主成分分析によるデータの可視化

使用するデータの説明

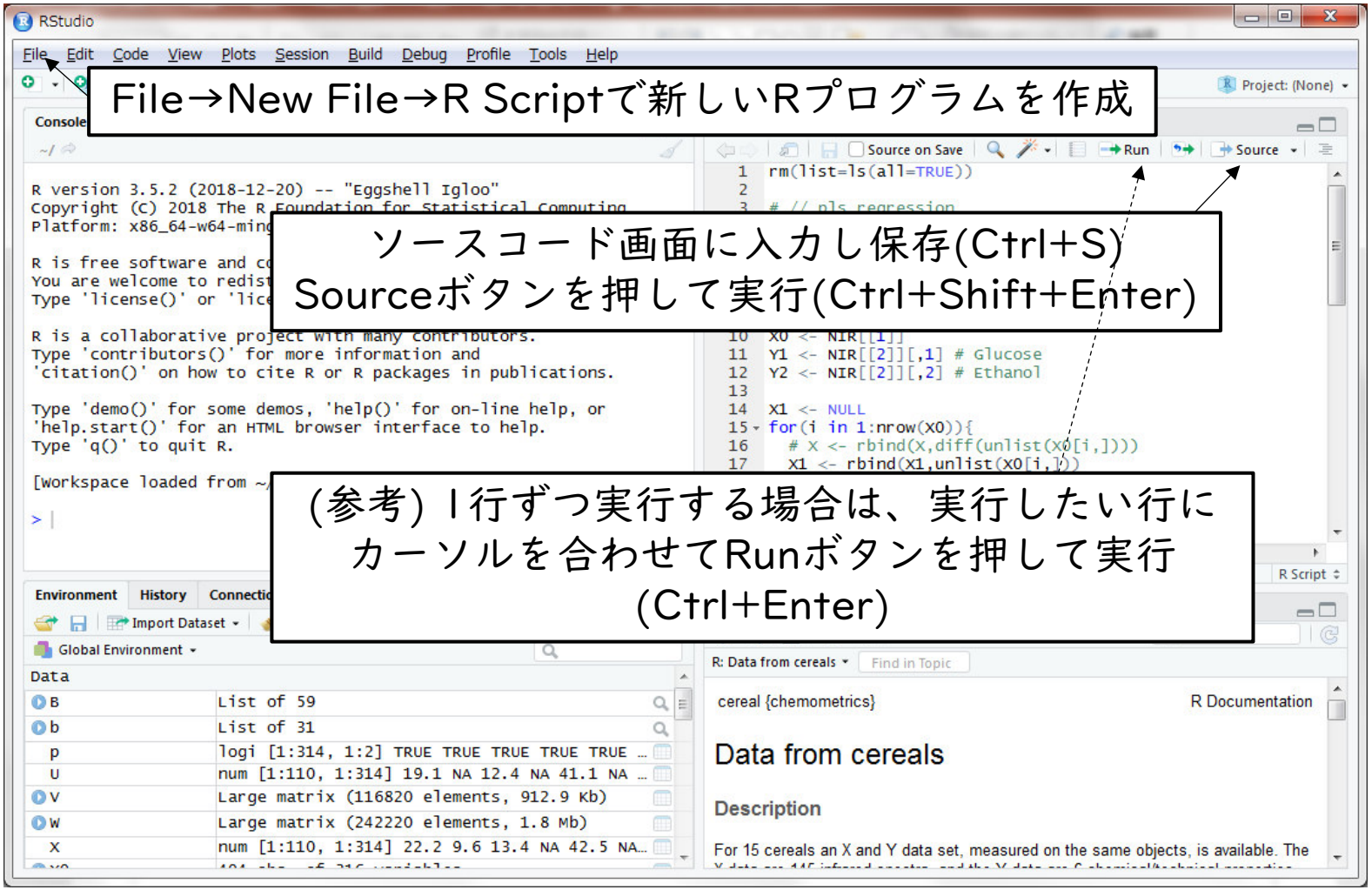
- MetaboAnalystのデモデータ

  Alan Saghatelian et al., "Assignment of endogenous substrates to enzymes by global metabolite profiling",43(45):14332-9 (2004).

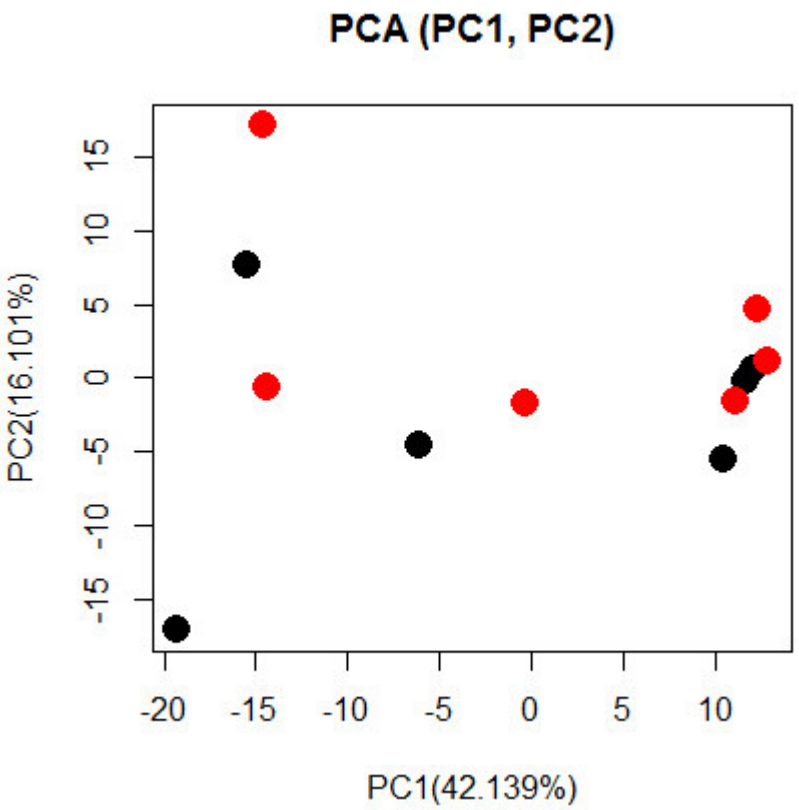
- データの説明
  - 野生型マウス、FAAH欠損マウスそれぞれN=6の脳(脊椎)サンプルのメタボロームデータ
    - Fatty acid amide hydrolase (FAAH)：脂肪酸アミドのアミド結合を切って脂肪酸とアミンに分解する酵素
    - 例：アナンダミド(脂肪酸アミド、脳内麻薬物質の一つ)

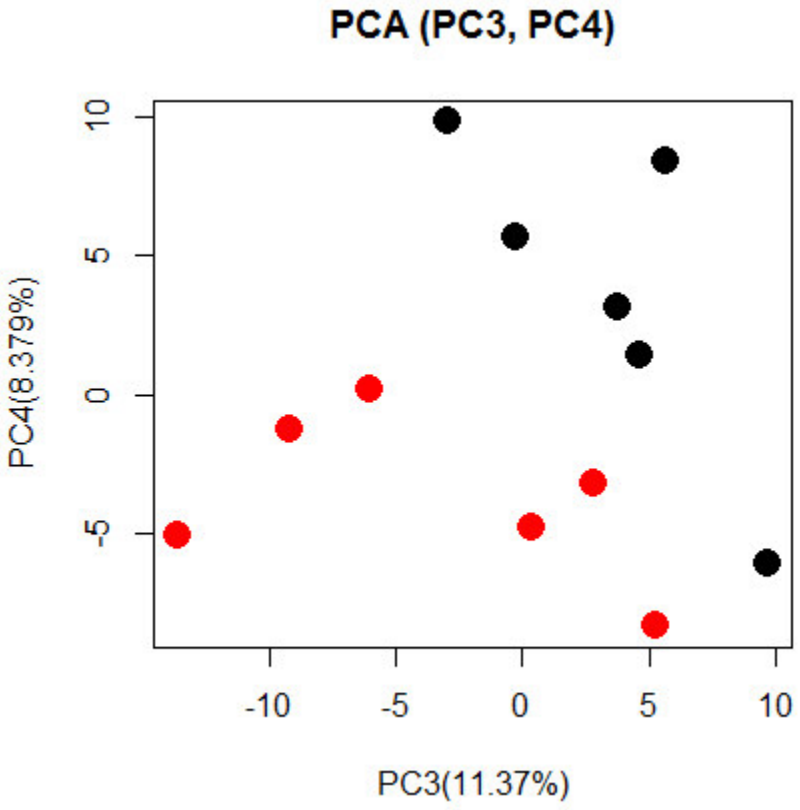
    →(阻害)→アラキドン酸(脂肪酸)＋エタノールアミン(アミン)
    - FAAH阻害薬は創薬ターゲットになっているが、様々な副作用が報告されており、市場に出た薬剤は無い
      - 参考 https://aasj.jp/news/watch/4746

Rstudioの画面


File→New File→R Scriptで新しいRプログラムを作成
ソースコード画面に入力し保存(Ctrl+S)
Sourceボタンを押して実行(Ctrl+Shift+Enter)
(参考) 1行ずつ実行する場合は、実行したい行に
カーソルを合わせてRunボタンを押して実行
(Ctrl+Enter)
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Project: (None)
Source on Save
Run
Source
rm(list=ls(all=TRUE))
X0 <- NIR[[1]]
Y1 <- NIR[[2]][,1] # Glucose
Y2 <- NIR[[2]][,2] # Ethanol
X1 <- NULL
for(i in 1:nrow(X0)){
Environment History
Import Dataset
Global Environment
Data
B List of 59
b List of 31
p logi [1:314, 1:2] TRUE TRUE TRUE TRUE TRUE …
U num [1:110, 1:314] 19.1 NA 12.4 NA 41.1 NA …
V Large matrix (116820 elements, 912.9 Kb)
W Large matrix (242220 elements, 1.8 Mb)
X num [1:110, 1:314] 22.2 9.6 13.4 NA 42.5 NA…
R: Data from cereals
Find in Topic
cereal {chemometrics}
R Documentation
Data from cereals
Description
For 15 cereals an X and Y data set, measured on the same objects, is available. The

最終的なゴール


PCA (PC1, PC2)
PC2(16.101%)
PC1(42.139%)
15
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10


PCA (PC3, PC4)
PC4(8.379%)
PC3(11.37%)
10
5
0
-5
-10
-5
0
5
10

データの準備

- csvファイルの読み込み(実行済み)

```
file <- "C:/R/mouse_data_original.csv"
X0 <- read.csv(file, skip=1) # 1行目をスキップ
```

- データの準備

```
X <- X0[,-1] # 1列目の群情報の文字列を除く
class <- c(1,1,1,1,1,1,2,2,2,2,2,2) # 群情報
X <- t(X)　# 行列の転置 (各行が変数→各行がサンプル)
X <- scale(X) # スケーリング
```

mouse_data_original.csvの中身

サンプル

変数

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		ko15	ko16	ko18	ko19	ko21	ko22	wt15	wt16	wt18	wt19	wt21	wt22
2	Label	KO	KO	KO	KO	KO	KO	WT	WT	WT	WT	WT	WT
3	200.1/2926	147887.5	451600.7	65290.38	56540.93	85146.33	162012.4	175177.1	82619.48	51951.61	69198.22	153273.5	98144.28
4	205/2791	1778569	1567038	1482796	1039130	1223132	1072038	1950287	1466781	1572679	1275313	1356014	1231442
5	206/2791	237993.6	269714	201393.4	150107.3	176989.7	156797	276541.8	222366.2	211717.7	186850.9	188285.9	172348.8
6	207.1/2719	380873	460629.7	351750.1	219288	286848.6	235022.6	417169.6	324892.5	277990.7	220972.4	252874	236728.2
7	219.1/2524	235544.9	173623.4	82364.59	79480.4	185792.4	174458.8	244584.5	161184.1	72029.38	75096.99	238194.4	173830
8	231/2516	117649.8	48960.63	222609.1	286232.2	435094.5	62168.71	465898	61234.44	96841.46	240261.2	201316.2	179437.7
9	233/3023	399145.3	356951.3	410550.7	198416.5	363381.7	317805.8	397107.8	271252.1	334459.9	181901.3	456900.5	294270.6
10	234/3024	76880.87	99526.27	97493.76	53461.71	88227.79	81072.23	65215.64	55952.44	73781.01	45211.66	83693.39	57516.2
11	235.1/2695	171995.2	128945.2	155442.5	115286.3	37769.45	7663.88	199981.5	30028.6	156968.3	52596.48	14641.59	27713.29
12	236.1/2524	252282	206031.9	71763.79	73602.47	186661	198804.3	253791.1	187225.7	79389.63	90012.64	256263.3	206487
13	240.2/3682	112440.6	90466.98	193768.7	170641.5	81223.74	146563	43974.38	270872.9	176425.2	187063	75235.94	86068.18
14	241.1/3679	1465989	1318747	1215369	632037.4	579968.2	561964.9	1468103	1594705	1006929	805533.2	731777.6	522916.1
15	242.1/3679	280767.6	248792.4	224467.7	109019.4	103841.9	101092.9	280260.2	299453.2	188328.2	139748.3	139968.5	95347.85
16	244.1/2832	612169.9	256316.5	90539.86	33269.29	54610.78	115505.4	627835	56883.71	168524.5	27260.37	110170.7	1352930
17	246.1/2517	27932.12	51557.09	38018.76	41747.84	56145.39	47901.55	105508.8	70508.82	46090.83	45630.55	77205.36	48624.55
18	249.1/3668	1435001	1228148	1193347	641881.8	536808.3	574005.7	1297986	1566269	1076654	747969.4	688496.3	485427.6
19	250.1/3668	347794.8	238893.4	248245.6	126623.8	107844.1	118627	281745	336058.8	215534.9	149872.8	139317.3	99814.5
20	254.1/3231	78911.16	564680.8	89307	127186.8	21055.46	24185.93	54645.64	124200.9	108479	30096	66315.96	27554.78
21	255.2/3679	1420043	1187752	1264222	673816.2	581177.9	651560.3	1211727	1550071	1024087	875296.3	759075.9	640885.7
22	256.2/3679	307708.4	263317.7	275104.9	138168.4	121284.4	128540.4	258452.2	338390.9	213618.1	180384.5	164675.7	91822.7

主成分分析、スコアプロット

- 主成分分析の計算

```
pca <- prcomp(X)
```

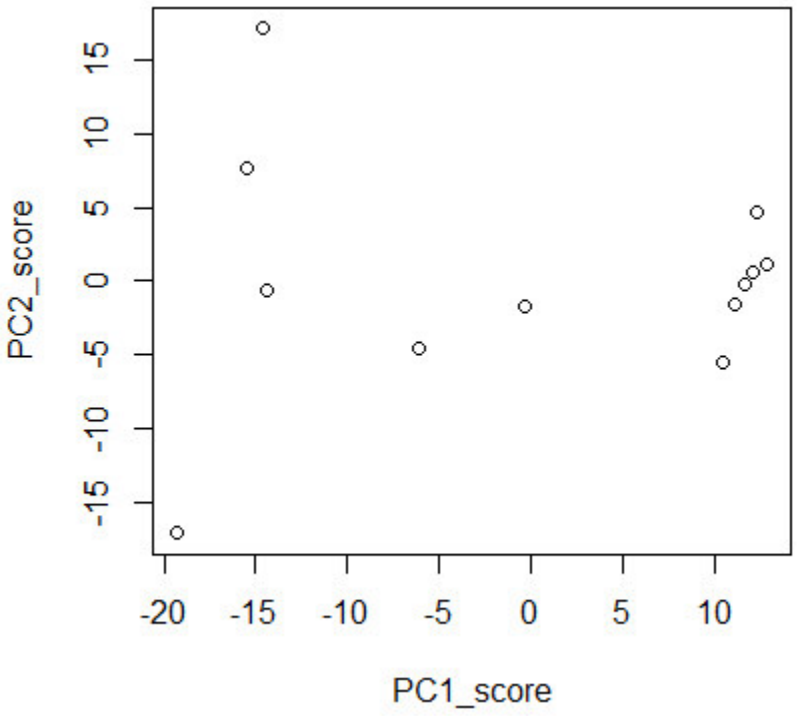
- 主成分スコア

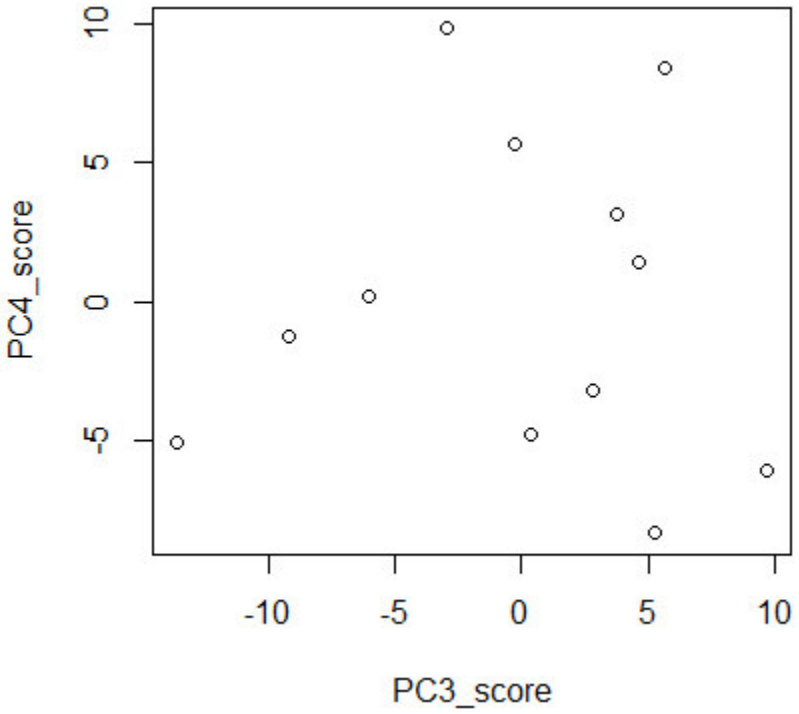
```
PC_score <- pca$x
PC1_score <- PC_score[,1] # 第1主成分
PC2_score <- PC_score[,2] # 第2主成分
```

- 主成分スコアプロット(PC1,PC2)

```
plot(PC1_score,PC2_score)
```

主成分スコアプロットの結果(1)





PC3,PC4のスコアプロットも同様に描画

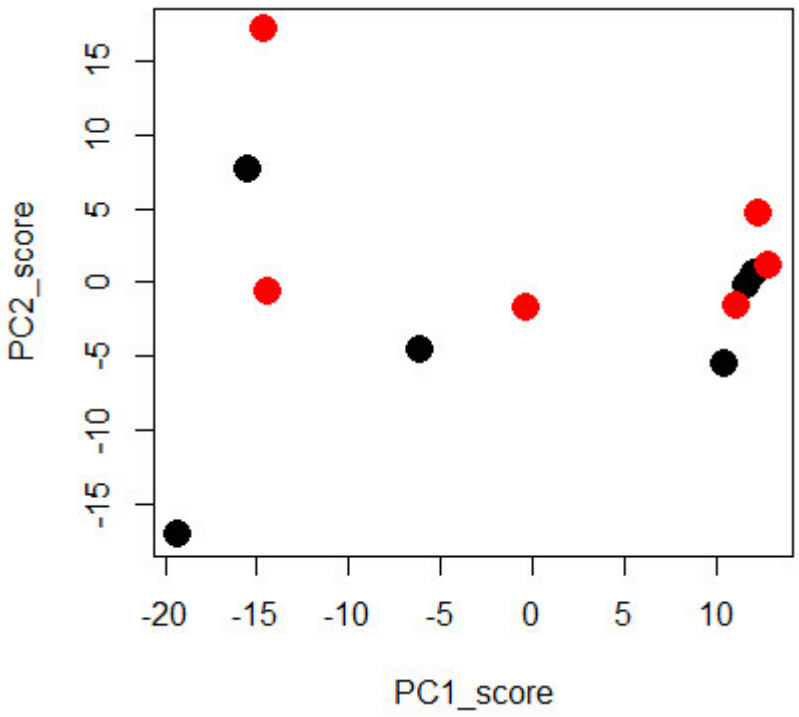
群情報の色付け

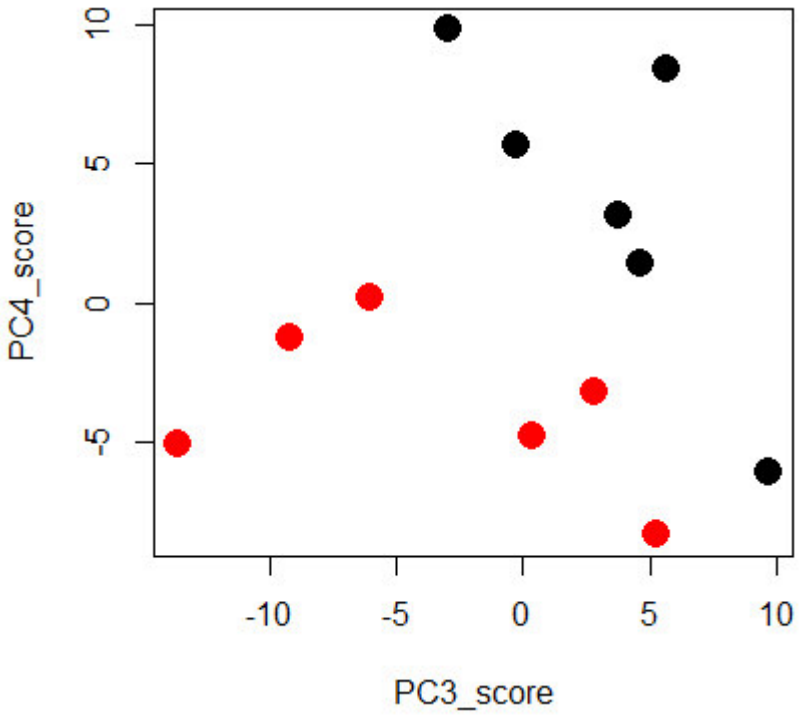
- 主成分スコアプロット(PC1,PC2)

```
plot(PC1_score, PC2_score, col=class) # 色付け
plot(PC1_score, PC2_score, col=class, pch=16) # 色塗り
plot(PC1_score, PC2_score, col=class, pch=16, cex=2)
                                        # サイズ変更
```

- 演習
  - 第3主成分と第4主成分についても同様に計算

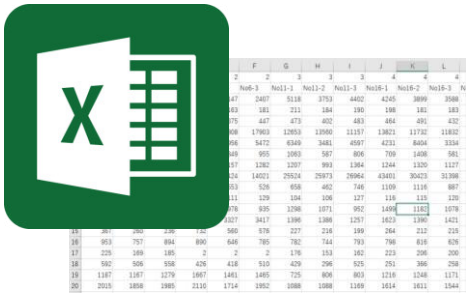
主成分スコアプロットの結果(2)





(Rのバージョンによって色は異なります)

主成分分析におけるスケーリングの影響



変数

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

各変数毎に平均値を計算し、
その値を引く

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

平均値が0

各変数毎に標準偏差を計算し、
その値で割る(スケーリング)

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

平均値が0、分散が1

主成分分析は変数の分散の値に影響を受ける

同じ傾向を示す変数

分散：40000

変数1の分散が大きく、
変数2～4は同じ傾向を示す
データを考える

$$\begin{bmatrix} 100 & 3 & 6 & 12 \\ 500 & 6 & 12 & 24 \\ 300 & 12 & 24 & 48 \end{bmatrix}$$

平均値を引く

$$\begin{bmatrix} 100 & 3 & 6 & 12 \\ 500 & 6 & 12 & 24 \\ 300 & 12 & 24 & 48 \end{bmatrix} \longrightarrow \begin{bmatrix} -200 & -4 & -8 & -16 \\ 200 & -1 & -2 & -4 \\ 0 & 5 & 10 & 20 \end{bmatrix}$$

第1主成分スコア

$$\longrightarrow \begin{bmatrix} -200.5 \\ 199.7 \\ 0.8 \end{bmatrix}$$

変数1の影響大

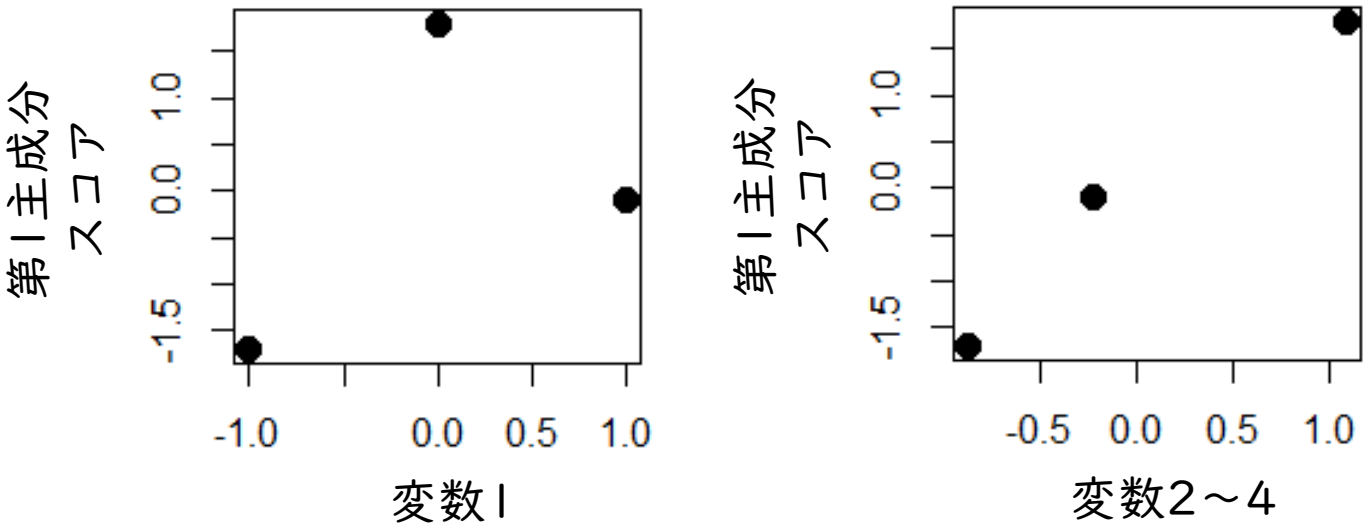
分散：1(全ての変数)

スケーリング

$$\begin{bmatrix} -1 & -0.87 & -0.87 & -0.87 \\ 1 & -0.22 & -0.22 & -0.22 \\ 0 & 1.09 & 1.09 & 1.09 \end{bmatrix} \longrightarrow \begin{bmatrix} -1.7 \\ -0.1 \\ 1.8 \end{bmatrix}$$

主成分分析は変数の分散の値に影響を受ける

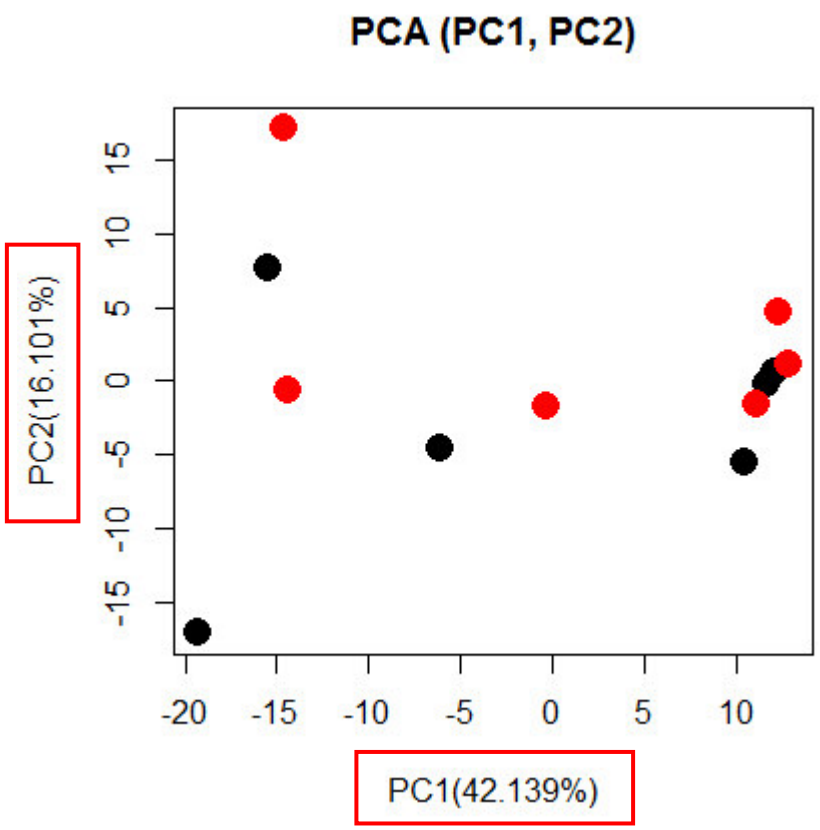
第1主成分スコアと各変数の関係(スケーリングした場合)

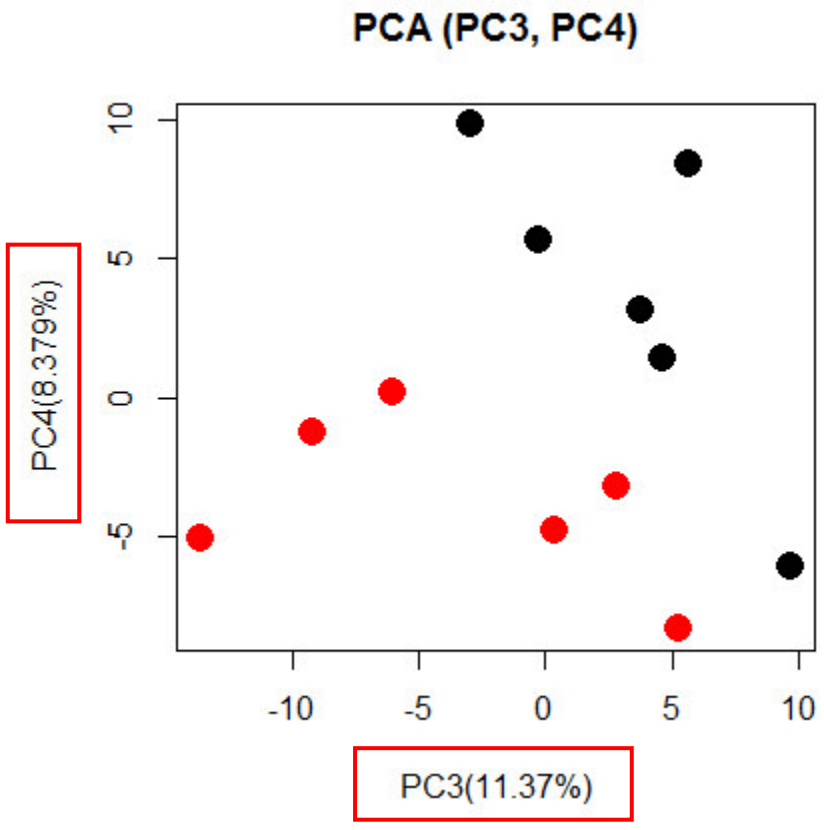


スケーリングと主成分分析の関係

- 特定の変数(変数1)の分散が大きいときは、
  その影響が主成分スコアに現れる
- 各変数の分散が同じ場合(スケーリングした場合)は、
  傾向が似た変数(変数2～4)の影響を受ける

最終的なゴール


PCA (PC1, PC2)
PC2(16.101%)
PC1(42.139%)
15
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10


PCA (PC3, PC4)
PC4(8.379%)
PC3(11.37%)
10
5
0
-5
-10
-5
0
5
10

寄与率

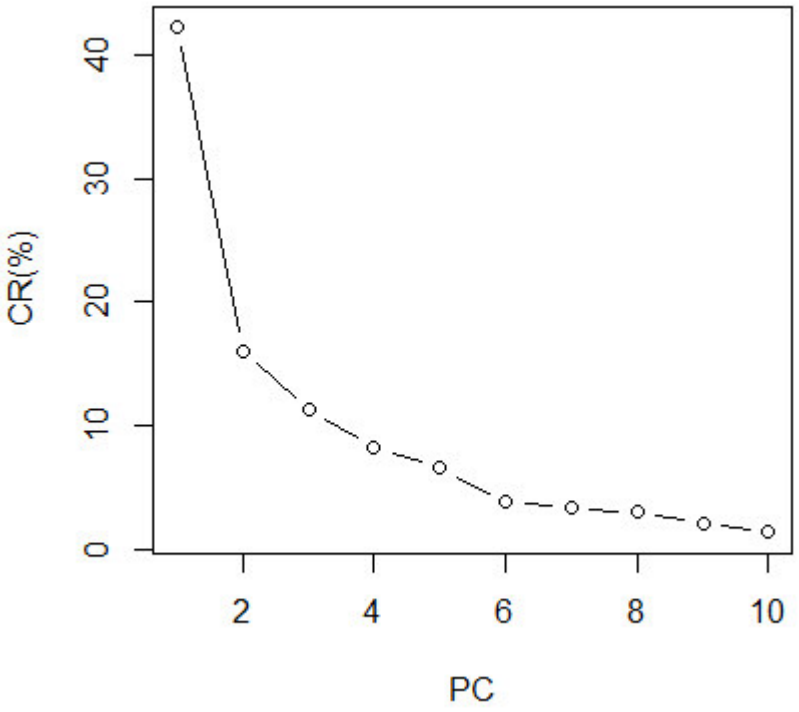
寄与率の計算

- 寄与率

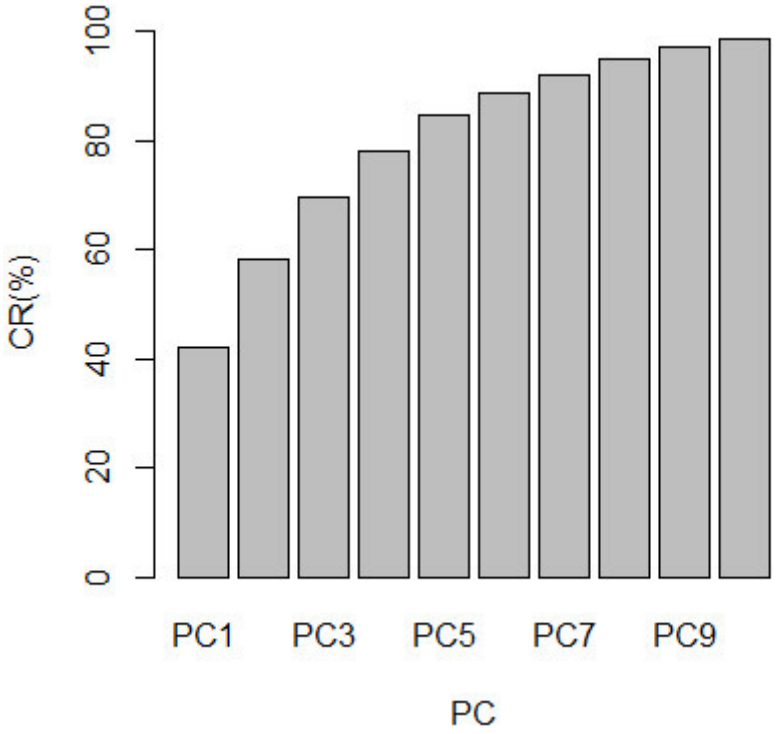
```
cr_var <- summary(pca)$importance[2,] # 寄与率
cumcr <- summary(pca)$importance[3,] # 累積寄与率
PC1_cr <- 100*cr_var[1] # PC1
PC2_cr <- 100*cr_var[2] # PC2
```

寄与率のグラフ表示

寄与率



累積寄与率



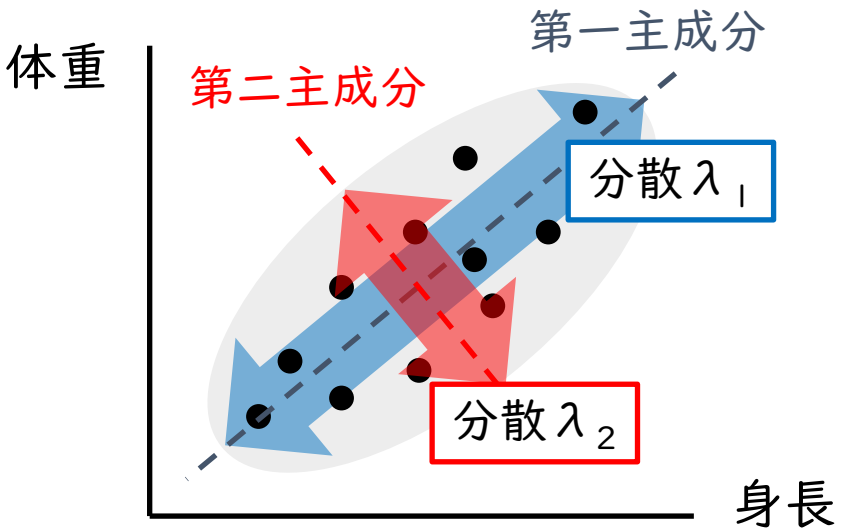
(左図)

```
plot(100*cr_var[1:10], type="b", xlab="PC", ylab="CR(%)")
```

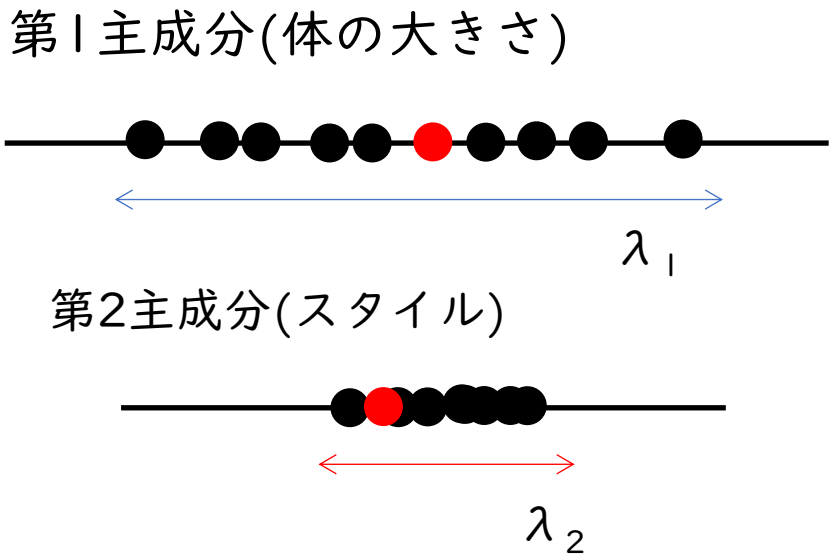
(右図)

```
barplot(100*cumcr[1:10], xlab="PC", ylab="CR(%)", ylim=c(0,100))
```

寄与率とは



第一主成分の寄与率(%)：
$100\times\lambda_1/(\lambda_1+\lambda_2)$



$$Var(\mathbf{t})=\frac{1}{n-1}\frac{\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w}}{\|\mathbf{w}\|^2}=\frac{\mathbf{w}'\lambda\mathbf{w}}{\|\mathbf{w}\|^2}=\lambda$$

$$\uparrow$$

$$\frac{1}{n-1}\mathbf{X}'\mathbf{X}\mathbf{w}=\lambda\mathbf{w}$$

最大固有値に対応する固有ベクトル
||
主成分スコアの分散が最大の
固有ベクトル(PC1)

第1主成分の寄与率の値が大きければ、
第1主成分でデータのばらつきの大半を説明できている

寄与率は変数の分散の値に影響を受ける

平均値を引く

$$\begin{bmatrix} 100 & 3 & 6 & 12 \\ 500 & 6 & 12 & 24 \\ 300 & 12 & 24 & 48 \end{bmatrix} \longrightarrow \begin{bmatrix} -200 & -4 & -8 & -16 \\ 200 & -1 & -2 & -4 \\ 0 & 5 & 10 & 20 \end{bmatrix}$$

第1主成分スコア

$$\longrightarrow \begin{bmatrix} -200.5 \\ 199.7 \\ 0.8 \end{bmatrix}$$

分散：1(全ての変数)

寄与率: 99.03%

スケーリング

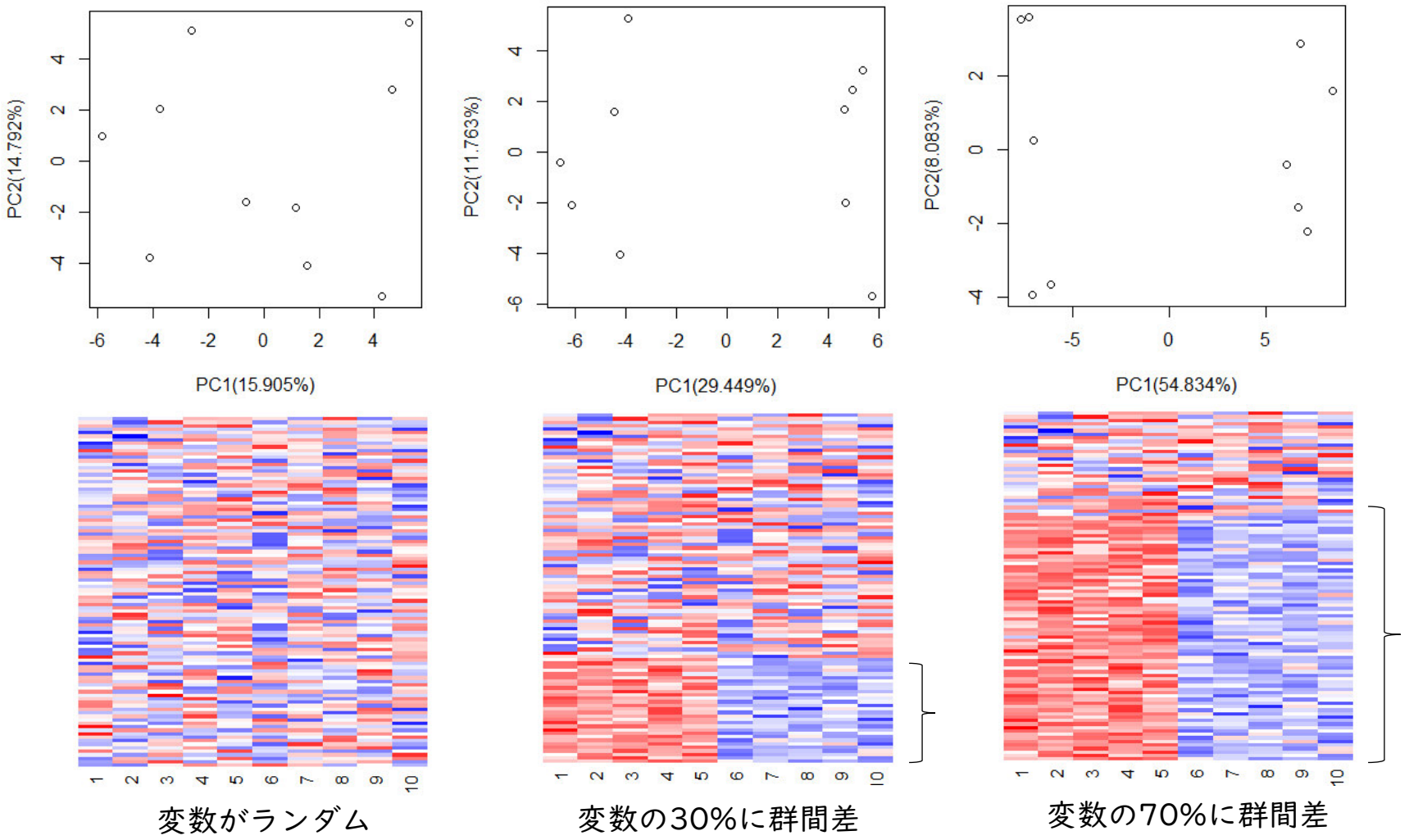
$$\begin{bmatrix} -1 & -0.87 & -0.87 & -0.87 \\ 1 & -0.22 & -0.22 & -0.22 \\ 0 & 1.09 & 1.09 & 1.09 \end{bmatrix} \longrightarrow \begin{bmatrix} -1.7 \\ -0.1 \\ 1.8 \end{bmatrix}$$

寄与率: 78.84%

寄与率は元の変数の分散の影響を受ける

特定の変数の分散が大きいときは、寄与率は大きくなる
スケーリングした時は、特定の元の変数の分散の影響ではなく、
傾向が似た変数が多ければ寄与率は大きくなる

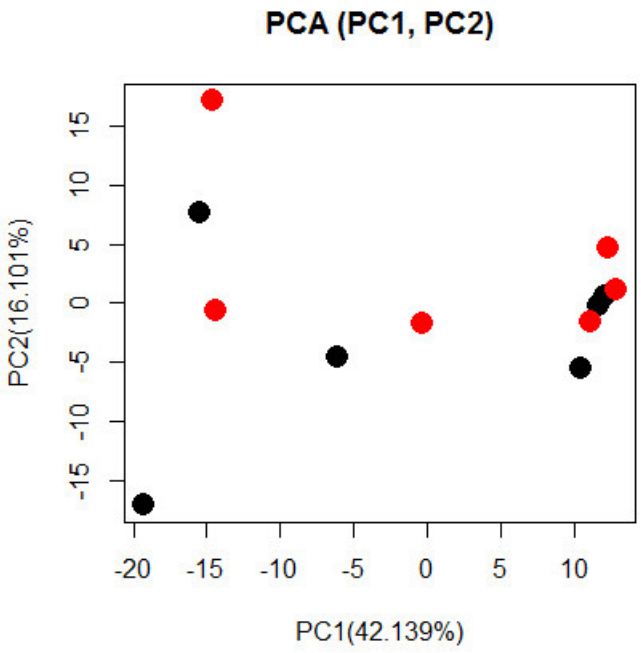
スケーリングした時の寄与率のイメージ

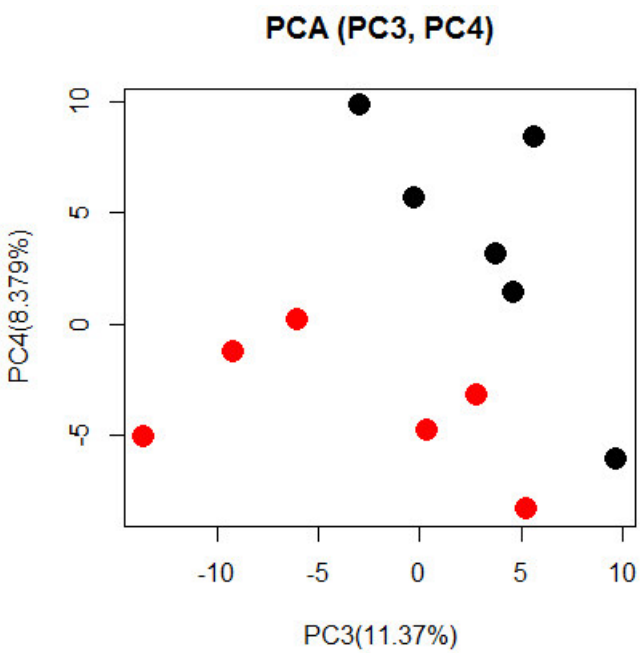


結果のグラフ表示

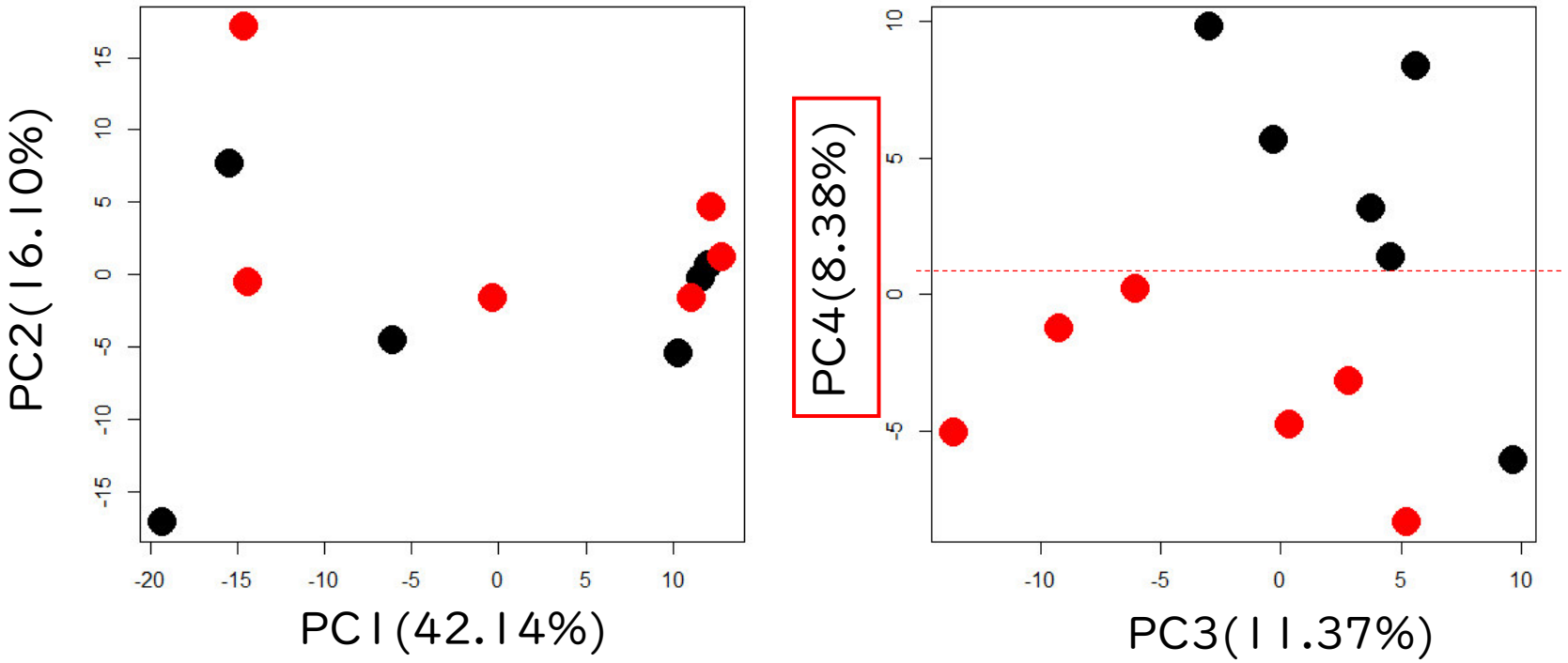
- スコアプロット、寄与率、タイトルの表示

```
plot(PC1_score,PC2_score,
  xlab="PC1(42.139%)", ylab="PC2(16.101%)",
  main="PCA (PC1, PC2)", col=class, pch=16, cex=2)
```



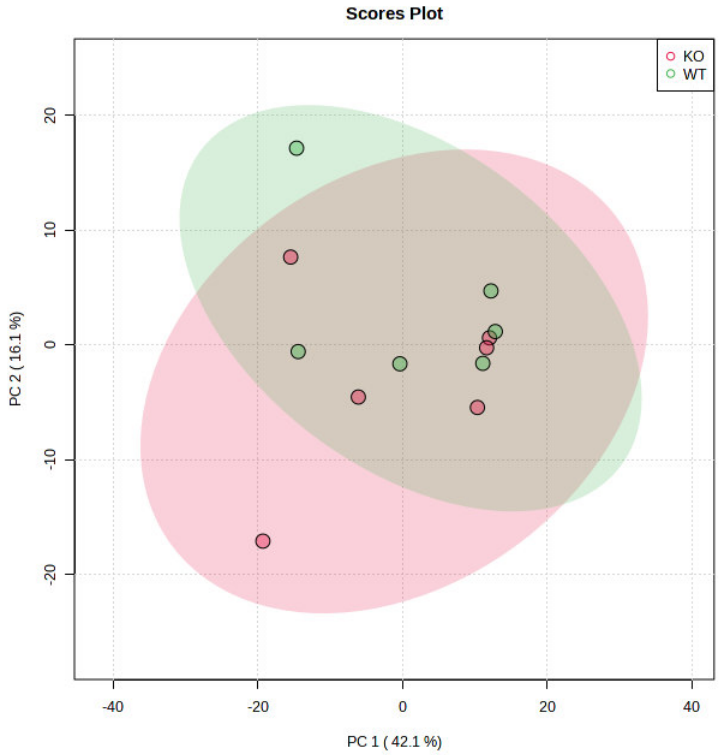


結果を改めて確認

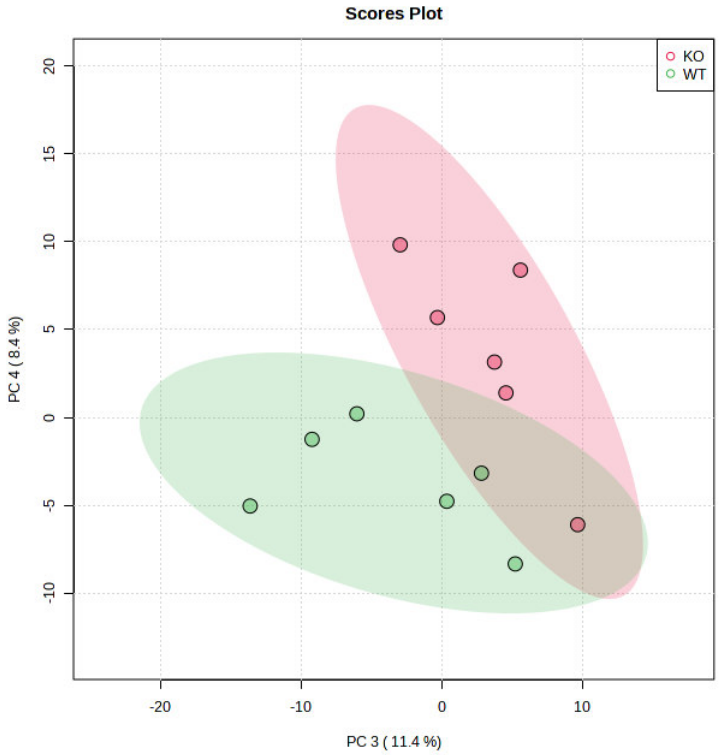


PC4で群間差がある程度確認できているので、
PCAの結果を利用する場合はPC4に着目して解析を進める

MetaboAnalystの結果



pca_score2d_0_dpi72



pca_score2d_1_dpi72