Bi 8600 Vícerozměrné statistické metody

Danka Haruštiaková

Podzim 2009

Inštitút bioštatistiky a analýz, Masarykova univerzita

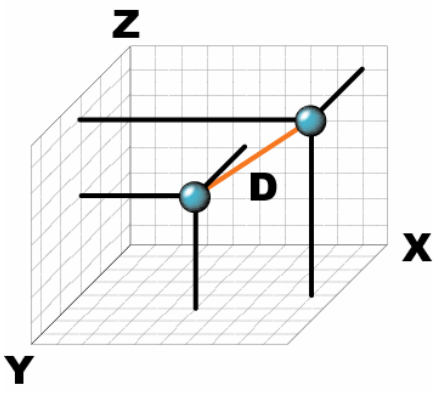
Plán kurzu

- **Rozvrh**:
  3.11.
  10.11.
  24.11.
  1.12.
  15.12.

- **Ukončenie**:
  písomná skúška zameraná na princípy a aplikáciu analýz

- **Cieľ kurzu**:
  vysvetliť princípy viacrozmerných analýz, ich aplikácie v biológii a ich interpretácie
  prehľad základného software
  príklady na reálnych dátach

Úvod



Vzťah klasickej a viacrozmernej štatistiky

- **Viacrozmerná analýza dát využíva prístupy klasickej štatistiky**
- Viacrozmerná analýza dát je **zároveň citlivá na problémy klasickej štatistiky**

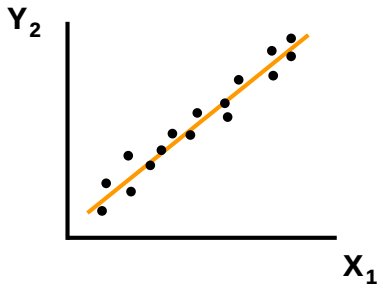
Kontingenčná tabuľka

Dôchodkový vek

Nákup	Áno	Nie	S
Áno	**20**	**82**	**102**
Nie	**10**	**54**	**64**
S	**30**	**136**	**166**

- Agregácia dát cez sumárnu štatistiku alebo **kontingenčné tabuľky** – korešpondenčná analýza

Korelácia



- **Korelácie** - analýza hlavných komponent, faktorová analýza, diskriminační analýza

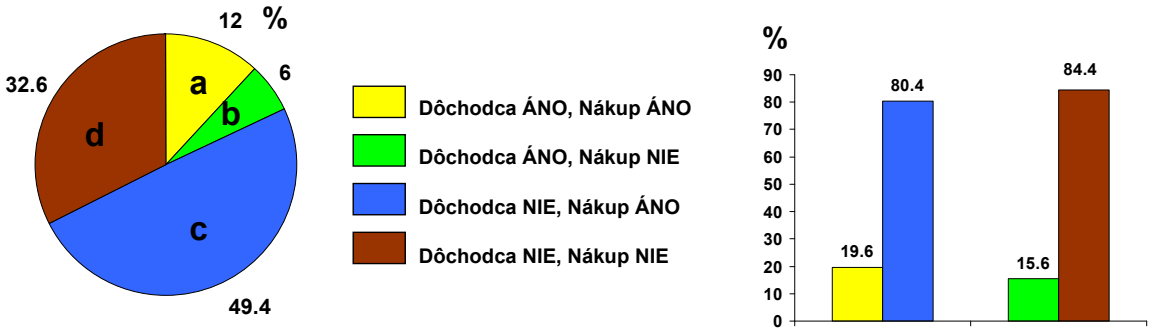
Kontingenčná tabuľka

Dôchodkový vek

Nákup	Áno	Nie	S
Áno	20	82	102
Nie	10	54	64
S	30	136	166

- Kontingenčná tabuľka je používaná pre hodnotenie vzťahu kategoriálnych premenných

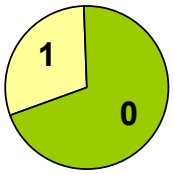
Kontingenčná tabuľka v obrázku



Kontingenčné tabuľky – princíp analýzy

Binomické javy (1/0)

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



I. jav 1 — II. jav 2

Príklad

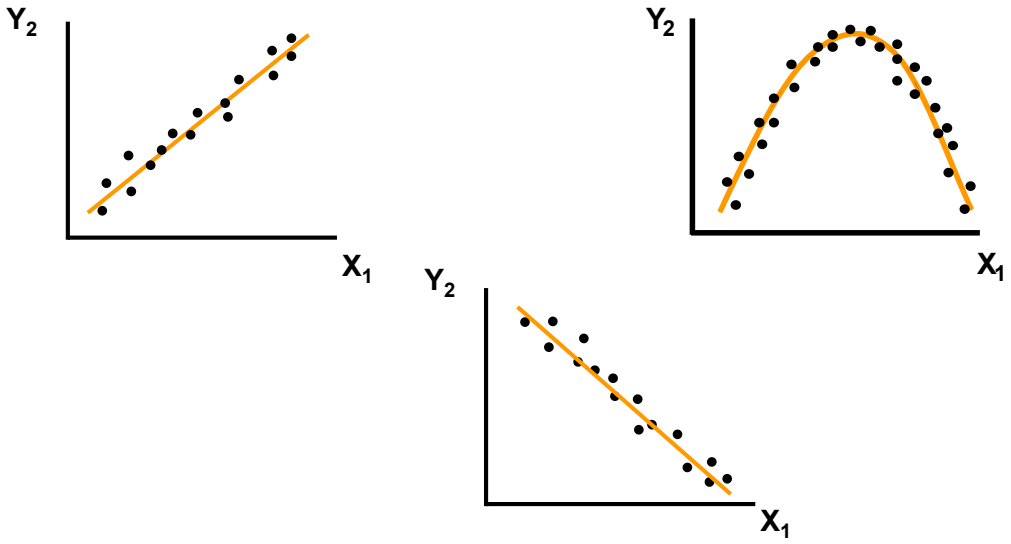
✓ 10 000 ľudí hádže mincou → rub: 4 000 prípadov (R); líc: 6 000 prípadov (L)

? Dá sa výsledok považovať za štatisticky významne odlišný (alebo neodlišný) od očakávaného pomeru R : L = 1 : 1 ?

Rovnakým spôsobom, teda hodnotením odchýlok od očakávaného vyrovnaného počtu prípadov hodnotí dáta i korešpondenčná analýza

Korelačná analýza

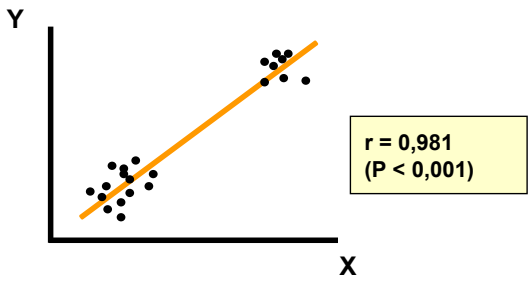
- **Korelácia** - vzťah (závislosť) dvoch znakov (parametrov)



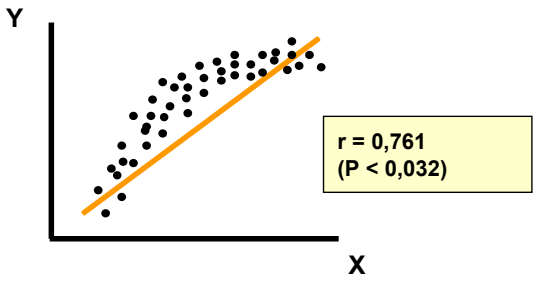
Korelácie medzi parametrami sú základom faktorovej analýzy a analýzy hlavných komponentov. Pokiaľ väzby medzi parametrami nie sú, tak tieto metódy strácajú zmysel.

Rizika korelačnej analýzy

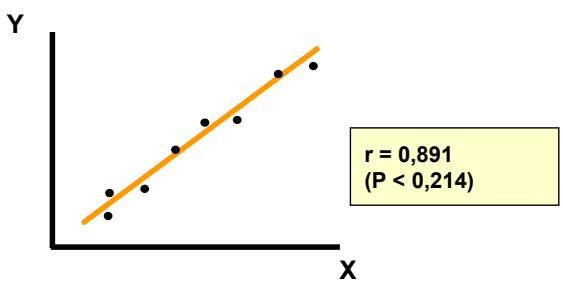
Problém rozloženia hodnôt

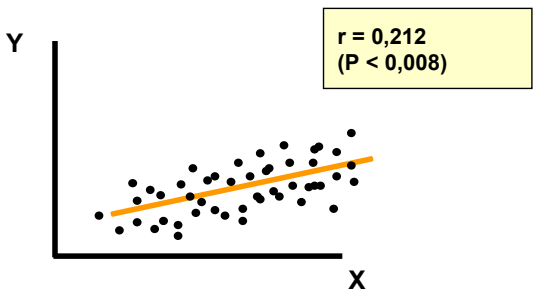


Problém typu modelu

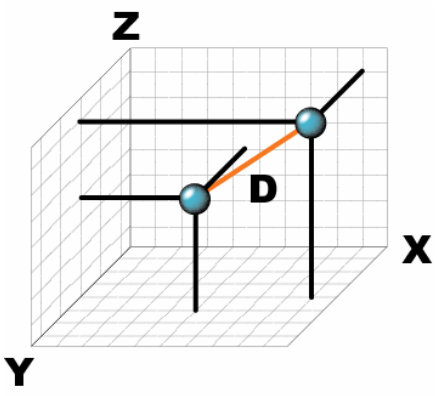


Problém veľkosti vzorky

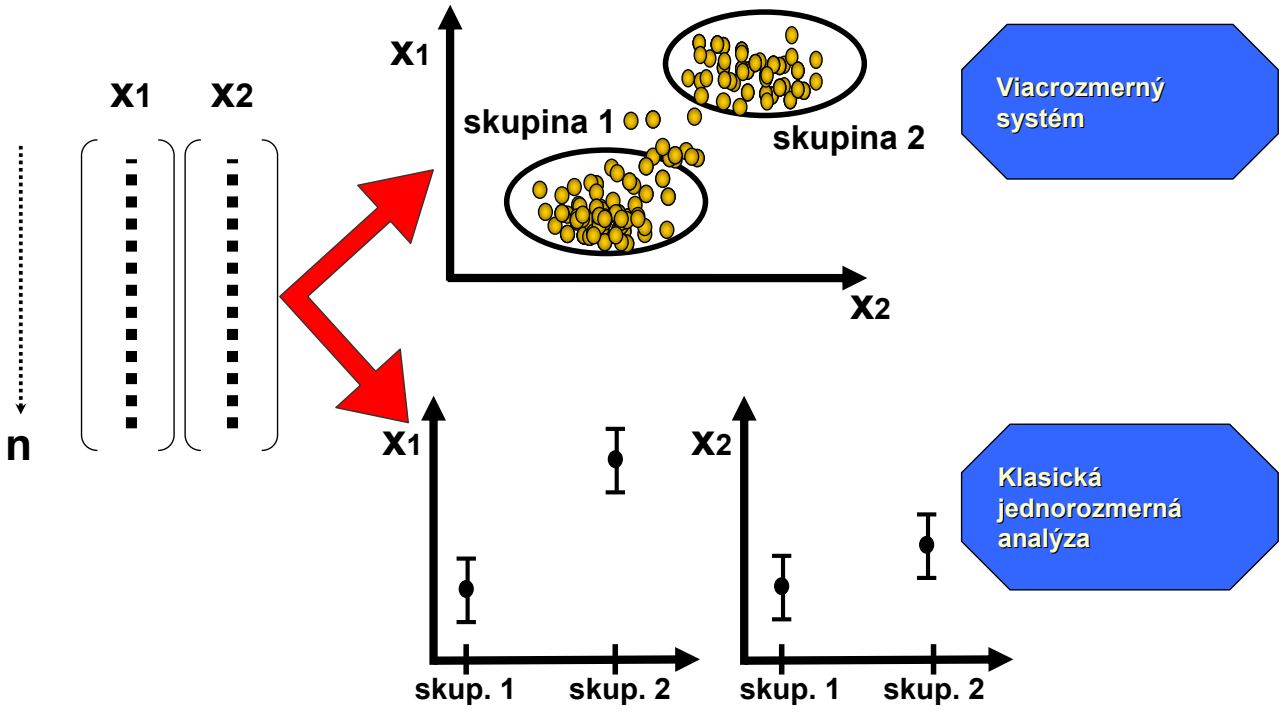




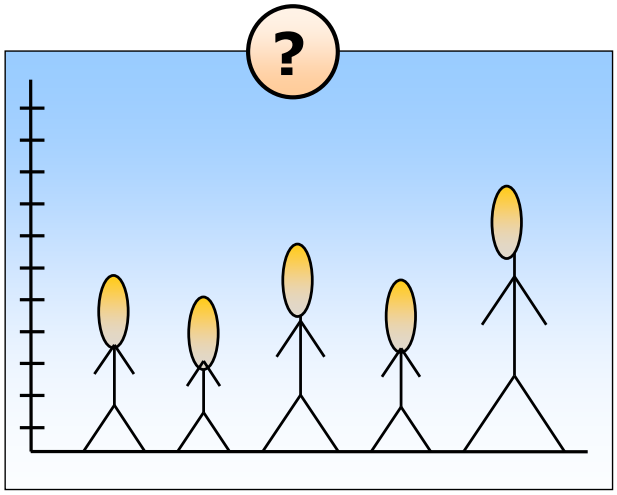
Význam viacrozmerného hodnotenia dát



Viacrozmerné vnímanie skutočnosti



Bežná sumarizácia dát „likviduje“ individualitu jedinca



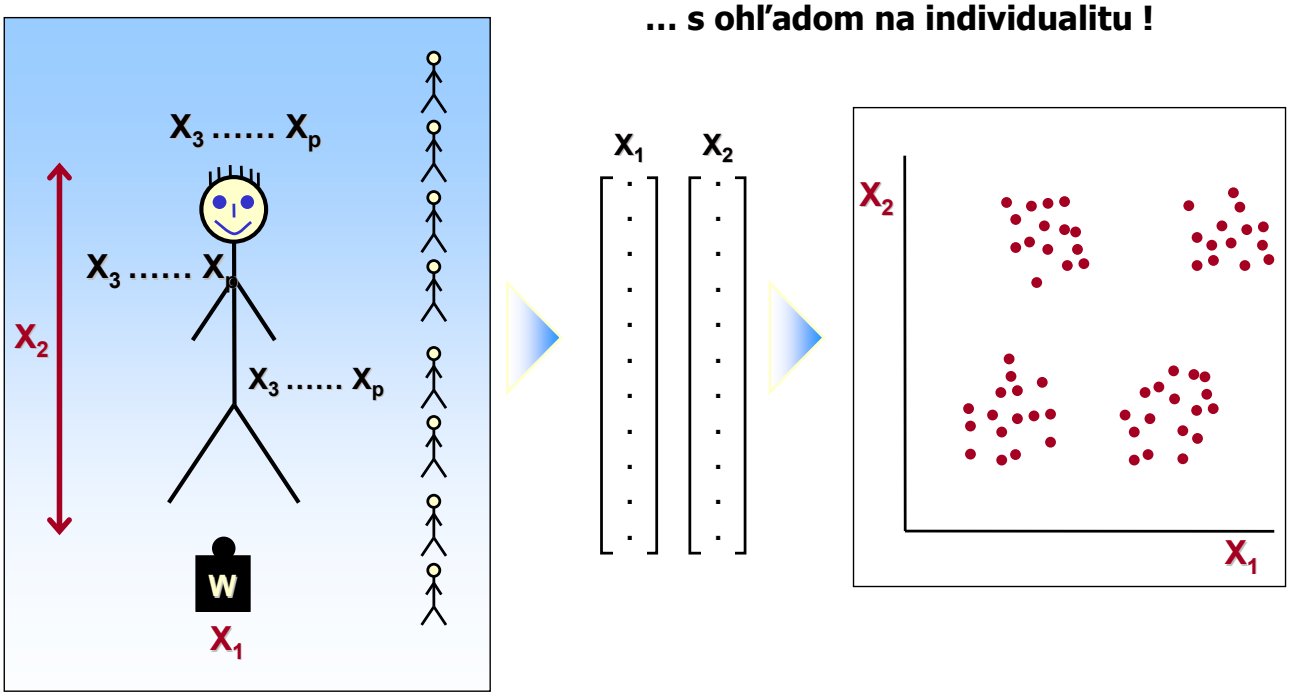
Priemer ± SE

BEŽNÁ ŠTATISTICKÁ SUMARIZÁCIA

- ✓ *Sprehľadnenie dát*
- ✓ *Neodlíši pôvodné meranie*

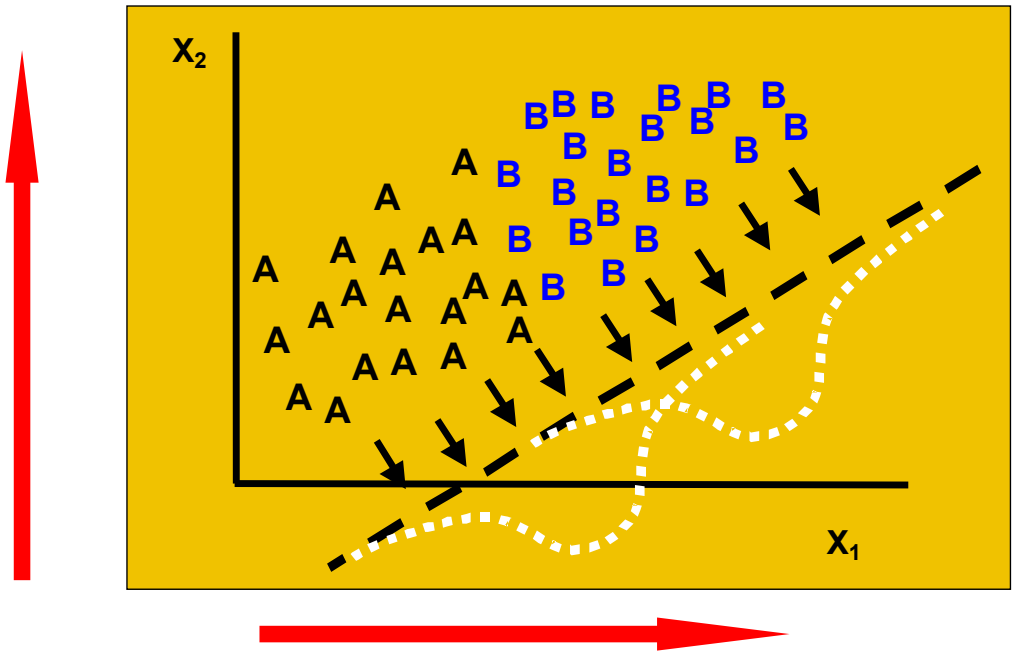
Viacrozmerné hodnotenie

… s ohľadom na individualitu !



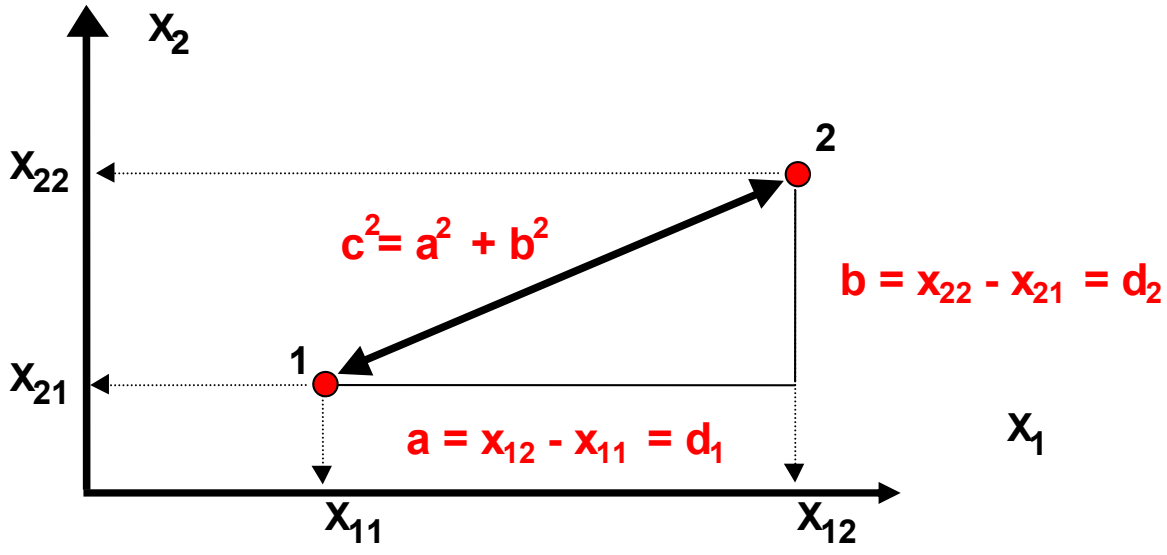
Viacrozmerné hodnotenie – nová kvalita

Len kombinované parametre majú odpovedajúcu informačnú silu



Viacrozmerné hodnotenie vychádza z jednoduchých princípov

Príklad: viacrozmerná vzdialenosť merania medzi dvoma objektami



Viacrozmerné modelovanie je strategickou disciplínou

$X_1 \ldots\ldots X_n$

technické parametre automobilu

$X_{n+1} \ldots\ldots X_p$

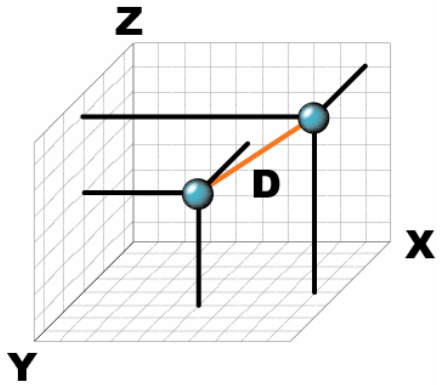
vodičove schopnosti a jeho stav

$X_{p+1} \ldots\ldots X_2$

rýchlosť, povrch, situácia

X_1 X_2 X_3 X_4 X_5 ..... X_p

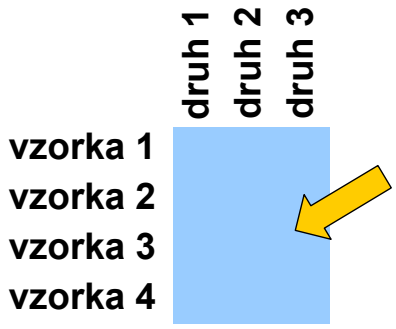
Základné princípy viacrozmerného hodnotenia dát



Pojmy vo viacrozmerných analýzach

- **Viacrozmerné metódy**:
  názov „viacrozmerné“, mnohorozmerné“ – vychádza z typu vstupných dát - dáta sú tvorené objektami (vzorky, lokality), každý z nich je charakterizovaný viacerými parametrami (druhmi)
  každý z týchto parametrov môžeme považovať za jeden rozmer objektu (vzorky)

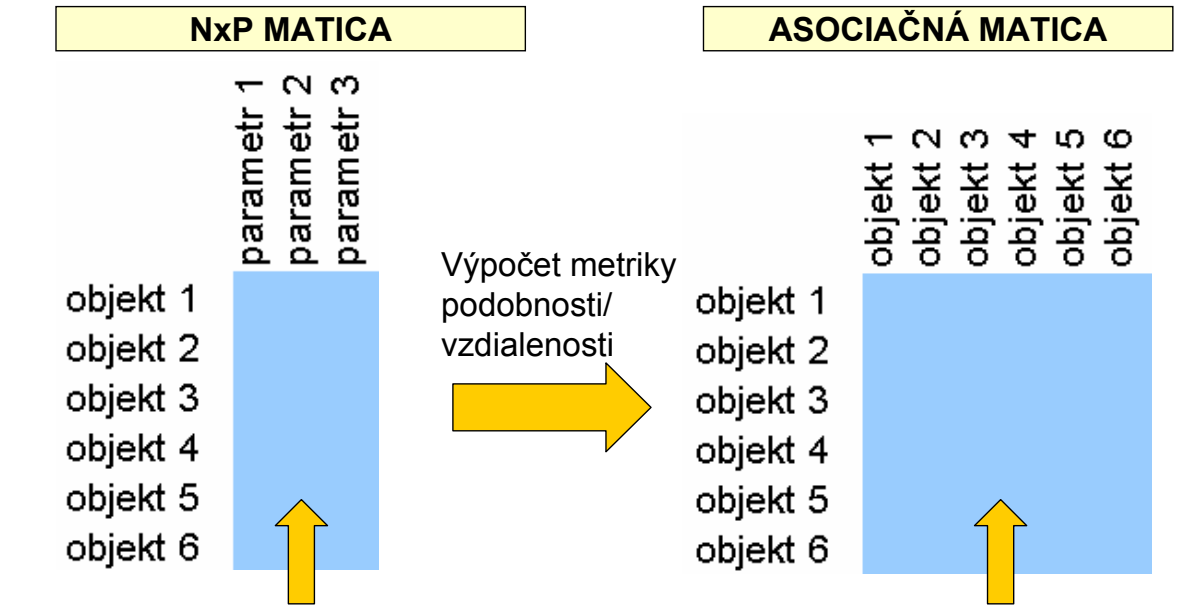
- **Maticová algebra**:
  Základom práce s dátami a výpočtami viacrozmerných metód je maticová algebra. Matice tvoria vstupné aj výstupné dáta a prebiehajú na nich výpočty.



Hodnoty pre druhy (presencia/absencia; abundancia; dominancia) pre každú vzorku

Vstupná matica viacrozmerných analýz

- **Dátová matica** – N objektov, P parametrov



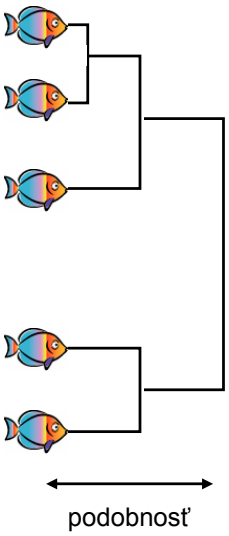
Hodnoty parametrov pre jednotlivé objekty

Korelácia, kovariancia, vzdialenosť, podobnosť

Typy viacrozmerných analýz

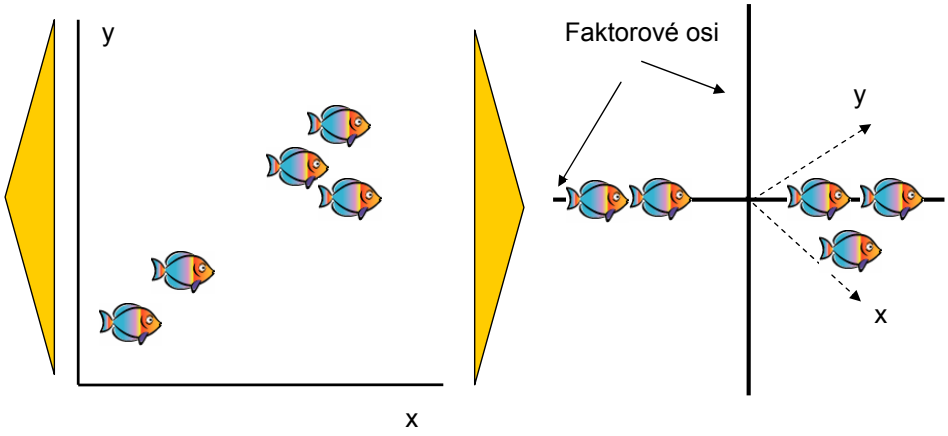
ZHLUKOVÁ ANALÝZA

- Vytvára zhluky objektov na základe ich podobnosti
- Identifikuje typy objektov

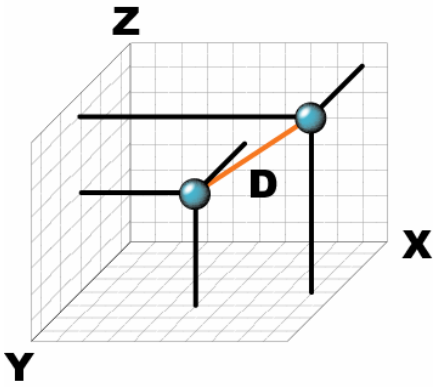


ORDINAČNÉ METÓDY

- Vytvárajú nové rozmery, ktoré lepšie vyčerpávajú variabilitu dát – zjednodušujú viacrozmerný priestor



Podobnosť a vzdialenosť objektov v mnohorozmernom priestore



Podobnosť a vzdialenosť

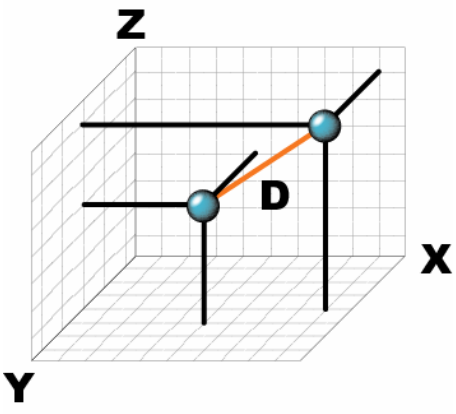
- Veľmi dôležitým pojmom je pojem podobnosti medzi jednotlivými objektami (miera podobnosti objektov).
- V literatúre sa možno stretnúť s troma základnými typmi popisu podobnosti – nepodobnosti objektov:

1. koeficienty asociácie
2. koeficienty korelácie
} **miery podobnosti**

3. metriky — **miery nepodobnosti**

Zoznam taxónov: viacrozmerný popis spoločenstva

- Na zoznam taxónov sa dá pozerať tiež ako na zoznam rozmerov spoločenstva
- Záznam o nájdených taxónoch tak vlastne tvorí viacrozmerný popis daného spoločenstva
- Spoločenstvá môžeme porovnávať podľa ich vzájomnej pozície v n-rozmernom priestore
- Pre porovnanie spoločenstiev sa dá teoreticky použiť ľubovoľný koeficient/metrika viacrozmernej podobnosti alebo vzdialenosti



Problém dvoch núl (double zero problem)

- V prípade binárnych koeficientov (druh sa vyskytuje/nevyskytuje) nie je možné uvažovať rovnakou váhou pre súhlas prítomnosti (11) a neprítomnosti (00) taxónov (symetrický koeficient)
- Problémom využitia všetkých typov metrík pre dáta abundancií spočíva v odlišnom význame prítomnosti a neprítomnosti taxónov
- Pokiaľ sa taxón nachádza v oboch porovnávaných spoločenstvách – znamená to, že spoločenstvá si budú v tomto ohľade podobné, pretože majú podmienky umožňujúce prítomnosť taxónu
- Pokiaľ sa taxón nenachádza ani v jednom z dvoch porovnávaných spoločenstiev – príčina môže byť najrôznejšia – **double zero problém**
- Pre odstránenie tohto problému sa používajú asymetrické koeficienty - hodnotenie súhlasnej prítomnosti (11) a neprítomnosti (00) taxónov nie je symetrické

Koeficienty podobnosti (indexy podobnosti)

- V ekológii sa využíva rada indexov podobnosti založených buď na prítomnosti/neprítomnosti taxónov alebo na abundanciách

Binárne koeficienty podobnosti

		Spoločenstvo 1	
		1	0
Spoločenstvo 2	1	a	b
	0	c	d

a, b, c, d = počet prípadov, kedy súhlasí binárna charakteristika spoločenstiev 1 a 2
a+b+c+d=p

Symetrické binárne koeficienty – nie je rozdiel medzi prípadom 1-1 a 0-0
Asymetrické binárne koeficienty - rozdiel medzi prípadom 1-1 a 0-0

Viac informácií a ďalšie merania vzdialenosti a podobnosti nájdete v knihe **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elseviere Science BV, Amsterdam.**

Symetrické binárne koeficienty

Symetrické binárne koeficienty

Simple matching coefficient (Sokal & Michener, 1958)

- Obvyklou metódou pre výpočet podobnosti medzi dvoma objektami je podiel počtu deskriptorov, ktoré kódujú objekt rovnako, a celkového počtu deskriptorov. Pri použití tohto koeficientu predpokladáme, že nie je rozdiel medzi nastaním 0 a 1 u deskriptorov.

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

Symetrické binárne koeficienty

Rogers & Tanimoto koeficient (1960)

- Dáva väčšiu váhu rozdielom ako podobnostiam.

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

Symetrické binárne koeficienty

Sokal & Sneath (1963)

- Ďalšie štyri navrhnuté koeficienty obsahujú double-zero, ale sú navrhnuté tak, aby sa znížil vplyv double-zero:

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

- tento koeficient dáva dvakrát väčšiu váhu zhodným deskriptorom než rozdielnym

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

- porovnáva zhody a rozdiely prostým podielem v merítku, ktoré ide od 0 do nekonečna

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

- porovnáva zhodné deskriptory so súčtami okrajov tabuľky

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

- je vytvorený z geometrických priemerov členov vzťahujúcich sa k a a d, podľa koeficientu S5.

Asymetrické binárne koeficienty

Asymetrické binárne koeficienty

Jaccardov koeficient (1900, 1901, 1908)

- Všetky členy majú rovnakú váhu

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

Asymetrické binárne koeficienty

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

- Varianta predchádzajúceho koeficientu dáva dvojnásobnu váhu dvojitým prezenciám, pretože sa môže zdať, že prítomnosť druhov je viac informatívna než ich absencia, ktorá môže být spôsobená rôznymi faktormi a nemusí nutne odrážať rozdielnosť prostredia. Prezencia druhu na oboch lokalitách je silným ukazovateľom ich podobnosti. S7 je monotónna k S8, preto podobnosť pre dve dvojice objektov vypočítaná podľa S7 bude podobná rovnakému výpočtu S8. Oba koeficienty sa líšia len v merítku. Tento index bol prvýkrát použitý Dicem v R-mode štúdii asociací druhov. Iná varianta tohto koeficientu dáva duplicitným prezenciám trojnásobnu váhu.

$$S_8(x_1, x_2) = \frac{2a}{2a+b+c} \qquad S_9(x_1, x_2) = \frac{3a}{3a+b+c}$$

Asymetrické binárne koeficienty

Russel & Rao (1940)

- navrhnutá miera umožňuje porovnanie počtu duplicitných prezencií (v čitateli) proti celkovému počtu druhov, nájdených na všetkých lokalitách, zahŕňajúcich druhy, ktoré chýbajú (d) na oboch uvažovaných lokalitách.

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

Asymetrické binárne koeficienty

Kulczynski (1928)

- koeficient porovnávající duplicitné prezencie s diferenciami

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

Kvantitatívne koeficienty

Kvantitatívne koeficienty

„Klasické“ indexy podobnosti

- Sørensenov kvantitatívny koeficient, kde aN a bN sú celkové počty jedincov v spoločenstvách A a B, jN je potom suma abundancií pokiaľ sa druh nachádza v oboch spoločenstvách, je počítaná vždy z nižšej abundancie daného druhu v spoločenstve

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

- Morisita-Horn index, kde aN je celkový počet jedincov v spoločenstve A a ani počet jedincov druhu i v spoločenstve A (obdobne platí pre spoločenstvo B)

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

Kvantitatívne koeficienty

Jednoduchý porovnávací koeficient (Sokal & Michener, 1958)

- modifikovaný simple matching coefficient môže byť použitý pre multistavové deskriptory - čitateľ obsahuje počet deskriptorov, pre ktoré sú dva objekty v rovnakom stave – napr. ak je dvojica objektov popísaná nasledujúcimi desiatimi multistavovými deskriptormi: hodnota S_1, vypočítaná pre 10 multistavových deskriptorov bude S_1 (x_1,x_2) = 4 / 10 = 0.4
- Podobným spôsobom je možné rozšíriť všetky binárne koeficienty pre multistavové deskriptory.

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

	Druhy										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	
zhoda	0	+1	+0	+0	+0	+1	+0	+0	+1	+1	4

Gowerov všeobecný koeficient podobnosti (1971) I.

- Gower navrhol všeobecný koeficient podobnosti, ktorý môže kombinovať rôzne typy deskriptorov. Podobnosť medzi dvoma objektami je vypočítaná ako priemer podobností, vypočítaných pre všetky deskriptory. Pre každý deskriptor j je hodnota parciálnej podobnosti s_{12j} medzi objektami x1 a x2 vypočítaná nasledovne:

$$S_{15}(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} s_{j12}$$

- Pre binárne deskriptory s_j=1 (zhoda) alebo 0 (nezhoda). Gower navrhol dve formy tohto koeficientu. Nasledujúca forma je symetrická, dáva s_j=1 double-zero. Druhá forma, Gowerov asymetrický koeficient dáva pro double-zero s_j=0
- Kvalitatívne a semikvantitatívne deskriptory sú upravené podľa jednoduchého zameňovacieho pravidla, s_j=1 pri súhlase a s_j = 0 pri nesúhlase deskriptorov. Double zero sú ošetrené rovnako ako v predchádzajúcom odstavci.
- Kvantitatívne deskriptory (reálne čísla) sú spracované nasledovne: pre každý deskriptor sa nejprv vypočíta rozdiel medzi stavmi oboch objektov, ktorý je potom vydelený najväčším rozdielom (R_j), nájdeným pre daný deskriptor medzi všetkými objektami v štúdii (alebo v referenčnej populácii – doporučuje sa vypočítať najväčšiu diferenciu R_j každého deskriptoru *j* pro celú populáciu, aby bola zistená konzistencia výsledkov pre všetky parciálne štúdie).

Gowerov všeobecný koeficient podobnosti (1971) II.

- normalizovaná vzdialenosť môže byť odpočítaná od 1 aby bola transformovaná na podobnosť.

$$s_j = 1 - \left[\frac{|y_{j1} - y_{j2}|}{R_j}\right]$$

- Gowerov koeficent môže byť nastavený tak, aby zahŕňal prídavný flexiblilný prvok: žiadne porovnanie nie je vypočítané u deskriptorov, u ktorých chýba informácia buď u jedného alebo u druhého objektu. Toto zaisťuje člen w_j, nazývaný Kroneckerovo delta, popisujúcí prítomnosť/neprítomnosť informácie v oboch objetkoch: ak je informácia o deskriptore y_j prítomná u oboch objektov (w_j=1), inak (w_j=0), tento koeficient nadobúda hodnôt podobnosti medzi 0 a 1 (najväčšia podobnosť objektov). Ďalšiou možnosťou je váženie rôznych deskriptorov prostým priradením čísla v rozsahu 0-1 w_j.

$$S_{15}(x_1, x_2) = \frac{\sum_{j=1}^{p} w_{j12} s_{j12}}{\sum_{j=1}^{p} w_{j12}}$$

Metriky vzdialenosti

Viacrozmerné metriky vzdialenosti

Metriky všeobecne

Na miery nepodobnosti, t.j. metriky kladieme spravidla určité požiadavky:

- Mali by rešpektovať rozdielnu variabilitu jednotlivých štatistických znakov a prisudzovať väčší vplyv tým jednorozmerným vzdialenosťam, ktoré vykazujú nižšiu variabilitu.

- Súčasne by mala zvolená metrika rešpektovať štruktúru dát a to tak, aby väčší vplyv na viacrozmernú vzdialenosť mali tie vzdialenosti, ktoré boli zistené u nekorelovaných či len slabo korelovaných štatistických znakov.

- Metrika musi spĺňať 4 vlastnosti:
    1. $d(A,B) = d(B,A)$
    2. $A \neq B \Rightarrow d(A,B) > 0$
    3. $A = B \Leftrightarrow d(A,B) = 0$
    4. $d(A,B) \leq d(A,C) + d(C,B)$

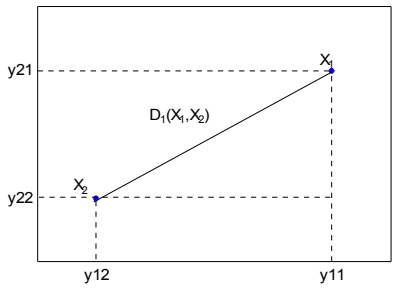
Viacrozmerné metriky vzdialenosti

Euklidovská vzdialenosť

- Ide o základné metrické merítko vzdialenosti a počíta vzdialenosť objektov obdobne ako Pythagorova veta (počíta preponu pravouhlého trojuholníka). Metóda je citlivá na rozdielny rozsah hodnôt vstupujúcich premenných (vhodným riešením môže byť štandardizácia) a double zero problém. Nemá hornú hranicu hodnôt.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{j1} - y_{j2})^2}$$

- Ako ďalšie merítko sa používá tiež štvorec tejto vzdialenosti. Jeho nevýhodou sú semimetrické vlastnosti.



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

Viacrozmerné metriky vzdialenosti

- Euklidovská vzdialenosť je využívaná častokrát úplne neoprávnene. Pri použití tejto metriky by sme mali byť veľmi obozretní, lebo jej využitím môžeme podstatne skresliť výsledky analýzy. Euklidovská metrika totiž neberie do úvahy korelovanosť jednotlivých parametrov (štatistických znakov).

Vážená euklidovská vzdialenosť

- Varianta euklidovskej vzdialenosti – pripisuje jednotlivým premenným rôzne váhy a zohľadňuje tak ich význam. Problémom však zostáva správne určenie vektoru váh.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} w_j (y_{j1} - y_{j2})^2}$$

Viacrozmerné metriky vzdialenosti

Priemerná vzdialenosť

- Euklidovská vzdialenosť je prepočítaná na počet parametrov (druhov v prípade vzdialenosti spoločenstiev odberov).

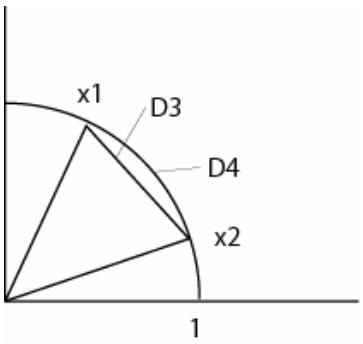
$$D_2^2(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p}(y_{j1} - y_{j2})^2$$

$$D_{_2}(x_1, x_2) = \sqrt{D_{_2}^2}$$

Viacrozmerné metriky vzdialenosti

Chord distance (Orlóci, 1967)

- Odstraňuje double zero problém a vplyv rozdielneho počtu jedincov druhov vo vzorcoch pri výpočte Euklidovskej vzdialenosti. Jej maximálna hodnota je druhá odmocnina z počtu druhov a minimum 0. Pri výpočte počíta len s pomermi druhov v rámci jednotlivých vzoriek. Ide vlastne o Euklidovskú vzdialenosť počítanú pre vektory vzoriek štandardizovaných na dĺžku 1, alebo je možný priamy výpočet už zahŕňujúci štandardizáciu. Vnútorná časť výpočtu je vlastne kosínus uhla zvieraného vektormi, zápis vzorca je možný i v tejto forme.



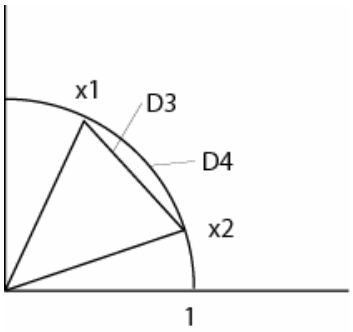
$$D_3(x_1, x_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{j1}y_{j2}}{\sqrt{\sum_{j=1}^{p} y_{j1}^2 \sum_{j=1}^{p} y_{j2}^2}}\right)}$$

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

Viacrozmerné metriky vzdialenosti

Geodetická metrika

- Počíta dĺžku výseče jednotkovej kružnice medzi normalizovanými vektormi (viz. Chord distance).



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

Viacrozmerné metriky vzdialenosti

Manhattanská vzdialenosť

- Ide vlastne o súčet rozdielov jednotlivých parametrov popisujúcich objekty

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

Viacrozmerné metriky vzdialenosti

Minkowského metrika

- Je všeobecnou formou výpočtu vzdialenosti – podľa zadaného koeficientu môže odpovedať napr. Euklidovskej alebo Manhattanskej metrike. So stúpajúcim koeficientom umocňovania stúpa významnosť väčších rozdielov. Existuje ešte obecnejšia forma, kedy koeficient umocňovánia a odmocňovania je zadávaný zvlášť.

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

Viacrozmerné metriky vzdialenosti

Mahalanobisova vzdialenosť (Mahalanobis 1936)

- Zohľadňuje vzájomné vzťahy medzi premennými, teda berie do úvahy ich skorelovanosť. Je nezávislá na rozsahu hodnôt premenných. Počíta tak vzdialenosť medzi objektami v systéme súradníc, kt. osi nemusia byť na seba kolmé.
- Je potrebné však upozorniť, že pri použití Mahalanobisovej vzdialenosti potlačujeme vplyv rozdielov vo variabilite premenných na výsledky, čo nemusí byť vždy žiadúce.
- Ak sú premenné nekorelované, párové korelačné koeficienty sú nulové a premenné vstupujúce do výpočtu sú prevedené na normovaný tvar, tak Mahalanobisova vzdialenosť odpovedá štvorcu euklidovskej vzdialenosti.

Viacrozmerné metriky vzdialenosti

Mahalanobisova vzdialenosť (Mahalanobis 1936)

- V praxi sa používa pre zistenie vzdialenosti medzi skupinami objektov. Sú dané dve skupiny objektov w1 a w2 o n1 a n2 počte objektov a popísané p parametrami:

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

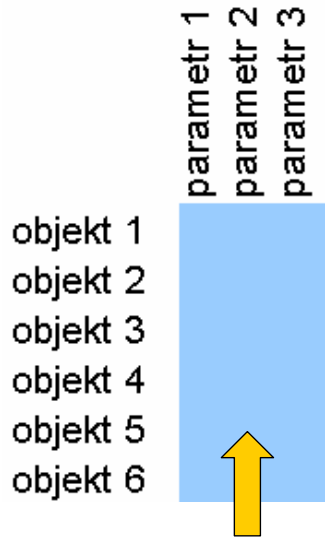
- kde $\overline{d_{12}}$ je vektor o dĺžke p rozdielov medzi priemermi p parametrov v oboch skupinách. V je vážená disperzná matica (matica kovariancií parametrov) vnútri skupín objektov.

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

- kde S1 a S2 sú disperzné matice jednotlivých skupín. Vektor $\overline{d_{12}}$ meria rozdiel medzi p- rozmernými priemermi skupín a V vkladá do rovnice kovarianciu medzi parametrami.

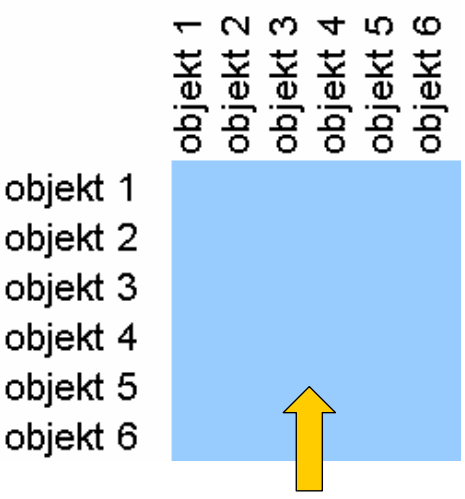
Vstupná matica viacrozmerných analýz

NxP MATICE



Hodnoty parametrov pre jednotlivé objekty

ASOCIAČNÁ MATICA



Korelácia, kovariancia, vzdialenosť, podobnosť