Bi8600: Vícerozměrné metody
2. cvičení

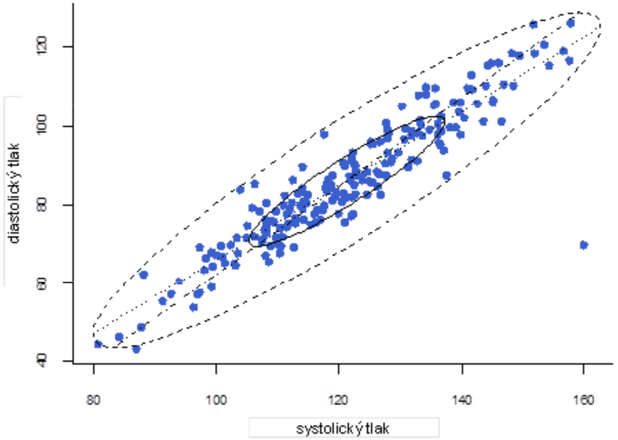
Vícerozměrné rozdělení dat
Koeficienty podobnosti a vzdálenosti
Asociační matice

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Brožová

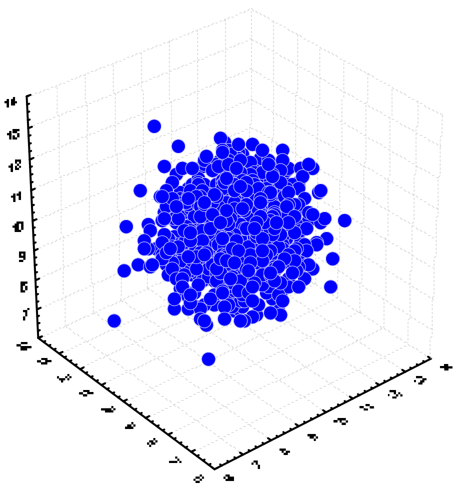
Jak vizualizujeme vícerozměrný prostor?

Jak vizualizujeme vícerozměrný prostor I

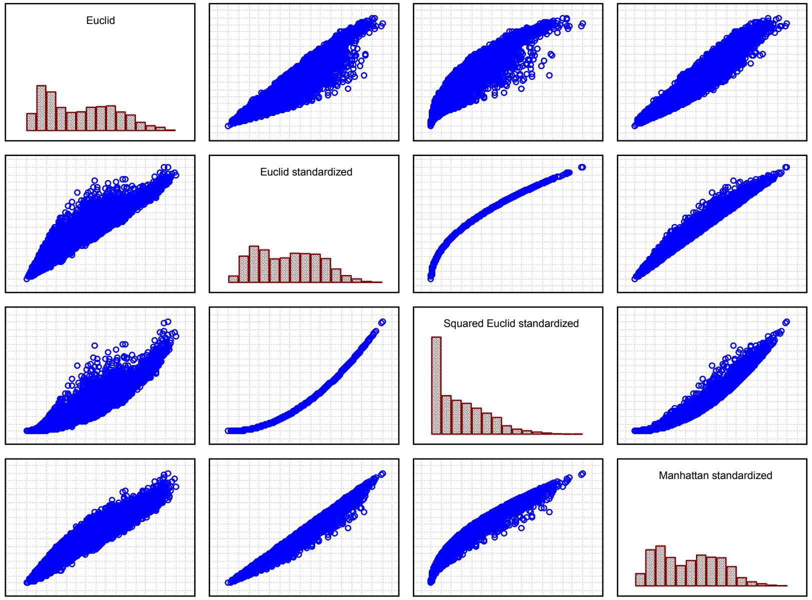
2D



3D

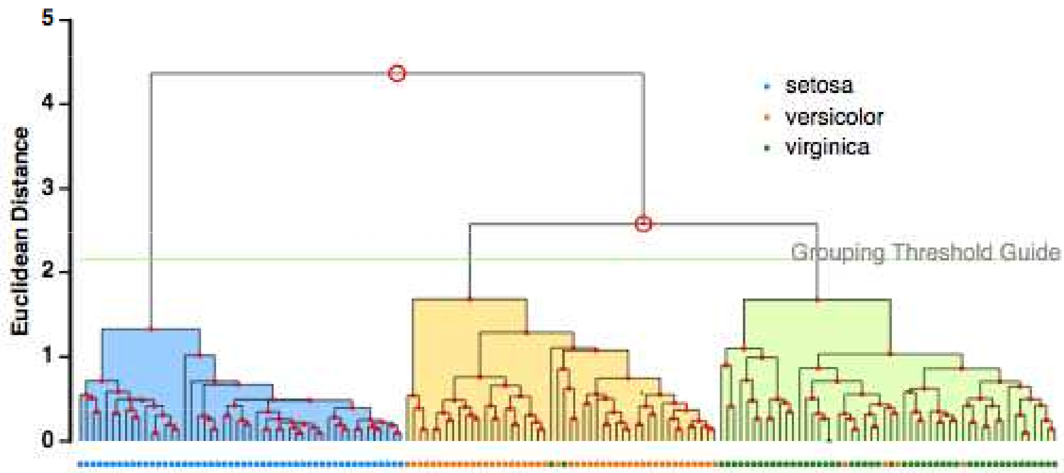


Maticové grafy

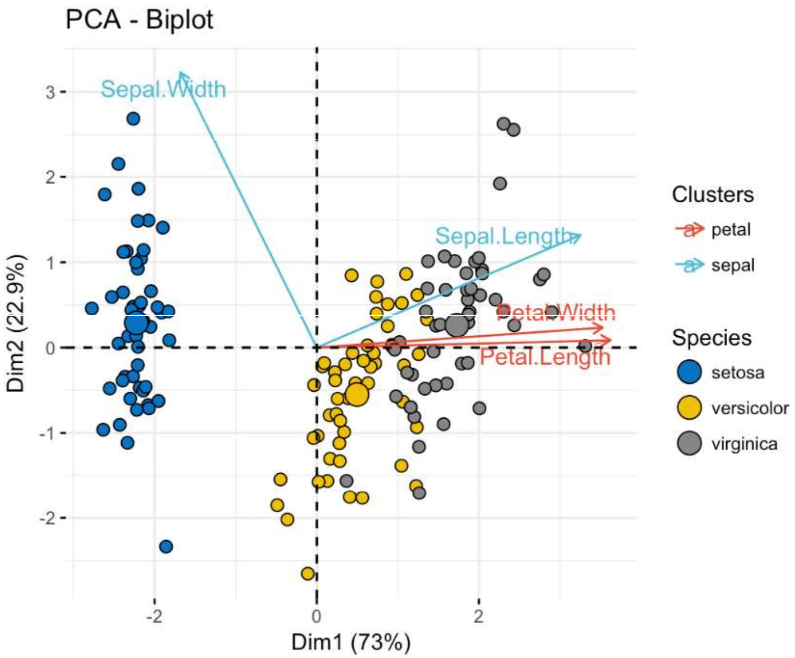


Jak vizualizujeme vícerozměrný prostor II

Dendrogram

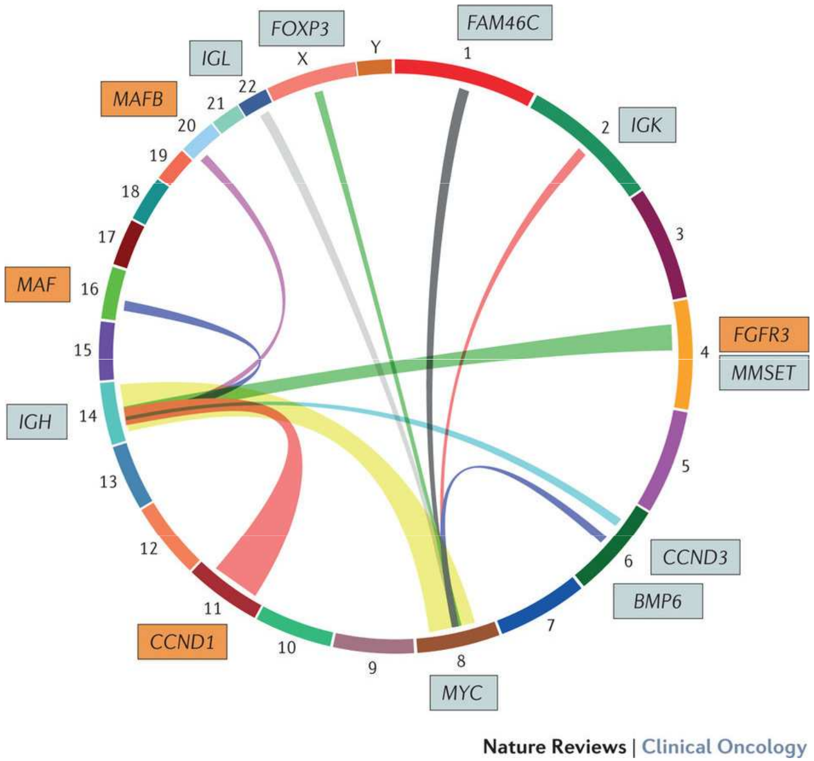


Biplot korelací a vzdáleností

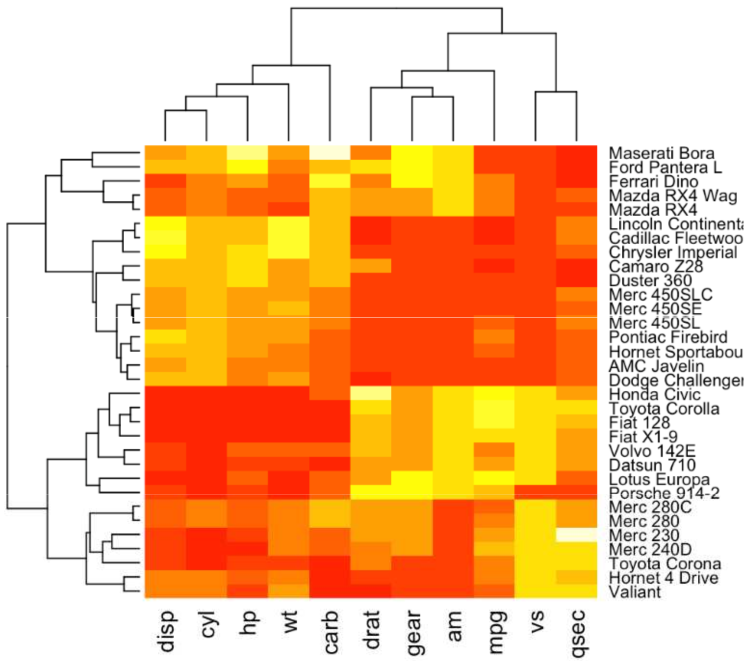


Jak vizualizujeme vícerozměrný prostor - jiné

Circos



Heatmap

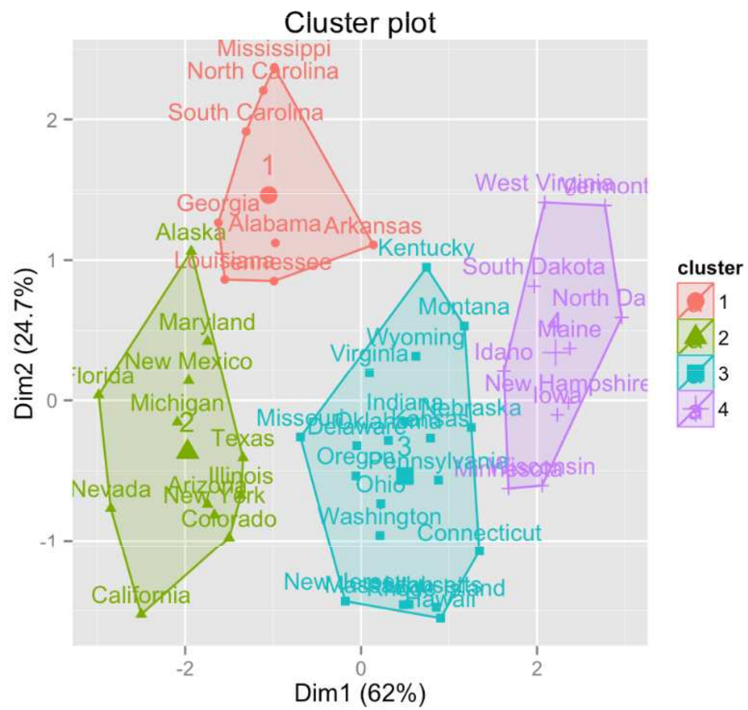


Jak popíšeme vícerozměrný prostor?

Popisné statistiky vícerozměrných dat

Charakteristiky polohy středu

- Udávají, kolem jaké hodnoty se data centrují.
- Centroid = vektor průměrných hodnot, reprezentuje virtuální střed.
- Medoid = reprezentuje reálný objekt.



Charakteristiky variability

- Zachycují rozptýlení hodnot v souboru.
- Kovarianční matice.
- Korelační matice.

Jaký je vztah mezi kovariancí a korelací?

Jaký je vztah mezi kovariancí a korelací?

- **Kovariance** popisuje vztah dvou proměnných; její rozsah závisí na variabilitě dat.

$$C(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{n-1}; C \in (-\infty; \infty)$$

- **Korelace** = kovariance standardizovaná na rozptyl proměnných.

$$r(x_1, x_2) = \frac{C(x_1, x_2)}{\sqrt{D(x_1)}\sqrt{D(x_2)}}; r \in \langle -1; 1 \rangle$$

- Jaké hodnoty se nachází na diagonále korelační a kovarianční matice?
- Má smysl použít metody redukce dimenzionality dat v situaci, kdy jsou hodnoty kovariance/korelace blízké nule?
- Čemu odpovídá kovariance na standardizovaných datech?

Chybějící data

Určete celkovou velikost souboru, která bude vstupovat do analýzy:

- A) Je průměrná hodnota systolického tlaku rovna 120 mmHg?
- B) Lze pacienty klasifikovat do skupin na základě systolického tlaku, tepové frekvence a saturace krve kyslíkem?

ID	Systolický tlak (mmHg)	Tepová frekvence (/min)	Saturace krve kyslíkem (%)
Xx_001	110	68	92
Xx_002	135	71	95
Xx_003	170	66	83
Xx_004	110	92	92
Xx_005	130		98
Xx_006	145	90	93
Xx_007	160	68	

Chybějící data - řešení

1) **„Complete case analysis“** – do analýzy zahrnujeme pouze pacienty, kteří mají kompletně vyplněná data → můžeme přijít o velké množství dat a tedy i jejich reprezentativnost.
2) **Imputace chybějících hodnot** – pomocí statistických přístupů odhadneme chybějící data → vnášíme do dat chybu, ale zachováváme jejich reprezentativnost.

   Např. balíček „mice“ (Multivariate Imputation by Chained Equations)

„Complete case analysis“

ID	Systolický tlak (mmHg)	Tepová frekvence (/min)	Saturace krve kyslíkem (%)
Xx_001	110	68	92
Xx_002	135	71	95
Xx_003	170	66	83
Xx_004	110	92	92
~~Xx_005~~	~~130~~		~~98~~
Xx_006	145	90	93
~~Xx_007~~	~~160~~	~~68~~	

Imputace chybějících hodnot

ID	Systolický tlak (mmHg)	Tepová frekvence (/min)	Saturace krve kyslíkem (%)
Xx_001	110	68	92
Xx_002	135	71	95
Xx_003	170	66	83
Xx_004	110	92	92
Xx_005	130	80	98
Xx_006	145	90	93
Xx_007	160	68	95

Co je to asociační matice?

- Jaké dimenze nabývá asociační matice?
- Co se nachází na diagonále asociační matice?
- Je matice symetrická kolem diagonály?

Asociační matice – Q mode analýza

NxP datová tabulka

NxN asociační matice

	p_1	p_2	p_3
n_1			
n_2			
n_3			
n_4			
n_5			

Hodnota subjektu n_5 v parametru p_1

Výpočet metriky vzdálenosti

	n_1	n_2	n_3	n_4	n_5
n_1	0				
n_2		0			
n_3			0		
n_4				0	
n_5					0

Vzdálenost subjektu n_5 od subjektu n_1.

	p_1	p_2	p_3
n_1			
n_2			
n_3			
n_4			
n_5			

Hodnota subjektu n_5 v parametru p_1

Výpočet metriky podobnosti

	n_1	n_2	n_3	n_4	n_5
n_1	1				
n_2		1			
n_3			1		
n_4				1	
n_5					1

Podobnost subjektu n_5 se subjektem n_1.

Asociační matice – R mode analýza

NxP datová tabulka

	p_1	p_2	p_3
n_1			
n_2			
n_3			
n_4			
n_5			

Hodnota subjektu n_5 v parametru p_1

Výpočet korelační/ kovarianční matice

PxP asociační matice

	p_1	p_2	p_3
p_1	**1**		
p_2		**1**	
p_3			**1**

Vztah parametru p_3 a parametru p_1.

Obecně:

- Základní výběr koeficientu je často spjat s metodou/algoritmem.
- Dále je potřeba zohlednit typ vstupních dat: spojitá/kategoriální/mix.
- Výběrem metriky ovlivníme výsledky analýz.

Koeficienty vzdálenosti

Kvantitativní data

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Brožová

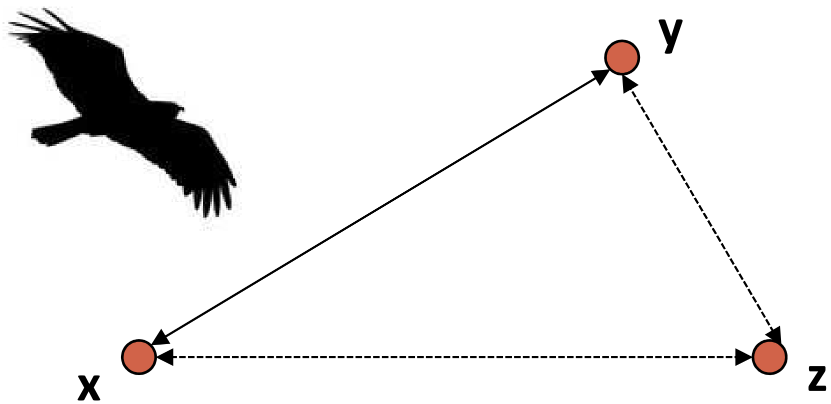
Koeficienty vzdálenosti

$$(i) \quad d(x,x) = 0;$$

$$(ii) \quad d(x,y) > 0, x \neq y;$$

$$(iii) \quad d(x,y) = d(y,x);$$

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

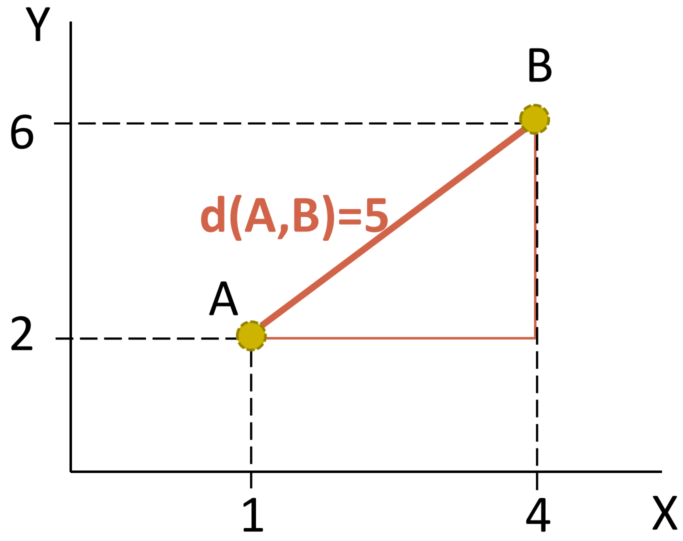


Podrobný přehled koeficientů vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998).** ***Numerical ecology*****. Elseviere Science BV, Amsterodam.**

Euklidova vzdálenost I

- Euklidova vzdálenost vychází z Pythagorovy věty:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$



- Jaká by byla vzdálenost bodů A a B dle Manhattanské metriky?

Euklidova vzdálenost II

	plat	počet cigaret/den
n_1	15 000	10
n_2	25 000	15
n_3	20 000	20
n_4	13 000	25
n_5	18 000	10

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

POZOR!

- Proměnné s číselně většími hodnotami budou mít větší váhu při shlukování!!!
- Např. pokud budeme hodnotit výšku (150–200 cm) a cholesterol (do 5 mmol/l), výška bude mít větší váhu při shlukování – objekty budou rozděleny do shluků podle jejich výšky.
- Data s nesrovnatelnými hodnotami proměnných je potřeba před analýzou **standardizovat**. Jak?
  - Standardizace na z-skóre.
  - Normalizace na rozsah 0–1.

Euklidova vzdálenost - příklad

- Pomocí MS Excel spočítejte Euklidovu vzdálenost subjektu n_2 a n_3 pro následující dva datové zdroje.

	BMI	váha	výška	cholesterol
n_2	24.9	72	170	5,1
n_3	25.8	98	195	5,2

$D(n_2,n_3) = ?$

	BMI	váha	výška	cholesterol
n_2	24.9	72	170	5,1
n_3	25.8	98	195	2,9

$D(n_2,n_3) = ?$

Euklidova vzdálenost - příklad

- Pomocí MS Excel spočítejte Euklidovu vzdálenost subjektu n_2 a n_3 pro následující dva datové zdroje.

	BMI	váha	výška	cholesterol
n_2	24.9	72	170	5,1
n_3	25.8	98	195	5,2

$D(n_2,n_3) = 36{,}08$

	BMI	váha	výška	cholesterol
n_2	24.9	72	170	5,1
n_3	25.8	98	195	2,9

$D(n_2,n_3) = 36{,}15$

Proč je vzdálenost téměř shodná, když se v druhém datovém souboru subjekty významně liší v hladině cholesterolu?

Euklidova vzdálenost III

	BMI	váha	výška
n_1	35.6	80	150
n_2	24.9	72	170
n_3	25.8	98	195
n_4	22.2	54	156
n_5	19.3	55	169

POZOR!

- U větších datových souborů, u kterých se často vyskytují korelované proměnné, dochází k nadhodnocení výsledků těmito korelovanými proměnnými = stejná informace je započtena více než jednou.
- Je potřeba zohlednit vztahy parametrů v datech → Mahalanobisova vzdálenost.

Mahalanobisova vzdálenost

$$D^2 = (x - \overline{x})^T S^{-1} (x - \overline{x})$$

Matice vzdáleností hodnot od průměru

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{bmatrix}$$

Inverze kovarianční matice

$$\begin{bmatrix} s_1^2 & \cdots & \mathrm{Cov}(s_n, s_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(s_1, s_n) & \cdots & s_n^2 \end{bmatrix}$$

- Odstraňuje vliv korelovaných parametrů.
- Dle volby $(x - \overline{x})$ lze hodnotit:
  1) **vzdálenosti objektů od centroidů** (vstupem je matice rozdílů původních hodnot od průměru).
  2) **vzdálenosti skupin objektů** (vstupem je matice rozdílů průměrných hodnot).
  3) **párové vzdálenosti jednotlivých subjektů** (vstupem je matice rozdílů srovnávaných subjektů).

Koeficienty podobnosti

Binární data

Koeficienty podobnosti

- Pokud proměnné popisují výskyt/nevýskyt = jsou tedy binárního typu, lze podobnost/odlišnost subjektů hodnotit dle tabulky níže:

	1	**0**	
1	a	b	a + b
0	c	d	c + d
	a + c	b + d	p = a + b + c + d

Co je to problém „double zero“?

Koeficienty podobnosti příklad I

- Úkol: Na základě datové matice doplňte tabulku.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
n_1	1	1	0	0	0	0	1
n_2	1	0	1	1	1	0	0

		n_1: 1	0	
n_2	1	a =	b =	a + b =
	0	c =	d =	c + d =
		a + c =	b + d =	p = a + b + c + d =

Koeficienty podobnosti příklad I

- Úkol: Na základě datové matice doplňte tabulku.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
n_1	1	1	0	0	0	0	1
n_2	1	0	1	1	1	0	0

		n_1: 1	n_1: 0	
n_2	1	a = 1	b = 3	a + b = 4
	0	c = 2	d = 1	c + d = 3
		a + c = 3	b + d = 4	p = a + b +c +d = 7

Jaké znáte koeficienty podobnosti?

Koeficienty podobnosti příklad II

- Úkol:
  1) Přiřaďte uvedené vzorce ke koeficientům podobnosti: „simple matching“, Jaccardův a Sørensenův koeficient podobnosti.
  2) Na základě získané tabulky spočítejte uvedené koeficienty.
  3) Podobnosti převeďte na vzdálenosti.

		n_1: 1	n_1: 0	
n_2	1	a = 1	b = 3	a + b = 4
	0	c = 2	d = 1	c + d = 3
		a + c = 3	b + d = 4	p = a + b +c +d = 7

- $S_{simple\ matching} = \rightarrow D =$
- $S_{Jaccard} = \rightarrow D =$
- $S_{Sørensen} = \rightarrow D =$

1. $$\frac{2a}{2a+b+c}$$

2. $$\frac{a+d}{a+b+c+d}$$

3. $$\frac{a}{a+b+c}$$

Koeficienty podobnosti příklad II

- Úkol:
  1) Přiřaďte uvedené vzorce ke koeficientům podobnosti: „simple matching“, Jaccardův a Sørensenův koeficient podobnosti.
  2) Na základě získané tabulky spočítejte uvedené koeficienty.
  3) Podobnosti převeďte na vzdálenosti.

		n_1: 1	n_1: 0	
n_2	1	a = 1	b = 3	a + b = 4
	0	c = 2	d = 1	c + d = 3
		a + c = 3	b + d = 4	p = a + b +c +d = 7

1. $$\frac{2a}{2a+b+c}$$

2. $$\frac{a+d}{a+b+c+d}$$

3. $$\frac{a}{a+b+c}$$

- $S_{simple\ matching}$= (1+1)/(1+2+3+1) = 0.3 → D = 0.7
- $S_{Jaccard}$= 1/(1+2+3) = 0.2 → D = 0.8
- $S_{Sørensen}$= 2*1/(2*1+2+3) = 0.3 → D = 0.8

Sørensenův asymetrický koeficient podobnosti pro kvantitavní data

- Tabulka popisuje abundance živočichů na dvou lokalitách.
- **Úkol**: Pomocí Sørensenova koeficientu vyhodnoťte, zda si jsou uvedené lokality podobné.

	Výskyt/nevýskyt živočicha									
Lokalita	**žralok**	**velryba**	**had**	**ještěrka**	**velbloud**	**varan**	**tučňák**	**aN**	**bN**	**jN**
Vysočina	0	0	2	3	0	0	0	5		
Sahara	0	0	4	1	5	6	0		16	
Minimum	0	0	2	1	0	0	0			3

$$C_N = \frac{2\,jN}{(aN + bN)} = \frac{2 \cdot 3}{(5 + 16)} = 0.3$$

Gowerův obecný koeficient podobnosti

Mix kategoriálních a kvantitativních dat

Gowerův obecný koeficient podobnosti

- Kombinuje různé typy deskriptorů.
- Podobnost mezi dvěma objekty je vypočítána jako průměr podobností, vypočítaných pro všechny deskriptory:

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

  - ✓ Pro kategoriální deskriptory s = 1 (shoda) nebo 0 (neshoda).
  - ✓ Kvantitativní deskriptory (reálná čísla): rozdíl mezi stavy obou objektů je vydělen největším rozdílem (R_j), nalezeným pro daný deskriptor mezi všemi objekty ve studii.

Asociační matice

Asociační matice vzdáleností

STATISTICA - [Data: Irisdat* (5v by 150c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Win

Add to Workbook Add to Rep

Arial 10

Fisher (1936) iris data: length & width of sepals and petals, 3 types of I

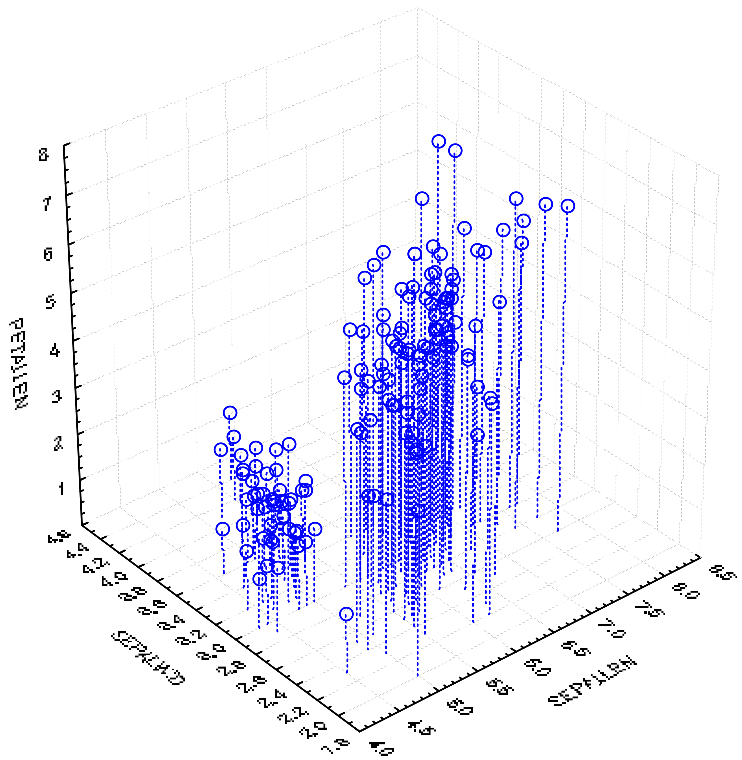
	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5.0	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6	[illegible]	[illegible]
3	6.5	2.8	4.6	[illegible]	[illegible]
4	6.7	3.1	5.6	[illegible]	[illegible]
5	6.3	2.8	5.1	1.5	VIRGINIC
6	4.6	3.4	1.4	0.3	SETOSA
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSICO
9	5.9	3.2	4.8	1.8	VERSICO
10	4.6	3.6	1.0	0.2	SETOSA
11	6.1	3.0	4.6	1.4	VERSICO
12	6.0	2.7	5.1	1.6	VERSICO
13	6.5	3.0	5.2	2.0	VIRGINIC
14	5.6	2.5	3.9	1.1	VERSICO
15	6.5	3.0	5.5	1.8	VIRGINIC
16	5.8	2.7	5.1	1.9	VIRGINIC
17	6.8	3.2	5.9	2.3	VIRGINIC
18	5.1	3.3	1.7	0.5	SETOSA
19	5.7	2.8	4.5	1.3	VERSICO
20	6.2	3.4	5.4	2.3	VIRGINIC
21	7.7	3.8	6.7	2.2	VIRGINIC
22	6.3	3.3	4.7	1.6	VERSICO
23	6.7	3.3	5.7	2.5	VIRGINIC
24	7.6	3.0	6.6	2.1	VIRGINIC
25	4.9	2.5	4.5	1.7	VIRGINIC
26	5.5	3.5	1.3	0.2	SETOSA
27	6.7	3.0	5.2	2.3	VIRGINIC
28	7.0	3.2	4.7	1.4	VERSICO
29	6.4	3.2	4.5	1.5	VERSICO
30	6.1	2.8	4.0	1.3	VERSICO
31	4.8	3.1	1.6	0.2	SETOSA
32	5.9	3.0	5.1	1.8	VIRGINIC
33	5.5	2.4	3.8	1.1	VERSICO
34	6.3	2.5	5.0	1.9	VIRGINIC

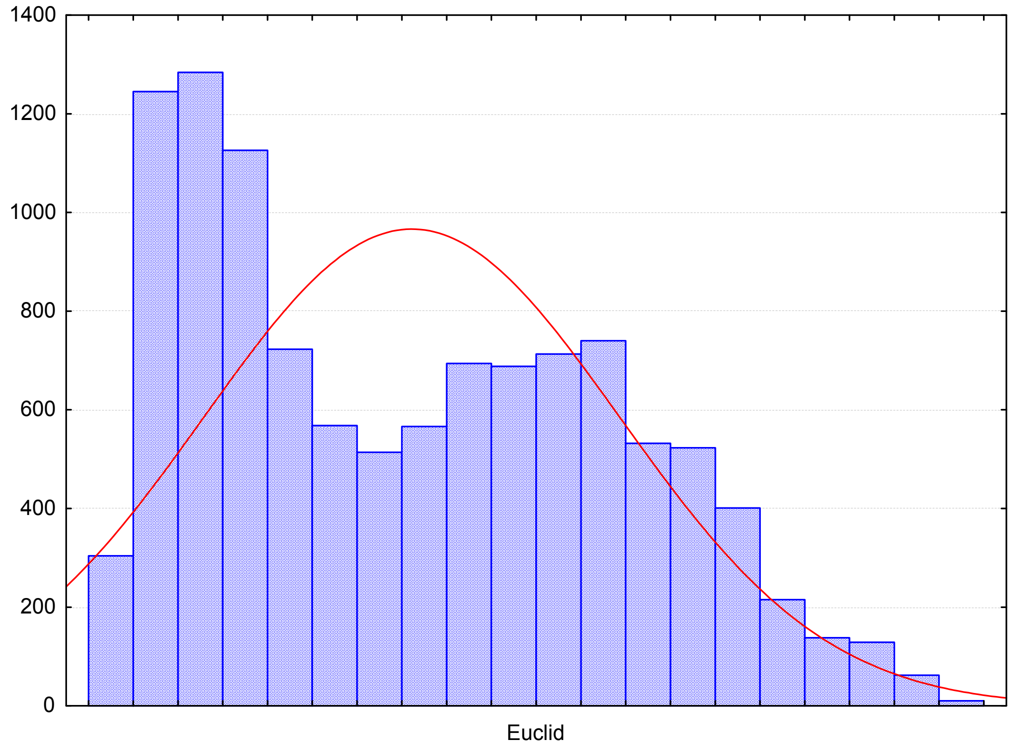
Euclidean distances (Irisdat)

Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17
C_1	0.00	4.88	3.80	5.04	4.16	0.42	4.66	3.73	3.87	0.64	3.60	4.12	4.47	2.84	4.66	4.19	5.28
C_2	4.88	0.00	1.22	0.47	0.87	4.98	0.77	1.45	1.10	5.39	1.33	0.88	0.50	2.20	0.47	0.84	0.65
C_3	3.80	1.22	0.00	1.39	0.54	3.96	1.07	0.68	0.81	4.35	0.46	0.72	0.81	1.24	0.97	0.95	1.61
C_4	5.04	0.47	1.39	0.00	1.14	5.15	0.55	1.75	1.28	5.54	1.54	1.24	0.61	2.48	0.65	1.21	0.35
C_5	4.16	0.87	0.54	1.14	0.00	4.29	1.04	0.85	0.71	4.69	0.58	0.33	0.58	1.48	0.57	0.65	1.30
[illegible]	[illegible]	4.98	3.96	5.15	4.29	0.00	4.80	3.88	3.94	0.46	3.72	4.22	4.59	2.95	4.78	4.26	5.40
[illegible]	[illegible]	0.77	1.07	0.55	1.04	4.80	0.00	1.52	1.16	5.17	1.31	1.21	0.52	2.22	0.76	1.24	0.81
C_8	[illegible]	1.45	0.68	1.75	0.85	3.88	1.52	0.00	1.13	4.30	0.82	0.81	1.21	0.98	1.35	0.96	1.99
C_9	3.87	1.10	0.81	1.28	0.71	3.94	1.16	1.13	0.00	4.34	0.53	0.62	0.77	1.37	0.94	0.60	1.51
C_10	0.64	5.39	4.35	5.54	4.69	0.46	5.17	4.30	4.34	0.00	4.12	4.64	4.98	3.38	5.17	4.69	5.78
C_11	3.60	1.33	0.46	1.54	0.58	3.72	1.31	0.82	0.53	4.12	0.00	0.62	0.94	1.04	1.06	0.82	1.74
C_12	4.12	0.88	0.72	1.24	0.33	4.22	1.21	0.81	0.62	4.64	0.62	0.00	0.71	1.37	0.73	0.36	1.42
C_13	4.47	0.50	0.81	0.61	0.58	4.59	0.52	1.21	0.77	4.98	0.94	0.71	0.00	1.89	0.36	0.77	0.84
C_14	2.84	2.20	1.24	2.48	1.48	2.95	2.22	0.98	1.37	3.38	1.04	1.37	1.89	0.00	2.03	1.47	2.71
C_15	4.66	0.47	0.97	0.65	0.57	4.78	0.76	1.35	0.94	5.17	1.06	0.73	0.36	2.03	0.00	0.87	0.73
C_16	4.19	0.84	0.95	1.21	0.65	4.26	1.24	0.96	0.60	4.69	0.82	0.36	0.77	1.47	0.87	0.00	1.43
C_17	5.28	0.65	1.61	0.35	1.30	5.40	0.81	1.99	1.51	5.78	1.74	1.42	0.84	2.71	0.73	1.43	0.00
C_18	0.44	4.48	3.41	4.63	3.77	0.62	4.25	3.36	3.46	0.96	3.21	3.73	4.07	2.47	4.26	3.79	4.88
C_19	3.40	1.58	0.83	1.87	0.87	3.49	1.70	0.81	0.73	3.91	0.47	0.74	1.29	0.71	1.39	0.86	2.08
C_20	4.68	0.67	1.32	0.62	1.05	4.75	0.82	1.70	0.86	5.14	1.27	1.05	0.62	2.20	0.71	0.95	0.81

Asociační matice euklidovských vzdáleností mezi rostlinami

Histogram jako popis asociační matice





Základní funkce v R pro výpočet asociační matice

Funkce v R pro výpočet asociační matice

Koeficienty vzdálenosti

`dist(data, method='euclidean')`= Euklidova vzdálenost

`mahalanobis(X, X.mean, X.cov)`= Mahalanobisova vzdálenost

`pairwise.mahalanobis (X,grouping)`= Mahalanobisova vzdálenost mezi skupinami `{HDMD}`

Koeficienty podobnosti

`dist.binary(data, method = 2) =` „simple matching“ koeficient `{ade4}`

`vegdist(data, "jac", binary=T)=` Jaccardův koeficient `{vegan}`

`vegdist(data, "jac", binary=F) =` Sørensenův asymetrický koeficient

Podobnosti jsou pomocí funkce 1-podobnost automaticky převáděny na vzdálenosti!!!

Výpočet **Mantelova testu** (testuje korelaci dvou asociačních matic – např. průběh onemocnění vs. genetická výbava pacientů, charakteristiky lokalit vs. abundance druhů na lokalitách): `mantel{vegan}`