Vícerozměrná analýza dat

Jiří Jarkovský

Plán kurzu

- ☑ **Každých 14 dní 4 vyučovací hodiny**

- ☑ **Ukončení zkouškou**
  - **Písemná**
  - **Zaměřená na principy a aplikace analýz**

- ☑ **Cíl kurzu**
  - **Vysvětlit principy vícerozměrných analýz, jejich aplikaci v biologii a jejich interpretaci**
  - **Přehled základního software**
  - **Příklady na reálných datech**

Náplň kurzu I

- Vícerozměná analýza dat – smysl a cíle
  - Příklady užití vícerozměrných analýz
  - Výhody a nevýhody vícerozměrné analýzy dat
  - Parametrická a neparametrická vícerozměrná statistika
  - Statistické SW pro vícerozměrnou analýzu dat

- Podobnost a vzdálenost objektů ve vícerozměrném prostoru
  - Metriky podobnosti a vzdálenosti a jejich úskalí
    - Obecné metriky podobnosti a vzdálenosti
    - Metriky podobnosti pro biologická společenstva – problém double zero
  - Asociační matice
    - Struktura asociační matice
    - Práce s asociační maticí
    - Mantelův test

- Vícerozměrné statistické testy a rozložení
  - Vícerozměrné normální rozložení
  - Vícerozměrné charakteristiky - medoid
  - Hottelingovo T, Wishartovo rozdělení

- Základy maticové algebry
  - Typy matic a jejich využití při vícerozměrné analýze dat
  - Matematické operace s maticemi
  - Eigenvalues (vlastní čísla) a eigenvectory (vlastní vektory) matic

Náplň kurzu II

- Shluková analýza
  - Kriteria posuzování výsledků shlukovacích metod
    - Minimální vnitroshluková varibilita
    - Maximální mezishluková variabilita
    - Silhouette width
  - Hierarchické aglomerativní shlukování
    - Shlukovací algoritmy
      - nearest neighbour (single linkage)
      - farthest neighbour (complete linkage)
      - UPGMA
      - WPGMA
      - UPGMC
      - WPGMC
      - Ward’s method
  - Hierarchické divizivní shlukování
    - TWINSPAN
  - Nehierarchické divizivní shlukování
    - K-means clustering
    - X-means clustering
    - Partitioning around medoids (PAM)

Náplň kurzu III

- Ordinační analýzy
  - Principy ordinačních analýz - redukce dimenzionality
    - Eigenvektor
    - Eigenvalue
  - Základní typy ordinační analýzy a jejich užití
    - PCA
    - CA
    - DCA
    - CCA
    - DCCA
    - RDA
    - MDS
    - PCoA
    - Kanonická korelace
- Analýza hlavních komponent
  - PCA na základě euklidovské vzdálenosti
  - PCA na základě korelací a kovariancí
  - Normalised PCA
  - Biplot a jeho interpretace
- Korespondenční analýza a její varianty
  - CA, DCA, CCA, DCCA
- MDS a PCoA – ordinační analýza na libovolné asociační matici

Software pro vícerozměrnou analýzu

- „Klikací všeobecné SW“
  - Statistica
  - SPSS
  - SAS
- Specializované SW
  - PcORD
  - CANOCO
  - PAST
  - WEKA
  - ORANGE
  - SW pro microarray analýzu
  - Nejrůznější utility na netu
  - ……
- Univerzální SW
  - R - ADE4 atd.

Vícerozměrná analýza dat

Základní statistické výpočty s vazbou na vícerozměrnou analýzu

Vztah klasické a vícerozměrné statistiky

- Vícerozměrná analýza dat využívá přístupů klasické statistiky
- Zároveň je citlivá i na jejich problémy !

- Agregace dat přes sumární statistiku nebo kontingenční tabulky – korespondenční analýza
- Korelace – analýza hlavních komponent, faktorová analýza, diskriminační analýza

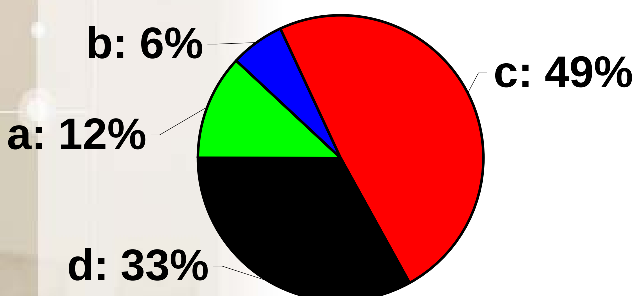
Kontingenční tabulka

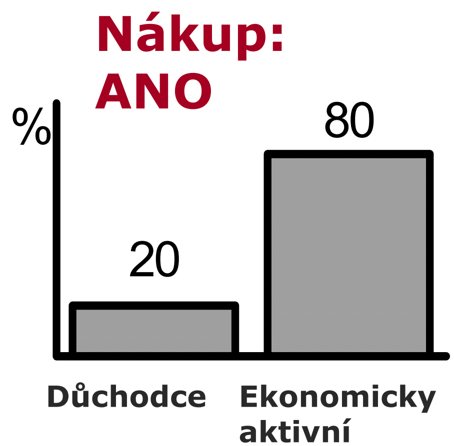
Důchodový věk

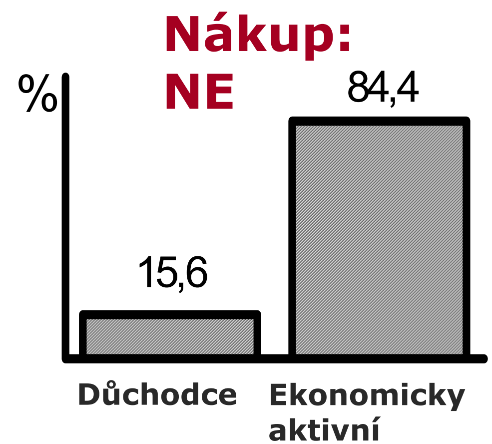
Nákup	Ano	Ne	Σ
Ano	20	82	102
Ne	10	54	64
Σ	30	136	166

- ☑ Kontingenční tabulka je používána pro hodnocení vztahu kategoriálních proměnných

Kontingenční tabulka v obrázku



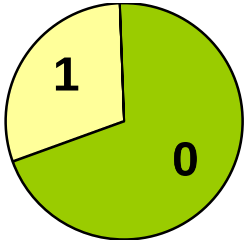




Kontingenční tabulky – princip analýzy

Binomické jevy (1/0)

$$\chi^2_{(1)} = \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}} + \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$



I. jev 1 | **II. jev 2**

Příklad

✓ **10 000 lidí hází mincí**

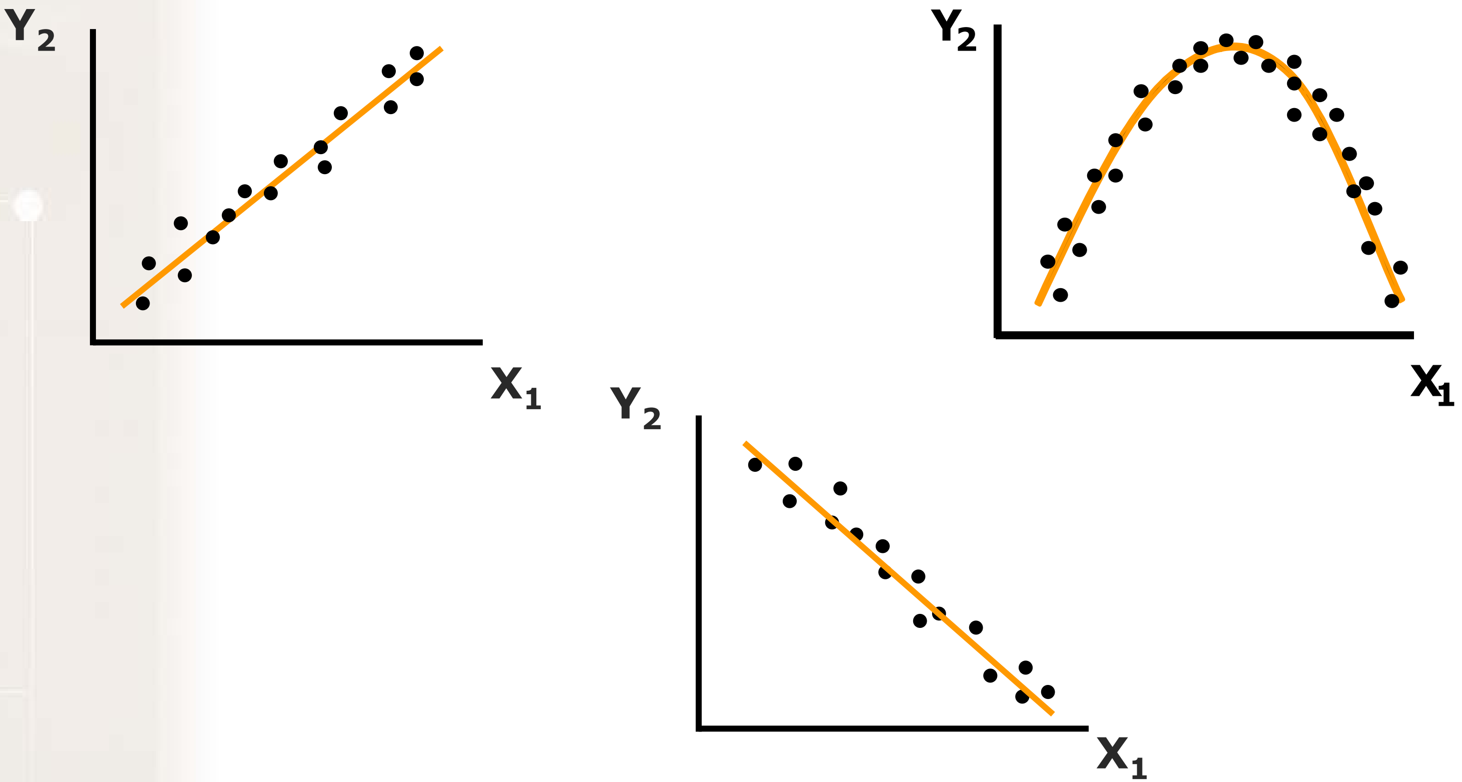
rub: 4 000 případů (R)
líc: 6 000 případů (L)

? **Lze výsledek považovat za statisticky významně odlišný (nebo neodlišný) od očekávaného poměru R : L = 1 : 1 ?**

Stejným způsobem, tedy hodnocením odchylek od očekávaného vyrovnaného počtu případů hodnotí data i korespondenční analýza

Korelační analýza

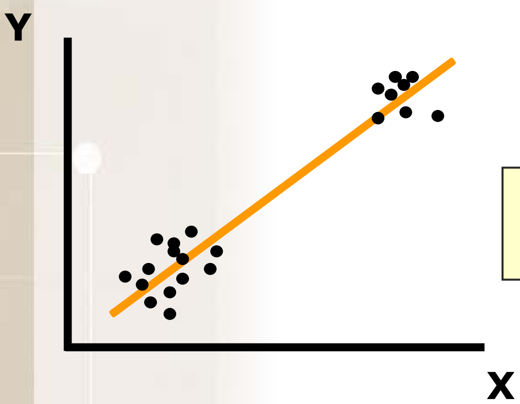
- Korelace - vztah (závislost) dvou znaků (parametrů)



Korelace mezi parametry jsou základem faktorové analýzy a analýzy hlavních komponent, pokud vazby mezi parametry nejsou tyto metody postrádají smysl.

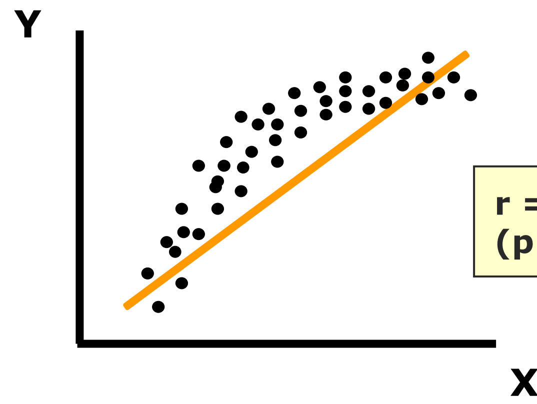
Rizika korelační analýzy

Problém rozložení hodnot



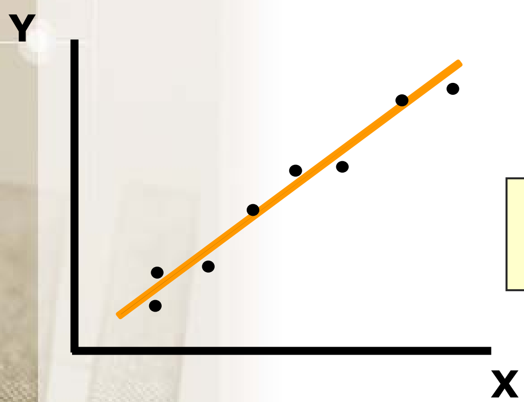
r = 0,981
(p < 0,001)

Problém typu modelu

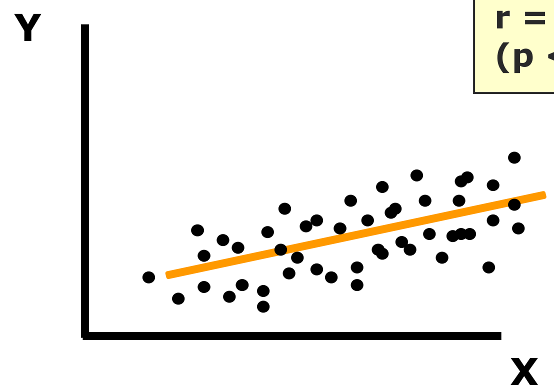


r = 0,761
(p < 0,032)

Problém velikosti vzorku



r = 0,891
(p < 0,214)

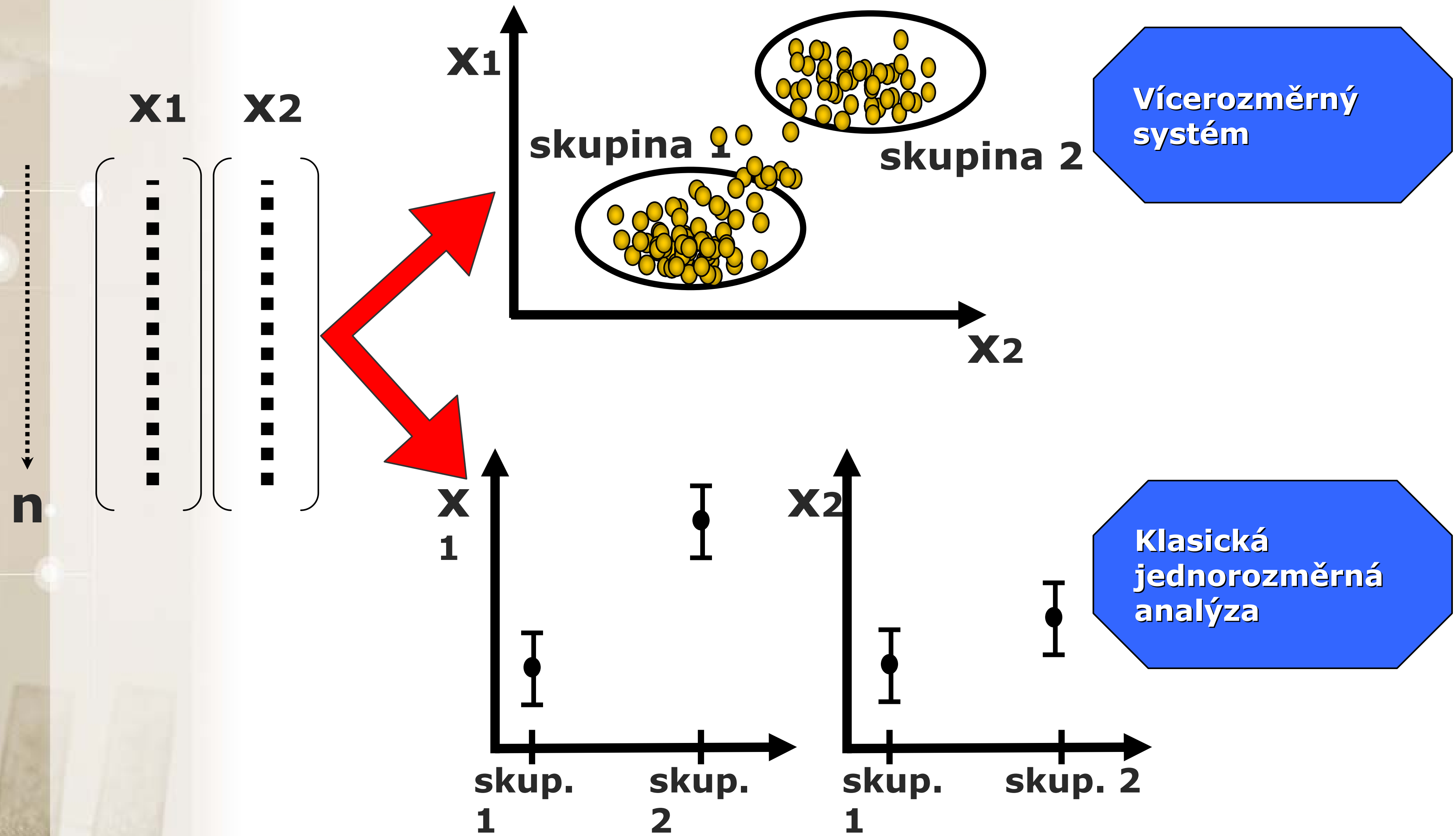


r = 0,212
(p < 0,008)

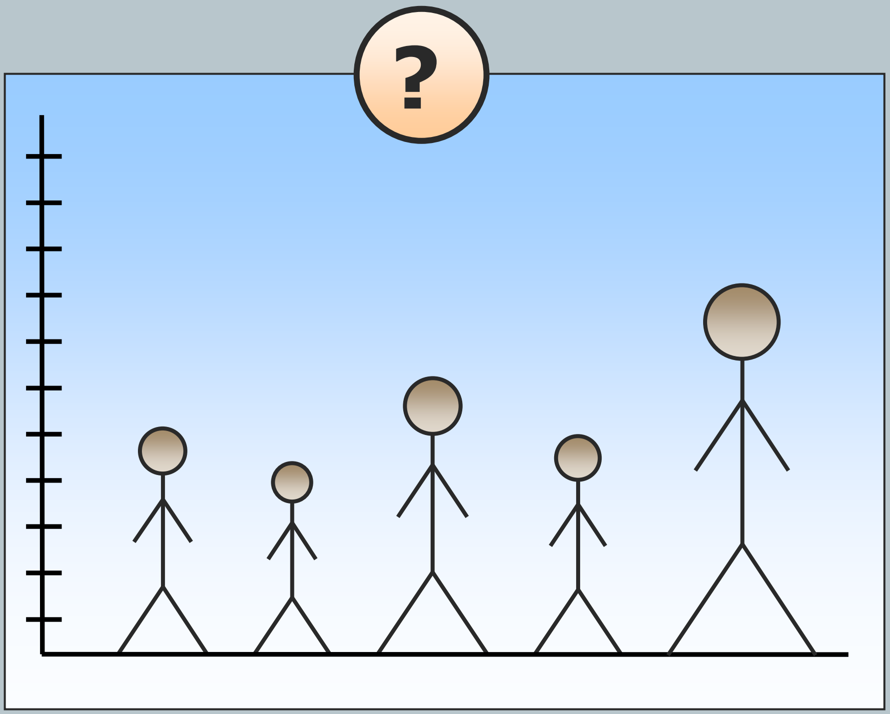
Vícerozměrná analýza dat

Význam vícerozměrného hodnocení dat

Vícerozměrné vnímání skutečnosti – nová kvalita analýzy dat



Běžná sumarizace dat „likviduje“ individualitu jedince



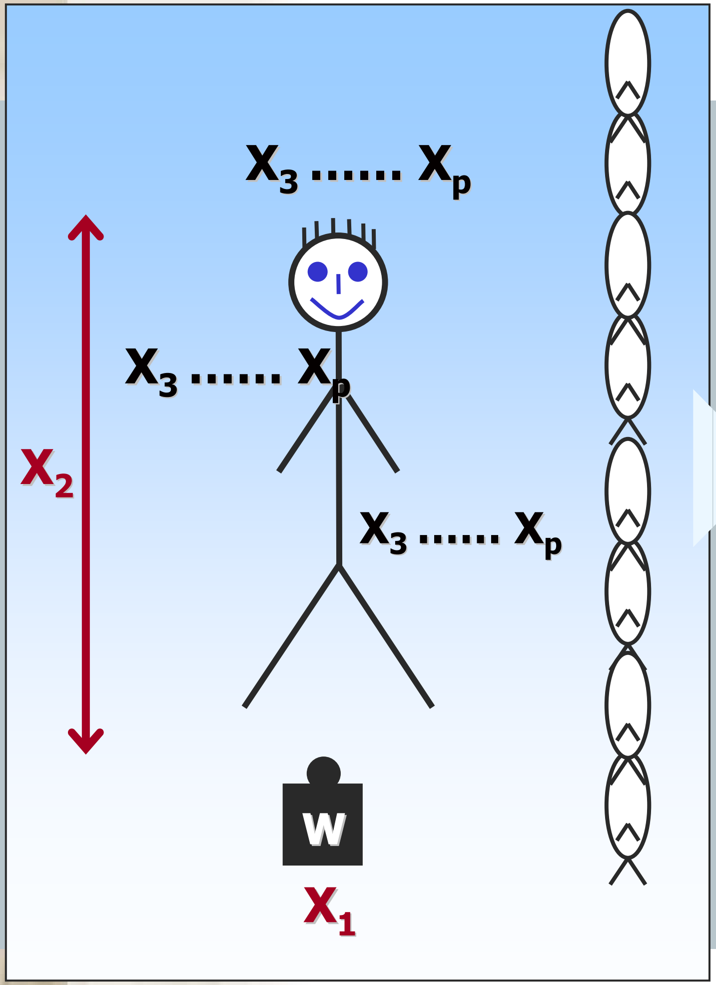
Průměr ± SE

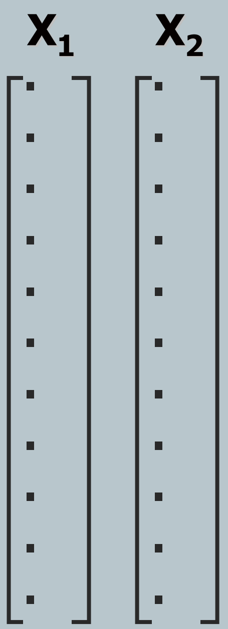
BĚŽNÁ STATISTICKÁ SUMARIZACE

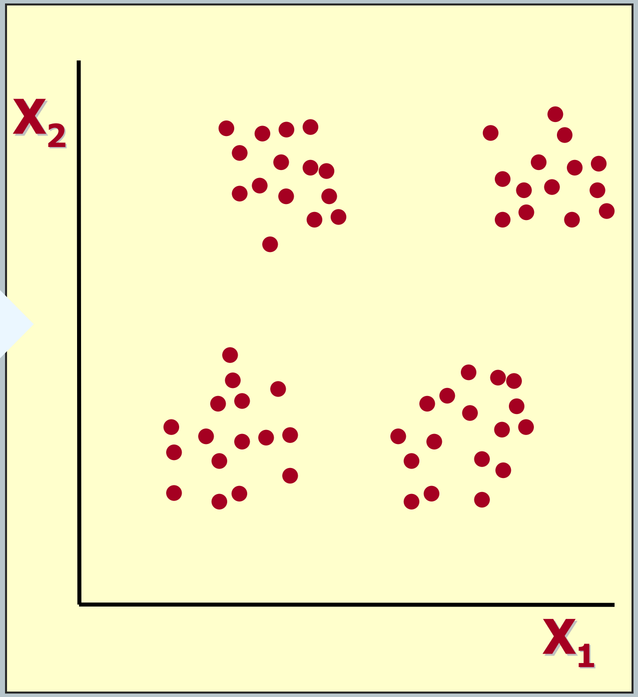
- ✓ *Zpřehlednění dat*
- ✓ *Neodliší původní měření*

Vícerozměrné hodnocení

… s ohledem na individualitu !

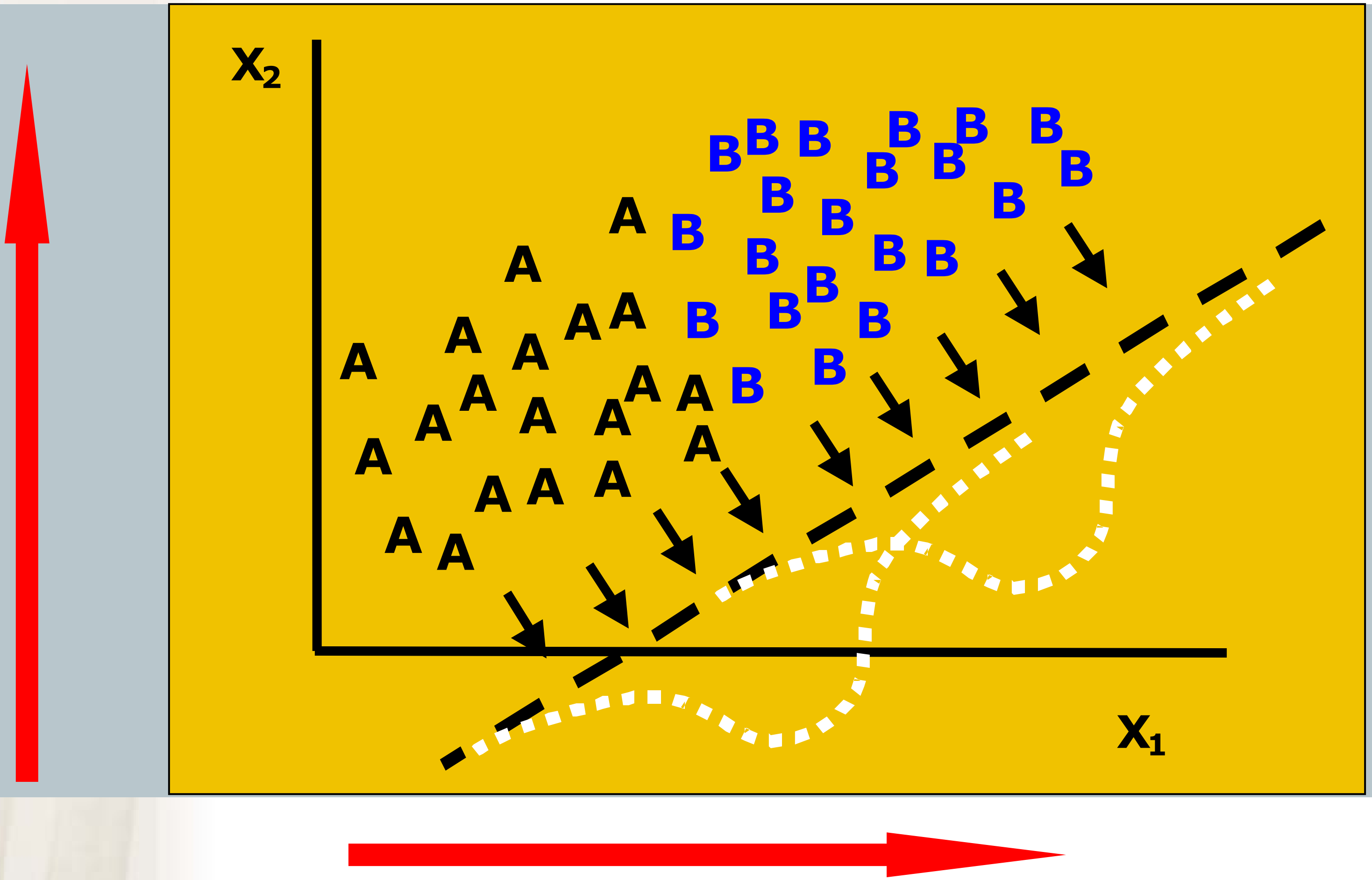






Vícerozměrné hodnocení – nová kvalita

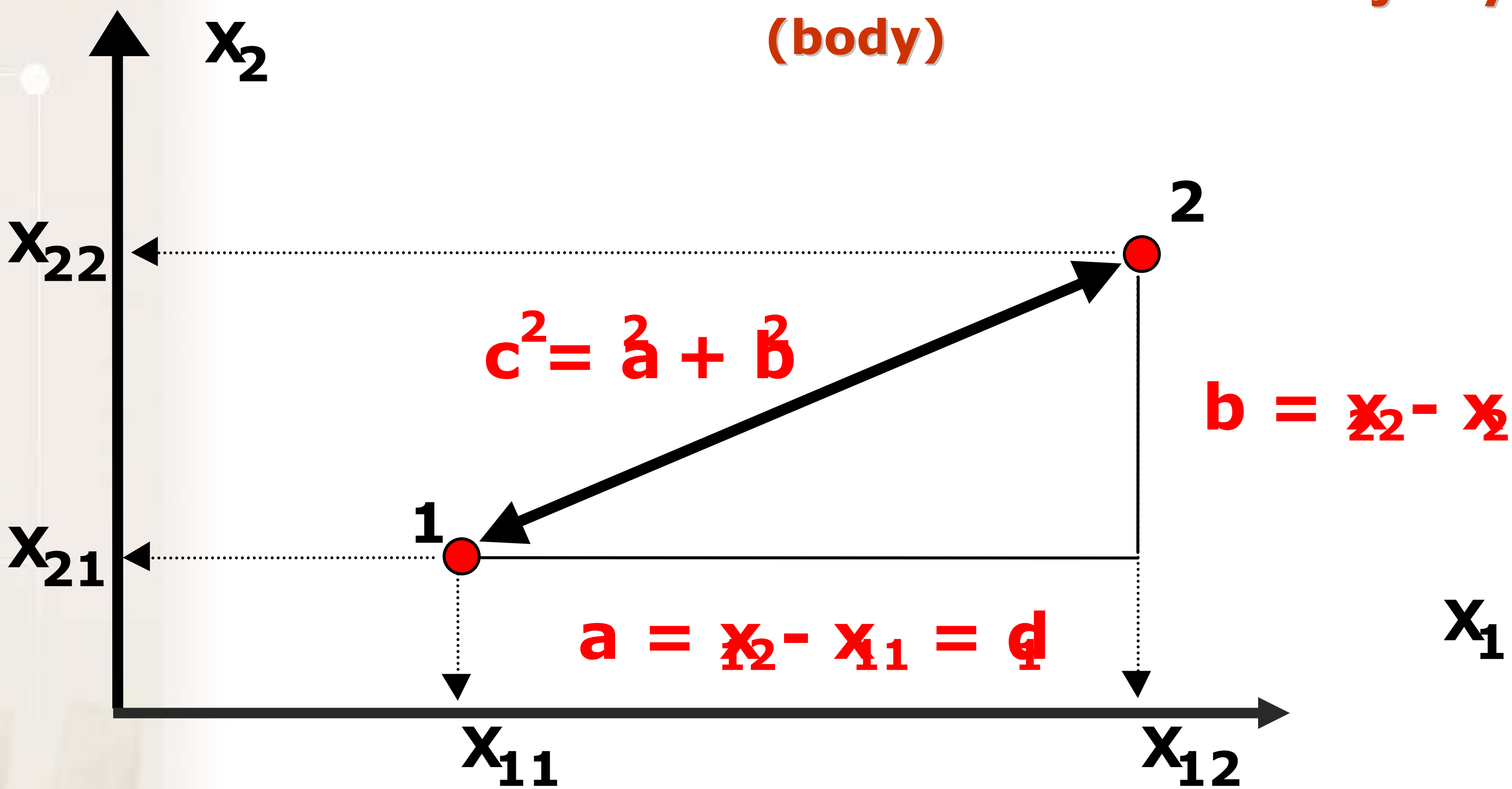
Pouze kombinované parametry mají odpovídající informační sílu



příklad: X1 =

Vícerozměrné hodnocení vychází z jednoduchých principů

příklad: vícerozměrná vzdálenost měření mezi dvěma objekty (body)



Vícerozměrné modelování je strategickou disciplínou

$X_1 \ldots\ldots X_n$	$X_{n+1} \ldots\ldots X_p$	$X_{p+1} \ldots\ldots X_2$
technické parametry automobilu	**řidičovy schopnosti a jeho stav**	**rychlost, povrch, situace**

X_1 X_2 X_3 X_4 X_5 .... X_p

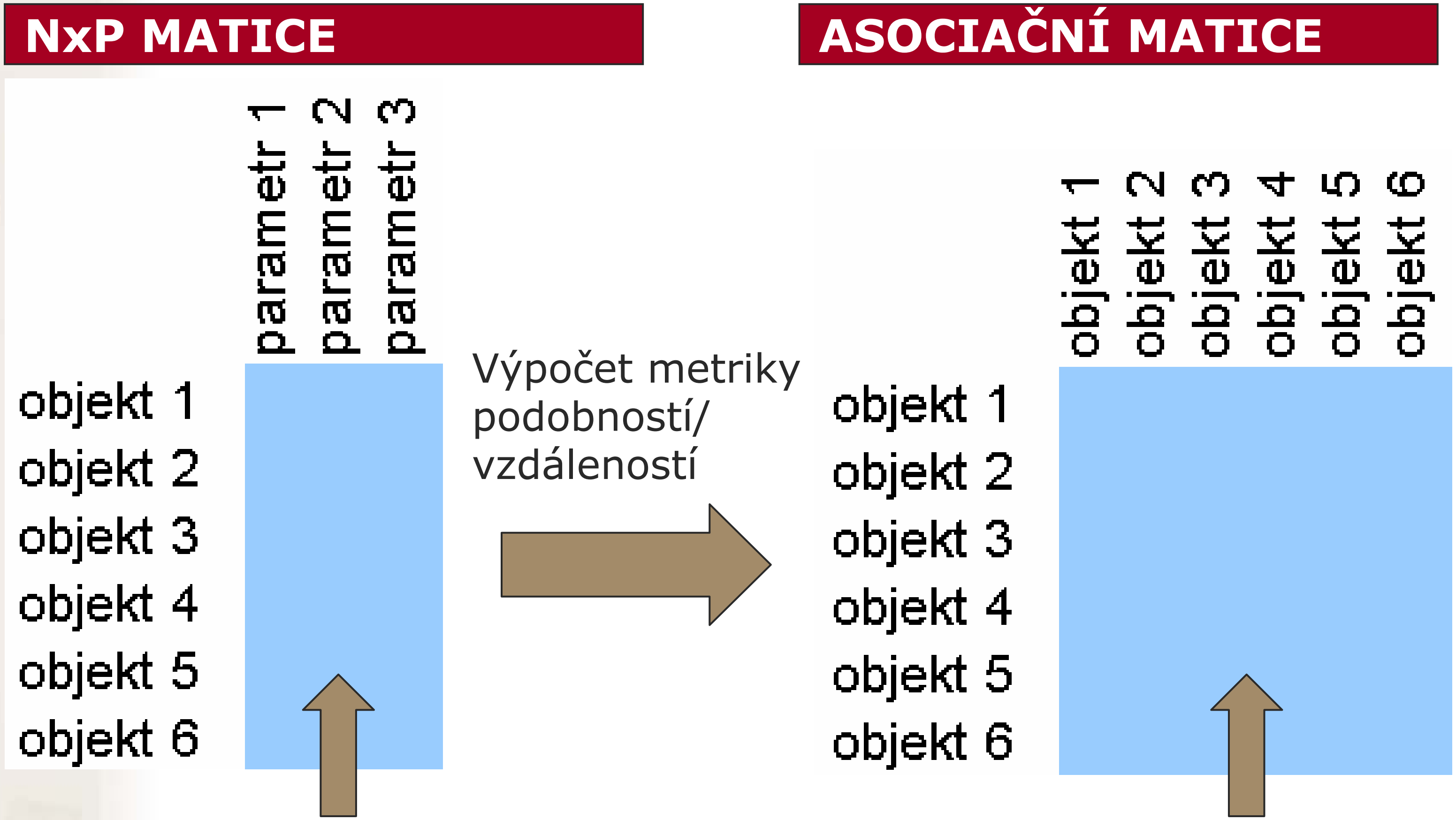
Vícerozměrná analýza dat

Základní principy vícerozměrného hodnocení dat

Pojmy vícerozměrných analýz

- Vícerozměrné metody: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty (i.e. klienti) a každý z nich je charakterizován svými parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- Maticová algebra: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- NxP matice: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- Asociační matice: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. metriky) buď objektů (Q mode analýza) nebo parametrů (R mode analýza).Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

Vstupní matice vícerozměrných analýz



Hodnoty parametrů pro jednotlivé objekty

Korelace, kovariance, vzdálenost, podobnost

Základní typy vícerozměrných analýz

SHLUKOVÁ ANALÝZA

- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

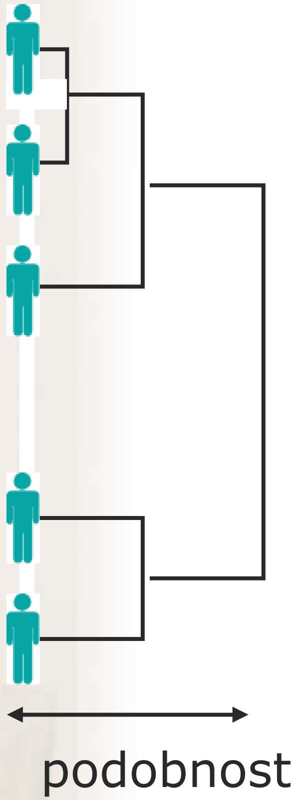
ORDINAČNÍ METODY

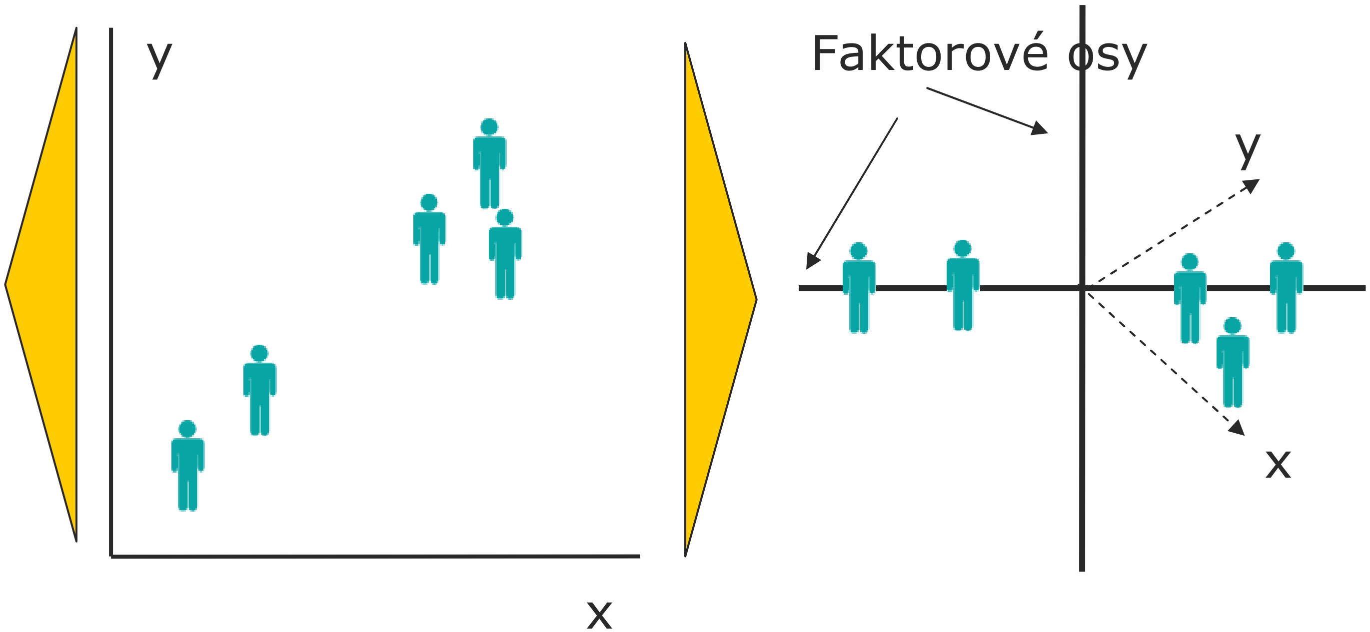
- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

Typy vícerozměrných analýz

SHLUKOVÁ ANALÝZA

ORDINAČNÍ METODY



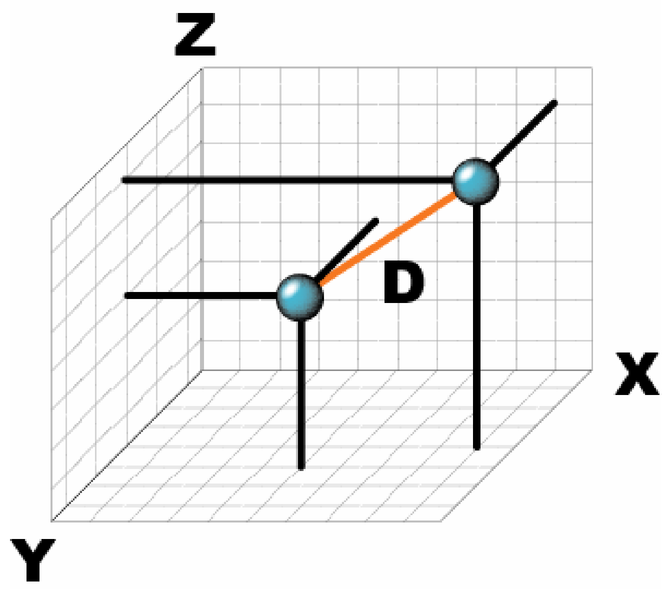


Vícerozměrná analýza dat

Asociační matice
Vícerozměrná vzdálenost a podobnost

Seznam taxonů – vícerozměrný popis společenstva

- Na seznam taxonů lze pohlížet také jako seznam rozměrů společenstva
- Záznam o nalezených taxonech tak vlastně tvoří vícerozměrný popis daného společenstva
- Společenstva můžeme srovnávat podle jejich vzájemné pozice v n-rozměrném prostoru
- Pro srovnání společenstev lze teoreticky využít libovolnou metriku vícerozměrné podobnosti nebo vzdálenosti



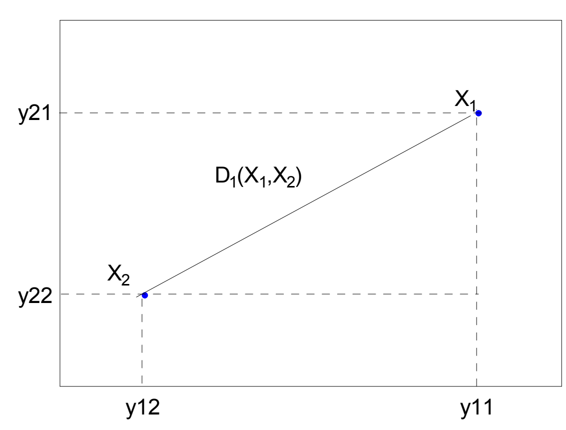
Euklidovská vzdálenost

- **Jde o základní metrické měřítko vzdálenosti a počítá vzdálenost objektů obdobně jako Pythagorova věta počítá přeponu pravoúhlého trojúhelníku. Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a double zero problém. Nemá horní hranici hodnot.**

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

- **Jako další měřítko se používá také čtverec této vzdálenosti. . Jeho nevýhodou jsou semimetrické vlastnosti.**

$$D_1^2(x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$



Double zero problém !!!

- V případě binárních metrik (druh se vyskytuje/nevyskytuje) není možné uvažovat stejnou váhu pro souhlas přítomnosti (11) a nepřítomnosti (00) taxonů (symetrický koeficient)
- Problémem využití všech typů metrik pro data abundancí spočívá v odlišném významu přítomnosti a nepřítomnosti taxonů
- Pokud se taxon nachází v obou srovnávaných společenstvech – znamená to že společenstva si budou v tomto ohledu podobná, protože mají podmínky umožňující přítomnost taxonu
- Pokud se taxon nenachází ani v jednom ze dvou srovnávaných společenstev – příčina může být nejrůznější – double zero problem
- Pro odstranění tohoto problému je použito asymetrické hodnocení souhlasné přítomnosti (11) a nepřítomnosti (00) taxonů (asymetrické koeficienty)

Koeficienty podobosti (indexy podobnosti)

- V ekologii se využívá řada indexů podobnosti založených buď na přítomnosti/nepřítomnosti taxonů nebo na abundancích

Binární koeficienty podobnosti

		Společenstvo 1	
		1	0
Společenstvo 2	1	a	b
	0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika společenstev 1 a 2
a+b+c+d=p

Symetrické binární koeficienty - není rozdíl mezi případem 1-1 a 0-0
Asymetrické binární koeficienty - rozdíl mezi případem 1-1 a 0-0

Více informací a další měření vzdáleností a podobností najdete v knize
LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elseviere Science BV, Amsterodam.

Vícerozměrná analýza dat

Symetrické binární koeficienty

Simple matching coefficient (Sokal & Michener, 1958)

- **Obvyklou metodou pro výpočet podobnosti mezi dvěma objekty je podíl počtu deskriptorů, které kódují objekt stejně, a celkového počtu deskriptorů. Při použití tohoto koeficientu předpokládáme, že není rozdíl mezi nastáním 0 a 1 u deskriptorů.**

$$S_1(x_1, x_2) = \frac{a+d}{p}$$

Rogers & Tanimoto koeficient (1960)

- Dává větší váhu rozdílům než podobnostem.

$$S_2(x_1, x_2) = \frac{a+d}{a+2b+2c+d}$$

Sokal & Sneath (1963)

- Další čtyři navržené koeficienty obsahují double-zero, ale jsou navrženy tak, aby se snížil vliv double-zero:

$$S_3(x_1, x_2) = \frac{2a + 2d}{2a + b + c + 2d}$$

- tento koeficient dává dvakrát větší váhu shodným deskriptorům než rozdílným;

$$S_4(x_1, x_2) = \frac{a + d}{b + c}$$

- porovnává shody a rozdíly prostým podílem v měřítku jdoucím od 0 do nekonečna;

$$S_5(x_1, x_2) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$$

- porovnává shodné deskriptory se součty okrajů tabulky;

$$S_6(x_1, x_2) = \frac{a}{\sqrt{(a+b)(a+c)}}\frac{d}{\sqrt{(b+d)(c+d)}}$$

- je vytvořen z geometrických průměrů členů vztahujících se k ***a*** a ***d*,** podle koeficientu S5.

Hammannův koeficient

$$S = \frac{a + d - b - c}{p}$$

Yuleho koeficient

$$S = \frac{ad - bc}{ad + bc}$$

Pearsonovo Φ (phi)

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Vícerozměrná analýza dat

Asymetrické binární koeficienty

Jaccardův koeficient (1900, 1901, 1908)

- ☑ **Všechny členy mají stejnou váhu**

$$S_7(x_1, x_2) = \frac{a}{a+b+c}$$

Sørensenův koeficient (1948) (Coincidence index, Dice(1945))

- varianta předchozího koeficientu dává dvojnásobnou váhu dvojitým prezencím , protože se může zdát, že přítomnost druhů je více informativní než jejich absence, která může být způsobena různými faktory a nemusí nutně odrážet rozdílnost prostředí. Prezence druhu na obou lokalitách je silným ukazatelem jejich podobnosti. S7 je monotónní k S8, proto podobnost pro dvě dvojice objektů vypočítaná podle S7 bude podobná stejnému výpočtu S8. Oba koeficienty se liší pouze v měřítku. Tento index byl poprvé použit Dicem v R-mode studii asociací druhů. Jiná varianta tohoto koeficientu dává duplicitním prezencím trojnásobnou váhu.

$$S_8(x_1, x_2) = \frac{2a}{2a+b+c} \qquad S_8(x_1, x_2) = \frac{3a}{3a+b+c}$$

Sokal & Sneath (1963)

- ☑ **navržen jako doplněk Rogers & Tanimotova koeficientu (S2), dává dvojnásobnou váhu rozdílům ve jmenovateli.**

$$S_{10}(x_1, x_2) = \frac{a+d}{a+2b+2c}$$

Russel & Rao (1940)

- navržená míra umožňuje porovnání počtu duplicitních prezencí (v čitateli) proti celkovému počtu druhů, nalezených na všech lokalitách, zahrnujícím druhy, které chybějí (*d*) na obou uvažovaných lokalitách.

$$S_{11}(x_1, x_2) = \frac{a}{p}$$

Kulczynski (1928)

- ☑ **koeficient porovnávající duplicitní prezence s diferencemi**

$$S_{12}(x_1, x_2) = \frac{a}{b+c}$$

Binární verze asymetrického kvantitativního Kulczynski koeficientu (1928)

- Mezi svými koeficienty pro presence/absence data zmiňují Sokal & Sneath (1963) tuto verzi kvantitativního koeficientu S18, kde jsou duplicitní prezence srovnávány se součty okrajů tabulky (*a*+*b*) a (*a*+*c*).

$$S_{13}(x_1, x_2) = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$$

Ochiachi (1957)

- ☑ **použil jako míru podobnosti geometrický průměr poměrů *a* k počtu druhů na každé lokalitě, tj. se součty okrajů tabulky (*a*+*b*) a (*a*+*c*), tento koeficient je obdobou S6, bez části, týkající se double-zero (d).**

$$S_{14}(x_1, x_2) = \sqrt{\frac{a}{(a+b)}\frac{a}{(a+c)}} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Faith (1983)

- V tomto koeficientu je neshoda (přítomnost na jedné a absence na druhé lokalitě) vážena proti duplicitní prezenci. Hodnota S26 klesá s růstem double-zero

$$S_{26}(x_1, x_2) = \frac{a + d/2}{p}$$

Vícerozměrná analýza dat

Kvantitativní koeficienty

„Klasické“ indexy podobnosti

- Sørensenův kvantitativní koeficient, kde aN a bN jsou celkové počty jedinců v společenstvech A a B, jN je pak suma abundancí pokud se druh nachází v obou společenstvech, je počítána vždy z nižší abundance daného druhu ve společenstvu

$$C_N = \frac{2jN}{(aN + bN)}$$

- Morisita-Horn index, kde aN je celkový počet jedinců ve společenstvu A a an_i počet jedinců druhu i ve společenstvu A (obdobně platí pro společenstvo B)

$$C_{mH} = \frac{2\sum(an_i bn_i)}{(da + db).aN.bN} \qquad da = \frac{\sum an_i^2}{aN^2}$$

Jednoduchý srovnávací koeficient (Sokal & Michener, 1958)

- modifikovaný simple matching coefficient může být použit pro multistavové deskriptory - čitatel obsahuje počet deskriptorů, pro které jsou dva objekty ve stejném stavu – např. je-li dvojice objektů popsána následujícími deseti multistavovými deskriptory: hodnota S1,vypočítaná pro 10 multistavových deskriptorů bude S1,(x1,x2) = 4 agreements/ 10 descriptors = 0.4
- Podobným způsobem je možné rozšířit všechny binární koeficienty pro multistavové deskriptory.

$$S_1(x_1, x_2) = \frac{agreements}{p}$$

	Deskriptors										Σ
Object x_1	9	3	7	3	4	9	5	4	0	6	
Object x_2	2	3	2	1	2	9	3	2	0	6	
Agreements	0	+1	+0	+0	+0	+1	+0	+0	+1	+1	4

Gowerův obecný koeficient podobnosti (1971) I.

- Gover navrhl obecný koeficient podobnosti, který může kombinovat různé typy deskriptorů. Podobnost mezi dvěma objekty je vypočítána jako průměr podobností, vypočítaných pro všechny deskriptory. Pro každý deskriptor j je hodnota parciální podobnosti s_{12j} mezi objekty x1 a x2 vypočítána následovně:

$$S_{15}(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j}$$

- ✓ Pro binární deskriptory sj=1 (shoda) nebo 0 (neshoda). Gower navrhl dvě formy tohoto koeficientu. Následující forma je symetrická, dává sj=1 double-zero. Druhá forma, Gowerův asymetrický koeficient S19 dává pro double-zero sj=0
- ✓ Kvalitativní a semikvantitivní deskriptory jsou upraveny podle jednoduchého zaměňovacího pravidla, sj=1 při souhlasu a sj = 0 při nesouhlasu deskriptorů. Double zero jsou ošetřeny stejně jako v předchozím odstavci.
- ✓ Kvantitativní deskriptory (reálná čísla) jsou zpracovány následovně: pro každý deskriptor se nejprve vypočte rozdíl mezi stavy obou objektů který je poté vydělen největším rozdílem (Rj), nalezeným pro daný deskriptor mezi všemi objekty ve studii (nebo v referenční populaci – doporučuje se vypočítat největší diferenci Rj každého deskriptoru j pro celou populaci, aby byla zajištěna konzistence výsledků pro všechny parciální studie).

Gowerův obecný koeficient podobnosti (1971) II.

- normalizovaná vzdálenost může být odečtena od 1 aby byla transformována na podobnost:

$$s_{12j} = 1 - \left[\frac{\left| y_{1j} - y_{2j} \right|}{R_j} \right]$$

- Gowerův koeficent může být nastaven tak, aby zahrnoval přídavný flexibilní prvek: žádné porovnání není vypočítáno u deskriptorů, u nichž chybí informace buď u jednoho, nebo u druhého objektu. Toto zajišťuje člen *wj*, nazývaný Kroneckerovo delta, popisující přítomnost/nepřítomnost informace v obou objektech: je-li informace o deskriptoru yj přítomna u obou objektů (*wj*=1), jinak (*wj*=0), tento koeficient nabývá hodnot podobnosti mezi 0 a 1 (největší podobnost objektů). Další možností je vážení různých deskriptorů prostým přiřazením čísla v rozsahu 0-1 *wj*.

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}$$

Vícerozměrná analýza dat

Různé vícerozměrné metriky vzdáleností

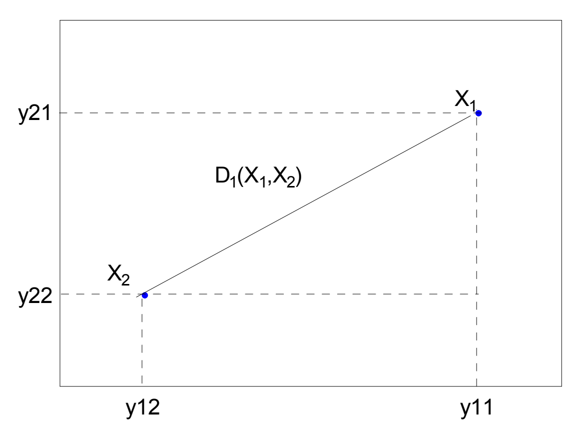
Euklidovská vzdálenost

- Jde o základní metrické měřítko vzdálenosti a počítá vzdálenost objektů obdobně jako Pythagorova věta počítá přeponu pravoúhlého trojúhelníku. Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a double zero problém. Nemá horní hranici hodnot.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

- Jako další měřítko se používá také čtverec této vzdálenosti. . Jeho nevýhodou jsou semimetrické vlastnosti.

$$D_1^2(x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$



Průměrná vzdálenost

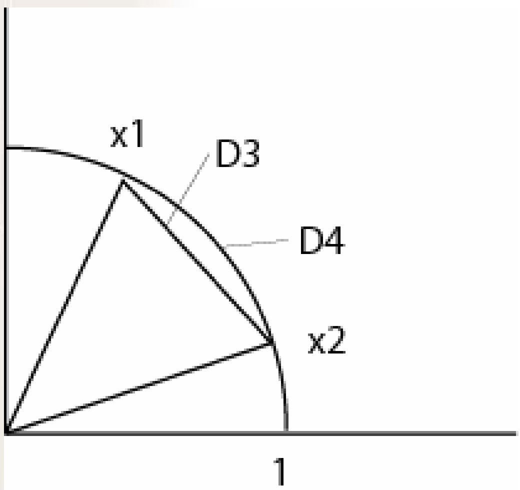
- Euklidovská vzdálenost je přepočítána na počet parametrů (druhů v případě vzdálenosti společenstev odběrů).

$$D_2^2(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p}(y_{1j} - y_{2j})^2$$

$$D_2(x_1, x_2) = \sqrt{D_2^2}$$

Chord distance (Orlóci, 1967)

- Odstraňuje double zero problém a vliv rozdílného počtu jedinců druhů ve vzorcích při výpočtu Euklidovské vzdálenosti. Její maximální hodnota je druhá odmocnina ze dvou a minimum 0. Při výpočtu počítá pouze s poměry druhů v rámci jednotlivých vzorků. Jde vlastně o Euklidovskou vzdálenost počítanou pro vektory vzorků standardizované na délku 1, nebo je možný přímý výpočet už zahrnující standardizaci. Vnitřní část výpočtu je vlastně cosinus úhlu svíraného vektory, zápis vzorce je možný i v této formě.

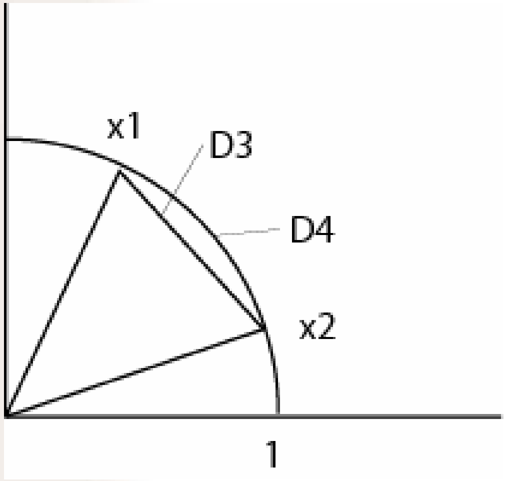


$$D_3(x_1, x_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)}$$

$$D_3 = \sqrt{2(1 - \cos\theta)}$$

Geodetická metrika

- Počítá délku výseče jednotkové kružnice mezi normalizovanými vektory (viz. Chord distance).



$$D_4(x_1, x_2) = \arccos\left[1 - \frac{D_3^2(x_1, x_2)}{2}\right]$$

Mahalanobisova vzdálenost (Mahalanobis 1936)

- Jde o obecné měřítko vzdálenosti beroucí v úvahu korelaci mezi parametry a je nezávislá na rozsahu hodnot parametrů. Počítá vzdálenost mezi objekty v systému souřadnic jehož osy nemusí být na sebe kolmé. V praxi se používá pro zjištění vzdálenosti mezi skupinami objektů. Jsou dány dvě skupiny objektů w1 a w2 o n1 a n2 počtu objektů a popsané p parametry:

$$D_5^2(w_1, w_2) = \overline{d_{12}} V^{-1} \overline{d_{12}^{`}}$$

- Kde $\overline{d_{12}}$ je vektor o délce p rozdílů mezi průměry p parametrů v obou skupinách. V je vážená disperzní matice (matice kovariancí parametrů) uvnitř skupin objektů.

$$V = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1 + (n - 2)S_2\right]$$

- kde S1 a S2 jsou disperzní matice jednotlivých skupin. Vektor měří rozdíl mezi p- rozměrnými průměry skupin a V vkládá do rovnice kovarianci mezi parametry.

Minkowskeho metrika

- Je obecnou formou výpočtu vzdálenosti – podle zadaného koeficientu může odpovídat např. Euklidovské nebo Manhattanské metrice. Se stoupající koeficientem umocňování stoupá významnost větších rozdílů. Existuje ještě obecnější forma, kdy koeficient umocňování a odmocňování je zadáván zvlášť.

$$D_6(x_1, x_2) = \left[\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|^r \right]^{1/r}$$

Manhattanská vzdálenost

- Jde vlastně o součet rozdílů jednotlivých parametrů popisujících objekty

$$D_7(x_1, x_2) = \sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|$$

Mean character difference (Czekanowski 1909)

- **Manhattanská vzdálenost přepočítaná na počet parametrů.**

$$D_8(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} \left| y_{1j} - y_{2j} \right|$$

Whittakerův asociační index (Whittaker 1952)

- Je dobře použitelný pro data abundancí, každý druh je nejprve transformován ve svůj podíl ve společenstvu, následující výpočet je opět obdobou Manhattanské vzdálenosti.

$$D_9(x_1, x_2) = \frac{1}{2}\sum_{j=1}^{p}\left|\frac{y_{1j}}{\sum_{j=1}^{p} y_{ij}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}}\right|$$

- Jeho hodnota je 0 v případě identických proporcí druhů. Stejný výsledek lze získat i jako součet nejmenších podílů v rámci obou vzorků.

$$D_9(x_1, x_2) = \left[1 - \min\left(\frac{y_j}{\sum_{j=1}^{p} y_j}\right)\right]$$

Canberra metric (Lance & Williams 1966)

- Varianta Manhattanské vzdálenosti (před výpočtem musí být odstraněny double zero a není jimy tedy ovlivněna). Stejný rozdíl mezi početnými druhy ovlivňuje vzdálenost méně než mezi druhy vzácnějšími.

$$D_{10}(x_1, x_2) = \sum_{j=1}^{p} \left[\frac{\left| y_{1j} - y_{2j} \right|}{\left(y_{1j} + y_{2j} \right)} \right]$$

- Stephenson et al. (1972) a Moreau & Legendre (1979) použili tuto metriku jako součást koeficientu podobnosti

$$S(x_1, x_2) = 1 - \frac{1}{p} D_{10}$$

Koeficient divergence

- ☑ **Obdobná metrika jako D10 ale založená na Euklidovské vzdálenosti a vztažená na počet parametrů.**

$$D_{11}(x_1, x_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}}\right)^2}$$

Coefficient of racial likeness (Pearson 1926)

- Umožňuje srovnávat skupiny objektů podobně jako Mahalanobisova vzdálenost, ale na rozdíl od ní neeliminuje vliv korelace parametrů. Dvě skupiny objektů w1 a w2 jsou charakterizovány $\bar{y}_{ij}$ (průměr parametrů ve skupinách) a s_{ij}^2 (rozptyl parametrů ve skupinách).

$$D_{12}(w_1, w_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left[\frac{\left(\bar{y}_{1j} - \bar{y}_{2j}\right)^2}{\left(\frac{s_{1j}^2}{n_1}\right) + \left(\frac{s_{2j}^2}{n_2}\right)}\right] - \frac{2}{p}}$$

$\chi 2$ metrika (Roux & Reyssac 1975)

- ☑ První ze skupiny metrik založených na $\chi 2$ pro výpočet vzdáleností odběrů založených na abundancích druhů nebo jiných frekvenčních datech (nejsou přípustné žádné záporné hodnoty). Data původní matice abundancí/frekvencí Y jsou nejprve přepočítána do matice poměrných frekvencí (součty frekvencí v řádcích (odběry) jsou rovny 1). Jako dodatečné charakteristiky uplatňované při výpočtu jsou spočteny součty řádků yi+ a sloupců y+j celé! matice n(i) odběrů x p(j) druhů.

$$Y = \begin{bmatrix} y_{ij} \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} y_{i+} \\ \\ \\ \end{bmatrix} \rightarrow \begin{bmatrix} y_{ij} / y_{i+} \end{bmatrix}$$

$$\lfloor y_{+j} \qquad \lfloor y_{++}$$

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

- ☑ Výpočet odstraňuje problém double zero. Nejjednodušším výpočtem je obdoba Euklidovské vzdálenosti
- ☑ která je dále vážena součty jednotlivých druhů

$$D_{15}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

χ2 vzdálenost (Lébart & Fénelon 1971)

- Výpočet je podobný χ2 metrice, ale vážení je prováděno relativní četností řádku v matici místo jeho absolutního součtu, při výpočtu se užívá parametr y++ (celkový součet matice). Je využívána také při výpočtu vztahů řádků a sloupců kontingenční tabulky.

$$D_{16}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j} / y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

Hellingerova vzdálenost (Rao 1995)

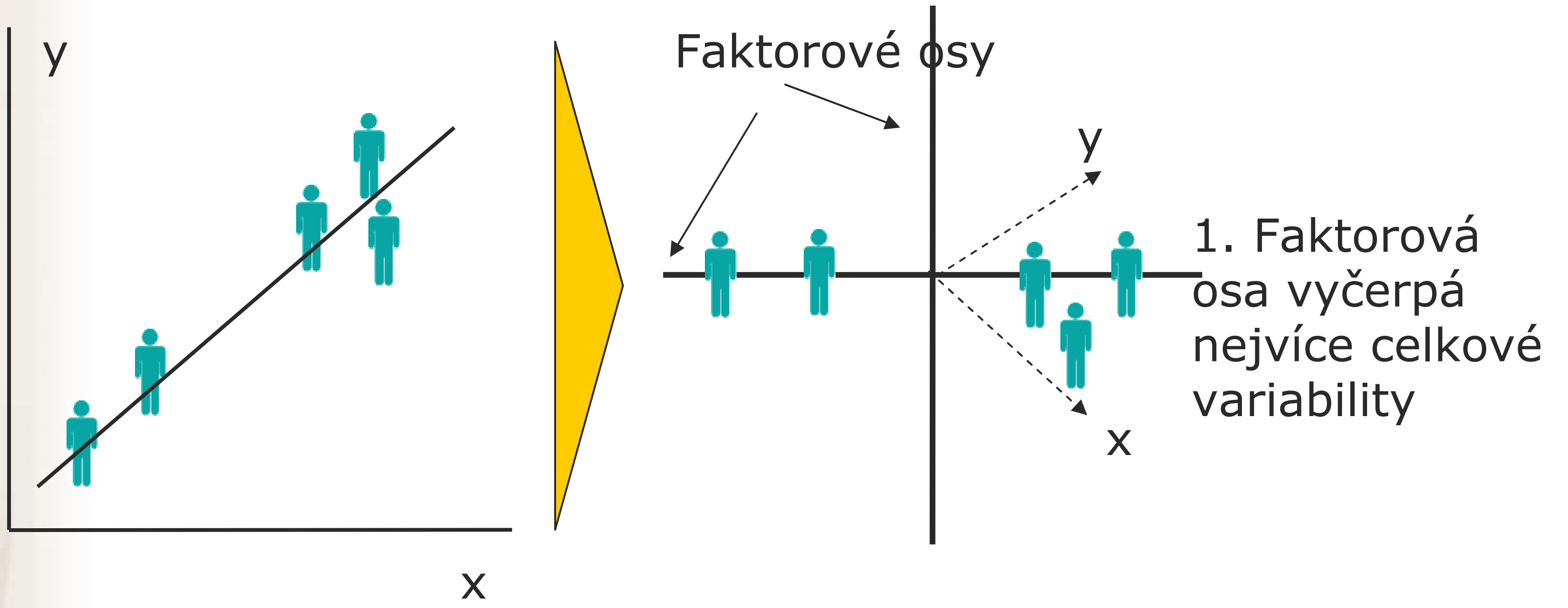
- ☑ **Koeficient související s D15 a D16.**

$$D_{17}(x_1, x_2) = \sqrt{\sum_{j=1}^{p}\left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}}\right]^2}$$

Analýza hlavních komponent
Faktorová analýza

- **Vstupní data:**
  - Spojité nebo dummy proměnné popisující jednotlivé respondenty
- **Výstupy analýzy**
  - Vztahy všech původních faktorů v jednoduchém xy grafu
  - Pozice respondentů v prostoru – jednoduchá identifikace segmentů a vlivů faktorů na různé skupiny
- **Kritické problémy analýzy**
  - Odlehlé hodnoty
  - Zcela nezávislé proměnné – není zde žádná duplicitní informace k vysvětlení

- **Proměnné jsou vzájemně korelovány, tedy část informace v souboru je duplicitní**
- **Analýza odstraní duplicitu z dat a zobrazí pouze unikátní informaci**



Vstupy výpočtu PCA

STATISTICA - [Data: Activities (12v by 28c)]

File Edit View Insert Format Statistics Graphs Tools Data Window Help

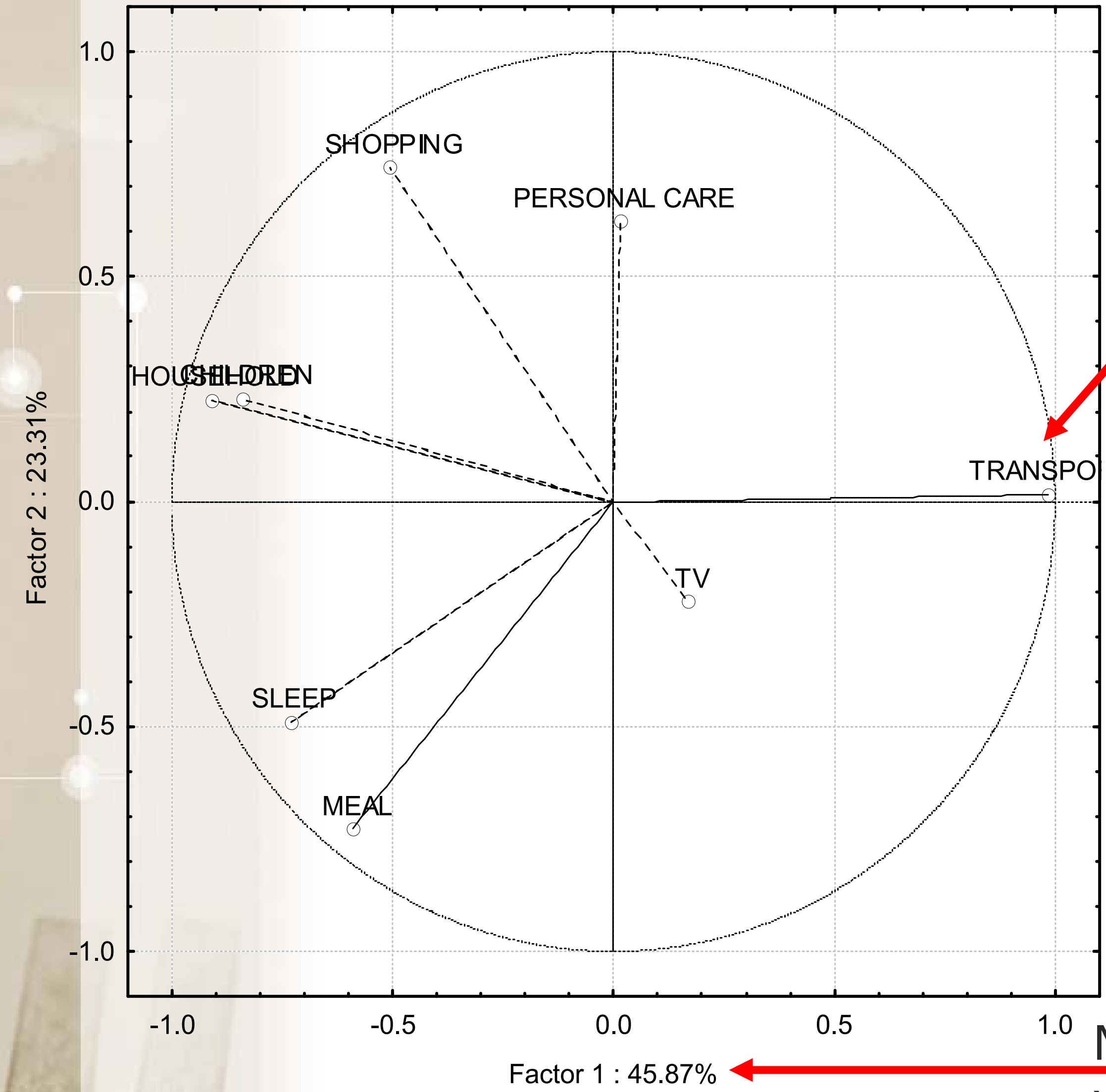
Activities timetable data for 28 population groups; modified example data reported in Exploratory and Multivariate Data

	1 WORK	2 TRANSPORT	3 HOUSEHOLD	4 CHILDREN	5 SHOPPING	6 PERSONAL CARE	7 MEAL	8 SLEEP	9 TV	10 LEISURE
EMU	610	140	60	10	120	95	115	760	175	315
EWU	475	90	250	30	140	120	100	775	115	[illegible]
UWU	10		495	110	170	110	130	785	160	[illegible]
MMU	615	141	65	10	115	90	115	765	180	305
MWU	179	29	421	87	161	112	119	776	143	373
SMU	585	115	50		150	105	100	760	150	385
SWU	482	94	196	18	141	130	96	775	132	336
EMW	652	100	95	7	57	85	150	807	115	330
EWW	510	70	307	30	80	95	142			
UWW	20	7	567	87	112	90	180			
MMW	655	97	97	10	52	85	152			
MWW	168	22	529	69	102	83	174			
SMW	642	105	72		62	77	140			
SWW	389	34	262	14	92	97	147			
EME	650	142	122	22	76	94	100			
EWE	578	106	338	42	106	94	92			
UWE	24	8	594	72	158	82	128			
MME	652	133	134	22	68	54	102			
MWE	434	77	431	60	117	88	105			
SME	627	148	68		88	92	86			
SWE	433	88	296	21	128	102	94			
EMY	650	140	120	15	85	90	105			
EWY	560	105	375	45	90	90	95			
UWY	10	10	710	55	145	85	130			
MMY	650	145	112	15	85	90	105			
MWY	260	52	576	59	116	85	117			
SMY	615	125	95		115	90	85			
SWY	433	89	318	23	112	96	102			

Vstupní tabulka spojitých dat

Nezbytnost analýzy vztahu proměnných – analýza předpokladů.

Výstupy analýzy hlavních komponent

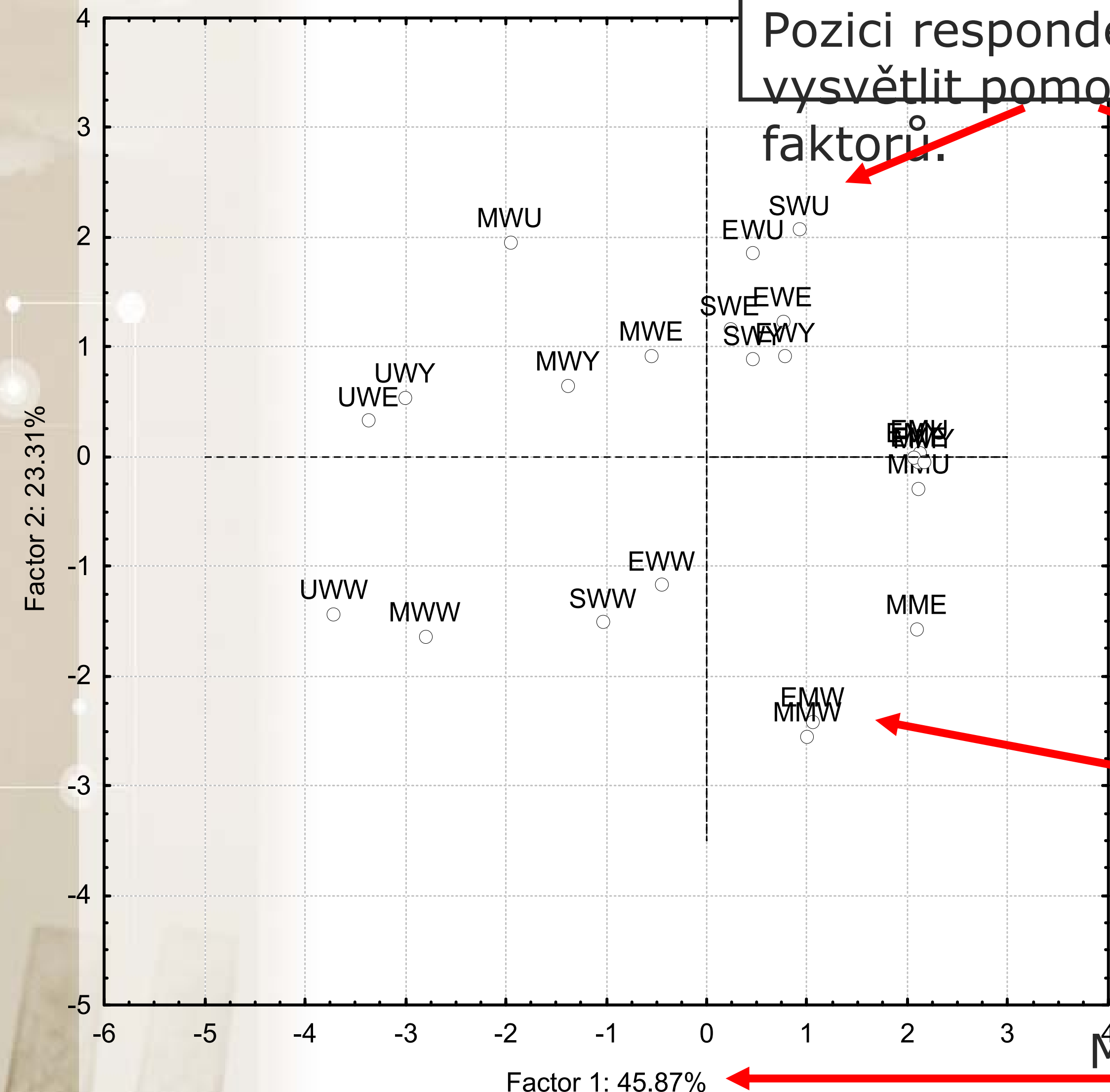


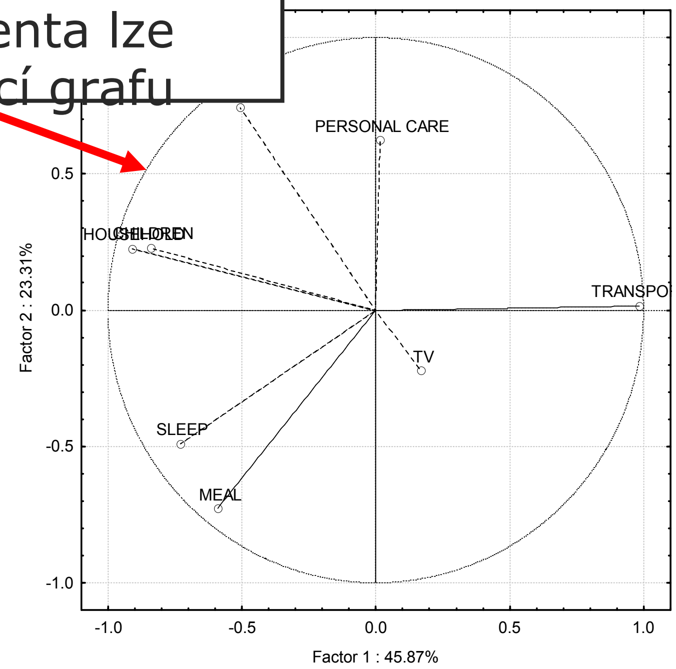
Pozice faktoru = míra vazby parametru s danou osou (-1,+1)
Důležitá pro interpretaci.

Množství vyčerpané variability (informační hodnota osy)

Výstupy analýzy hlavních komponent

Pozici respondenta lze vysvětlit pomocí grafu faktorů.



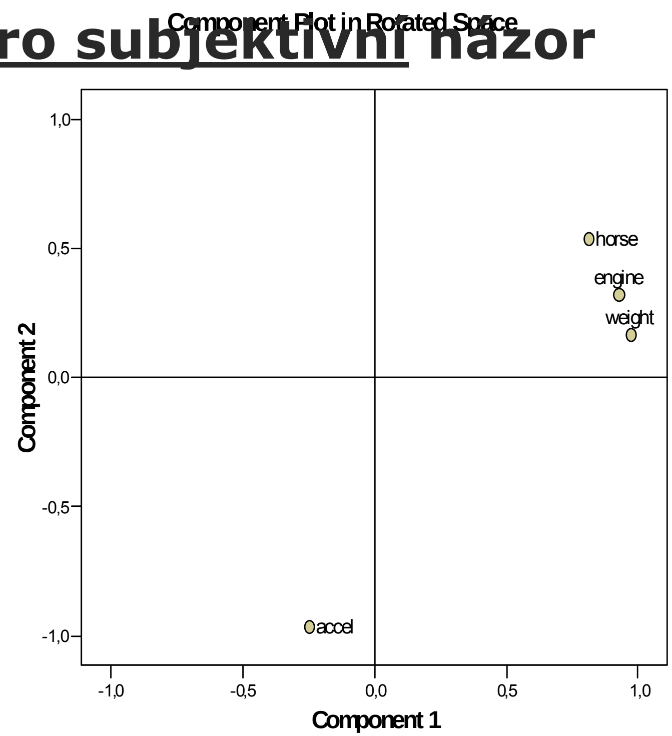


Pozice respondenta ve faktorovém prostoru

Množství vyčerpané variability (informační hodnota osy)

Čím se liší od analýzy hlavních komponent?

- Jediným rozdílem je rotace proměnných tak aby se vytvořené faktorové osy daly dobře interpretovat
- Výhodou je <u>lepší interpretace</u> vztahu původních proměnných
- Nevýhodou je <u>prostor pro subjektivní</u> názor analytika

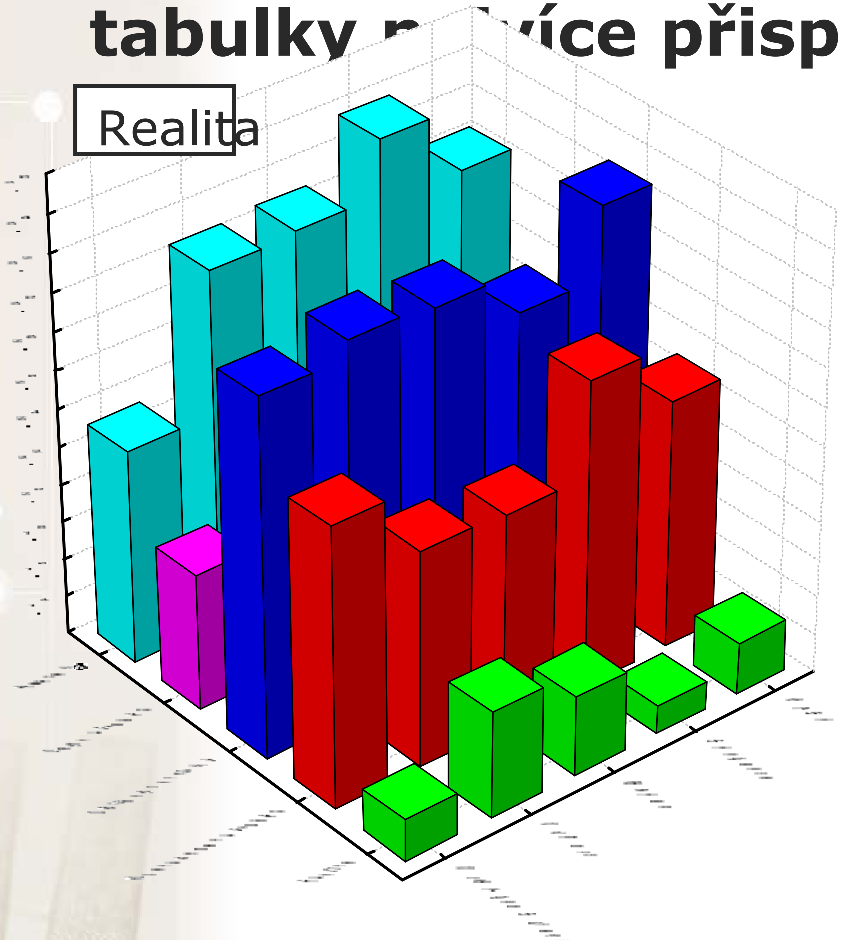


Korespondenční analýza

- **Vstupní data:**
  - **Tabulka obsahující souhrny proměnných (počty, průměry) za skupiny respondentů**
- **Výstupy analýzy**
  - **Vztahy všech původních faktorů a/nebo skupin respondentů v jednoduchém xy grafu**
- **Kritické problémy analýzy**
  - **Skupiny s malým počtem hodnot mohou být zatíženy značným šumem a náhodnou chybou**
  - **Obtížná interpretace velkého množství malých skupin respondentů**

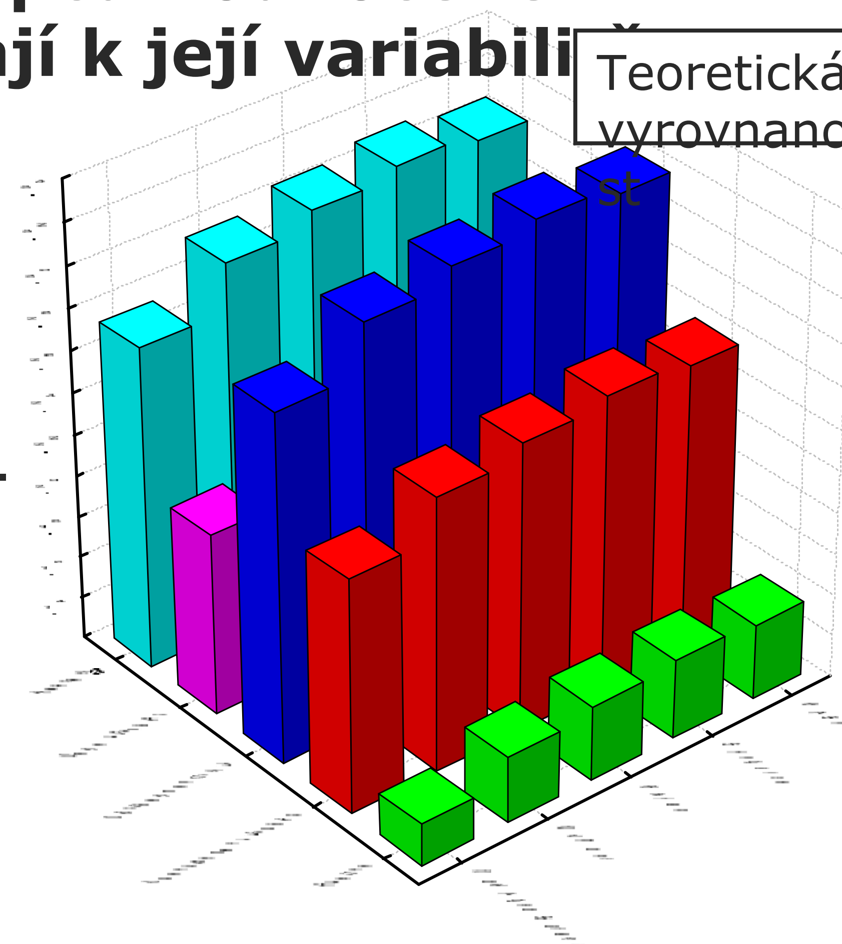
- Korespondenční analýza hledá, které kombinace řádků a sloupců hodnocené tabulky n… více přispívají k její variabili…

Realita



Vs.

Teoretická vyrovnanost

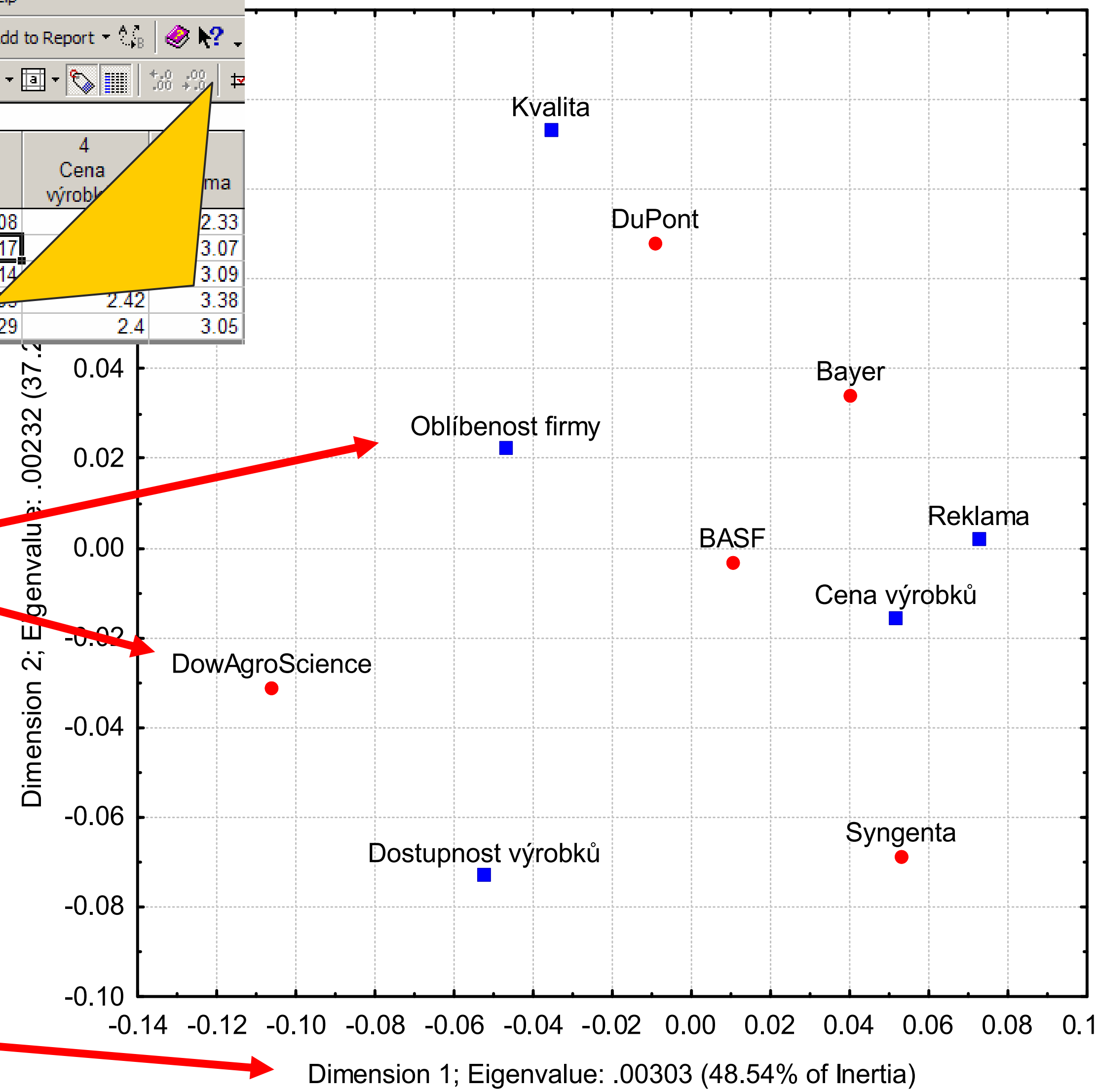


Výstupy korespondenční analýzy

STATISTICA - [Data: mark_pruzkum* (5v by 5c)]

File Edit View Insert Format Statistics Graphs Tools Data Window Help

	1 Kvalita	2 Dostupnost výrobků	3 Oblíbenost firmy	4 Cena výrobk	ma
DowAgro Science	1.42	2.67	3.08	[illegible]	2.33
DuPont	1.76	2.34	3.17	[illegible]	3.07
Bayer	1.62	2.32	3.14	[illegible]	3.09
Syngenta	1.35	2.81	[illegible]	2.42	3.38
BASF	1.47	2.51	3.29	2.4	3.05



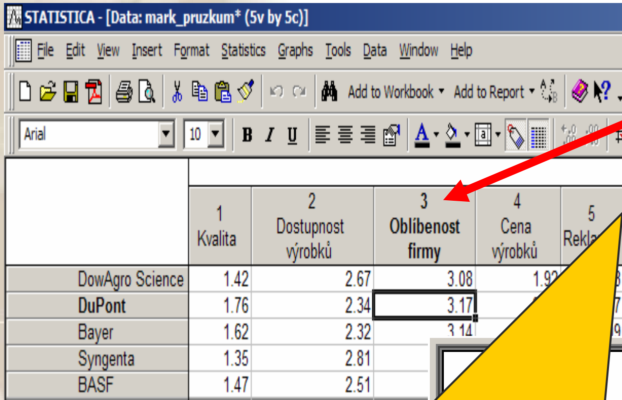
Vzájemná pozice faktorů a skupin respondentů: vzájemnou pozici lze interpretovat

Variabilita vyčerpaná danou faktorovou osou

Shluková analýza

- ☑ **Vstupní data:**
  - ➔ Tabulka spojitých nebo kategoriálních dat popisujících respondenty nebo jejich skupiny
- ☑ **Výstupy analýzy**
  - ➔ Tzv. dendrogram popisující vazby mezi respondenty nebo parametry
  - ➔ Rozdělení respondentů nebo parametrů do daného počtu skupin
- ☑ **Kritické problémy analýzy**
  - ➔ Velké množství parametrů nebo respondentů v dendrogramu je obtížně interpretovatelné
  - ➔ Analýza je silně závislá na zvolení vhodné metriky vzdáleností
  - ➔ Analýza je silně závislá na shlukovacím algoritmu

Postup výpočtu hierarchické shlukové analýzy



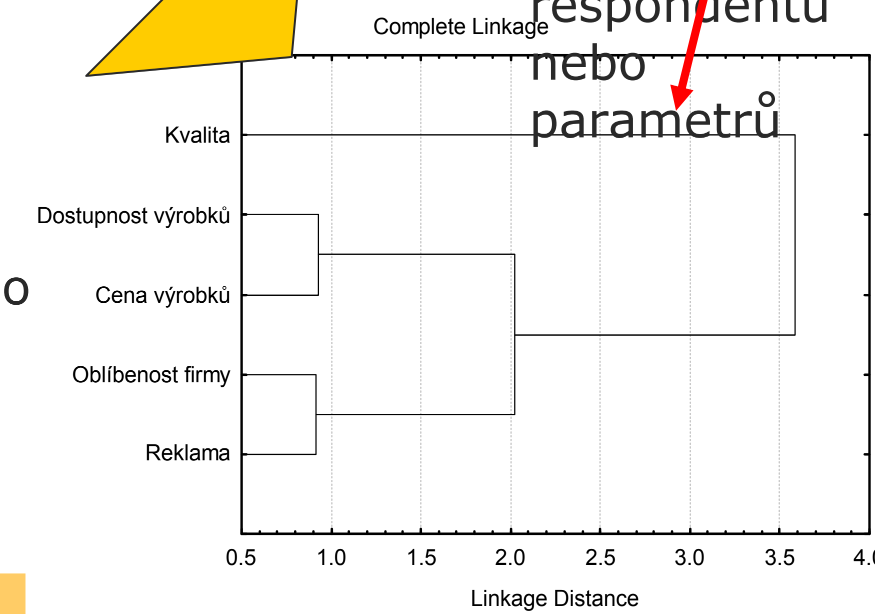
	1 Kvalita	2 Dostupnost výrobků	3 Oblíbenost firmy	4 Cena výrobků	5 Rekla
DowAgro Science	1.42	2.67	3.08	1.92	
DuPont	1.76	2.34	3.17		
Bayer	1.62	2.32	3.14		
Syngenta	1.35	2.81			
BASF	1.47	2.51			

Vstupní datová tabulka

Matice vzdáleností

Variable	Euclidean distances (mark_pruzkum) Kvalita	Dostupnost výrobků	Oblíbenost firmy	Cena výrobků	Reklama
Kvalita	0.00	2.37	3.59	1.77	3.36
Dostupnost výrobků	2.37	0.00	1.47	0.93	1.36
Oblíbenost firmy	3.59	1.47	0.00	2.02	0.91
Cena výrobků	1.77	0.93	2.02	0.00	1.71
Reklama	3.36	1.36	0.9	1.71	0.00

Dendrogram – schéma podobnosti respondentů nebo parametrů



Výběr vhodné metriky vzdáleností je klíčový pro výsledek shlukové analýzy – různé typy proměnných vyžadují různé metriky vzdáleností

Shlukovací pravidlo je dalším velmi důležitým krokem při shlukové analýze a může změnit její

Někter

Euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Vážená euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2}$$

i,j – označení objektů
d_{ij} – vzdálenost objektů i a j
p – počet parametrů
k – k-tý parametr
w_k – váha parametru k

Minkowski (power distance)

$$d_{ij} = \sqrt[\frac{1}{\lambda}]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\lambda}}$$

- - celé číslo
- =1 Manhattan (city block)
- = 2 Euklidovská vzdálenost

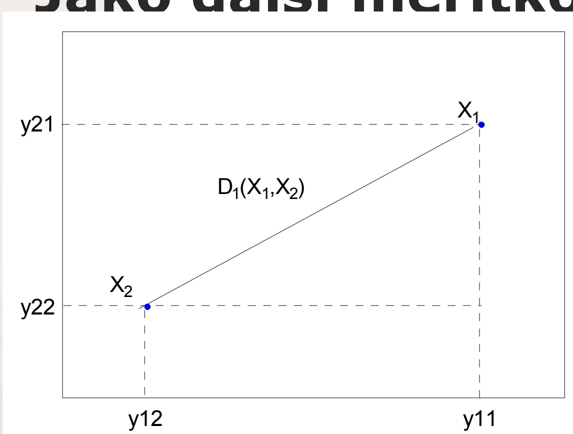
Chebychev

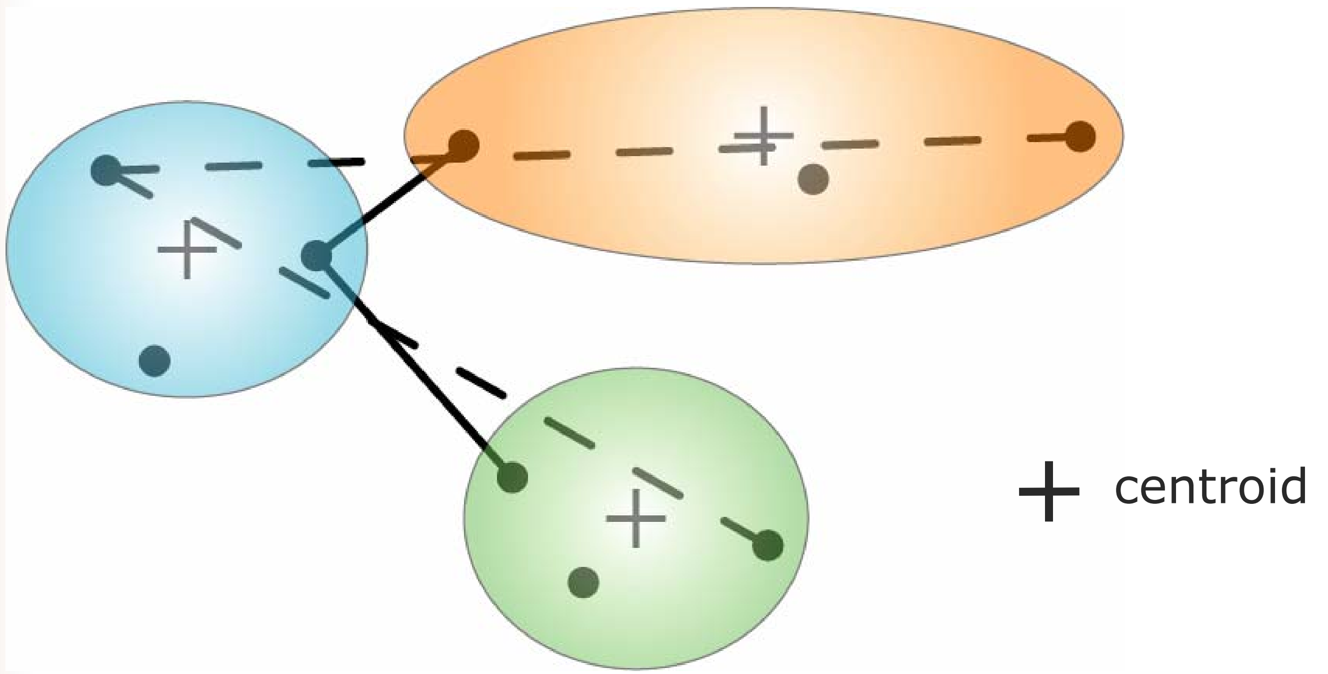
$$d_{ij} = \max \left| x_{ik} - x_{jk} \right|$$

Euklido
víceroz

- Jde o základní metrické měřítko vzdálenosti a počítá vzdálenost objektů obdobně jako Pythagorova věta počítá přeponu pravoúhlého trojúhelníku. Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a double zero problém. Nemá horní hranici hodnot. $D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$

- Jako další měřítko se používá také čtverec této … nevýhodou jsou semimetrické $D_1^2(x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$

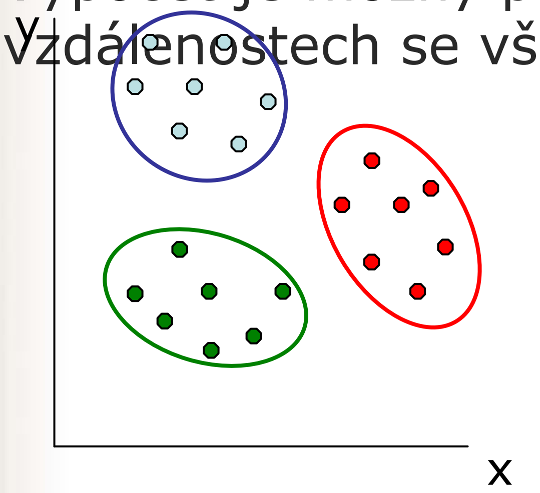


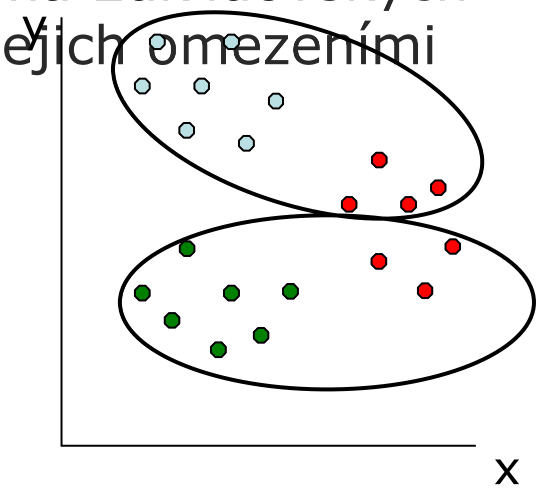


Na tuto vzdálenost se ptá **single linkage**
Na tuto vzdálenost se ptá **complete linkage**

Další metody počítají s **průměrnou vzdáleností** všech objektů shluků nebo vzdáleností **centroidů** (vzdálenost může být **vážena** velikostí shluků). **Wardova metoda** se snaží minimalizovat variabilitu uvnitř shluků.

- Respondenti jsou na základě zadaného počtu shluků rozděleni podle kritéria maximální homogenity shluků
- Rizika analýzy
  - Při špatném odhadu počtu shluků dává metoda chybné výsledky
  - Výpočet je možný pouze na Euklidovských vzdálenostech se všemi jejich omezeními

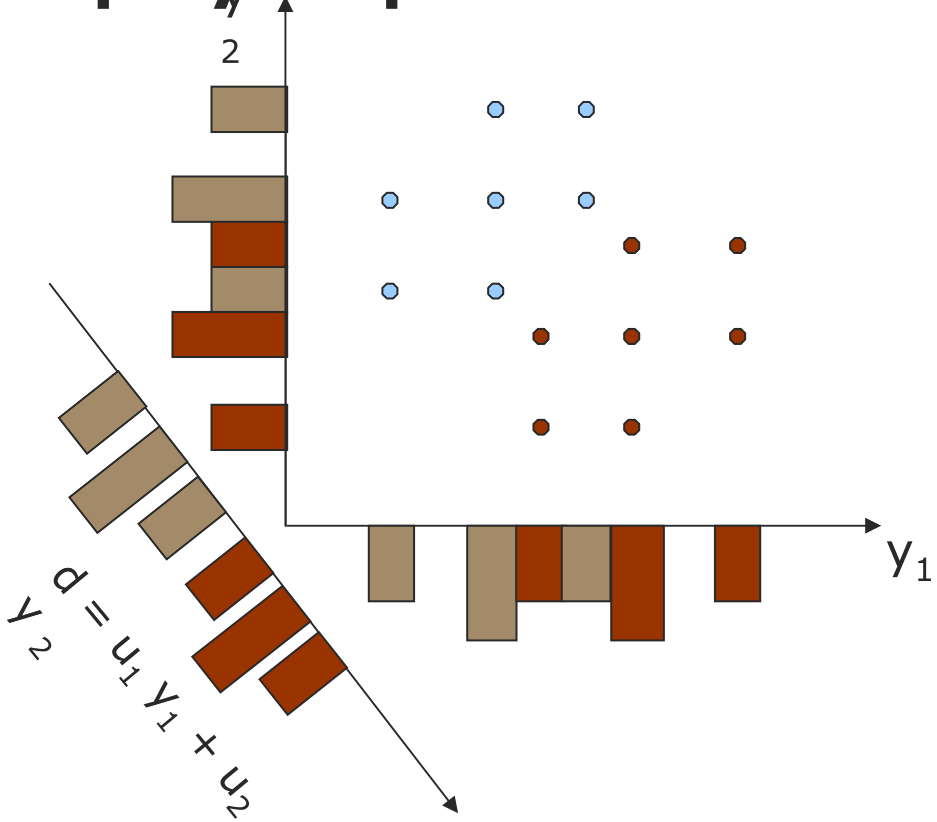




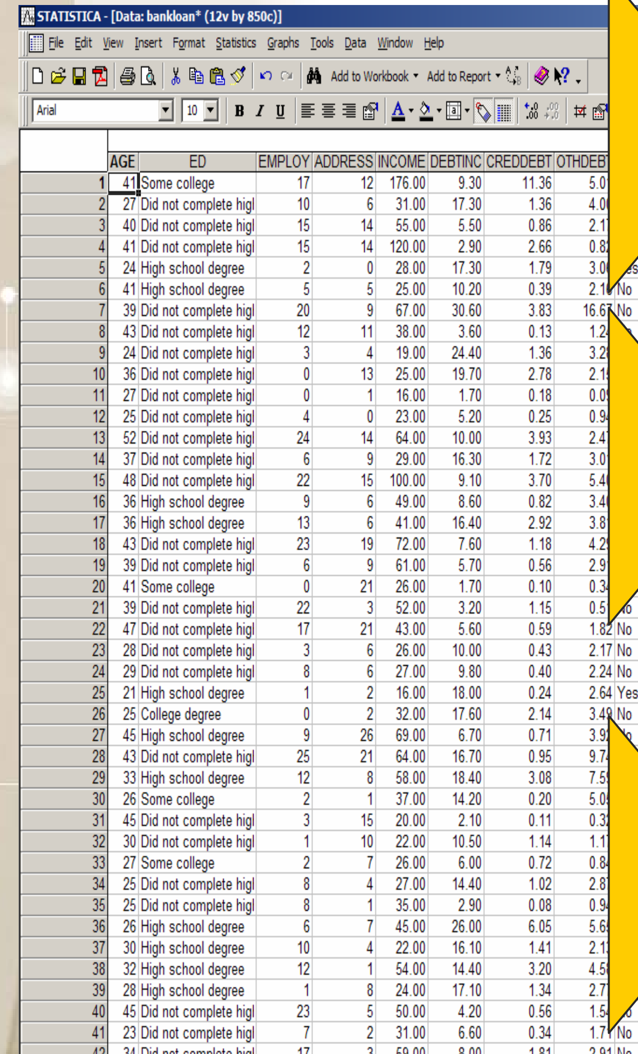
Diskriminační analýza

- **Vstupní data:**
  - **Tabulka spojitých dat popisujících respondenty**
  - **Respondenti jsou rozděleni do předem daných skupin**
- **Výstupy analýzy**
  - **Seznam parametrů významně rozlišujících různé skupiny respondentů**
  - **Zobrazení pozice respondentů v diskriminačním prostoru**
  - **Model pro zařazení nových respondentů do skupin**
- **Kritické problémy analýzy**
  - **Odlehlé hodnoty a asymetrické rozložení uvnitř skupin respondentů**
  - **Silná korelace mezi prediktory**
  - **Nutná expertní znalost významu parametrů**
  - **Pro tvorbu diskriminačních modelů pro praktické**

- ☑ **Analýza nachází takovou kombinaci vstupních parametrů, která odděluje od sebe skupiny respondentů**



Výstupy diskriminační analýzy



	AGE	ED	EMPLOY	ADDRESS	INCOME	DEBTINC	CREDDEBT	OTHDEBT	
1	41	Some college	17	12	176.00	9.30	11.36	[illegible]	
2	27	Did not complete higl	10	6	31.00	17.30	1.36	[illegible]	
3	40	Did not complete higl	15	14	55.00	5.50	0.86	[illegible]	
4	41	Did not complete higl	15	14	120.00	2.90	2.66	[illegible]	
5	24	High school degree	2	0	28.00	17.30	1.79	[illegible]	[illegible]
6	41	High school degree	5	5	25.00	10.20	0.39	[illegible]	No
7	39	Did not complete higl	20	9	67.00	30.60	3.83	[illegible]	No
8	43	Did not complete higl	12	11	38.00	3.60	0.13	[illegible]	
9	24	Did not complete higl	3	4	19.00	24.40	1.36	[illegible]	
10	36	Did not complete higl	0	13	25.00	19.70	2.78	[illegible]	
11	27	Did not complete higl	0	1	16.00	1.70	0.18	[illegible]	
12	25	Did not complete higl	4	0	23.00	5.20	0.25	[illegible]	
13	52	Did not complete higl	24	14	64.00	10.00	3.93	[illegible]	
14	37	Did not complete higl	6	9	29.00	16.30	1.72	[illegible]	
15	48	Did not complete higl	22	15	100.00	9.10	3.70	[illegible]	
16	36	High school degree	9	6	49.00	8.60	0.82	[illegible]	
17	36	High school degree	13	6	41.00	16.40	2.92	[illegible]	
18	43	Did not complete higl	23	19	72.00	7.60	1.18	[illegible]	
19	39	Did not complete higl	6	9	61.00	5.70	0.56	[illegible]	
20	41	Some college	0	21	26.00	1.70	0.10	[illegible]	
21	39	Did not complete higl	22	3	52.00	3.20	1.15	[illegible]	[illegible]
22	47	Did not complete higl	17	21	43.00	5.60	0.59	1.82	No
23	28	Did not complete higl	3	6	26.00	10.00	0.43	2.17	No
24	29	Did not complete higl	8	6	27.00	9.80	0.40	2.24	No
25	21	High school degree	1	2	16.00	18.00	0.24	2.64	Yes
26	25	College degree	0	2	32.00	17.60	2.14	3.49	No
27	45	High school degree	9	26	69.00	6.70	0.71	[illegible]	[illegible]
28	43	Did not complete higl	25	21	64.00	16.70	0.95	[illegible]	
29	33	High school degree	12	8	58.00	18.40	3.08	[illegible]	
30	26	Some college	2	1	37.00	14.20	0.20	[illegible]	
31	45	Did not complete higl	3	15	20.00	2.10	0.11	[illegible]	
32	30	Did not complete higl	1	10	22.00	10.50	1.14	[illegible]	
33	27	Some college	2	7	26.00	6.00	0.72	[illegible]	
34	25	Did not complete higl	8	4	27.00	14.40	1.02	[illegible]	
35	25	Did not complete higl	8	1	35.00	2.90	0.08	[illegible]	
36	26	High school degree	6	7	45.00	26.00	6.05	[illegible]	
37	30	High school degree	10	4	22.00	16.10	1.41	[illegible]	
38	32	High school degree	12	1	54.00	14.40	3.20	[illegible]	
39	28	High school degree	1	8	24.00	17.10	1.34	[illegible]	
40	45	Did not complete higl	23	5	50.00	4.20	0.56	[illegible]	[illegible]
41	23	Did not complete higl	7	2	31.00	6.60	0.34	[illegible]	No
42	34	Did not complete higl	17	3	59.00	8.00	1.81	2.91	No

Discriminant Function Analysis Summary (bankloan)
No. of vars in model: 4; Grouping: DEFAULT (2 grps)
Wilks' Lambda: .82129 approx. F (4,695)=37.808 p<0.0000

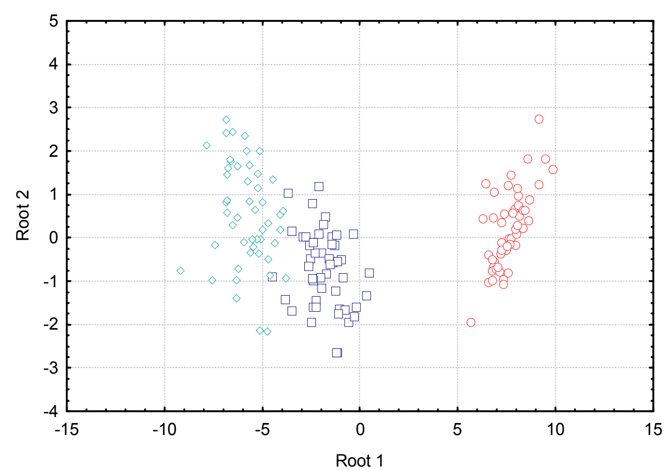
N=700	Wilks' Lambda	Partial Lambda	F-remove (1,695)	p-level	Toler.	1-Toler. (R-Sqr.)
INCOME	0.823311	0.997544	1.71119	0.191263	0.285507	0.714493
DEBTINC	0.865135	0.949319	37.10408	0.000000	0.351216	0.648784
CREDDEBT	0.837204	0.980990	13.46809	0.000261	0.422380	0.577620
OTHDEBT	0.826985	0.993111	4.82090	0.028446	0.267552	0.732448

Význam parametrů pro klasifikaci

Classification Matrix (Irisdat)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
SETOSA	100.0000	50	0	0
VERSICOL	96.0000	0	48	2
VIRGINIC	98.0000	0	1	49
Total	98.0000	50	49	51

Predikční schopnost modelu



Pozice v diskriminačním prostoru

- **Analýza hlavních komponent, faktorová analýza, korespondenční analýza a diskriminační analýza se snaží zjednodušit vícerozměrnou strukturu dat výpočtem souhrnných os**
- **Metody se liší v logice tvorby těchto os**
  - **Maximální variabilita (analýza hlavních komponent, korespondenční analýza)**
  - **Maximální interpretovatelnost os (faktorová analýza)**
  - **Maximální diskriminace skupin (diskriminační analýza)**