Partial least squaresとは

ケモメトリックスで最も良く用いられる Partial least squares (PLS)

- PLSとは
  - 日本語では、部分的最小二乗回帰（ぶぶんてきさいしょうじじょうかいき、略称: PLS回帰）と呼ばれる。偏最小二乗回帰または部分最小二乗回帰とも呼ばれる。

“https://ja.wikipedia.org/wiki/部分的最小二乗回帰”より改変

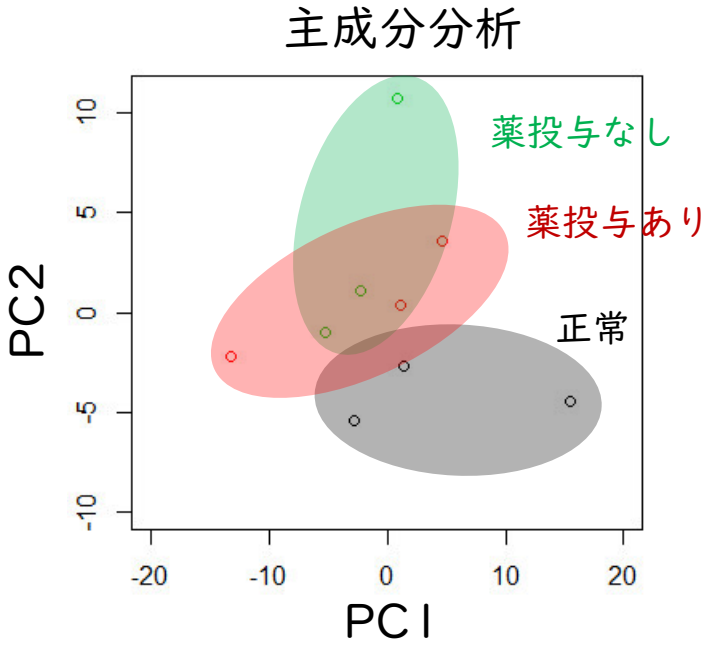
- PLSにはいくつかの種類がある
  - 回帰分析で用いられるPLS回帰
  - 判別分析で用いられるPLS-DA(Discriminant Analysis)

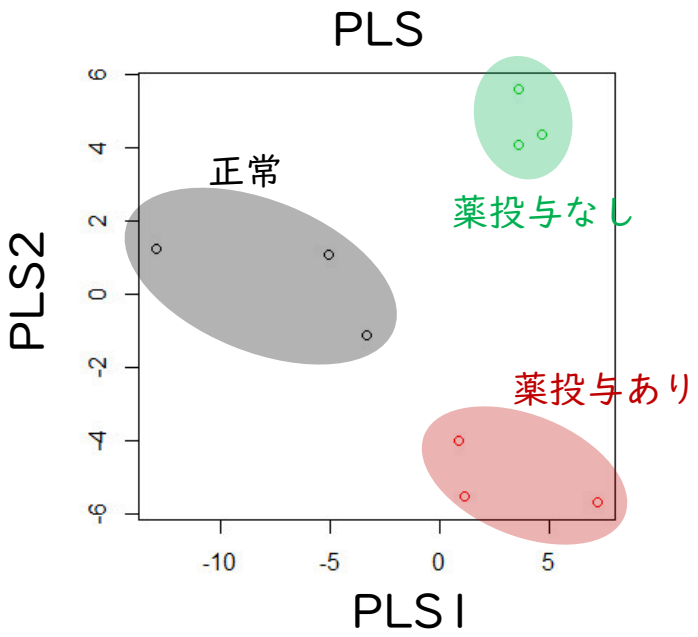
PLSの代表的な論文

- Herman Woldによって社会科学分野でPLSが提案
  - Wold, H. (1975) “Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach.” Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett. Academic Press, London, 117-142. [引用数1041(2023.10時点)]
- Svante Woldが化学分野、特に近赤外スペクトル(NIRS)データに適用し、ケモメトリックスが広く知られるようになった
  - Wold, S; Sjostrom, M.; Eriksson, L. (2001). “PLS-regression: a basic tool of chemometrics”. Chemometrics and Intelligent Laboratory Systems. 58 (2): 109-130. [引用数10038 (2023.10時点)]
- 生物分野、特にメタボロミクスで広く用いられる
  - Max Bylesjö, Mattias Rantalainen, Olivier Cloarec, Jeremy K Nicholson, Elaine Holmes, Johan Trygg (2006). “OPLS discriminant analysis: combining the strengths of PLS - DA and SIMCA classification”, Journal of Chemometrics. 20(8-10): 341-351. [引用数1419(2023.10時点)]

主成分分析とPLSの解析例

高脂血症ウサギの肝臓のメタボローム解析
3群比較：Wild type、高脂血症ウサギ、薬剤投与後の高脂血症ウサギ

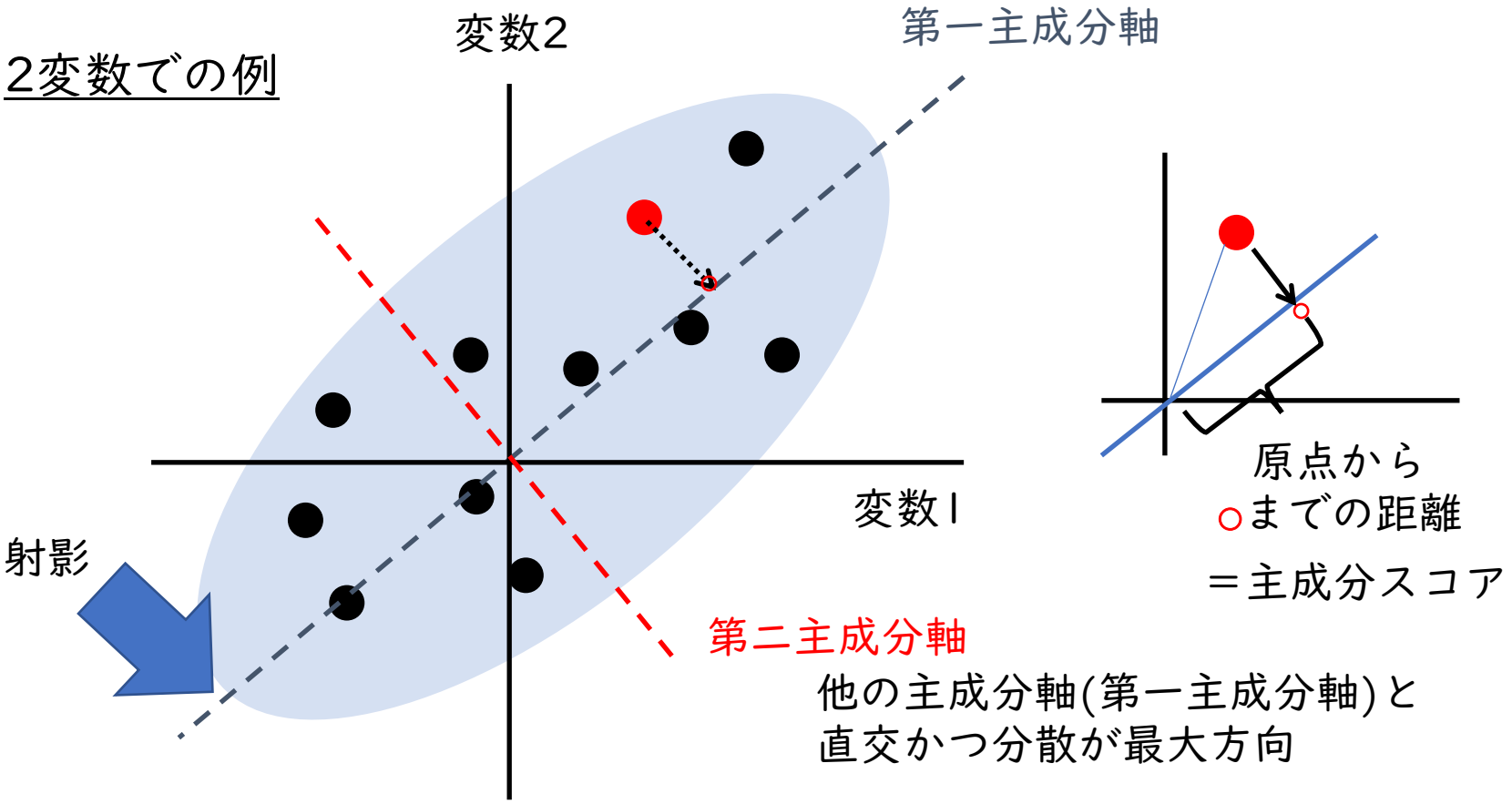




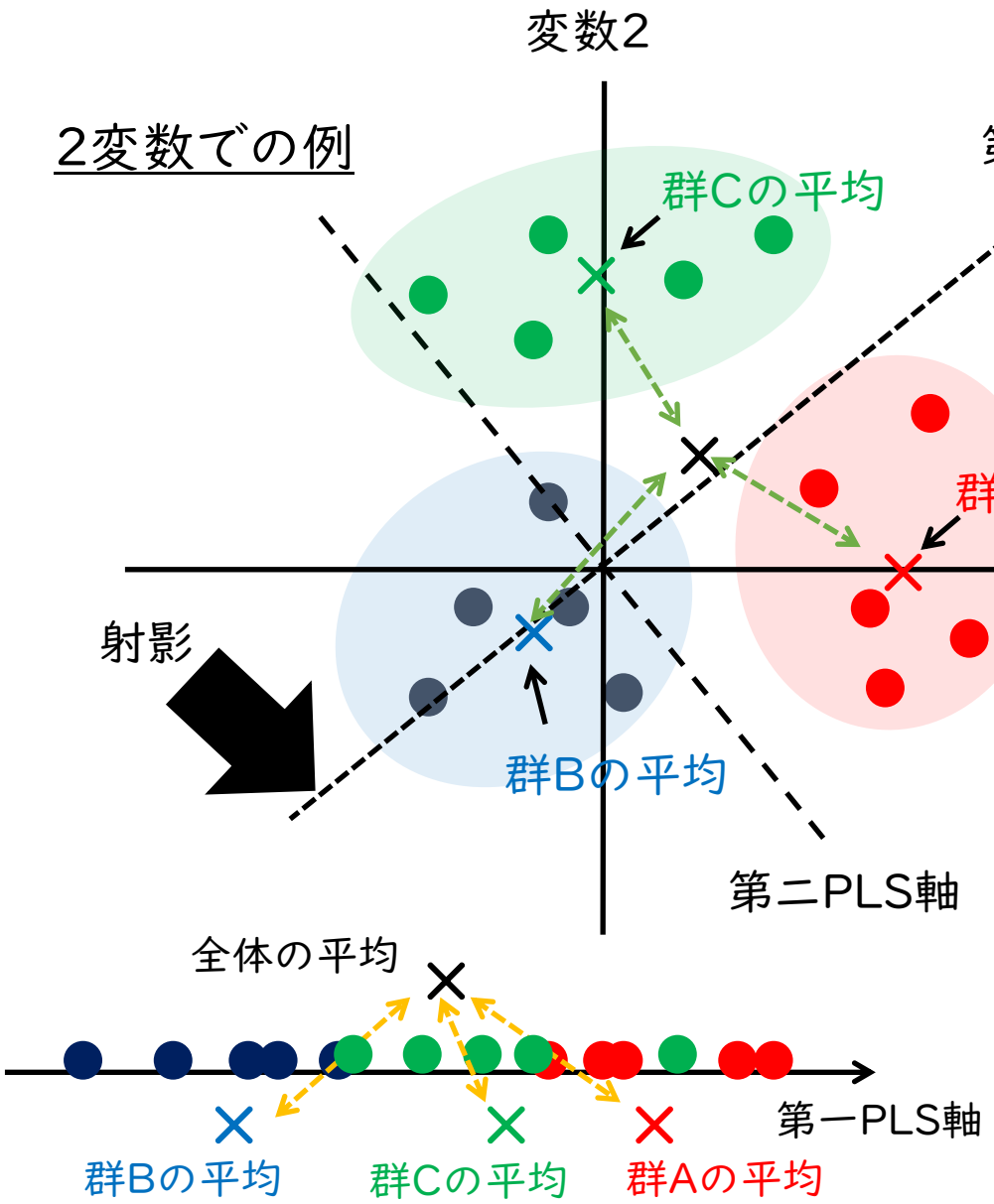
主成分分析の結果、主成分スコアで群間の差が表れなかったとき、
PLSが用いられることが多い

Ooga T, Sato H, Nagashima A, Sasaki K, Tomita M, Soga T, Ohashi Y., “Metabolomic anatomy of an animal model revealing homeostatic imbalances in dyslipidaemia.”, Mol Biosyst. 2011 Apr;7(4):1217-23.

主成分分析は、群情報を利用しない方法



Partial Least Squares (PLS)は、群情報を利用した方法



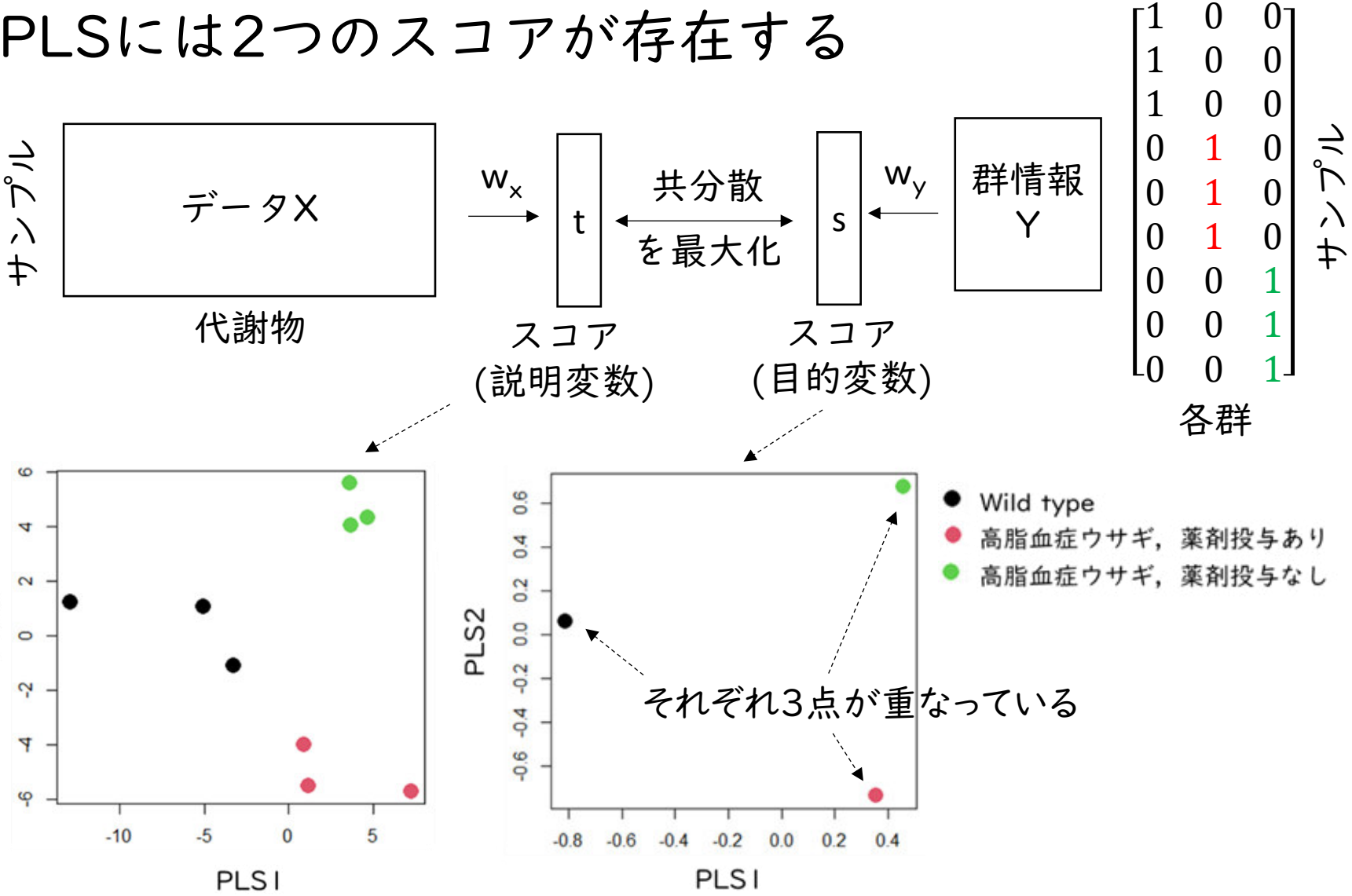
全体の平均(×)と、各群の平均(×××)の間の距離が最大の方向

他のPLS軸(第一PLS軸)と直交かつ群間の距離が最大方向
(”群数-1”個まで計算可能)

PLSスコアの群間の距離が最大になるようにPLS軸を決める

説明変数のスコアと目的変数のスコアの共分散最大化

PLSには2つのスコアが存在する



説明変数のスコア、目的変数のスコアいずれも同様の位置(左、右上、右下)に配置されており、傾向が一致していることが確認できる。

PLSは、スコアの共分散を最大化する方法(1)

$$\mathrm{cov}(\mathbf{t},\mathbf{s}) = \frac{1}{n-1}\mathbf{w}_x{}'\mathbf{X}'\mathbf{Y}\mathbf{w}_y \quad \text{を最大化}$$

$$\mathbf{t} = \mathbf{X}\mathbf{w}_\mathbf{x}$$

$$\mathbf{s} = \mathbf{Y}\mathbf{w}_\mathbf{y}$$

$$\mathbf{w}_\mathbf{x}'\mathbf{w}_\mathbf{x} = 1,\ \mathbf{w}_\mathbf{y}'\mathbf{w}_\mathbf{y} = 1$$

3群 N=3のとき

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

ラグランジュ乗数法より

$$J = \frac{1}{n-1}\mathbf{w}_x{}'\mathbf{X}'\mathbf{Y}\mathbf{w}_y + \lambda_x(1-\mathbf{w}_\mathbf{x}'\mathbf{w}_\mathbf{x}) + \lambda_y(1-\mathbf{w}_\mathbf{y}'\mathbf{w}_\mathbf{y})$$

$$\frac{\partial J}{\partial w_x} = \frac{1}{n-1}\mathbf{X}'\mathbf{Y}\mathbf{w}_y - 2\lambda_x\mathbf{w}_\mathbf{x} = 0 \qquad \frac{\partial J}{\partial w_y} = \frac{1}{n-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x - 2\lambda_y\mathbf{w}_\mathbf{y} = 0$$

$$\downarrow \qquad\qquad \downarrow$$

$$\mathbf{w}_\mathbf{x} = \frac{1}{2\lambda_x}\frac{1}{n-1}\mathbf{X}'\mathbf{Y}\mathbf{w}_y \qquad \mathbf{w}_\mathbf{y} = \frac{1}{2\lambda_y}\frac{1}{n-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x$$

PLSは、スコアの共分散を最大化する方法(2)

- 説明変数(データ)

$$\frac{1}{n-1}\mathbf{X'Y}\mathbf{w}_y - 2\lambda_x \mathbf{w}_\mathbf{x} = 0 \quad \text{に} \quad \mathbf{w}_\mathbf{y} = \frac{1}{2\lambda_y}\frac{1}{n-1}\mathbf{Y'X}\mathbf{w}_x \quad \text{を代入して}$$

$$\left(\frac{1}{n-1}\right)^2 \mathbf{X'YY'X}\mathbf{w}_\mathbf{x} = 4\lambda_x\lambda_y \mathbf{w}_\mathbf{x}$$

最大固有値に対応する固有ベクトルw_xが
第1PLSの固有ベクトルとなる

- 目的変数(群情報)

$$\frac{1}{n-1}\mathbf{Y'X}\mathbf{w}_x - 2\lambda_y \mathbf{w}_\mathbf{y} = 0 \quad \text{に} \quad \mathbf{w}_\mathbf{x} = \frac{1}{2\lambda_x}\frac{1}{n-1}\mathbf{X'Y}\mathbf{w}_y \quad \text{を代入して}$$

$$\left(\frac{1}{n-1}\right)^2 \mathbf{Y'XX'Y}\mathbf{w}_\mathbf{y} = 4\lambda_x\lambda_y \mathbf{w}_\mathbf{y}$$

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) # データの転置
```

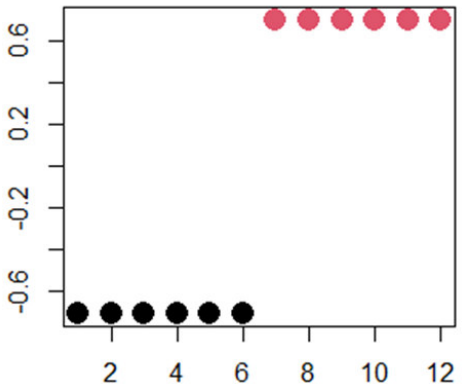
- 目的変数の準備

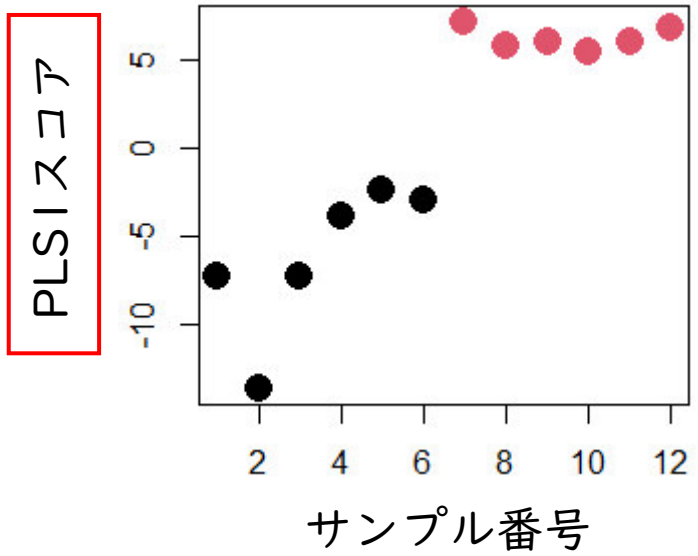
```
Y0 <- factor(class)
Y <- model.matrix(~ Y0 + 0)
```

Partial least squares (1)

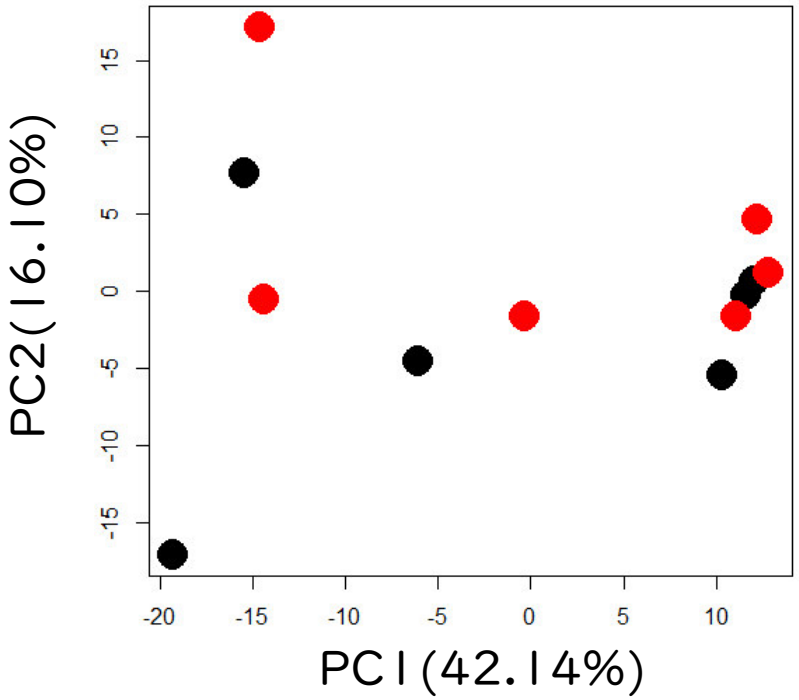
- PLSスコア(第1成分)

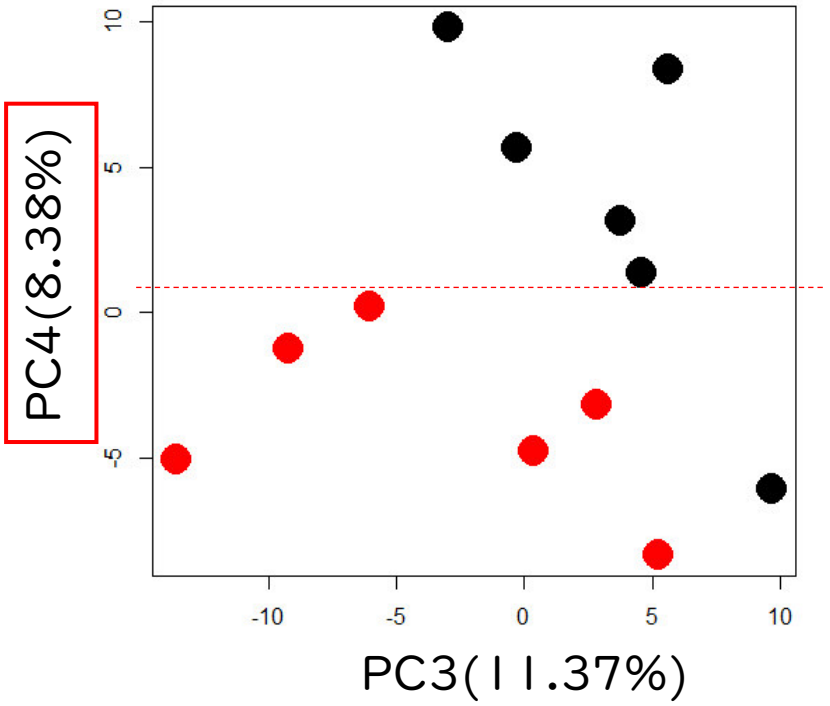
```
library(loadings) # 読み込み済みの場合は不要
pls <- pls_svd(X,Y)
score <- pls$T[,1] # PLS1 スコア
plot(score)
scoreY <- pls$U[,1] # PLS1 スコア
plot(scoreY)
```





主成分分析の結果



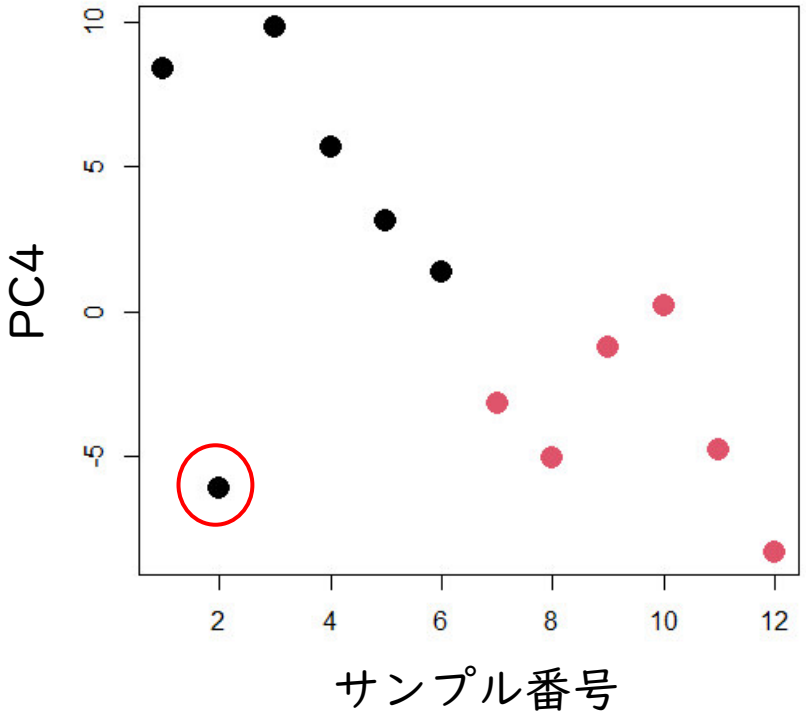


PC4で群間差がある程度確認できているので、
PCAの結果を利用する場合はPC4に着目して解析を進める

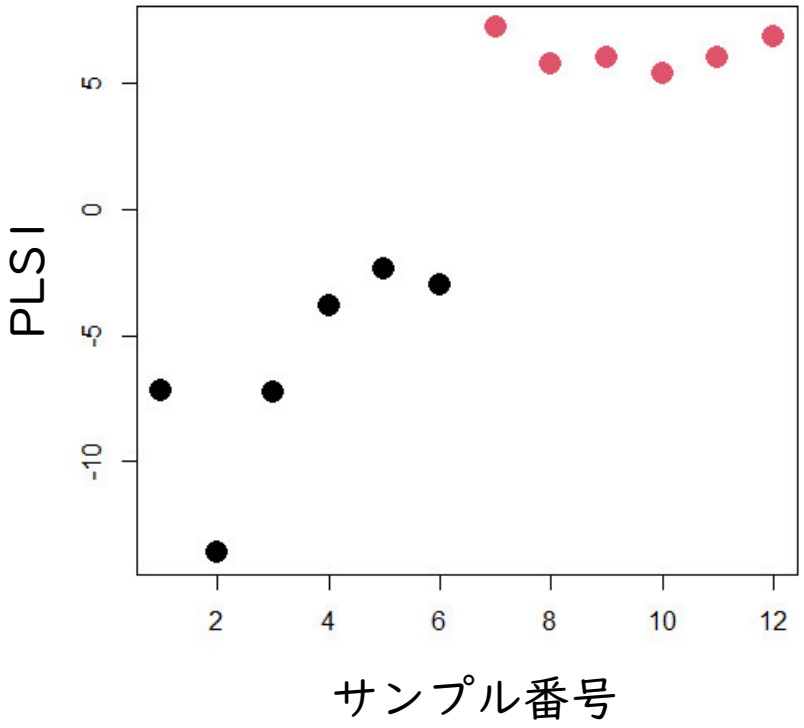
ただし、完全な群間差は確認できない

主成分分析とPLSの比較

主成分分析



PLS

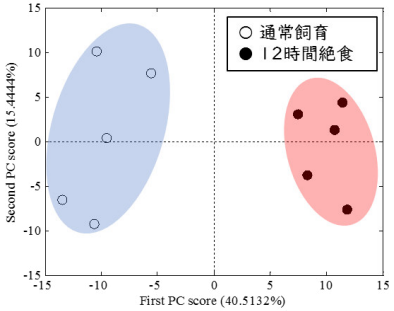


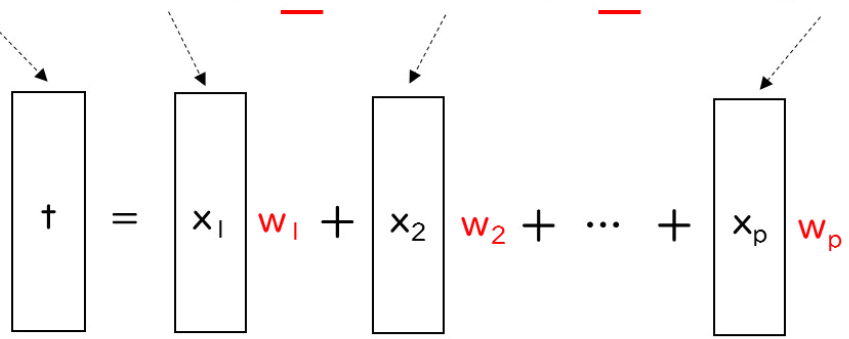
PLS Iスコアで完全な群間差が確認できる

主成分負荷量の復習

- 主成分係数による重要な変数の選び方
  - 主成分係数**w**は、各変数に対する重要度を示す重みであり、主成分係数**w**が大きい変数が重要な変数となる

主成分スコア=(代謝物1)×w_1+(代謝物2)×w_2+・・・+(代謝物P)×w_p





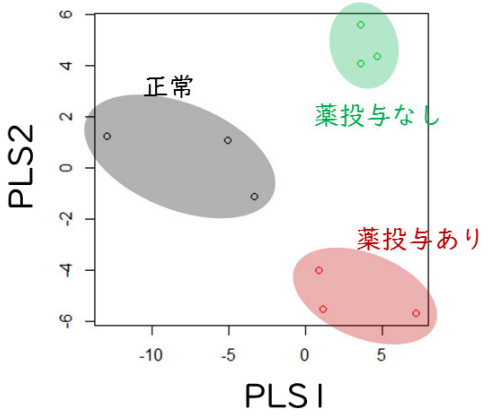
係数w_2の値が大きければ、第1(主)成分と、2番目の変数との関連が高い

- 主成分係数**w**は、「主成分スコアと各変数との相関係数」で定義される主成分負荷量と比例する
- 主成分負荷量の大きい変数を選ぶことで、主成分係数に比べ、統計的な基準(相関係数の値)で重要な変数を選ぶことが出来る

PLS負荷量の場合

- PLS係数による重要な変数の選び方
  - PLS係数**w**は、各変数に対する重要度を示す重みであり、PLS係数**w**が大きい変数が重要な変数となる

PLSスコア＝(代謝物1)×w_1＋(代謝物2)×w_2＋･･･＋(代謝物P)×w_p



$$t = x_1 w_1 + x_2 w_2 + \cdots + x_p w_p$$

[R] PLS係数wの値が大きい上位10物質

・正に大きい10物質　sort(pls$P[,1],decreasing=TRUE)[1:10]
・負に大きい10物質　sort(pls$P[,1])[1:10]

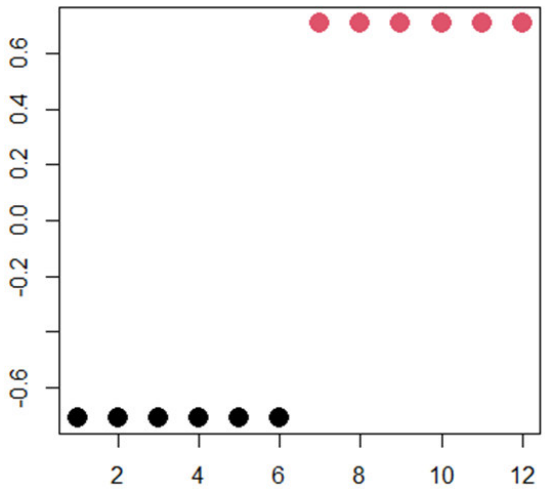
変数の順番を調べるときは、sortの代わりにorderを使用する

PLS係数の値が大きい2つの変数を確認

plot(X[,115], col=class, cex=2, pch=16)

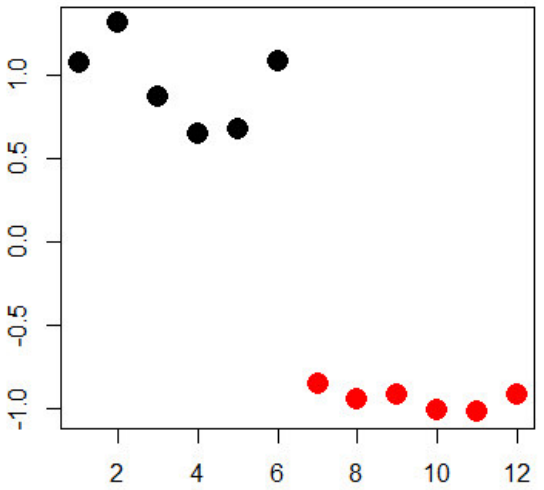
pls$P[115,1]

第1PLSスコア



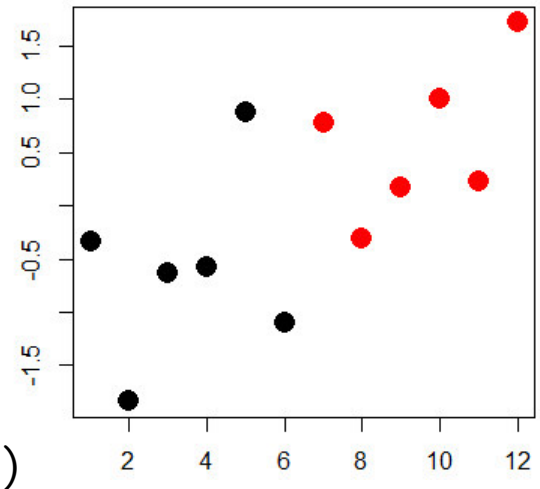
PLS係数の値
-0.1517
(負に最大)

348.2/3288



PLS係数の値
0.0969
(正に最大)

594.2/3395



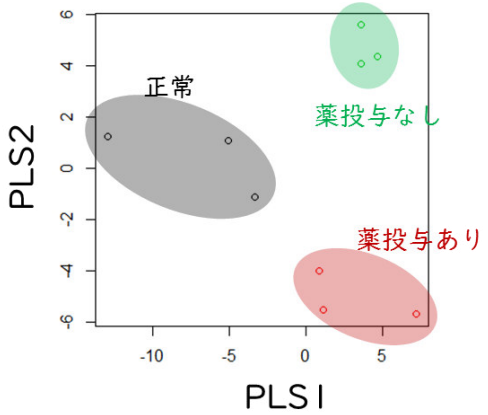
pls$P[401,1]

plot(X[,401],col=class, cex=2, pch=16)

PLS負荷量の場合

- PLS係数による重要な変数の選び方
  - PLS係数**w**は、各変数に対する重要度を示す重みであり、PLS係数**w**が大きい変数が重要な変数となる

PLSスコア
(説明変数)

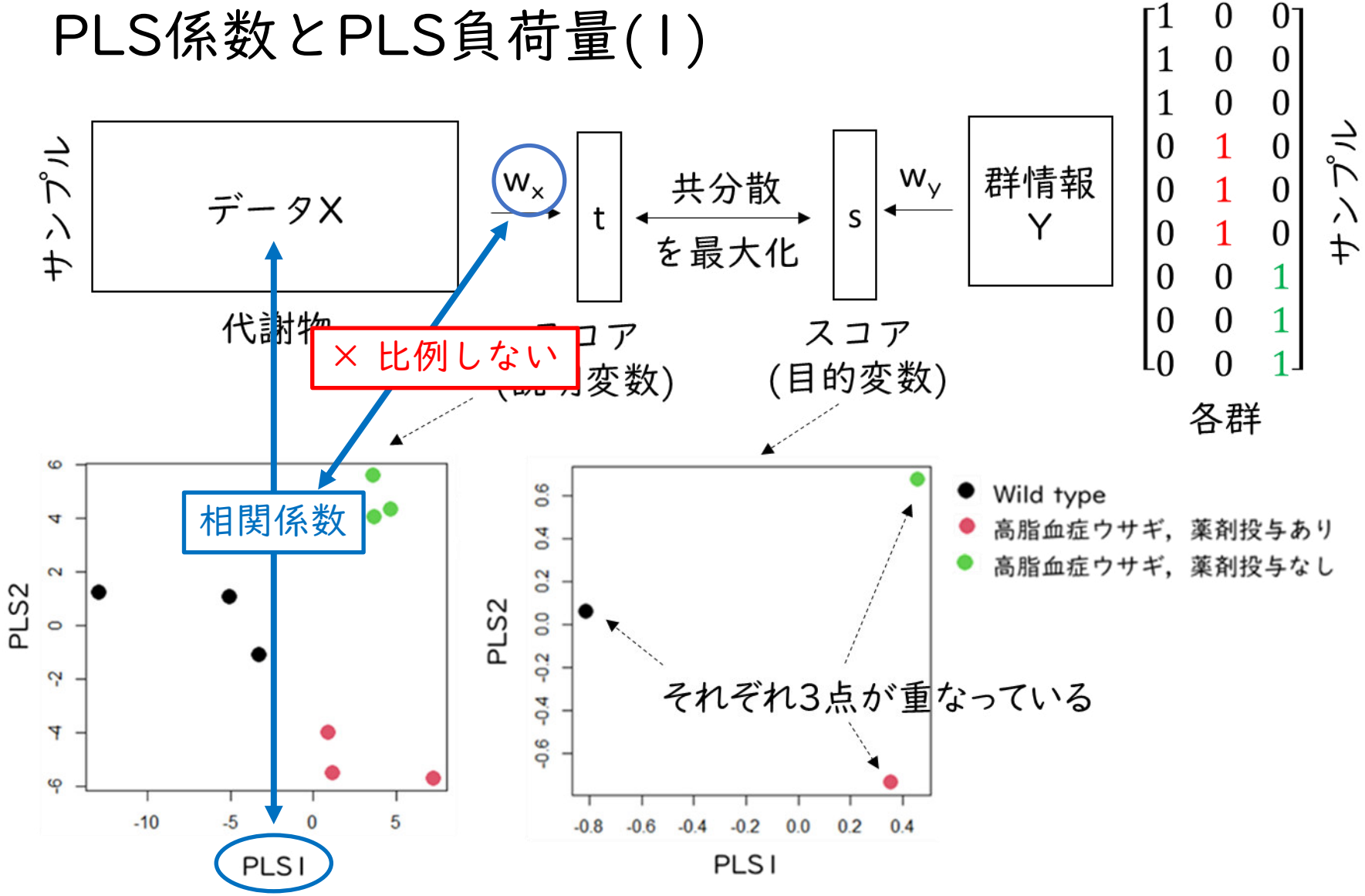


PLSスコア＝(代謝物1)×w_1＋(代謝物2)×w_2＋･･･＋(代謝物P)×w_p

$$t = x_1 w_1 + x_2 w_2 + \cdots + x_p w_p$$

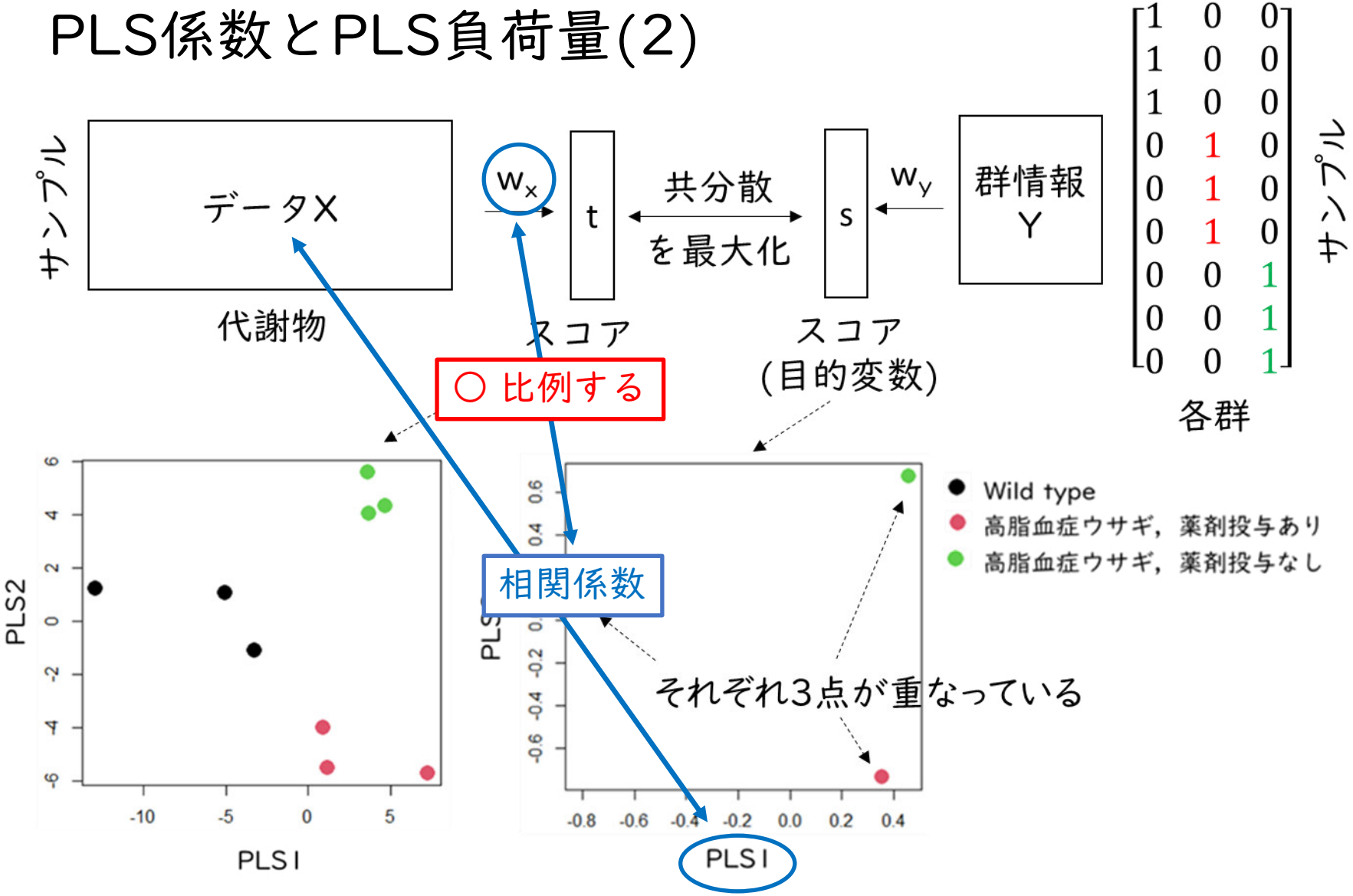
- PLS係数**w**は、PLSスコア(説明変数)と各変数との相関係数に比例しない
- このままでは、主成分負荷量のように統計的な基準(相関係数の値)で重要な変数を選ぶことが出来ない

PLS係数とPLS負荷量(1)



説明変数のスコア、目的変数のスコアいずれも同様の位置(左、右上、右下)に配置されており、傾向が一致していることが確認できる。

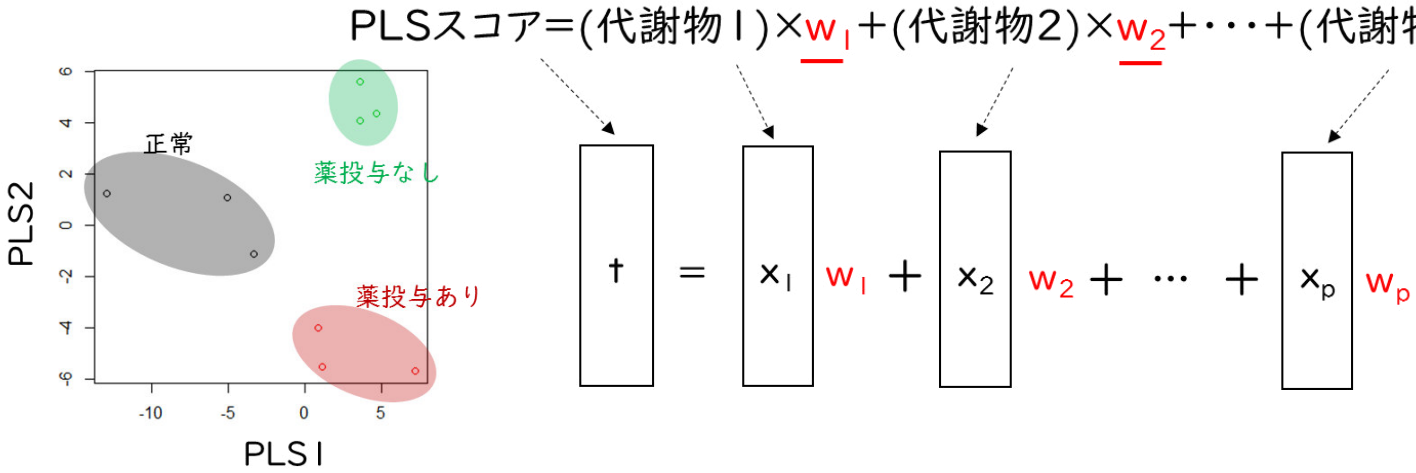
PLS係数とPLS負荷量(2)



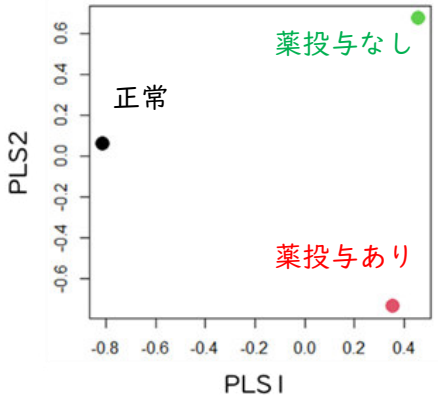
説明変数のスコア、目的変数のスコアいずれも同様の位置(左、右上、右下)に配置されており、傾向が一致していることが確認できる。

PLS負荷量の場合

- PLS係数による重要な変数の選び方
  - PLS係数**w**は、各変数に対する重要度を示す重みであり、PLS係数**w**が大きい変数が重要な変数となる



- PLS係数**w**は、目的変数のPLSスコアと各変数との相関係数と比例する
- PLS負荷量を「目的変数のPLSスコアと各変数との相関係数と定義」し、その値が大きい変数を選ぶことで、統計的な基準(相関係数の値)で重要な変数を選ぶことが出来る



Partial least squares(2)

- PLS負荷量とp-value(PLS1)

```
pls <- pls_loading(pls)
PLS_loading <- pls$loading$R
p_PLS <- pls$loading$p.value
```

- PLS負荷量の出力

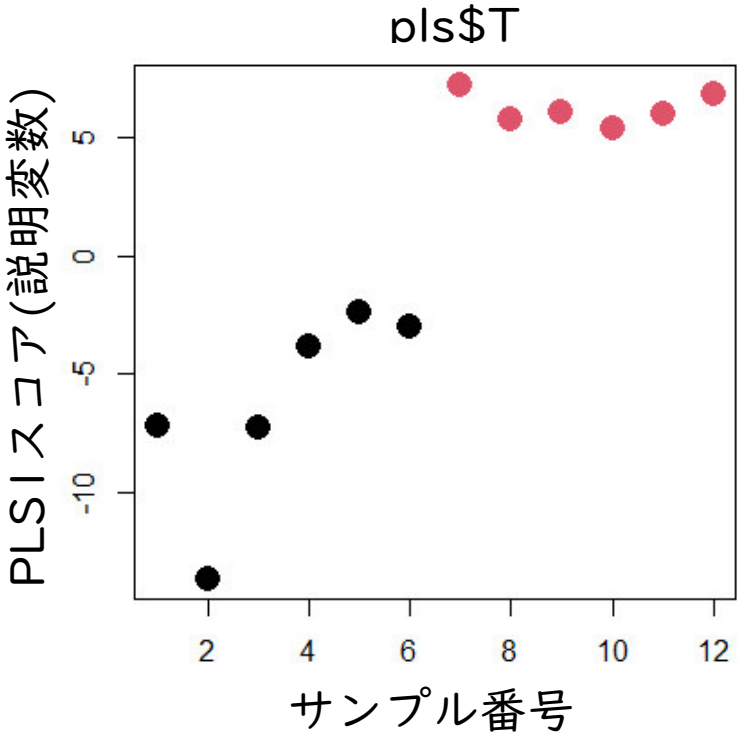
```
PLS1_loading <-  PLS_loading[,1]
p_PLS1 <- p_PLS [,1]
PLS1 <- cbind(PLS1_loading, p_PLS1)
rownames(PLS1) <- X0[,1]
write.csv(PLS1, file= "C:/R/loading_PLS1.csv")
```

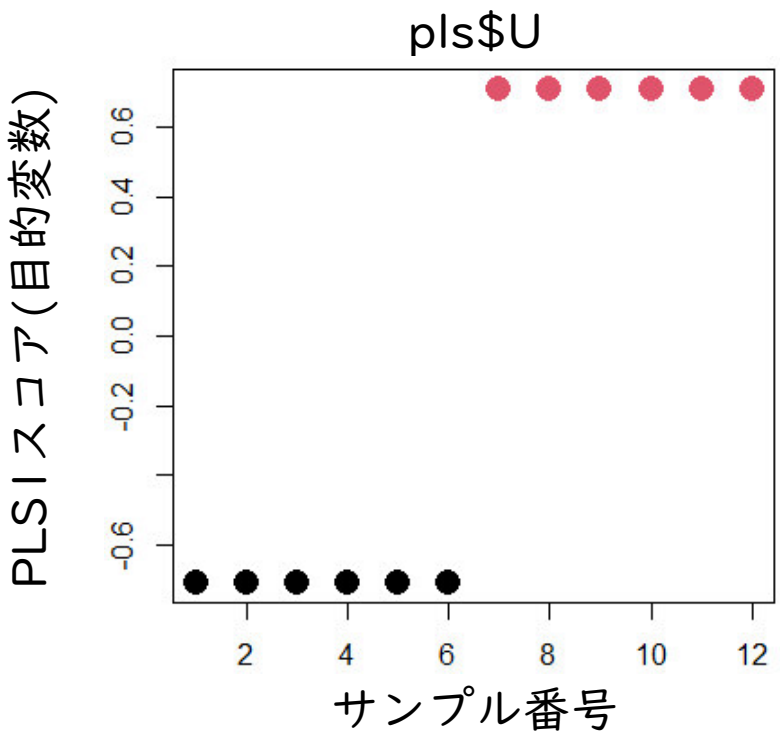
変数名

	A	B	C
1		loading	p
2	348.2/3288	-0.98358206	9.14E-09
3	491.2/3398	-0.979846428	2.53E-08
4	301.2/3389	-0.979721181	2.61E-08
5	300.2/3392	-0.976861564	5.02E-08

PLS負荷量　　p-value

PLS Iの説明変数と目的変数のスコア





PLS Iでいずれも群間差が確認されているが、
　PLS負荷量の値は右図のPLS Iスコア(目的変数)との相関係数であることに注意

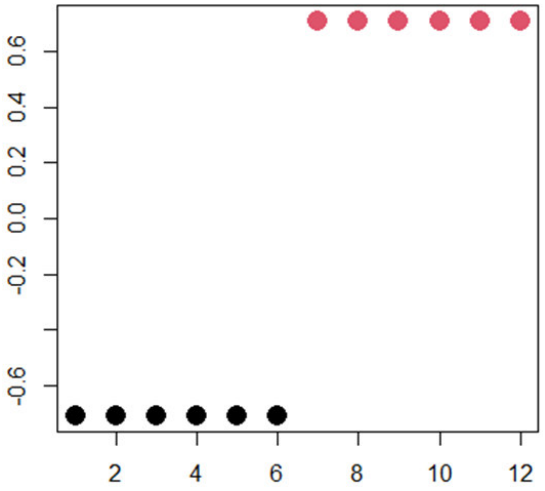
PLS負荷量の値が大きい2つの変数

plot(X[,115],col=class, cex=2, pch=16)

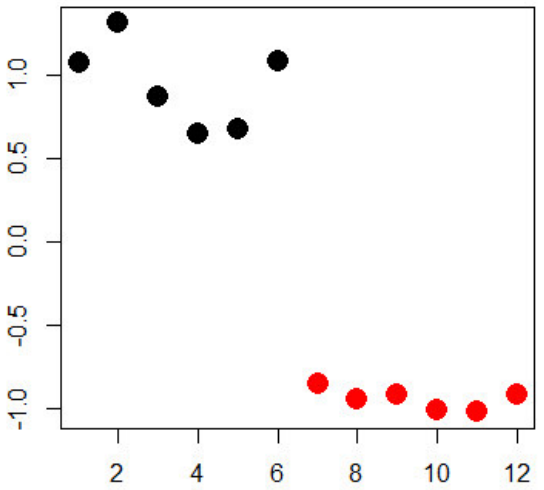
pls$loading$R[115,1]

PLS負荷量の値
-0.9836
(負に最大)

第1PLSスコア

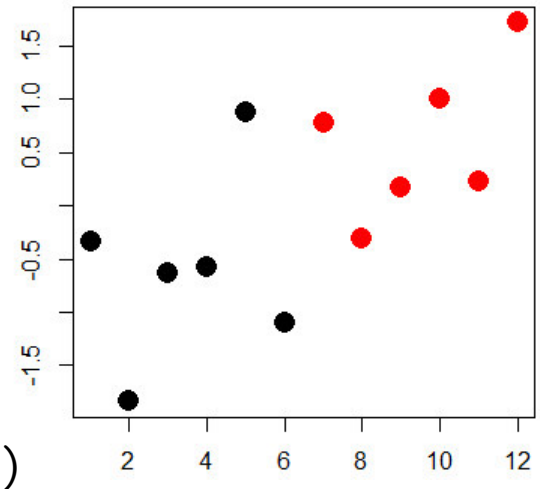


348.2/3288



PLS負荷量の値
0.6283
(正に最大)

594.2/3395



pls$loading$R[401,1]

plot(X[,401],col=class, cex=2, pch=16)

PLSとPLS-DAの違い

- PLS　　本セミナーで紹介してきたPLS

$$\frac{1}{(n-1)^2}\mathbf{X'YY'Xw} = \lambda\mathbf{w}$$

スコア　$\mathbf{t} = \mathbf{Xw}$

- スコアの直交化

$\mathbf{X} = \mathbf{X} - \mathbf{tp'}$

繰り返し

ただし　$\mathbf{p} = \frac{\mathbf{X't}}{\mathbf{t't}}$

PLS-DA

PLSとスコアの直交化の操作を繰り返し計算して得られるスコアを用いて、可視化や判別分析を行うのがPLS-DA

本セミナーでは、最も単純なPLSとPLS-DAを区別して説明して来たが、PLS-DAをPLSと呼ぶことも多い

[再掲] 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) # データの転置
```

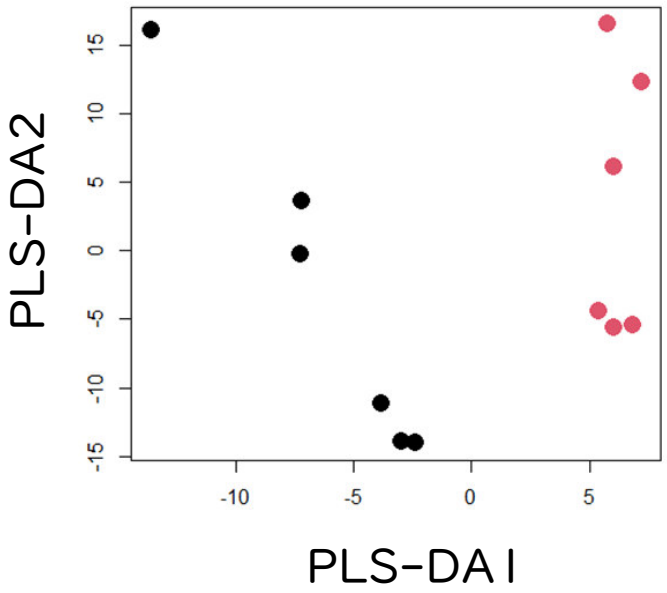
- 目的変数の準備

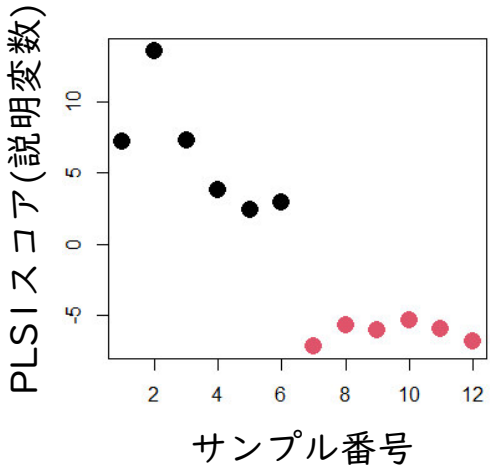
```
Y0 <- factor(class)
Y <- model.matrix(~ Y0 + 0)
```

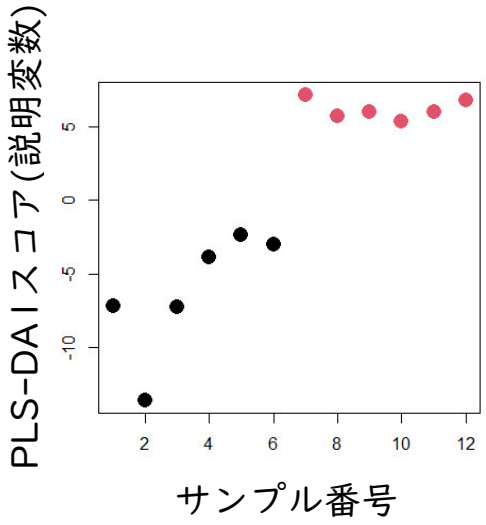
PLS-DAの計算(1)

- PLS-DAスコアの計算

```
library(loadings) # 読み込み済みの場合は不要
plsda <- pls_da(X,Y,2)
score <- plsda$T # PLS-DAスコア(第1、第2)
plot(score, col=class, pch=16, cex=2)
```







PLSとPLS-DAの第1成分のスコアは一致
(正負が逆でも意味は同じ)

PLS-DAのための各種パッケージ

- chemometricsパッケージ

```
library(chemometrics)
plsda1 <- pls2_nipals(X, Y, a=2, scale=TRUE)
```

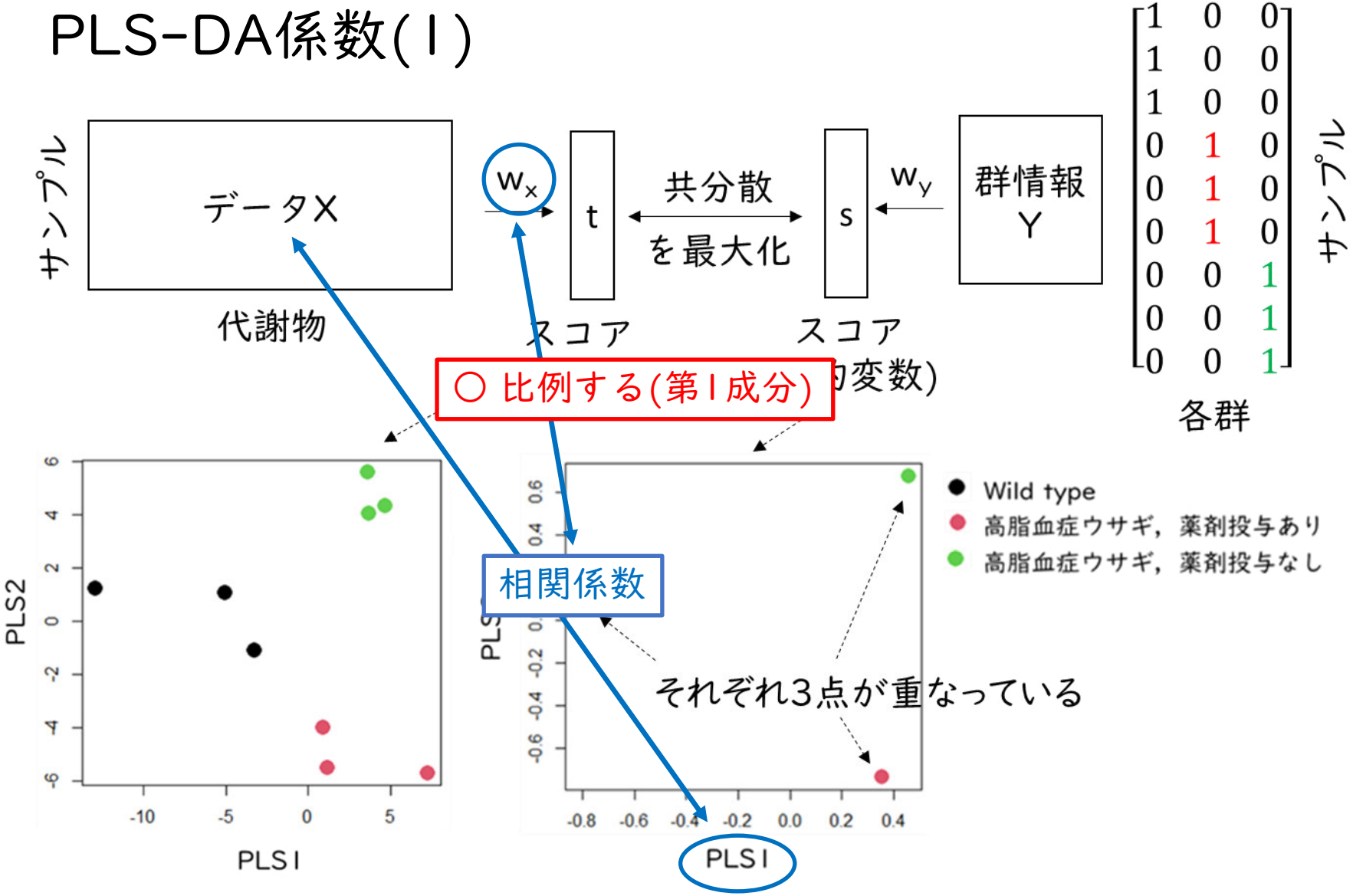
- plsパッケージ

```
library(pls)
plsda2 <- cppls(Y~X, 2, data=data.frame(X=X, Y=Y),
                                        scale=TRUE)
```

- mixOmicsパッケージ

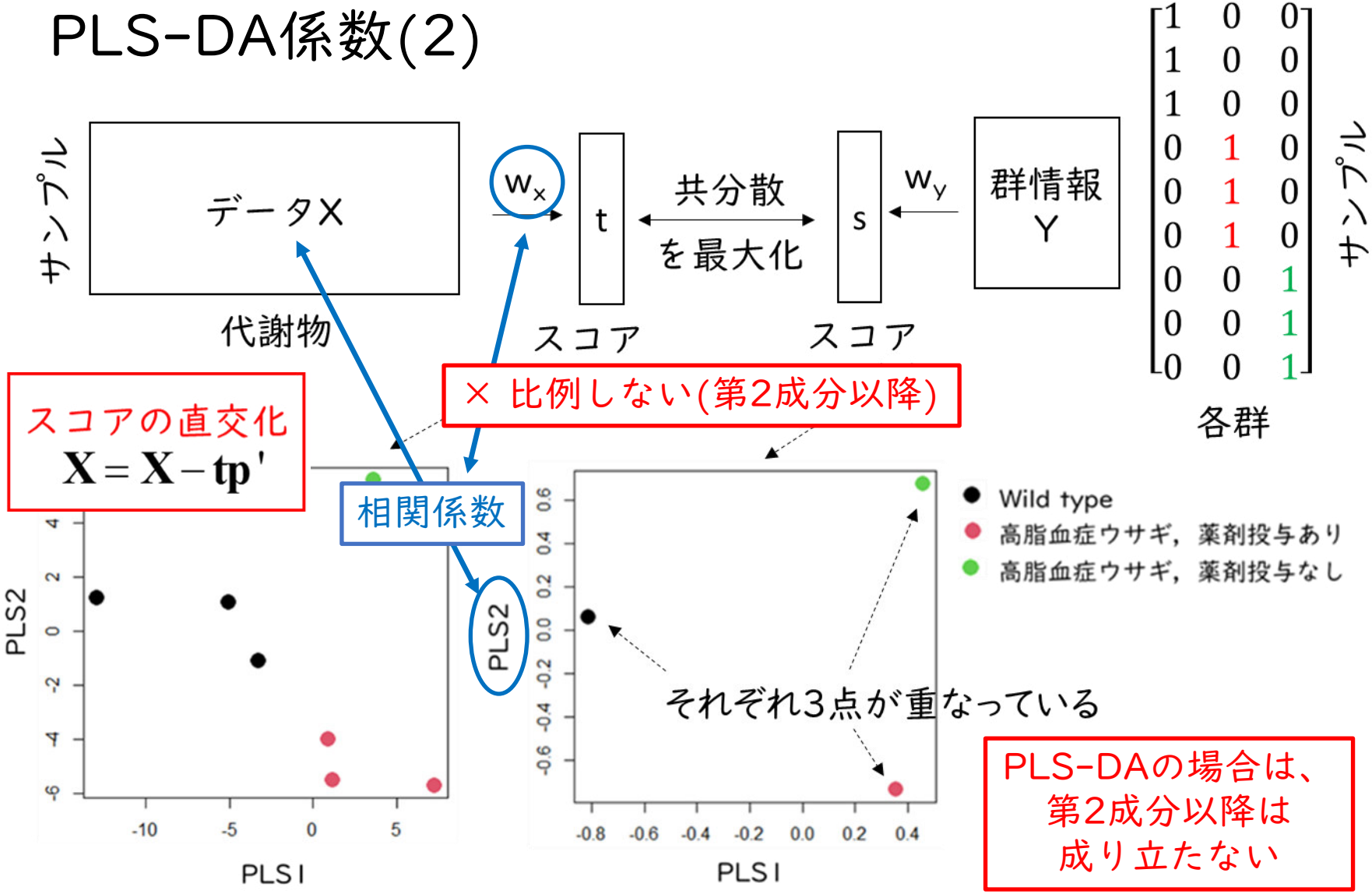
```
library(mixOmics)
plsda3 <- plsda(X, Y0) # デフォルトでscale=TRUE
```

PLS-DA係数(1)



説明変数のスコア、目的変数のスコアいずれも同様の位置(左、右上、右下)に配置されており、傾向が一致していることが確認できる。

PLS-DA係数(2)



説明変数のスコア、目的変数のスコアいずれも同様の位置(左、右上、右下)に配置されており、傾向が一致していることが確認できる。

PLS-DAにおけるローディング

- PLS　　本セミナーで紹介してきたPLS

$$\frac{1}{(n-1)^2}\mathbf{X'YY'Xw} = \lambda\mathbf{w}$$

スコア　$\mathbf{t} = \mathbf{Xw}$

- スコアの直交化

繰り返し

$\mathbf{X} = \mathbf{X} - \mathbf{tp'}$

ただし $\mathbf{p} = \dfrac{\mathbf{X't}}{\mathbf{t't}}$

PLS-DA

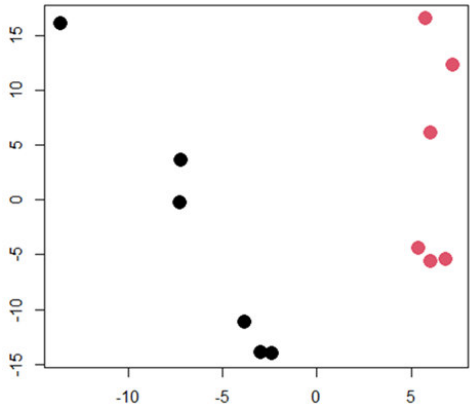
ローディング

PLSとスコアの直交化の操作を繰り返し計算して得られるスコアを用いて、可視化や判別分析を行うのがPLS-DA

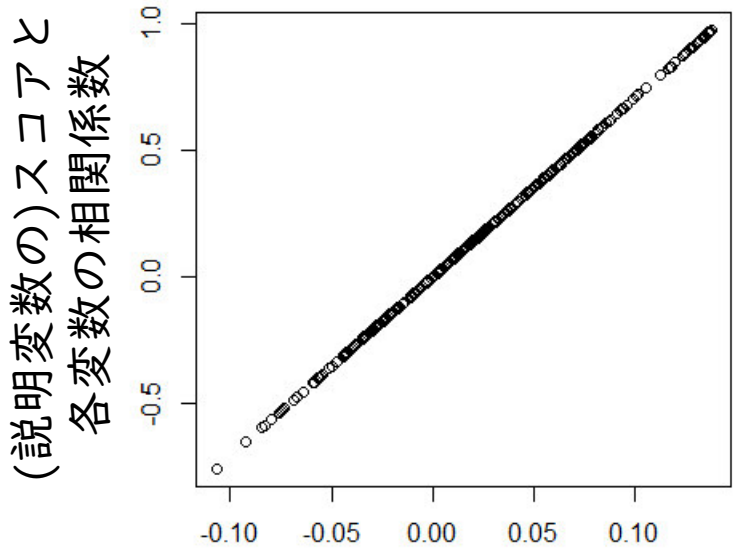
PLS-DAのローディングとは

PLS-DAのローディングは、目的変数のスコアと各代謝物の相関係数ではなく、説明変数のスコアと各代謝物の相関係数に比例する値(p)が用いられる

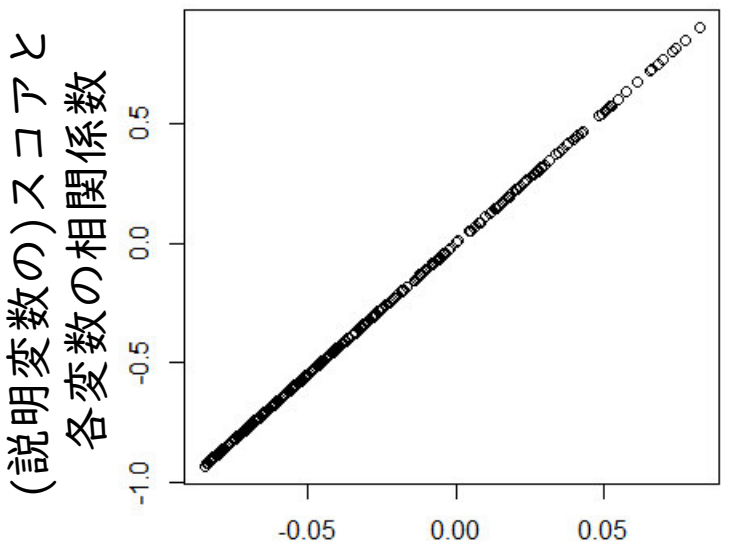
PLS-DAスコア(説明変数)



第1PLS-DA



第2PLS-DA



PLS-DA負荷量の性質

- PLS-DA負荷量
  - PLS-DAローディングとしてベクトルpを利用するよりも、PLS-DAスコアと各代謝物の相関係数をPLS-DA負荷量として定義する、と考えた方が理解しやすい

- PLS-DA負荷量に変換

$$\mathbf{R} = \mathbf{P} \times \sqrt{\frac{1}{N}\mathbf{t}'\mathbf{t}}$$

主成分分析やPLSと同様に、PLS-DA負荷量に変換することにより、相関係数として利用可能

PLS-DAでのローディングから負荷量への変換

第1PLS-DAのローディング

$$\mathbf{p}_1 = \frac{\mathbf{X}'\mathbf{t}_1}{\mathbf{t}_1'\mathbf{t}_1}$$

第1成分のローディングはPLS-DAスコア(説明変数)と各変数の相関係数に比例する統計量となっている

第2PLS-DAのローディング

$$\mathbf{p}_2 = \frac{\widetilde{\mathbf{X}}'\mathbf{t}_2}{\mathbf{t}_2'\mathbf{t}_2} \quad \text{に} \quad \widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1' \quad \text{を代入して}$$

$$\mathbf{p}_2 = \frac{(\mathbf{X} - \mathbf{t}_1\mathbf{p}_{1'})'\mathbf{t}_2}{\mathbf{t}_2'\mathbf{t}_2} = \frac{\mathbf{X}'\mathbf{t}_2 - \mathbf{p}_1'\mathbf{t}_1'\mathbf{t}_2}{\mathbf{t}_2'\mathbf{t2}}$$

ここで第1成分のスコア$\mathbf{t}_1$と第2成分のスコア$\mathbf{t}_2$は直交し$\mathbf{t}_1\mathbf{t}_2=0$となることから、次のように書ける

$$\mathbf{p}_2 = \frac{\mathbf{X}'\mathbf{t}_2}{\mathbf{t}_2'\mathbf{t2}}$$

以上より、第2成分(以降)のローディングも、PLS-DAスコア(説明変数)と各変数の相関係数に比例する統計量となることが確認できる

PLS-DA負荷量

- PLS-DA負荷量とp-value(PLS-DA l)

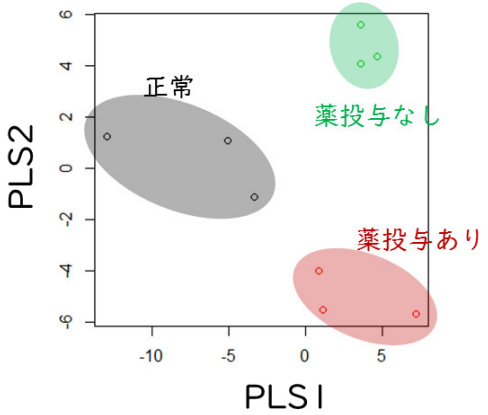
```
plsda <- plsda_loading(plsda)
PLSDA_loading <- plsda$loading$R
p_PLSDA <- plsda$loading$p.value
```

- PLS負荷量の出力

```
PLSDA l_loading <-  PLSDA_loading[, l ]
p_PLSDA l <- p_PLSDA [, l ]
PLSDA l <- cbind(PLSDA l_loading, p_PLSDA l )
rownames(PLSDA l ) <- X0[, l ]
write.csv(PLSDA l , file= "C:/R/loading_PLSDA l .csv")
```

PLS負荷量とPLS-DA負荷量の違い

- [再掲] PLS係数による重要な変数の選び方
  - PLS係数**w**は、各変数に対する重要度を示す重みであり、PLS係数**w**が大きい変数が重要な変数となる



PLSスコア＝(代謝物1)×w_1＋(代謝物2)×w_2＋･･･＋(代謝物P)×w_p

$$t = x_1 w_1 + x_2 w_2 + \cdots + x_p w_p$$

PLS-DAのpは、PLSのwとは異なり、
上式のような関係にはならない

pは、PLS-DAスコアと各代謝物の相関係数に比例することから、
pの値が大きな代謝物は、PLS-DAスコアに似た傾向を示す代謝物となる

PLSとPLS-DAの比較

- PLS
  - PLSスコア・PLS係数・PLS負荷量は、(群数-1)個までしか計算することが出来ない
  - PLS係数、PLS負荷量を利用する
    - PLS係数は、目的変数のスコアと各変数との相関係数に比例
    - PLS負荷量を相関係数で定義することができる

PLS-DA

- PLS-DAスコアは、より多くの数計算することが出来る
  - (サンプルサイズもしくは変数の数-1)個まで
- 第2成分以降のPLS-DA係数は利用できない
  - PLS-DA係数は、目的変数のスコアと各変数との相関係数に比例せず、PLSと同様に理解することは出来ない
- PLS-DAローディング、PLS-DA負荷量を利用する
  - PLS-DAのローディングは、説明変数のスコアと各変数との相関係数に比例し、PLS-DA負荷量は相関係数として定義できる