Fenetický přístup

shluková analýza
ordinační metody
diskriminační analýza

Stupnice

- **Nominální stupnice (*nominal scale*)**

  z matematických operátorů zde platí jen rovnost (=) nebo nerovnost (≠)

- **Pořadová stupnice (*ordinal scale*)**

  kromě rovnosti a nerovnosti zde platí také operátory < a >

- **Intervalová stupnice (*interval scale*)**

  kromě vlastností předcházejících dvou stupnic je zde možné také sčítání a odečítání (znaky mohou nabývat také hodnoty 0)

- **Poměrová stupnice (*ratio scale*)**

  dovoluje vyjádřit poměr mezi objekty (lze použít též operátor dělení

Klasifikace znaků

(1) **kvalitativní (*qualitative*):**

binární (*binary*, dvoustavové, dvouhodnotové, alternativní)
vícestavové (*multistate*, vícehodnotové)

(2) **semikvantitativní (*semiquantitative*)**

(3) **kvantitativní (*quantitative*)**

nespojité, diskrétní (*discontinuous*, *discrete*, *meristic*)
spojité, kontinuální (*continuous*)

Převedení čtyřstavového kvalitativního znaku do soustavy binárních znaků

stavy kvalitativního znaku	umělé binární proměnné			
a	**1**	**0**	**0**	0
b	**0**	**1**	**0**	0
c	**0**	**0**	**1**	0
d	**0**	**0**	**0**	1

Transformace dat

Některé mnohorozměrné metody nevyžadují **normální rozdělení dat**, případně jsou dostatečně robustní ve vztahu k odchylkám od normálního rozdělení dat (např. shlukové analýzy, PCA ...)

Jiné metody mnohorozměrné normální rozdělení dat vyžadují (např. diskriminační analýza).

Transformací lze někdy rozdělení dat přiblížit k normálnímu rozdělení.

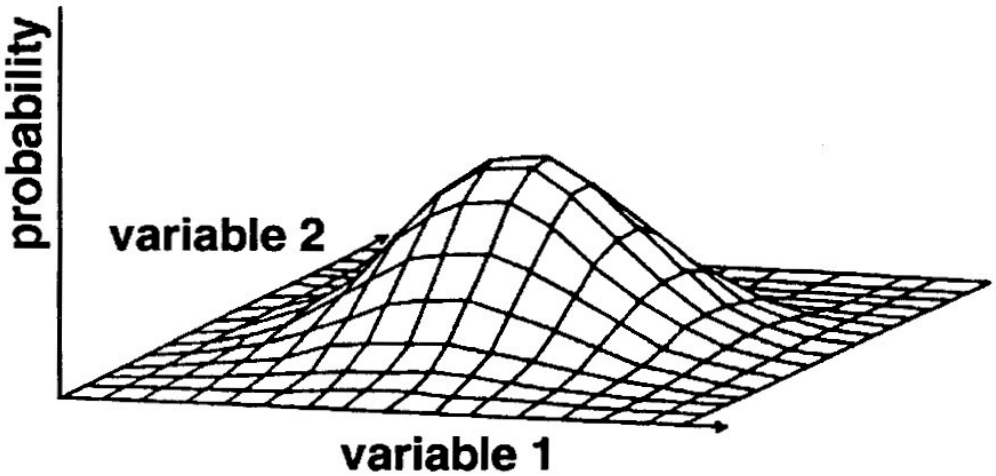


Diagram hustoty pravděpodobnosti pro dvourozměrné normální rozdělení

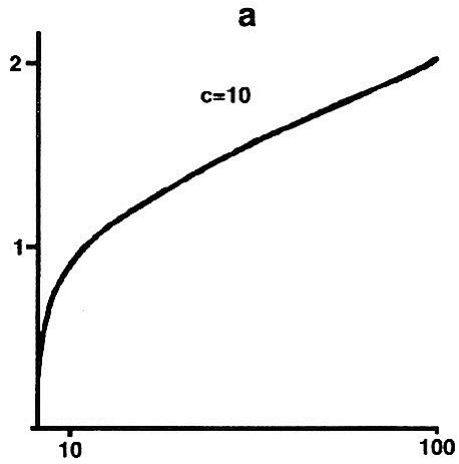
Transformace dat

K **transformaci** se používají **konstanty** a **funkce nezávislé** na analyzovaných datech

Lineární transformace (např. násobení znaků konstantou) pokud se aplikují u všech znaků – výsledky anlýzy se tím nemění; pokud se použijí jenom u jednoho nebo několika znaků, dochází tu k jejich vážení

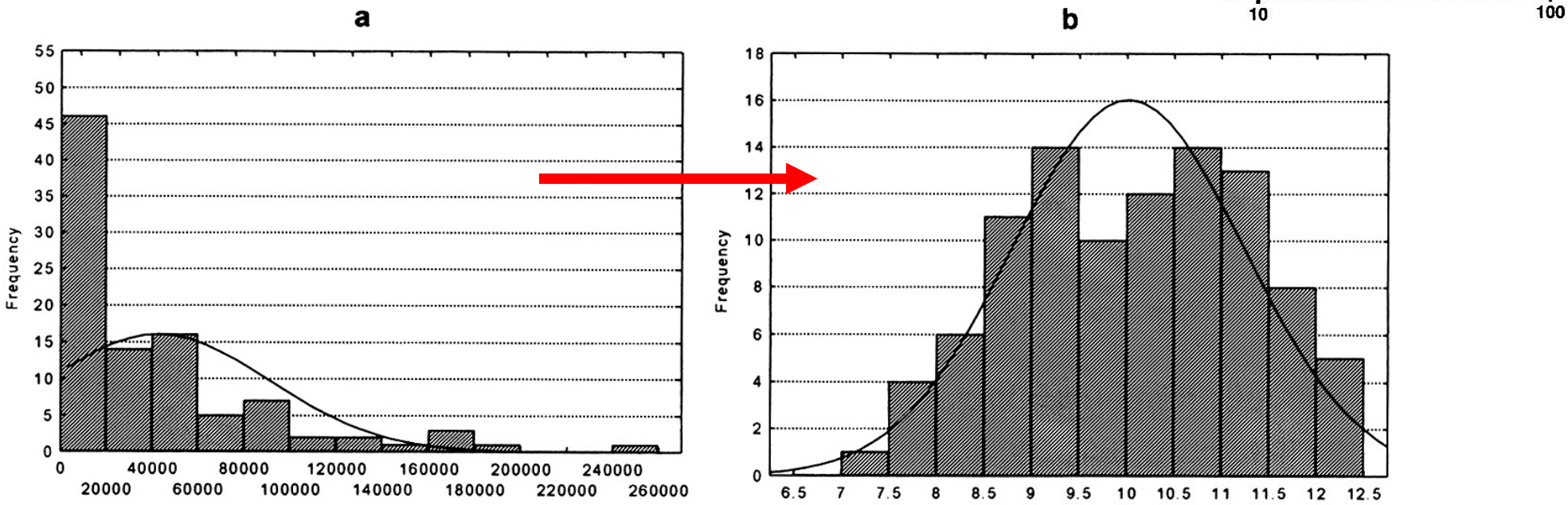
Nelineární transformace mění strukturu dat

Transformace dat



Logaritmická transformace (*logarithmic transformation*): Naměřené hodnoty se nahrazují jejich logaritmem

$$x'_{ij} = \log_c x_{ij}$$



Protože logaritmus nuly není definován, připočítává se v takových případech ke každé naměřené hodnotě daného znaku konstanta 1 anebo 0,5.

Vzorec má potom tvar $x'_{ij} = \log_c (x_{ij}+1)$

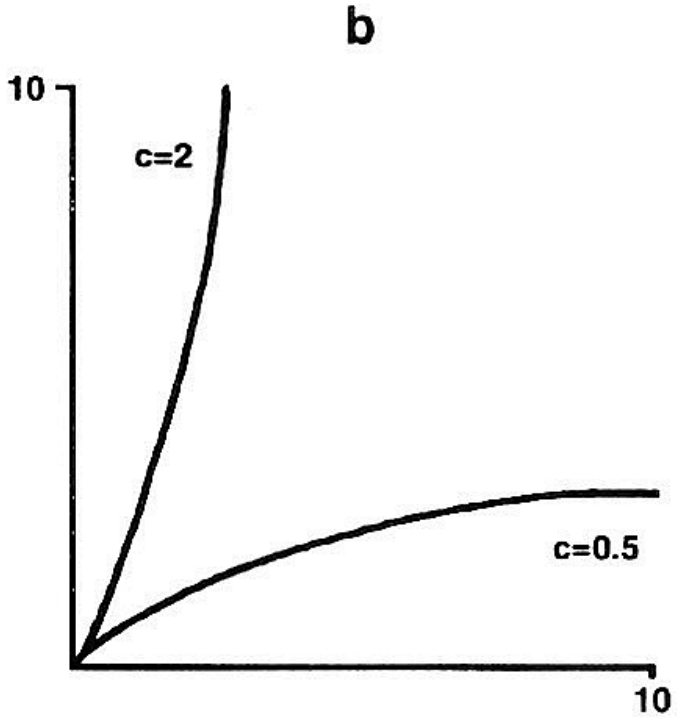
Transformace dat

Odmocninová transformace (*square root transformation*)

obecně $x'_{ij} = x^c_{ij}$

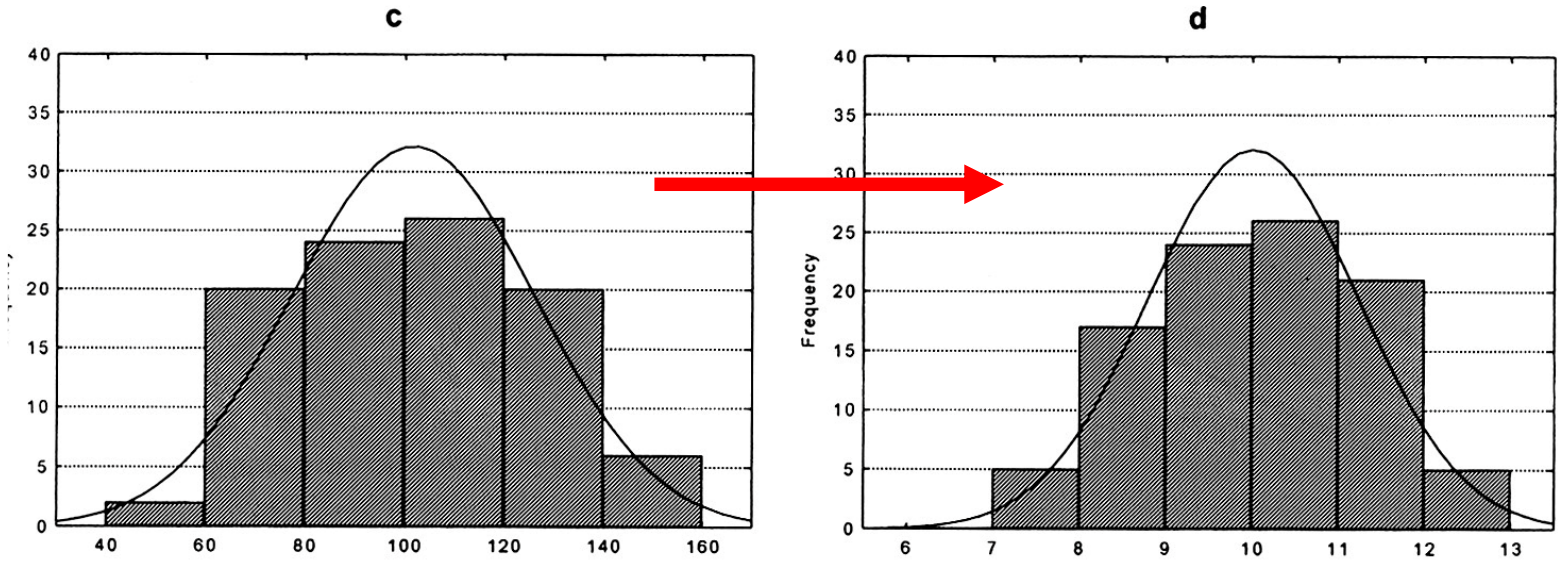
c>1 zdůrazňují se vysoké číselné hodnoty – používá se zřídka
c<1 vysoké číselné hodnoty se podhodnocují

c=0.5 **- odmocninová transformace**



Transformace dat

Odmocninová transformace (*square root transformation*)



Znaky nesmí dosahovat nulových hodnot, proto se někdy používá ve tvaru

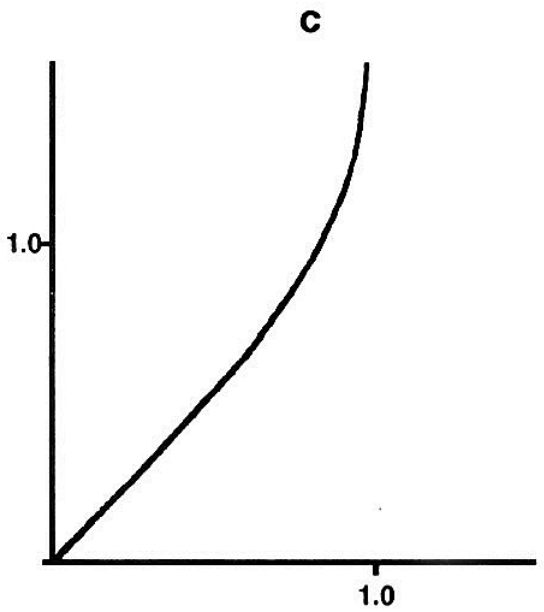
$$x'_{ij} = \sqrt{x_{ij} + 0.5}$$

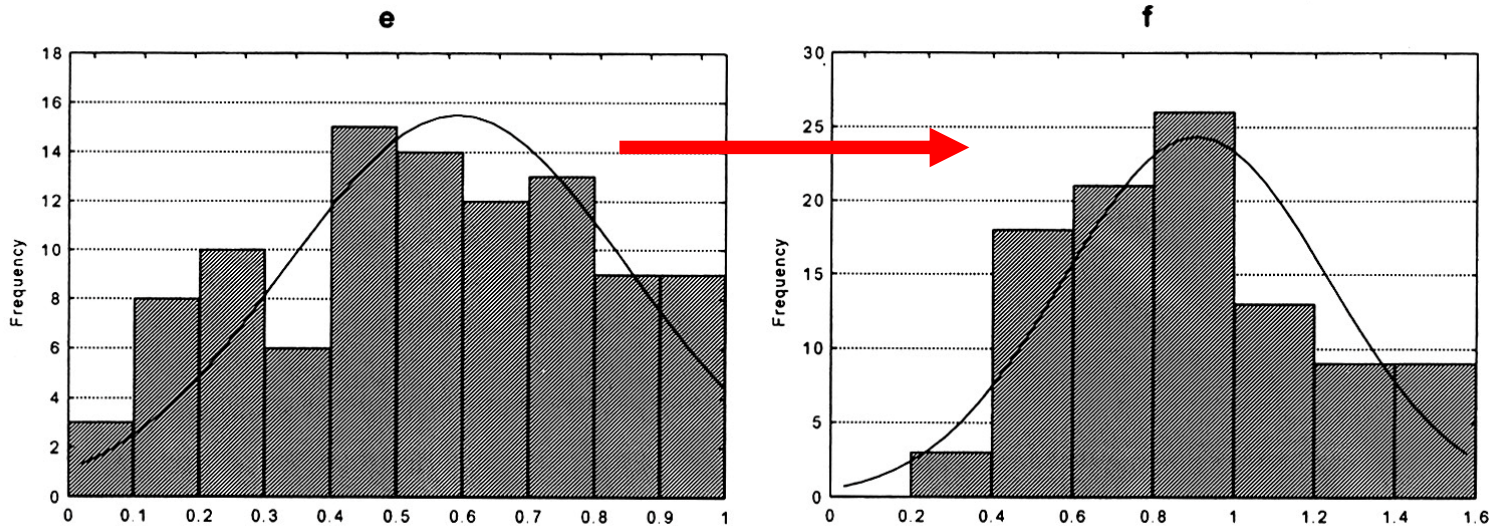
Transformace dat

Arkussinová transformace (Arc sin transformation)

$x'_{ij} = \text{arc sin } x_{ij}$

Používá se i v kombinaci s odmocninovou transformací;
arkussinová transformace předpokládá, že data jsou měřená v intervalu <0, 1>
pokud tomu tak není, je možné naměřené hodnoty vydělit konstantami 10, 100, 1000, atd.





Standardizace dat

Ke **standardizaci** se používají statistiky odvozené z analyzovaného souboru dat (rozpětí, směrodatná odchylka, průměr, maximum atd.)
Znaky se tímto postupem převádějí na stejné měřítko (jinými slovy přestává záležet na skutečném rozměru příslušného znaku)

Centrování (*centring*, standardizace na průměr rovný nule)

$$x'_{ij} = x_{ij} - \overline{x_i}$$

Centrování nemění jednotky, ve kterých jsou znaky měřené, mění se jen poloha nulového bodu v soustavě souřadnic.
Standardizace rozpětím (*standardization by range*, *ranging*)

$$x'_{ij} = \frac{x_{ij} - \min_j \{x_{ij}\}}{\max_j \{x_{ij}\} - \min_j \{x_{ij}\}}$$

Doporučuje se použít v případech, kdy jsou sice znaky měřeny ve stejném měřítku, ale mezi jejich hodnotami jsou velmi velké rozdíly, hodnoty znaků se převedou do intervalu [0,1]

Standardizace dat

Standardizace směrodatnou odchylkou (*standardization by standard deviation*)

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$$

kde s_i je směrodatná odchylka znaku i

doporučuje se použít v případech, kdy jsou znaky měřené v odlišných škálách a jednotkách

Koeficienty vyjadřující vztahy mezi objekty nebo znaky (*resemblance coefficients*)

(1) koeficienty vzdálenosti pro kvantitativní a binární znaky (*metric distances*)

(2) koeficienty podobnosti pro binární znaky (*binary similarity coefficients*)

(3) koeficienty pro smíšená data (*coefficients for mixed data*)

(4) korelační koeficienty (*correlation coefficients*)

Metriky (vzdálenosti)

Pokud koeficienty vzdálenosti splňují následující požadavky, považují se za **metriky** (metric):

(1) symetrie – pro vzdálenost dvou objektů (x, y) platí:

$$d(x,y) = d(y,x) \geq 0$$

(2) trojúhelníková (triangulární) nerovnost – pro vzdálenost třech objektů (x, y, z) platí:

$$d(x,y) \leq d(x,z) + d(y,z)$$

tj. vzdálenost dvou objektů je menší, nanejvýš rovna součtu jejich vzdáleností od objektu třetího;

(3) vzdálenost totožných objektů (a vzdálenost objektu od sebe samého) je 0:

$$d(x,y) = 0 \text{ v případě, že } x = y$$

(4) vzdálenost objektů, které nejsou totožné, je větší než 0 (je kladná):

$$d(x,y) > 0 \text{ v případě, že } x \neq y.$$

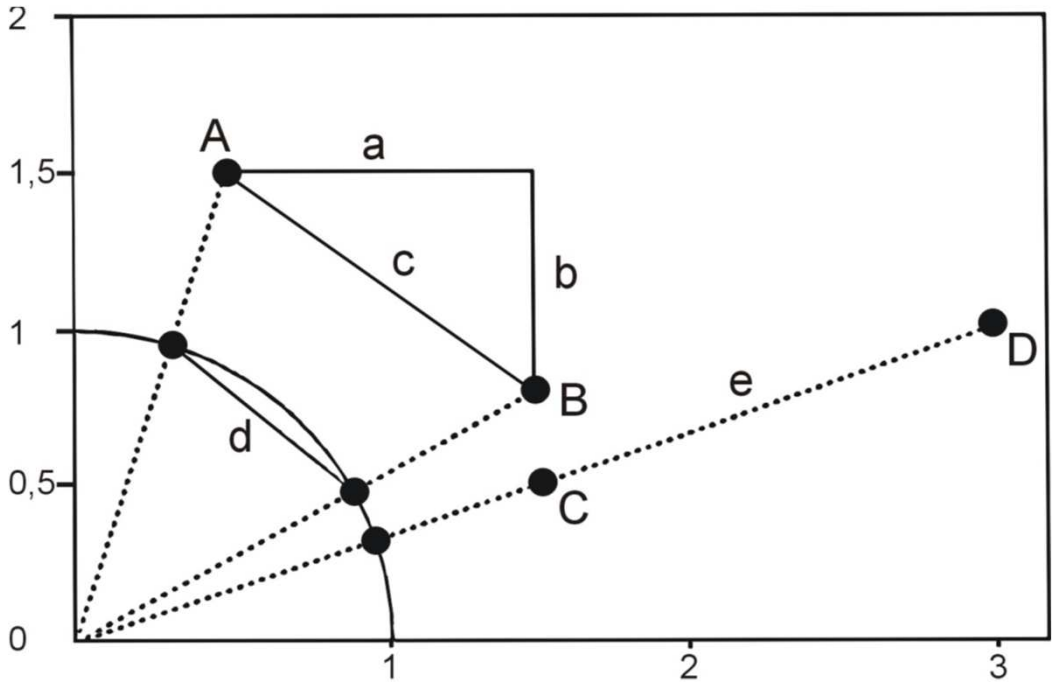
Pokud koeficienty vzdálenosti nesplňují kritérium trojúhelníkové nerovnosti, považují se za pseudometriky (*pseudometric*, *semimetric*).

Metriky (vzdálenosti)

Euklidovská vzdálenost (Euclidean distance):

$$EU=c \qquad EU_{jk} = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{ik}\right)^2}$$

kde x_{ij} je hodnota znaku i pro objekt j, x_{ik} je hodnota znaku i pro objekt k, n je celkový počet znaků



Euklidovská vzdálenost je závislá na škále znaků

	Váha v librách	Výška v stopách	Výška v palcích
A	60	3,0	36,0
B	65	3,5	42,0
C	63	4,0	48,0

$= (60 – 65)^2 + (3{,}0 – 3{,}5)^2 = 25.25$ $[(60 – 65)^2 + (36{,}0 – 42{,}0)^2 = 61]$

$= (60 – 63)^2 + (3{,}0 – 4{,}0)^2 = 10.00$ $[(60 – 63)^2 + (36{,}0 – 48{,}0)^2 = 153]$

$= (65 – 63)^2 + (3{,}5 – 4{,}0)^2 = 4.25$ $[(65 – 63)^2 + (42{,}0 – 48{,}0)^2 = 40]$

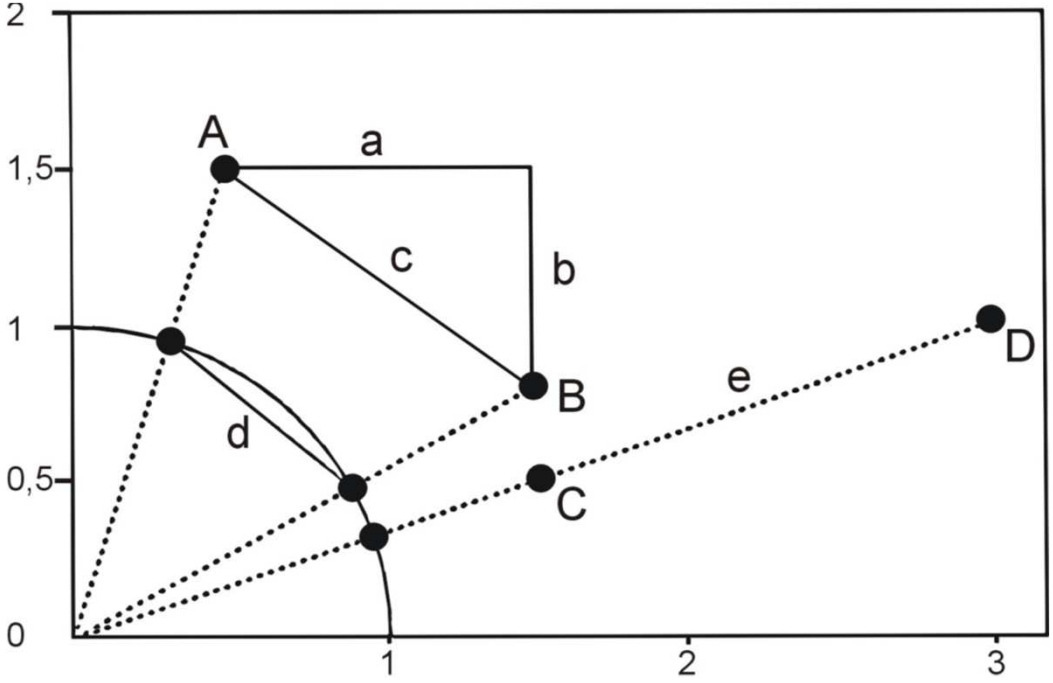
Metriky (vzdálenosti)

Manhattanská (city block) metrika:

$CB=a+b$

$$CB_{jk} = \sum_{i=1}^{n} \left| x_{ij} - x_{ik} \right|$$

Připomíná severoamerická města s kolmými ulicemi, kde se musí chodit kolem bloků



Minkowského metrika:

$$MNK_{jk} = \sqrt[r]{\sum_{i=1}^{n} \left(x_{ij} - x_{ik} \right)^{r}}$$

kde $r \geq 1$;
pro $r=1$ …. CB
***Pro* $r=2$ … EU**

Vzdálenosti

Tětivová vzdálenost (*chord distance*)

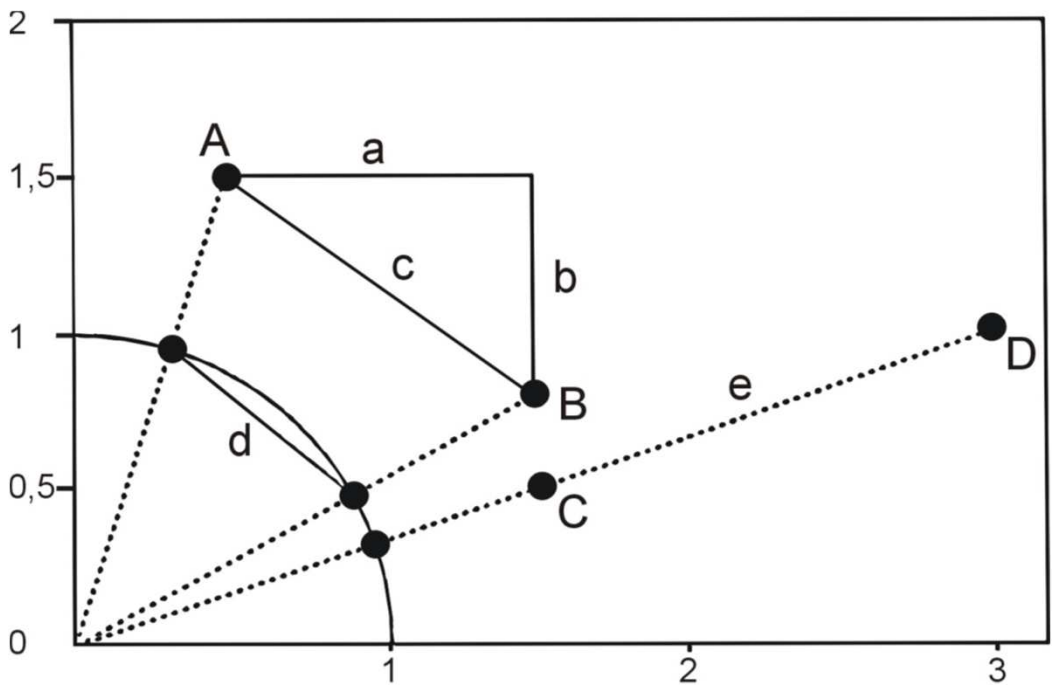
Pro dva znaky je tětivová vzdálenost přímou vzdáleností mezi projekcí bodů na kružnici s jednotkovým poloměrem

$$CH = d$$

$$CH_{jk} = \left(2 \left[1 - \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2}} \right] \right)^{1/2}$$

Tětivová vzdálenost dosahuje stejných hodnot v případě, že dva nebo více objektů vykazují ve všech znacích proporčně těch samých hodnot, aniž by konkrétní hodnoty těchto znaků musely být u všech objektů stejné (vzdálenost bodů C a D).

Není pravou metrikou.



Koeficienty podobnosti pro binární data

Jakákoliv funkce *d* je **nepodobností** pokud odpovídá alespoň prvním třem pravidlům o metrikách
(pokud $j=k$, pak $d_{jk}=0$; pokud $j\neq k$, pak $d_{jk}>0$; $d_{jk} = d_{kj}$);

- většina funkcí nepodobnosti má dolní hranici = 0, horní hranici = 1: $0 \leq d_{jk} \leq 1$
- většina funkcí nepodobnosti po transformaci $(d_{jk})^{1/2}$ vyhovuje všem pravidlům o metrikách a pak představují **vzdálenosti**

obvykle uvažujeme o **podobnosti**: $s_{jk} = 1 - d_{jk}$

pro identické objekty platí $s_{jk} = 1$

Koeficienty podobnosti pro binární data

Výběr koeficientu podobnosti

	objekt 2		
objekt 1		1	0
	1	*a*	*b*
	0	*c*	*d*

a – počet znaků, ve kterých mají oba objekty hodnotu + (resp. 1) (pozitivní shoda)

b – počet znaků, ve kterých má objekt i hodnotu – (resp. 0) a objekt j hodnotu + (resp. 1)

c – počet znaků, ve kterých má objekt i hodnotu + (resp. 1) a objekt j hodnotu – (resp. 0)

d – počet znaků, ve kterých mají oba objekty hodnotu – (resp. 0) (negativní shoda)

Volba mezi koeficienty závisí především na tom, jestli pro dané znaky má nebo nemá smysl **negativní shoda**, tj. zdali má nebo nemá smysl uvažovat, že nulová hodnota znaku má u porovnávaných objektů stejnou příčinu

Koeficienty podobnosti pro binární data

Koeficienty hodnotící *a* a *d* symetricky:

	object 2		
object 1		1	0
	1	*a*	*b*
	0	*c*	*d*

Koeficient jednoduché shody (*simple matching*):

koeficient je blízký ED:
$ED^2 = n(1\text{-}SM)$
$n=a+b+c+d$
$ED = \sqrt{b+c}$

$$SM = \frac{a+d}{a+b+c+d}$$

Koeficient Rogerse a Tanimota:

neshody jsou vážené dva krát;
hodnoty vždy nižší než u SM, s výjimkou b+c=0

$$RT = \frac{a+d}{a+2b+2c+d}$$

Hamannův index:

rozpětí [-1,1]

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

$$SM = \frac{HAM+1}{2}$$

Koeficienty podobnosti pro binární data

	objekt 2		
objekt 1		1	0
	1	*a*	*b*
	0	*c*	*d*

Koeficienty hodnotící *a* a *d* asymetricky:

d se sice bere do úvahy, *a* a *d* se však neváží stejně

Baroni-Urbani – Buser II:

$$BB2 = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$$

modifikovaný *SM*, *d* -> geometrický průměr *a* a *d*
rozpětí [0,1]

Baroni-Urbani – Buser I:

$$BB1 = \frac{\sqrt{ad} + a - b - c}{\sqrt{ad} + a + b + c}$$

modifikovaný *HAM*, *d* -> geometrický průměr *a* a *d*
rozpětí [0,1]

Russellův Raův koeficient:

$$RR = \frac{a}{a + b + c + d}$$

zvýšenie hodnoty *d* snižuje hodnotu nepodobnosti

Koeficienty podobnosti pro binární data

Koeficienty, které neberou do úvahy negativní shodu:

	object 2		
object 1		1	0
	1	*a*	*b*
	0	*c*	*d*

Jaccardův koeficient: $JAC = \frac{a}{a+b+c}$

rozpětí [0,1]

konverze $d_{jk} = \sqrt{1 - s_{jk}}$

má za výsledek Euklidovskou vzdálenost

Sorensenův koeficient: $SOR = \frac{2a}{2a+b+c}$

pozitivní shoda se váží dva krát

genetické vzdálenosti podle Nei & Li (1979), Link et al. (1995) využívané při NJ, PCoA odpovídají také tomuto typu koeficientů

Nei & Li (1979):

$$NL = 1 - \frac{2a}{2a+b+c}$$

Link et al. (1995):

$$L = \frac{b+c}{b+c+a}$$

Koeficienty pro smíšená data

Do této kategorie patří Gowerův koeficient a vzdálenost pro smíšená data. Používají se v případech, kdy jsou v matici současně zastoupeny kvalitativní znaky a znaky kvantitativní nebo binární (případně všechny tři druhy znaků).

Gowerův koeficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objekty charakterizované znakem k,
n – celkový počet znaků,
s_{ijk} – skóre znaku k

w_{ijk} je váha, která může nabývat hodnot 1 nebo 0 podle toho, jestli je nebo není možné srovnání hodnot znaku k u objektů i a j (krome binárních znaku může mít nulovou hodnotu jenom tehdy, pokud hodnota znaku k není u jednoho nebo obou objektů známá); s_{ijk} je skóre (hodnota) pro príslušný znak k.

a) pro binární znaky:
$w_{ijk} = 1$ a $s_{ijk} = 0$ pokud $x_{ik} \neq x_{jk}$ (hodnoty znaku k pro objekty i a j)
$w_{ijk} = s_{ijk} = 1$ pokud $x_{ik} = x_{jk} = 1$ nebo pokud $x_{ik} = x_{jk} = 0$ a negativní shoda se bere do úvahy (odpovídá koeficientu jednoduché shody)
$w_{ijk} = s_{ijk} = 0$ pokud $x_{ik} = x_{jk} = 0$ a negativní shoda se nebere do úvahy (odpovídá Jaccardovu koeficientu)

Koeficienty pro smíšená data

Gowerův koeficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objekty charakterizované znakem k,
n – celkový počet znaků,
s_{ijk} – skóre znaku k

b) pro nominální znaky:
$w_{ijk} = 1$ pokud x_{ik} a x_{jk} jsou známé; pak
$s_{ijk} = 0$ pokud $x_{ik} \neq x_{jk}$; $s_{ijk} = 1$ pokud $x_{ik} = x_{jk}$ (počet stavů se nebere do úvahy)

Koeficienty pro smíšená data

Gowerův koeficient:

$$GOWjk = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}}$$

i,j – objekty charakterizované znakem k,
n – celkový počet znaků,
s_{ijk} – skóre znaku k

c) pro kvantitativní znaky:
w_{ijk} = 1 pokud x_{ik} a x_{jk} jsou oba známé, a $s_{ijk} = 1 - \{|x_{ik} - x_{jk}| / (\text{rozpětí znaku } i)\}$
(odpovídá Manhattanské metrice s daty stadardizovanými na rozpětí)

Koeficienty pro smíšená data

příklad:

Taxon / znak	Větvení lodyhy	barva korunních lístků	charakter listů	průměrná výška rostliny (cm)	Průměrná délka korunních lístků (mm)
1	1	bílá (1)	jednoduché (1)	30	2,6
2	1	červená (2)	lichozpeřené (2)	25	2,3
3	0	modrá (3)	lichozpeřené (2)	10	8,5
4	0	modrá (3)	dlanitodílné (3)	80	8,2

$$GOW(1,2) = \frac{[1\times 1]+[1\times 0]+[1\times 0]+\left[1\times\left(1-\frac{|30-25|}{80-10}\right)\right]+\left[1\times\left(1-\frac{|2.6-2.3|}{8.5-2.3}\right)\right]}{1+1+1+1+1} = 0.576$$

Korelační koeficienty

Pearsonův korelační koeficient

$$r_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}}$$

n počet objektů,
hodnota znaku 1 pro objekt *i*

lineární korelace, predpokládá normální rozdělení dat

Spearmanův korelační koeficient (rank koeficient, koeficient pořadí):

$$r_{12} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

do úvahy se neberou konkrétní hodnoty znaků ale pořadí objektů,
kde d_i je rozdíl v pořadí mezi objekty;

Pearsonův korelační koeficient a Spearmanův korelační koeficient:
rozpětí [-1, +1], +1 ´přímá závislost, -1 nepřímá závislost, 0 absence vztahu