Wielowymiarowe modele w analizie danych biologicznych

Monika Mokrzycka

Instytut Genetyki Roślin PAN w Poznaniu

Warsztaty, Politechnika Warszawska

Zadanie

Identyfikacja struktury macierzy kowariancji dla

- danych metabolomicznych
- danych fenotypowych

Grupy: 1-4 osoby

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 8 podzbiorów:

- $\mathbf{X}_1$ - cechy od 1 do 100
- $\mathbf{X}_2$ - cechy od 101 do 200
- $\mathbf{X}_3$ - cechy od 201 do 300
- ...
- $\mathbf{X}_7$ - cechy od 601 do 700
- $\mathbf{X}_8$ - cechy od 701 do 781

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

Identyfikacja struktury

Struktury kowariancyjne $m \times m$:

- kompletnej symetrii (CS)
- trójdiagonalne macierze Toeplitza (T_1)
- pięciodiagonalne macierze Toeplitza (T_2)
- autoregresji stopnia pierwszego (AR)
- iloczyn Kroneckera $\mathbf{\Psi} \otimes \mathbf{\Sigma}$
- iloczyn Kroneckera $\mathbf{\Psi}_{\mathrm{CS}} \otimes \mathbf{\Sigma}$
- iloczyn Kroneckera $\mathbf{\Psi}_{\mathrm{AR}} \otimes \mathbf{\Sigma}$

Metody identyfikacji struktury:

- norma Frobeniusa
- entropijna funkcja straty
- kwadratowa funkcja straty

$\mathscr{G}$: $\mathscr{S}_{CS}$, $\mathscr{S}_{T_1}$, $\mathscr{S}_{T_2}$, $\mathscr{S}_{AR}$, $\mathscr{S}_{\otimes}$, $\mathscr{S}_{\mathrm{CS}\otimes}$, $\mathscr{S}_{\mathrm{AR}\otimes}$

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

$\mathbf{\Omega}$ - nieznana

MLE($\mathbf{\Omega}$)

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

gdzie

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

$$
\begin{aligned}
\mathbf{S}_1 &= \tfrac{1}{n}\mathbf{X}_1\mathbf{Q}_{1_n}\mathbf{X}_1' \\
\mathbf{S}_2 &= \tfrac{1}{n}\mathbf{X}_2\mathbf{Q}_{1_n}\mathbf{X}_2' \\
&\vdots \\
\mathbf{S}_8 &= \tfrac{1}{n}\mathbf{X}_8\mathbf{Q}_{1_n}\mathbf{X}_8'
\end{aligned}
$$

$\log \mathbf{X}_i$

$$\det \mathbf{S}_i > 0, \qquad i = 1, \cdots, 8$$

Twierdzenia

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

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:

- $\xi_F = \zeta_F / ||\mathbf{S}||_F$
- $\xi_E = 1 - 1/(1+\zeta_E)$
- $\xi_Q = 1 - 1/(1+\zeta_Q)$

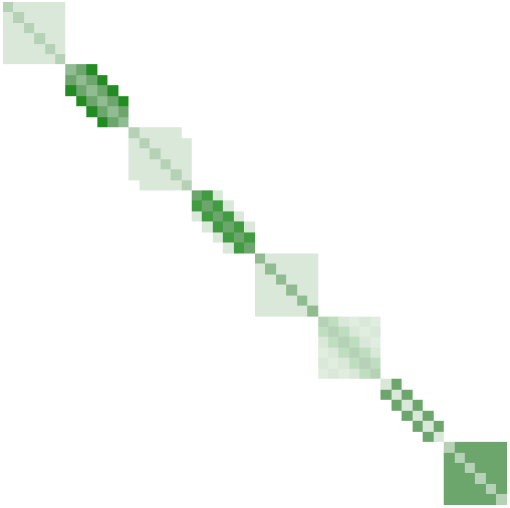
- $\xi_k = 1 - 1/(1+\log \zeta_k), \quad k \in \{F, E, Q\}$

Rozbieżności i oceny

struktura	podzbiór	ζ_F	ζ_E	ζ_Q	ξ_F	ξ_E	ξ_Q
CS	1						
	2						
	⋮						
	8						
AR	1						
	2						
	⋮						
	8						
⋮	1						
	2						
	⋮						
	8						

estymator	podzbiór	σ^2	ρ	ρ_1	ρ_2
$\widehat{\mathbf{\Gamma}}_{CS}^{(F)}$	1			-	-
	2			-	-
	⋮			-	-
	8			-	-
$\widehat{\mathbf{\Gamma}}_{CS}^{(E)}$	1			-	-
	2			-	-
	⋮			-	-
	8			-	-
$\widehat{\mathbf{\Gamma}}_{CS}^{(Q)}$	1			-	-
	2			-	-
	⋮			-	-
	8			-	-
⋮	1			-	-
	2			-	-
	⋮			-	-
	8			-	-

Wizualizacja



Zadanie

Dane fenotypowe z wysokoprzepustowej platformy

Fenotypowanie wysokoprzepustowe: film

- dane jęczmienne (*Hordeum vulgare* L.)
- kamery górne i boczne
- zakres widma światła, sygnały fluorescencyjne i widmo bliskiej podczerwieni

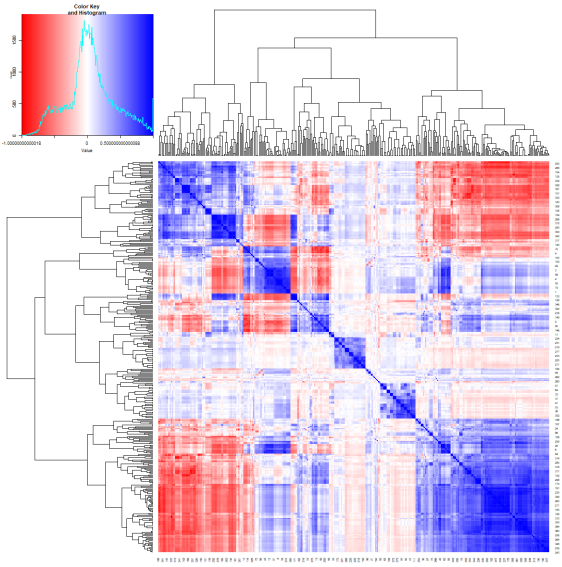
- surowe dane: 376 cech dla 12672 próbek
- 396 pomiarów dla 366 cech
- różne odmiany

Zadanie

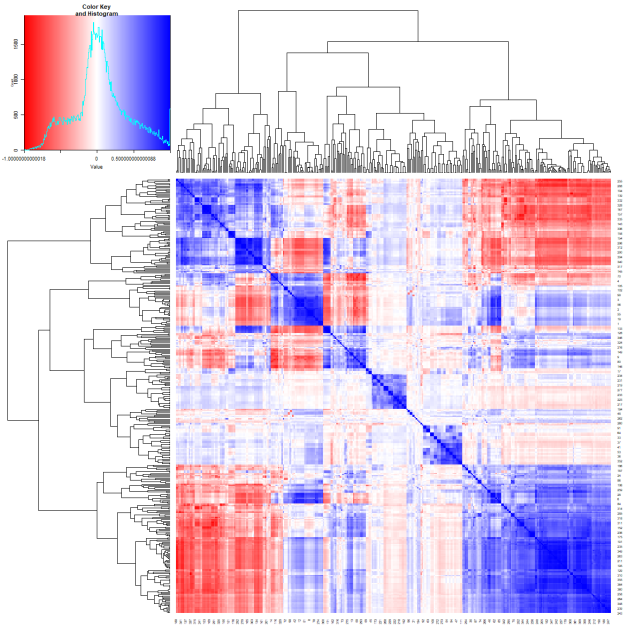
Treat	txline	txcarr	Day!	trait1	trait2	trait3	trait4	trait5	trait6	trait7	trait8	trait9	trait10	tra
Drought	Cam/B1/Cl	1436PK1222	4	33616,3201	3000,20424	3830,5	877	54615,75	13	398,5911	312,405495	0,0145067	398	7
Drought	Cam/B1/Cl	1436PK1252	4	29352,6953	2717,96668	3400,75	803,75	45952,5	16,25	347,430527	277,675393	0,0164566	344,75	6
Drought	Maresi	1436PK1091	4	32191,8311	2484,99239	3109,25	575,5	50397,25	12,75	337,536545	269,767869	0,02238344	334	6
Drought	Maresi	1436PK1293	4	28572,6236	2548,45251	3358,75	693	49674,625	17	355,410703	269,673352	0,01883057	354,5	6
Drought	MCam001	1436PK1109	4	30045,8581	2634,43059	3406,75	610,75	50516,75	15	378,637539	292,26162	0,02084656	378	6
Drought	MCam001	1436PK1242	4	34893,3859	2869,59791	3706,5	665,25	57913	16,5	358,422531	277,043904	0,01968387	352,75	6
Drought	MCam002	1436PK1019	4	31362,3379	2843,04818	3669,25	635	52256,375	16,25	336,50587	260,287523	0,02084116	333	6
Drought	MCam002	1436PK1111	4	32632,8162	2620,25558	3389,5	538,25	54804,5	16,75	390,523762	300,99323	0,02478887	381,5	7
Drought	MCam004	1436PK1258	4	29990,6613	2567,78891	3436,75	592,25	53723,375	12	367,741188	274,760091	0,02134318	360,25	6
Drought	MCam004	1436PK1353	4	25613,3231	2404,55633	3055,5	591,5	41396,875	12,75	313,059065	245,68159	0,02221782	306,25	5
Drought	MCam005	1436PK1166	4	29512,4341	2320,63721	2958	469	48062,375	10,25	316,141752	247,67824	0,02913421	315,25	6
Drought	MCam005	1436PK1234	4	29782,5595	2663,66956	3401	667,25	48553,5	14,5	333,557144	261,235287	0,01951676	331	6
Drought	MCam006	1436PK1043	4	33819,316	2655,22746	3322,25	581	52945,125	17,75	361,278865	288,74334	0,02271297	359	6
Drought	MCam006	1436PK1175	4	29365,6529	2328,84911	2967,5	519,25	47639,875	10,5	337,90239	264,94982	0,02481685	326	6
Drought	MCam007	1436PK1100	4	34331,08	2837,58303	3621,5	618	55882,625	14,5	371,09271	290,216593	0,02192412	366,25	6
Drought	MCam007	1436PK1312	4	33132,6472	2755,53028	3447,75	597,75	51870,125	18,75	363,391549	290,431852	0,02139566	363	6
Drought	MCam008	1436PK1084	4	30248,3509	2586,62817	3463,75	604,75	54243,5	19	360,709264	269,365398	0,02086042	358,5	6
Drought	MCam008	1436PK1287	4	31950,9558	2780,16722	3480,5	652,25	50075,5	17,25	355,878948	284,270359	0,02024113	347,25	6
Drought	MCam009	1436PK1151	4	31870,1069	2666,50225	3410,5	699,75	51943	12,5	334,155334	262,010876	0,01927656	324,25	6
Drought	MCam009	1436PK1398	4	31297,964	2739,94538	3428,25	738,5	48997,875	14,25	358,012642	286,132891	0,01756219	351,25	6
Drought	MCam010	1436PK1039	4	37720,1536	3054,64049	3822	743,5	59052	22,25	353,040472	282,159005	0,01762941	352,5	7
Drought	MCam010	1436PK1220	4	30752,303	2591,7533	3243,75	515	48168,375	14,75	337,043298	269,29805	0,02476127	336,25	6
Drought	MCam014	1436PK1002	4	26123,816	2429,95609	3227,25	568,5	46047,875	13,25	335,42769	252,72732	0,02369377	334,25	6
Drought	MCam014	1436PK1201	4	29296,8448	2638,86171	3392,75	649,75	48495,875	17	366,704028	285,170572	0,02027304	365,25	
Drought	MCam017	1436PK1145	4	30702,8868	2681,23009	3526,75	630	53216,625	17	371,443027	282,393793	0,02161094	369,5	5
Drought	MCam017	1436PK1360	4	30056,2901	2554,78364	3251,25	596,25	48639	15,25	352,660224	276,909488	0,02219455	351,25	6
Drought	MCam021	1436PK1182	4	33332,2416	2917,29958	3719,75	752,75	54392,625	16,25	388,754895	304,843271	0,01772222	387,75	
Drought	MCam021	1436PK1344	4	24797,926	2214,7752	2896	570,5	42426,625	20	358,804436	274,199267	0,02357285	356,5	5

Zadanie

- $\mathbf{X}_{396\times366} \sim N_{n,m}(\mu\mathbf{1}_n', \mathbf{I}_n, \mathbf{\Omega})$
- $\mathbf{S}_{366\times366} = \frac{1}{n}\mathbf{X}'\mathbf{Q}_{1_n}\mathbf{X}$
- $\det(\mathbf{S}) = 0$
- Macierz korelacji $\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$
- $\mathbf{D}^{-1} = diag(\frac{1}{\sqrt{S_{11}}}, \frac{1}{\sqrt{S_{22}}}, ..., \frac{1}{\sqrt{S_{mm}}})$



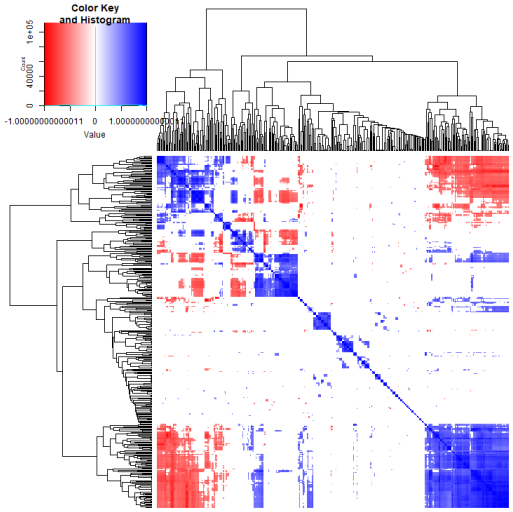
Zadanie



- bloki diagonalne
- wysoka/ bardzo wysoka korelacja
- grupowanie hierarchiczne
  - funkcje odległości
  - kryteria połączenia
- estymacja
- bloki pozadiagonalne

Zadanie

- if $|\mathbf{R}_{ij}| < 0.5$ then $\mathbf{R}_{ij} = 0$



Załóżmy:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{33} \end{pmatrix}$$

gdzie

$$\begin{aligned} \mathbf{\Sigma}_{11} &= \sigma_1^2(\rho_1 \mathbf{J} + (1-\rho_1))\mathbf{I} \\ \mathbf{\Sigma}_{22} &= \sigma_2^2 \mathbf{I} \\ \mathbf{\Sigma}_{33} &= \sigma_3^2(\rho_3 \mathbf{J} + (1-\rho_3))\mathbf{I} \end{aligned}$$

Wyniki

$$\mathbf{S} \longrightarrow \mathbf{PSP}' = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{33} \end{pmatrix}$$

$$\min \|\mathbf{S}_{ii} - \boldsymbol{\Sigma}_{ii}\|_F, \quad i = 1, 2, 3$$

$\widetilde{\boldsymbol{\Sigma}}_{11}$

$\sigma^2 = 31.4724$

$\rho = 0.5746$

$\widetilde{\boldsymbol{\Sigma}}_{22}$

$\sigma^2 = 29.14777$

$\widetilde{\boldsymbol{\Sigma}}_{33}$

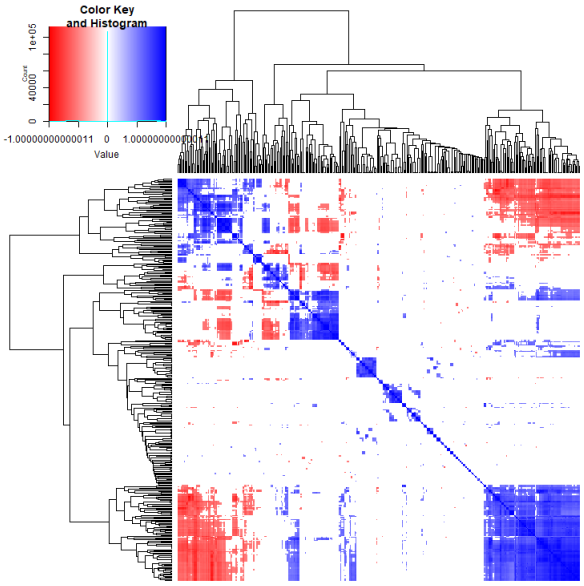
$\sigma^2 = 26.4708$

$\rho = 0.5868$

$$\zeta_F = \min \|\mathbf{PSP}' - \widetilde{\boldsymbol{\Sigma}}_1\|_F = 12817.32$$

$$\xi_F = \zeta_F / ||\mathbf{PSP}'||_F = 0.8325$$

Zadanie



Załóżmy:

$$\widetilde{\mathbf{\Sigma}}_2 = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{13} \\ \mathbf{\Sigma}'_{12} & \mathbf{\Sigma}_{22} & \mathbf{\Sigma}_{23} \\ \mathbf{\Sigma}'_{13} & \mathbf{\Sigma}'_{23} & \mathbf{\Sigma}_{33} \end{pmatrix}$$

gdzie

$$\begin{aligned} \mathbf{\Sigma}_{11} &= \sigma_1^2(\rho_1 \mathbf{J} + (1-\rho_1))\mathbf{I} \\ \mathbf{\Sigma}_{22} &= \sigma_2^2 \mathbf{I} \\ \mathbf{\Sigma}_{33} &= \sigma_3^2(\rho_3 \mathbf{J} + (1-\rho_3))\mathbf{I} \end{aligned}$$

$$\begin{aligned} \mathbf{\Sigma}_{ij} &= \delta_k \mathbf{J} \\ & i = 1,2 < j = 2,3, \quad k = 1,2,3 \end{aligned}$$

Wyniki

$$\mathbf{S} \longrightarrow \mathbf{PSP}' = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}'_{12} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}'_{13} & \mathbf{S}'_{23} & \mathbf{S}_{33} \end{pmatrix}$$

$$\min \|\mathbf{S}_{ij} - \mathbf{\Sigma}_{ij}\|_F, \quad i = 1, 2 < j = 2, 3$$

$\widetilde{\mathbf{\Sigma}}_{11}$

$\sigma_1^2 = 31.4724$

$\rho_1 = 0.5746$

$\widetilde{\mathbf{\Sigma}}_{22}$

$\sigma_2^2 = 29.14777$

$\widetilde{\mathbf{\Sigma}}_{33}$

$\sigma_3^2 = 26.4708$

$\rho_3 = 0.5868$

$\widetilde{\mathbf{\Sigma}}_{ij}$

$\delta_1 = 14.71865$

$\delta_2 = 18.57934$

$\delta_3 = 16.74399$

$$\zeta_F = \min \|\mathbf{PSP}' - \widetilde{\mathbf{\Sigma}}_2\|_F = 7046.1770$$

$$\xi_F = \zeta_F / ||\mathbf{PSP}'||_F = 0.4577$$

Zadanie

$$\widetilde{\mathbf{\Sigma}}_2 = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{13} \\ \mathbf{\Sigma}'_{12} & \mathbf{\Sigma}_{22} & \mathbf{\Sigma}_{23} \\ \mathbf{\Sigma}'_{13} & \mathbf{\Sigma}'_{23} & \mathbf{\Sigma}_{33} \end{pmatrix}$$

$\widetilde{\mathbf{\Sigma}}_2 \notin \mathbb{R}^{>}$

$$\widetilde{\mathbf{\Sigma}}_{ij}$$

$$\begin{aligned} \delta_{1p} &= a_1 \cdot \delta_1 \\ \delta_{2p} &= a_2 \cdot \delta_2 \\ \delta_{3p} &= a_3 \cdot \delta_3 \end{aligned}$$

$$a_i \in (0,1], \qquad i = 1,2,3$$

$\widetilde{\mathbf{\Sigma}}_2 \in \mathbb{R}^{>}$

Zadanie

Na zaliczenie:

- prezentacja
- pliki R

Każda grupa:

- prezentacja wyników
- odpowiedzi na pytania