodit povahu společných faktorů tak, aby tyto (hypotetické) veličiny objasňovaly pozorované závislosti co nejjednodušeji a aby počet nalezených faktorů byl co nejmenší.

Faktorová analýza vznikla v oblasti psychologie a byla po dlouhou dobu používána téměř výhradně v tomto oboru.

Faktorovou analýzu lze považovat za rozšíření metody hlavních komponent. Na rozdíl od PCA vychází ze snahy vysvětlit závislosti proměnných. Mezi nedostatky PCA patří zejména fakt, že není invariantní vůči změnám měřítka proměnných. Přístup faktorové analýzy umožňuje odstranit tento nedostatek, trpí však na druhé straně nejednoznačností odhadů faktorových parametrů (problém rotace). Předností faktorové analýzy je větší úspornost a obecnost, je však nutné předpokládat, že pozorování pocházejí z vícerozměrného normálního rozdělení, a specifikovat počet společných faktorů před provedením analýzy.

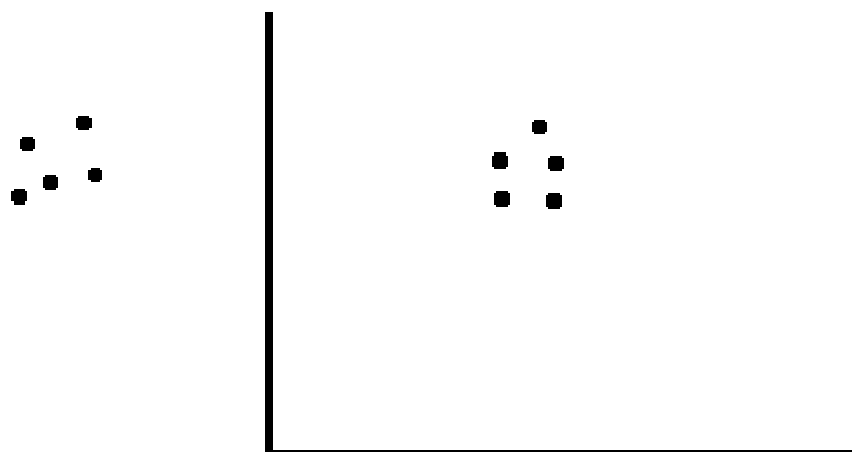
Ve faktorové analýze se vysvětluje vzájemná lineární závislost pozorovaných proměnných  $X_1, X_2, \dots, X_p$  existencí menšího počtu nepozorovatelných veličin  $f_1, f_2, \dots, f_m$  (zvaných společné faktory) a  $p$  dalších zdrojů variability  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$  (zvaných chybové či specifické faktory nebo též rušivé či reziduální složky). Společné faktory vyvolávají korelace mezi proměnnými, zatímco chybové faktory pouze přispívají k rozptylu jednotlivých pozorovaných proměnných. Předmětem zájmu faktorové analýzy jsou především společné faktory.

Faktorová analýza pracuje podobně jako PCA. Rovněž jako PCA pracuje s korelační nebo kovarianční maticí a nalézá první hlavní faktor tak, aby vysvětloval největší část variability datové matice. Další faktory jsou konstruovány tak, aby byly nezávislé. Rozdíl mezi faktorovou analýzou a analýzou hlavních komponent je v dalším kroku analýzy. Tady jsou hlavní faktory rotovány tak, aby co nejjednodušeji popisovaly proměnné, tj. aby byly co nejbližší situovány co nejvíce původním proměnným. To je dosaženo v situacích, kdy hlavní faktory jsou nejbližší skupině silně korelovaných proměnných.

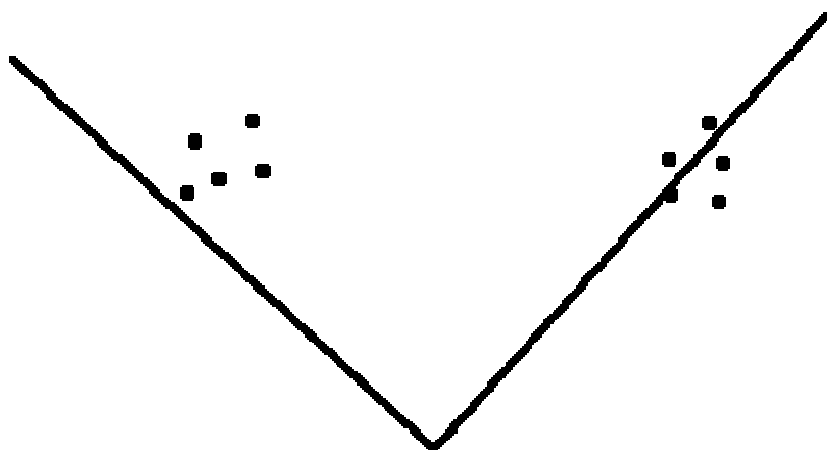
Pro rotaci faktorů existuje několik možností. Ortogonální rotace zachovávají nezávislost faktorů, kým při neortogonální rotaci dochází k tomu, že faktory se stávají do určité míry korelovány. Nejznámější metody ortogonální rotace jsou *varimax* a *quartimax*.

I když první fáze faktorové analýzy probíhá stejně jako PCA, interpretace výsledků je jiná než při PCA, co je způsobeno právě rotací faktorů v druhé fázi analýzy.

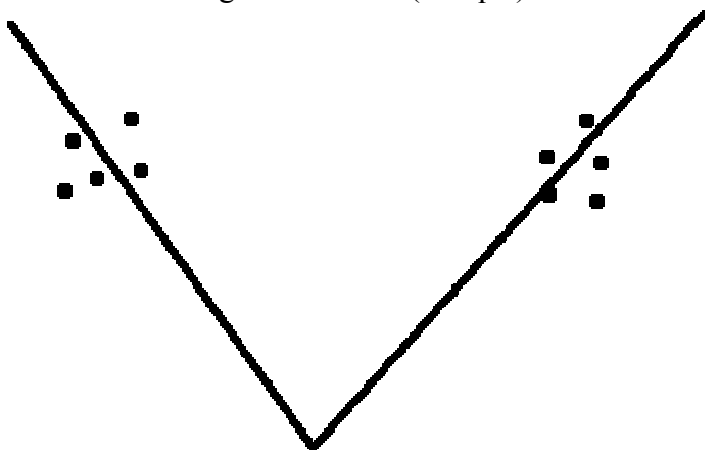
**Obrázek.** Nerotovaný prostor.



**Obrázek.** Ortogonální rotace.



**Obrázek.** Neortogonální rotace (oblique).



**Vstup analýzy hlavních komponent a faktorové analýzy:**  
Matice korelací nebo kovariancí původních proměnných

**Výstup analýzy hlavních komponent a faktorové analýzy:**  
Ordinační diagram  
Korelace původních proměnných s hlavními komponenty, resp. s faktorovými osami

**Při použití analýzy hlavních komponent a faktorové analýzy je nutno pamatovat na níže uvedená omezení:**

- parametrická metoda
- problém odlehlých hodnot
- závislé na rozložení proměnných
- nelze použít když jsou faktory úplně nezávislé (jejich korelace je 0)
- zabývá se pouze variabilitou, nezabývá se příčinnými vztahy

## 5 Diskriminační analýza (Canonical variate analysis)

Častým cílem v sociálních i přírodních vědách je diskriminovat již známé skupiny objektů na základě skupiny kvantitativních parametrů (deskriptorů). Důvodem pro to může být přiřazení nového objektu do jedné ze skupin (identifikace) nebo jednoduše určení vlastností zástupců jednotlivých skupin (diskriminace).

Diskriminační analýza se zabývá závislostí jedné kvalitativní proměnné na několika kvantitativních proměnných. Objekty jsou charakterizovány sérií deskriptorů (parametrů). Tyto parametry musí být kvantitativní. Dále je známá příslušnost všech objektů do jedné ze skupin. Diskriminační analýza se používá k určení parametrů, které jsou diskriminující mezi dvěma nebo více skupinami.

Diskriminační analýza je parametrická metoda lineárního modelování. Nejprve testuje rozdíly mezi prediktory v jednotlivých předdefinovaných skupinách daných kategoriální závislou proměnnou a následně hledá lineární kombinace (diskriminační funkce) prediktorů, které nejlépe diskriminují mezi jednotlivými skupinami. Výpočet tak směřuje k nalezení diskriminační funkce a k zjištění relativního příspěvku jednotlivých parametrů (deskriptorů) k celkové diskriminaci skupin.

**Příklad.** Ukázka tabulky objektů příslušících do dvou skupin. Objekty jsou charakterizovány dvěma různými deskriptory.

Skupina	$y_1$	$y_2$
A	3	5
A	3	7
A	5	5
A	5	7
A	5	9
A	7	7
A	7	9
B	6	2
B	6	4
B	8	2
B	8	4
B	8	6
B	10	4
B	10	6

Kvalitativní  
proměnná

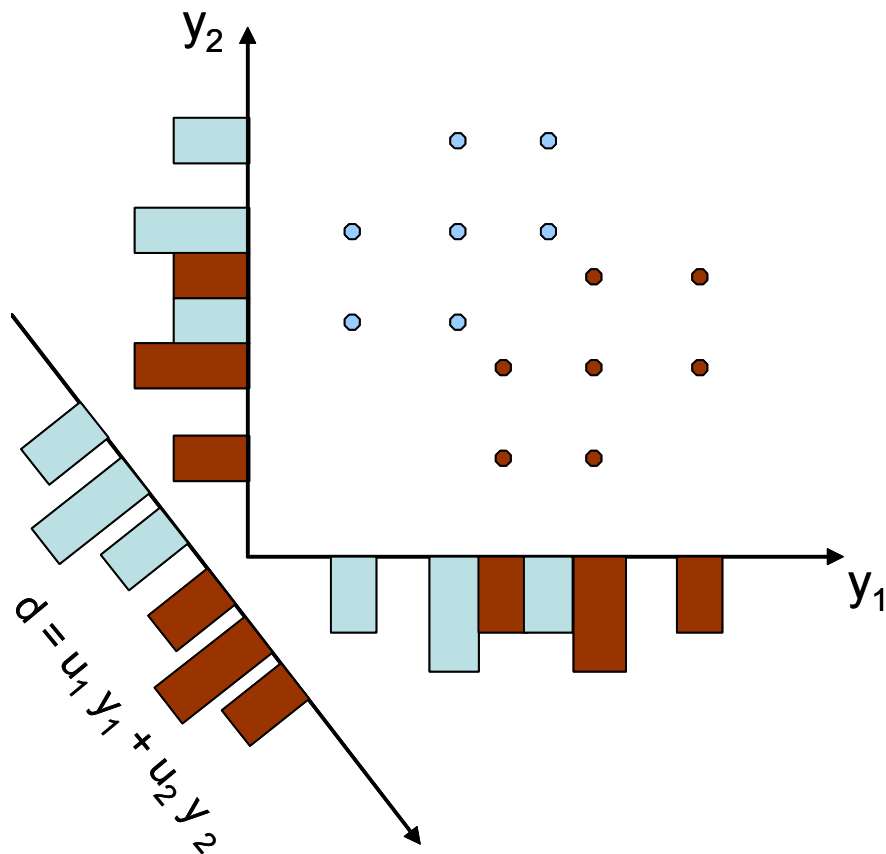
Kvantitativní proměnné  
(deskriptory)

Výsledkem diskriminační analýzy je diskriminační funkce (koeficienty deskriptorů). Proměnné s největšími (standardizovanými) koeficienty nejvíce přispívají k predikci příslušnosti do skupin. Výsledkem našeho příkladu jsou tyto koeficienty:

Skupina	Raw coefficients	Standardized coefficients
$y_1$	-0.6124	-1.0
$y_2$	0.6124	1.0
konstanta	0.6124	
Vlastní hodnota	3.9375	3.9375

Oba deskriptory přispívají stejnou mírou k diskriminaci skupin.

**Obrázek.** Dvě skupiny, každá se sedmi objekty, nemůžeme oddělit pomocí deskriptorů  $y_1$  nebo  $y_2$  (histogramy na osách). Ovšem tyto skupiny lze ideálně oddělit pomocí diskriminační funkce  $d$ .



Obrázek ukazuje ideální příklad dvou skupin popsanych pouze dvěma deskriptory. Skupiny nemůžou být odděleny žádným ze dvou deskriptorů. Řešením je nový diskriminační deskriptor  $d$ , který je lineární kombinací původních deskriptorů. Diskriminační osa  $d$  přechází směrem největší *meziskupinové* variability.

Tento analytický model může být zobecněn na několik skupin a mnoho deskriptorů. Výsledkem je nalezení diskriminačních funkcí (jejich počet je rovný počtu skupin snížený o jednu; v případě dvou skupin je výsledkem jedna diskriminační funkce).

**Vstup diskriminační analýzy:**

Tabulka objektů charakterizovaných několika kvantitativními parametry a jednou kvalitativní proměnnou (která přiřazuje objektům příslušnost ke skupině)

**Výstup diskriminační analýzy:**

Ordinační diagram (osy = kořeny = diskriminační funkce)  
Kořeny diskriminační analýzy (diskriminační funkce)

**Při použití diskriminační analýzy je nutno pamatovat na níže uvedená omezení:**

- parametrická metoda
- problém odlehlých hodnot
- závislé na rozložení proměnných
- výsledky udává v pravděpodobnostech
- není schopna zachytit nelineární vztahy mezi prediktory
- nelze použít na silně korelované prediktory

## 6 Shluková analýza

Jednou z možností využití informace obsažené ve vícerozměrných pozorováních je roztřídění objektů do několika poměrně homogenních shluků. Různými možnostmi a aspekty tvorby homogenních skupin objektů se zabývá shluková analýza. Shlukovou analýzou se sníží počet dimenzí objektů tak, že řadu uvažovaných proměnných zastoupí jediná proměnná, vyjadřující příslušnost objektu k definované skupině.

Uplatnění metod shlukové analýzy vede k příznivým výsledkům zejména tam, kde se studovaný soubor reálně rozpadá do tříd, tj. objekty mají tendenci se seskupovat do přirozených shluků. Použitím vhodných algoritmů se pak podaří odhalit strukturu studované množiny objektů a jednotlivé objekty klasifikovat. Zbývá pak již pouze najít vhodnou interpretaci pro popsání rozkladu, tj. charakterizovat vzniklé třídy.

Shlukovou analýzu používáme i v případech, kdy objekty nejeví tendenci k tvoření přirozených skupin, ale spíše připomínají víceméně homogenní chaos. Cíle analýzy zde musí být ovšem skromnější. Formálně může být cíl shlukové analýzy popsán následovně: Máme k dispozici datovou matici  $X$  typu  $n \times p$ , kde  $n$  je počet objektů a  $p$  je počet proměnných. Uvažujeme různé rozklady  $S^{(k)}$  množiny  $n$  objektů do  $k$  shluků a hledáme takový rozklad, který by byl z určitého hlediska nejvýhodnější. Zde připouštíme pouze rozklady s disjunktivními shluky, tj. jeden objekt patří pouze jednomu shluku. Cílem je dosáhnout, aby si objekty uvnitř shluku byly co nejvíce podobné a s objekty z různých shluků co nejméně. Existuje několik typů shlukové analýzy, které se liší postupem shlukování. Shlukování může být hierarchické nebo nehierarchické. Při hierarchické shlukové analýze každá skupina může obsahovat několik podskupin nižšího řádu a sama může být součástí skupiny vyššího řádu. Výsledek se dá graficky znázornit dendrogramem. Naproti tomu nehierarchická klasifikace rozdělí soubor na několik skupin stejného řádu.

### 6.1 Hierarchické shlukování

Hierarchické shlukovací metody uspořádají skupiny do hierarchické struktury. Jsou dvě možnosti k vytvoření hierarchické klasifikace: aglomerativní a divizivní. Při aglomerativních metodách spojujeme objekty navzájem nejpodobnější a poté s každou skupinou pracujeme jako se samostatným objektem až do okamžiku, kdy zůstane pouze jedna skupina.

Při divizivní klasifikaci dělíme celý soubor objektů nejčastěji na dvě části – a každou z nich potom považujeme za samostatný soubor, který znovu dělíme. Metody jsou konstruovány tak, aby podobnost uvnitř skupin a rozdíl mezi skupinami byly co největší.

#### Hierarchické aglomerativní shlukování

Aglomerativní shluková analýza pracuje se samostatnými objekty, které jsou shlukovány do větších shluků. V mnohých vědných disciplínách jsou aglomerativní techniky používány častěji než divizivní metody. Existuje mnoho aglomerativních metod, každá z nich využívá jiný pohled na data.

Základním krokem této metody je výpočet podobností mezi všemi dvojicemi objektů.

V různých etapách algoritmů posuzujeme podobnost dvou objektů, podobnost objektu a shluku a podobnost dvou shluků. Způsob výpočtu podobnosti zásadním způsobem ovlivňuje výsledek klasifikace. V následujícím textu jsou uvedeny různé míry podobnosti; většinou požadujeme, aby nabývaly hodnot od nuly pro maximální rozdílnost po jedničku pro totožnost. Často se však z praktických důvodů používají různé míry vzdálenosti, tentýž jev je

tedy měřen v opačném směru. Nevyplývají z toho žádné problémy; ostatně každou míru vzdálenosti  $D$  ( $D \geq 0$ ) lze převést na míru podobnosti  $A$ ,  $0 \leq A \leq 1$ , např.  $A = e^{-D}$  a naopak.

#### Vzdálenost a podobnost objektů

Po provedení výběru proměnných, které budou charakterizovat vlastnosti shlukovaných objektů, a po zjištění jejich hodnot rozhodneme o způsobu hodnocení vzdálenosti či podobnosti objektů. Velmi často je první etapou realizace shlukovacího algoritmu právě výpočet příslušných měr pro všechny páry objektů. Vzniká tak symetrická čtvercová matice typu  $n \times n$ , která má na diagonále nuly, jde-li o matici měr vzdálenosti  $D$ , nebo jedničky, jde-li o matici měr podobnosti  $A$ . Existuje mnoho měr vzdálenosti a indexů podobnosti.

Výsledek shlukové analýzy je silně ovlivněn výběrem metriky vzdálenosti, resp. indexu podobnosti.

Dalším krokem hierarchického aglomerativního shlukování je volba aglomerativní metody. Všechny aglomerativní metody jsou založeny na shlukování jednotlivých objektů nebo shluků do větších skupin. Skupiny, které jsou si nejvíc podobné, jsou sloučeny. Definice podobnosti mezi skupiny se u jednotlivých metod liší.

- Single-linkage clustering (metoda nejbližšího souseda; jednospojná metoda)

Historicky nejstarší metoda. Vzdálenost mezi dvěma shluky je daná jako minimální vzdálenost mezi všemi možnými zástupci shluků. Při použití této metody se často i značně vzdálené objekty mohou sejít ve stejném shluku, pokud větší počet dalších objektů mezi nimi vytvoří jakýsi most. Toto charakteristické řetězení objektů se považuje za nevýhodu, zvláště když máme důvod požadovat, aby shluky měly obvyklý eliptický tvar se zhuštěným jádrem.

- Complete-linkage clustering (metoda nejvzálenějšího souseda; všespojná metoda)

Tato metoda je založena na opačném principu než jednospojná metoda. Vzdálenost mezi dvěma shluky je daná maximální vzdáleností mezi všemi možnými zástupci obou shluků. Tato metoda produkuje shluky, které jsou mezi sebou dobře odděleny. Nežádoucí řetězový efekt zde odpadá, naopak je tu tendence ke tvorbě kompaktních shluků, nikoli mimořádně velkých.

- Average-linkage clustering (metoda průměrné vazby; středospojná metoda)

Při této metodě shlukování je meziskupinová (ne)podobnost definována jako průměrná (ne)podobnost mezi všemi možnými páry členů. Metoda vede často k podobným výsledkům jako metoda nejvzálenějšího souseda.

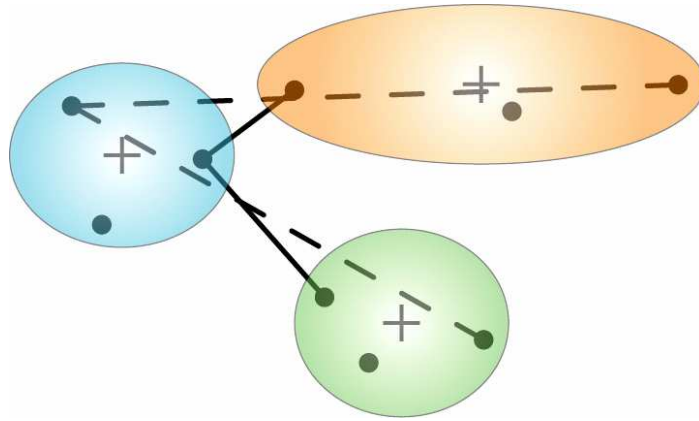
- Centroid clustering (centroidní metoda; Gowerova metoda)

Tato metoda nevyhází již ze shrnování informací o mezishlukových vzdálenostech objektů. Kritérium je euklidovská vzdálenost centroidů. Při této metodě je vzdálenost mezi shluky počítána jako vzdálenost mezi centroidy těchto shluků.

- Minimum variance clustering (Wardova metoda):

Wardova metoda je podobná středospojné a centroidní metodě. Kritérium pro spojování shluků je přírůstek celkového vnitroskupinového součtu čtverců odchylek pozorování od shlukového průměru. Přírůstek je vyjádřený jako součet čtverců v nově vznikajícím shluku, zmenšený o součty čtverců v obou zanikajících shlucích. Wardova metoda má tendenci odstraňovat malé shluky, tedy tvořit shluky zhruba shodné velikosti, což je často vítaná vlastnost.





+ centroid

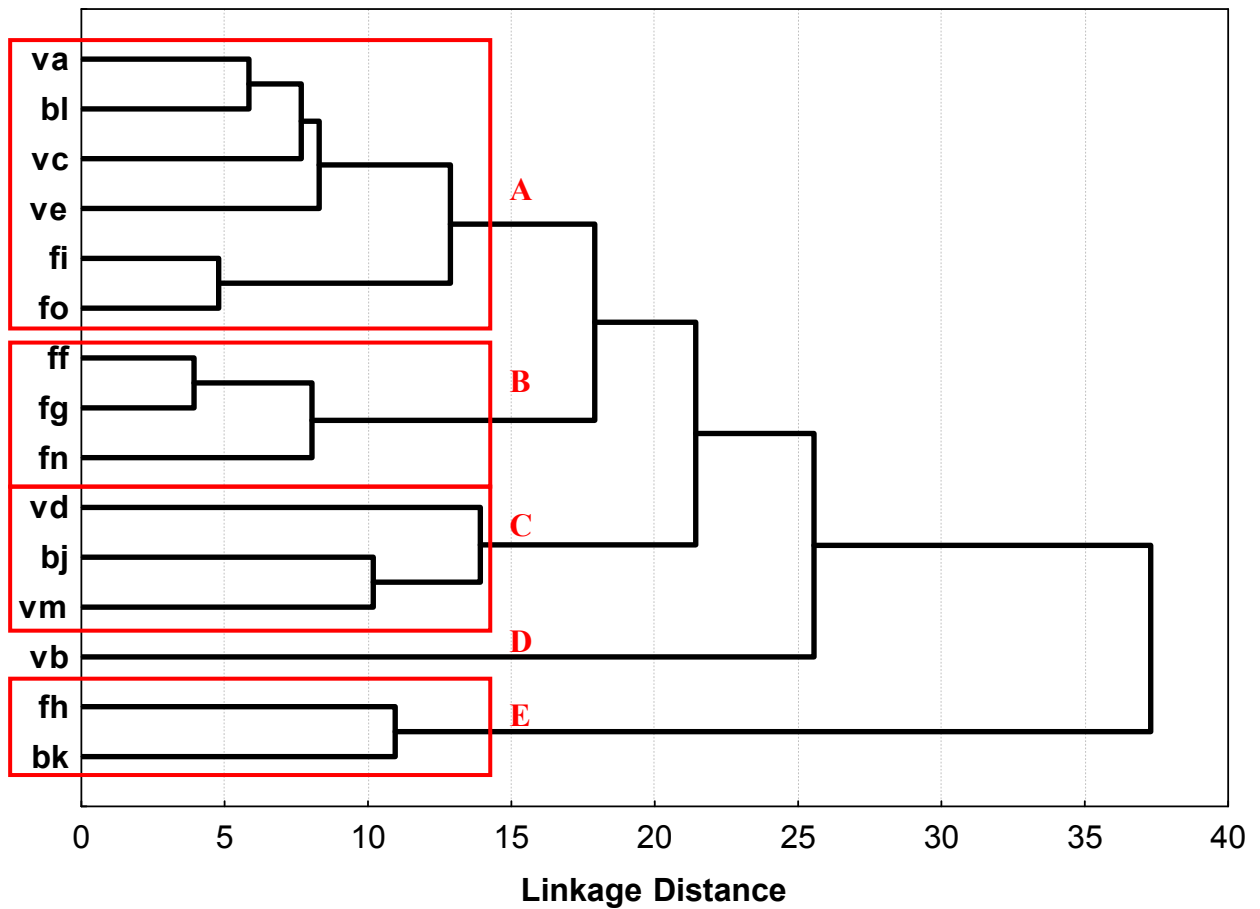
— Na tuto vzdálenost se ptá **single linkage**

- - - Na tuto vzdálenost se ptá **complete linkage**

Výsledek shlukové analýzy je velmi ovlivněn také výběrem měřítka vzdálenosti, resp. indexu podobnosti.

Výsledkem hierarchického shlukování je **dendrogram**.

**Obrázek.** Ukázka výsledku shlukové analýzy. Použita byla shlukovací metoda complete linkage, míra vzdálenosti: Euklidovské vzdálenosti.



Dendrogram znázorňuje podobnost novin, které pravidelně čtou obyvatelé 10 různých oblastí. (Příklad je totožný s příkladem při korespondenční analýze). Celkem 15 druhů novin bylo ze tří kategorií: vlámské (předpona „v“), francouzské (předpona „f“) a oba jazyky (předpona „b“). Interpretace dendrogramu je následující: Na určené hladině vzdálenosti se vytvořilo 5 shluků novin. První shluk (A) obsahoval noviny označené jako va, bl, vc, ve, fi, fo. V rámci tohoto shluku jsou si nejpodobnější noviny fi a fo (jsou sloučeny při nižší hladině vzdálenosti). Druhý shluk (B) zahrnuje pouze francouzsky psané noviny (ff, fg, fn; tady jsou si podobnější noviny ff a fg), třetí shluk vlámské noviny (vd, vm) a také noviny psané oběma jazyky (bj). Čtvrtý shluk tvoří pouze jeden objekt – vlámsky psané noviny vb. Poslední pátý shluk je tvořen novinami označenými fh a bk.

### Hierarchické divizivní shlukování

Divizivní metody mohou být monotetické – dělení souboru probíhá podle jediného parametru, nebo polytetické – dělení probíhá podle komplexní charakteristiky, získané na základě všech parametrů v rámci souboru.

Divizivní metody pracují ze začátku se všemi objekty jako s jednou skupinou. Nejdříve je tato skupina rozdělena do dvou menších skupin, které jsou opakovaně děleny na dvě podskupiny, až dokud není splněno kritérium, které ukončí analýzu (např. předem definovaný počet kroků). Principem tohoto způsobu shlukování je, že větší rozdíly přetrvávají nad méně důležitými rozdíly: celková struktura shluku determinuje podskupiny.

Jsou často používány v ekologii, konkrétně ke klasifikaci biologických společenstev, nicméně nic nebrání použití těchto metod i na marketingová data, zejména v případě trhu dominovaného jedním nebo malým počtem faktorů.

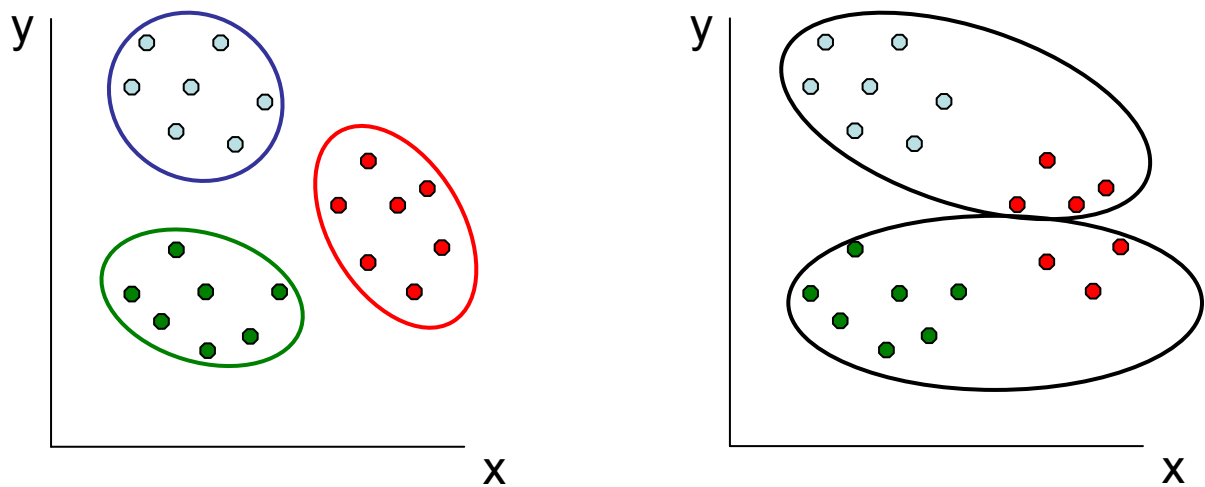
## **6.2 Nehierarchické shlukování**

Často existují případy, kdy není výhodné používat hierarchickou shlukovou analýzu z toho důvodu, že data nevykazují hierarchickou strukturu. V těchto případech může být preferováno nehierarchické shlukování, při kterém jsou skupiny stejného řádu.

Nejběžnější nehierarchickou metodou je metoda  $k$ -průměru ( $K$ -means clustering). Hlavním cílem metody je nalezení takových skupin v mnohorozměrném prostoru, kdy vnitroskupinová podobnost je co největší. Princip vytvoření shluků je stejný jako při Wardově metodě: minimalizace celkové sumy čtverců vzdáleností uvnitř skupin. Výsledkem je vytvoření  $K$  skupin, které jsou co nejvíce odděleny.

Algoritmus metody je následující: nejdříve zvolíme počet shluků, které v souboru objektů očekáváme. Dále jsou určeny centroidy pro všechny shluky. V posledním kroku jsou zařazeny všechny objekty do skupin na základě jejich vzdálenosti od centroidu skupiny. Proces je iterativní. Nevýhodou metody je nutnost definovat počet skupin  $K$  předem. Proto je vhodné provést analýzu pro několik dělení a následně určit poměr vnitroskupinové a meziskupinové variability pro všechny analýzy (všechny  $k$ ). Nakonec bude jako nejlepší určen takový počet shluků  $k$ , při kterém je poměr vnitroskupinové a meziskupinové variability nejmenší.

**Obrázek.** Ukázka rozdělení objektů do shluků nehierarchickou metodou K-means shlukování. Výsledek je ovlivněn volbou počtu shluků. Vlevo: počet shluků 3 je dobrá volba; vpravo: počet shluků 2 je špatná volba.



**Vstup shlukové analýzy:**  
 Matice podobnosti nebo vzdáleností objektů  
 nebo  
 tabulka objektů charakterizovaných několika parametry

**Výstup diskriminační analýzy:**  
 Strom (dendrogram) - při hierarchické shlukové analýze  
 Zařazení objektů do dopředu definovaného počtu shluků – při nehierarchické shlukové analýze

**Při použití shlukové analýzy je nutno pamatovat na níže uvedená problémy:**

- aglomerativní shlukování není efektivní pro velmi velká data
- při hierarchické aglomerativní analýze je výsledek silně ovlivněn výběrem indexu podobnosti, resp. metrikou vzdálenosti a shlukovacím algoritmem  
*! neexistuje správný shlukovací algoritmus !!!*
- při hierarchické divizivní analýze: monotetická metoda není robustní; polytetická metoda neposkytuje jednoduchý klíč ke zařazení nového objektu do skupiny
- při nehierarchickém shlukování je nutné určit počet skupin předem