Wielowymiarowe modele w analizie danych biologicznych

Monika Mokrzycka

Instytut Genetyki Roślin PAN w Poznaniu

Warsztaty, Politechnika Warszawska

Dane metabolomiczne

- dane jęczmienne (*Hordeum vulgare* L.)
- chromatografia gazowa ze spektrometrią mas (GC-MS)
- odporność na suszę

- surowe dane: 51135 charakterystyk dla 422 próbek
- 211 prób biologicznych po uśrednieniu po powtórzeniach technicznych
- 781 charakterystyk po wstępnym przetworzeniu

Dane metabolomiczne

Dane dostępne:
https://github.com/adammieldzioc/Barley-data

Dane metabolomiczne

Dane: 781×213

Scan_Nr	Ret_umin	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
111	8643700	301.5	190.5	196.5	227.0	203.0	272.5	228.0	209.5	230.0	177.0	184.5	190.0	224.5	271.5	251.0
112	8677050	1447.0	2767.5	1440.5	3730.0	2291.5	1994.5	1732.5	1428.5	2690.0	1799.5	2393.0	1333.0	1570.0	2106.0	3526.5
113	8710383	1233.5	1178.5	1248.0	1155.5	1000.5	556.5	487.0	682.0	636.0	1247.0	539.0	470.5	487.0	1166.5	1162.0
114	8743733	693.0	556.0	557.5	701.5	559.5	601.0	546.0	487.0	566.5	427.5	456.0	472.0	465.5	628.0	631.0
115	8777083	5855.5	6019.5	5083.0	3746.0	3499.0	4323.5	4251.5	5773.5	6521.0	3390.0	4059.0	4729.0	2091.5	5718.0	5994.0
116	8810416	632.0	534.0	573.0	2324.5	3099.0	2371.5	2323.0	535.0	608.5	1826.0	497.5	607.5	1300.0	537.0	686.5
117	8843766	871.5	1127.5	1114.0	905.0	829.5	655.0	608.5	591.0	654.0	665.0	593.5	587.0	566.0	832.0	901.0
118	8877017	3553.5	3025.5	2962.5	2051.5	1774.0	1475.5	1754.5	1482.0	5078.5	1227.5	1478.0	2475.5	1744.0	2613.5	2730.5
119	8910367	1074.5	1219.5	1193.0	1241.0	1275.0	822.5	806.0	680.0	822.5	794.0	666.5	697.0	1267.5	1427.0	1053.5
120	8943700	1161.5	450.0	461.5	464.5	1857.0	482.5	555.5	1780.5	932.5	737.5	1017.0	359.5	365.0	491.5	4652.0
121	8977050	615.0	1997.0	1690.0	2646.5	3066.0	2169.5	423.5	517.0	602.5	612.5	424.5	872.5	512.0	600.0	659.0
122	9010400	1548.0	1550.0	1468.0	1408.0	1315.0	894.5	1442.5	878.0	889.5	1002.0	752.0	719.5	971.5	1571.5	1492.0
123	9043734	1965.5	1728.5	1508.0	1570.5	1344.5	1311.5	1401.0	1319.5	1424.5	1175.5	1125.5	1086.5	730.0	1974.5	2151.5
124	9077084	1492.5	959.0	1063.0	1327.5	1491.0	1151.5	1045.0	913.5	989.5	1001.0	842.0	966.5	839.5	1488.5	1658.5
125	9110434	2663.5	1911.5	2208.5	2318.0	2387.0	2202.0	1999.5	1833.0	1964.5	1757.5	1803.5	1715.0	1797.5	2528.5	2384.0
126	9143766	3924.0	4451.5	3814.0	3774.5	3830.5	7101.5	4062.5	3824.0	4952.5	2679.0	4081.5	5620.5	3775.5	3679.0	4110.0
127	9177016	30424.0	8609.0	27016.5	22143.5	31923.0	25187.5	7609.0	5695.0	5447.0	4580.0	9635.5	7056.0	27737.0	8295.5	19328.0
128	9210366	6372.0	8307.5	3229.0	3252.5	8479.5	2740.5	2679.5	2661.0	2704.0	2362.0	2465.0	2355.0	5350.5	3564.5	3411.0
129	9243716	3057.0	2809.0	2731.5	8631.5	2778.5	1972.5	1911.0	2045.5	2042.5	1840.0	1817.0	1732.0	1735.0	3074.0	2903.5
130	9277050	2512.5	6596.0	4838.5	9338.0	9527.0	6540.0	4744.0	3534.5	6167.5	3522.0	5224.5	5608.5	2554.0	2466.5	4566.5
131	9310400	10837.0	4422.5	5479.5	8993.5	23851.5	7205.5	9219.0	3293.5	3348.0	7356.0	3195.5	3117.0	4308.5	6165.0	17068.0

Dane metabolomiczne

Podziel dane na podzbiory:

- $\mathbf{X}_1$ - cechy od 1 do 250
- $\mathbf{X}_2$ - cechy od 251 do 500
- $\mathbf{X}_3$ - cechy od 501 do 750

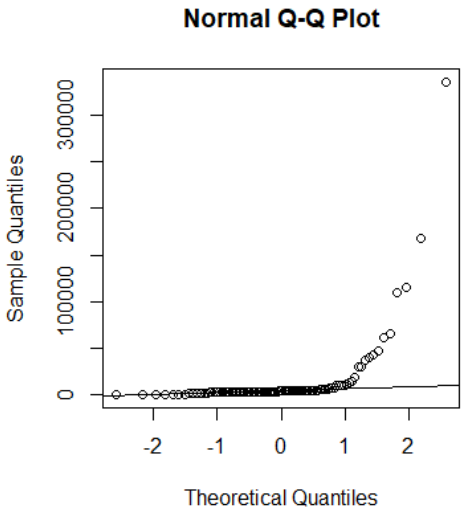
Dane metabolomiczne

- dane jęczmienne (*Hordeum vulgare* L.)
- chromatografia gazowa ze spektrometrią mas (GC-MS)
- odporność na suszę

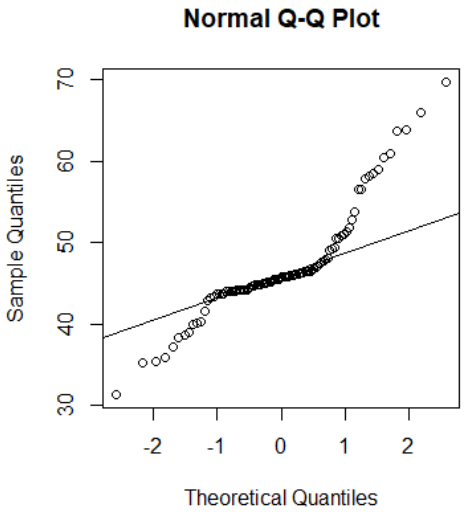
- surowe dane: 51135 charakterystyk dla 422 próbek
- 211 prób biologicznych po uśrednieniu po powtórzeniach technicznych
- 781 charakterystyk po wstępnym przetworzeniu
- przekształcone przez logarytm

Dane metabolomiczne

$\mathbf{X}_1[,1]$



$\log \mathbf{X}_1[,1]$



Identyfikacja struktury

Struktury kowariancyjne $m \times m$:

- kompletnej symetrii (CS)

Metody identyfikacji struktury:

- norma Frobeniusa
- entropijna funkcja straty

$\mathscr{S}_{CS}$

$$\zeta = \min_{\mathbf{\Gamma} \in \mathscr{S}_{CS}} f(\mathbf{\Omega}, \mathbf{\Gamma})$$

$\mathbf{\Omega}$ - nieznana

MLE($\mathbf{\Omega}$)

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{Q}_{1_n}\mathbf{X}'$$

gdzie

- $\mathbf{X}$ - macierz obserwacji $\quad \mathbf{X} \sim N_{q,n}(\mu \mathbf{1}_n', \mathbf{\Omega}, \mathbf{I}_n)$
- $\mathbf{Q}_{1_n} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$

$$\begin{aligned}\mathbf{S}_1 &= \tfrac{1}{n}\mathbf{X}_1\mathbf{Q}_{1_n}\mathbf{X}_1' \\ \mathbf{S}_2 &= \tfrac{1}{n}\mathbf{X}_2\mathbf{Q}_{1_n}\mathbf{X}_2' \\ \mathbf{S}_3 &= \tfrac{1}{n}\mathbf{X}_3\mathbf{Q}_{1_n}\mathbf{X}_3'\end{aligned}$$

$\log \mathbf{X}_i$

$$\det \mathbf{S}_i > 0, \qquad i = 1, \cdots, 3$$

Struktury kowariancyjne

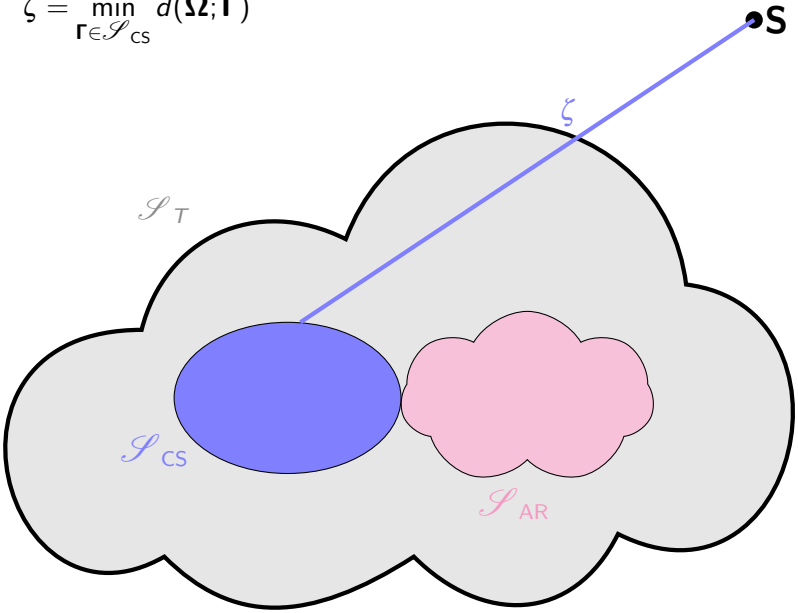
- kompletnej symetrii (CS)

$$\mathbf{\Gamma}_{CS} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix} = \sigma^2 \left[(1-\rho)\mathbf{I}_m + \rho \mathbf{1}_m \mathbf{1}'_m \right]$$

$$\sigma^2 > 0, \qquad \rho \in \left(-\frac{1}{m-1}; 1 \right)$$

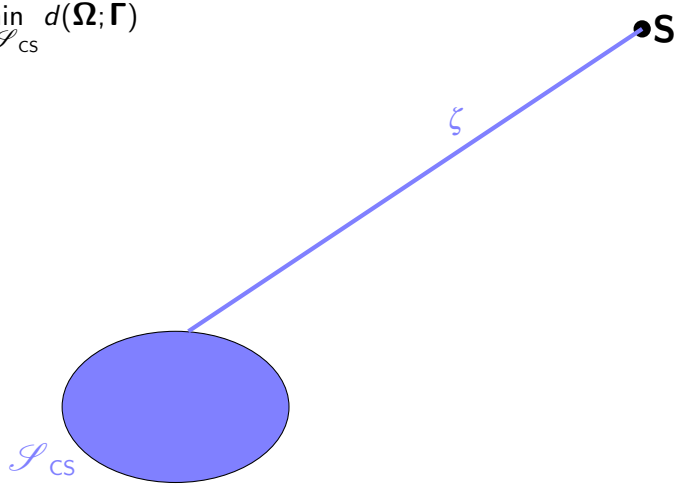
Identyfikacja struktury

$$\zeta = \min_{\mathbf{\Gamma} \in \mathscr{S}_{\mathrm{CS}}} d(\mathbf{\Omega}; \mathbf{\Gamma})$$



Identyfikacja struktury

$$\zeta = \min_{\mathbf{\Gamma} \in \mathscr{S}_{\mathrm{CS}}} d(\mathbf{\Omega}; \mathbf{\Gamma})$$



$\mathbf{S}, \mathbf{\Gamma} \in \mathbb{R}_m^>$

Frobenius norm

$$f_F(\mathbf{S}, \mathbf{\Gamma}) = ||\mathbf{S} - \mathbf{\Gamma}||_F = \sqrt{\mathrm{tr}\left[(\mathbf{S} - \mathbf{\Gamma})^2\right]}$$

Entropy loss function

$$f(\mathbf{S}, \mathbf{\Gamma}) \quad = \quad \mathrm{tr}\left(\mathbf{S}^{-1}\mathbf{\Gamma}\right) - \ln\left(\det\left[\mathbf{S}^{-1}\mathbf{\Gamma}\right]\right) - m$$

$$\zeta_F = \min_{\mathbf{\Gamma} \in \mathscr{S}_{\mathrm{CS}}} f_F(\mathbf{S}, \mathbf{\Gamma}) \qquad \text{i} \qquad \zeta_E = \min_{\mathbf{\Gamma} \in \mathscr{S}_{\mathrm{CS}}} f_E(\mathbf{S}, \mathbf{\Gamma})$$

Aproksymacja $\mathbf{\Gamma} \in \mathscr{S}_{CS}$

norma Frobeniusa:

$$\begin{cases} \rho & = & \frac{\alpha}{(m-1)\,\mathrm{tr}(\mathbf{S}')} \\ \sigma^2 & = & \frac{\mathrm{tr}(\mathbf{S}'+\rho\alpha)}{m+m(m-1)\rho^2} \end{cases} \qquad \alpha = \mathrm{tr}\left[\mathbf{S}'(\mathbf{1}_m\mathbf{1}_m' - \mathbf{I}_m)\right]$$

entropijna funkcja straty:

$$\begin{cases} \rho & = & -\frac{\alpha}{(m-1)\,\mathrm{tr}(\mathbf{S}^{-1})+(m-2)\alpha} \\ \frac{m}{\sigma^2} & = & \mathrm{tr}(\mathbf{S}^{-1}) + \rho\alpha \end{cases} \qquad \alpha = \mathrm{tr}\left[\mathbf{S}^{-1}(\mathbf{1}_m\mathbf{1}_m' - \mathbf{I}_m)\right]$$

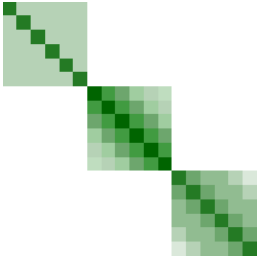
Rozbieżności

Skorygowane wartości:

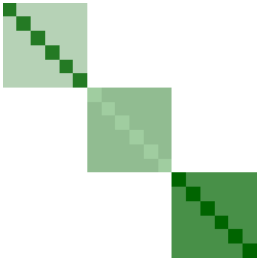
- $\xi_F = \zeta_F / ||\mathbf{S}||_F$
- $\xi_E = 1 - 1/(1+\zeta_E)$
- $\xi_k = 1 - 1/(1+\log \zeta_k), \quad k \in \{F, E\}$

Rozbieżności i oceny

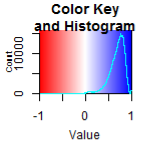
struktura	podzbiór	ζ_F	ζ_E	ξ_F	ξ_E
CS	1	1399.1336	655.5110	0.3271	0.7380
	2	1387.0465	588.9344	0.2992	0.7348
	3	1512.8434	619.9874	0.2852	0.7364



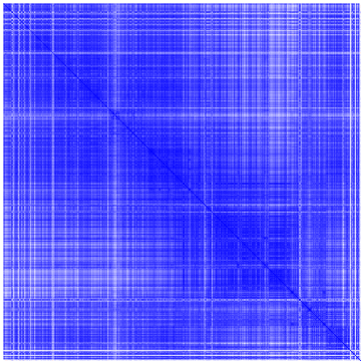
estymator	podzbiór	σ^2	ρ	ρ_1	ρ_2
$\widehat{\mathbf{\Gamma}}_{CS}$	1	30.7322	0.6555	-	-
	2	29.5329	0.7475	-	-
	3	33.2362	0.7635	-	-
$\widetilde{\mathbf{\Gamma}}_{CS}$	1	0.0632	0.5145	-	-
	2	0.0599	0.4558	-	-
	3	0.0443	0.3737	-	-



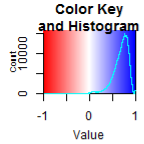
All data

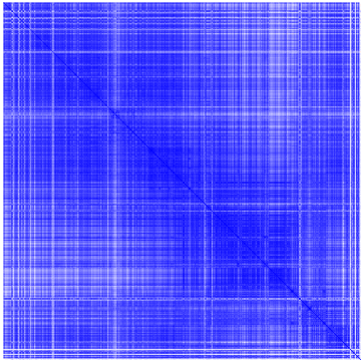


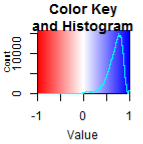
- $\mathbf{X}_{781\times211} \sim N_{m,n}(\mu \mathbf{1}_n', \mathbf{\Omega}, \mathbf{I}_n)$
- $\mathbf{S}_{781\times781}$ - MLE of $\mathbf{\Omega}$
- $\det(\mathbf{S}) = 0$
- Macierz korelacji $\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$
- $\mathbf{D}^{-1} = diag(\frac{1}{\sqrt{S_{11}}}, \frac{1}{\sqrt{S_{22}}}, ..., \frac{1}{\sqrt{S_{mm}}})$

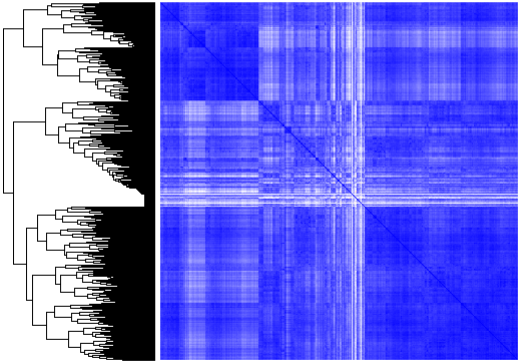


All data


Color Key
and Histogram
Count
10000
0
-1
0
1
Value

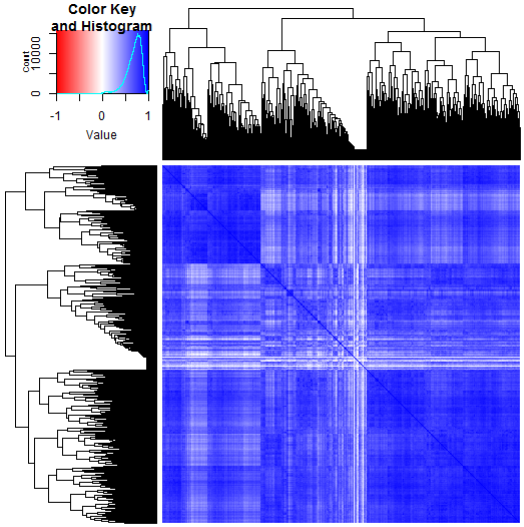



Color Key
and Histogram
Count
10000
0
-1
0
1
Value



Heatmapy

- *heatmap.2* w pakiecie *gplots*
- *hclust* w pakiecie *stats*
- dendrogramy



Grupowanie hierarchiczne

Miara niepodobieństwa

m obiektów

- na starcie każdy obiekt jest w swoim klastrze
- następnie najbardziej podobne grupy są łączone iteracyjnie do momentu, gdy wszystkie elementy będą należały do jednego klastra
- odległości między elementami - funkcja odległości
- odległości między klastrami - kryterium połączenia

Funkcje odległości

Odległości między elementami

- Euclidean - $\sqrt{\sum_{i=1}^{m}(\mathbf{u}_i - \mathbf{v}_i)^2}$
- Maximum - $\max_i |\mathbf{u}_i - \mathbf{v}_i|$
- Manhattan - $\sum_{i=1}^{m} |\mathbf{u}_i - \mathbf{v}_i|$
- Canberra - $\sum_{i=1}^{m} |\mathbf{u}_i - \mathbf{v}_i| / |\mathbf{u}_i + \mathbf{v}_i|$
- Binary - Jaccard index $J(A,B) = \left(\overline{\overline{A \cap B}}\right) / \left(\overline{\overline{A \cup B}}\right)$
- Minkowski - $\left(\sum_{i=1}^{m} |\mathbf{u}_i - \mathbf{v}_i|^k\right)^{\frac{1}{k}},\ k \geqslant 1$

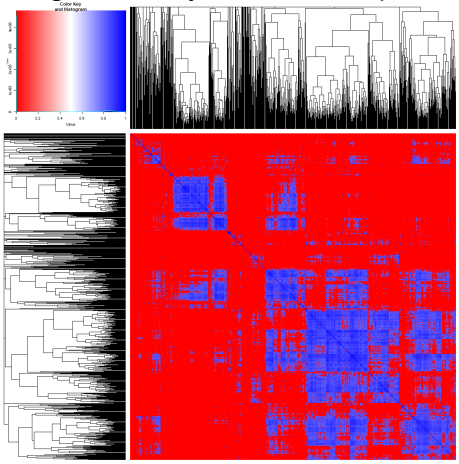
Kryteria połączenia

Odległości między klastrami, $D(A, B)$
A, B - klastry

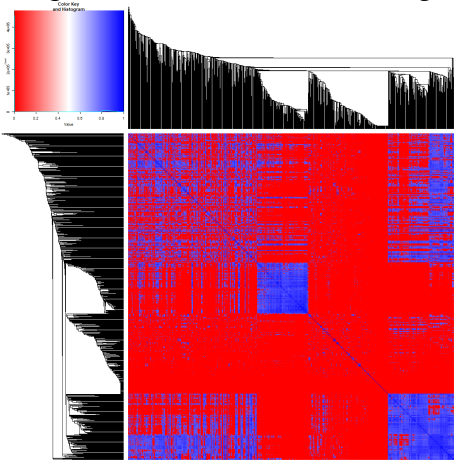
- Ward.D and Ward.D2 - używają analizy wariancji
- Single - $D(A, B) = \min_{a \in A, b \in B} d(a, b)$
- Complete - $D(A, B) = \max_{a \in A, b \in B} d(a, b)$
- Average (UPGMA - Unweighted Pair-Group Method using Arithmetic Averages) - $D(A, B) = \text{mean}_{a \in A, b \in B}\, d(a, b)$
- Mcquitty (WPGMA - Weighted Pair Group Method with Arithmetic Mean) - bazuje na UPGMA używając rozmiarów klastrów jako wag
- Centroid (UPGMC - Unweighted Pair-Group Method using the Centroid Average) - $D(A, B) = d(centroid(A), centroid(B))$
- Median (WPGMC - Weighted Pair-Group Method using the Centroid Average) - bazuje na UPGMC używając rozmiarów klastrów jako wag

All data

odległość: *binary*, metoda: *complete*

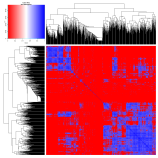


odległość: *manhattan*, metoda: *single*

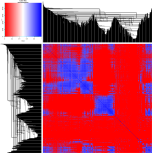


All data

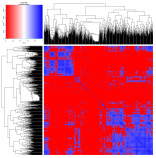
odległość: *minkowski*
metoda: *average*



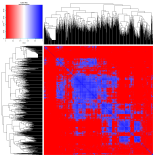
odległość: *minkowski*
metoda: *centroid*



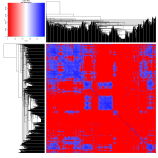
odległość: *minkowski*
metoda: *complete*



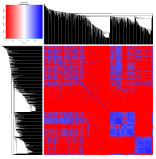
odległość: *minkowski*
metoda: *mcquitty*



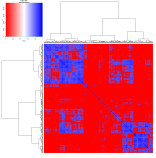
odległość: *minkowski*
metoda: *median*



odległość: *minkowski*
metoda: *single*



odległość: *minkowski*
metoda: *ward.D*



odległość: *minkowski*
metoda: *ward.D2*