BioC Asia 2021 Workshop
(2021.11.1)

Rを用いたメタボロームデータ解析: Metabolome data analysis using R

ヒューマン・メタボローム・テクノロジーズ株式会社
山本　博之

メタボロミクスにおけるデータ解析

質量分析装置

質量分析装置から得られた信号を解析

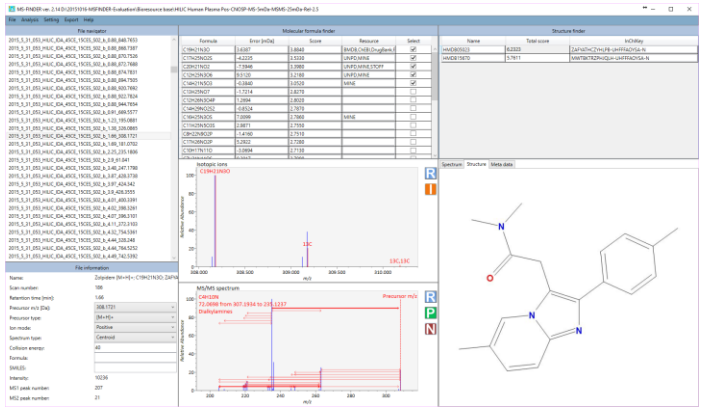
解析ソフトウェア　MS-DIAL

Excelデータ

MetaboAnalyst

代謝物名が不明の未知ピークの
構造を推定

構造推定ソフトウェア
MS-FINDER



リポジトリへの登録

MetaboBank

メタボロミクスにおけるデータ解析

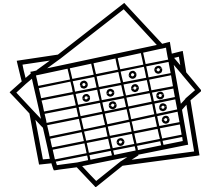
Rで全て実行

質量分析装置

質量分析装置から得られた信号を解析

解析ソフトウェア　MS-DIAL

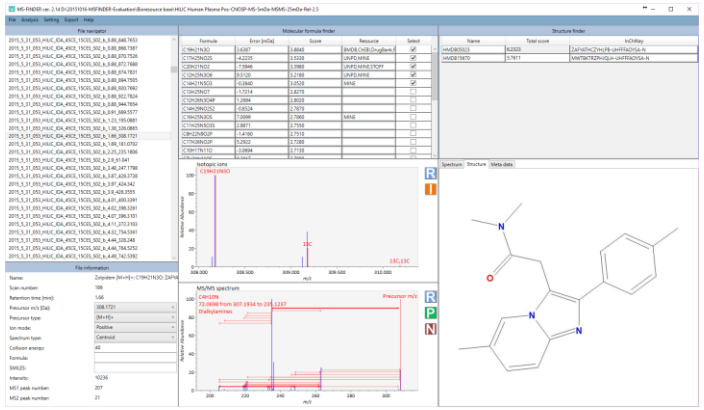
Excelデータ



MetaboAnalyst

代謝物名が不明の未知ピークの
構造を推定

構造推定ソフトウェア
MS-FINDER



リポジトリへの登録

MetaboBank

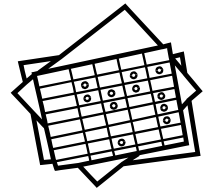
メタボロミクスにおけるデータ解析

質量分析装置

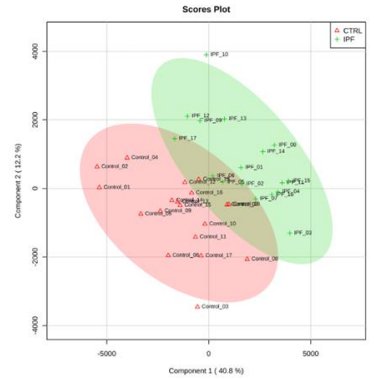
質量分析装置から得られた信号を解析

解析ソフトウェア　MS-DIAL

Excelデータ



MetaboAnalyst



MS Data Handling
mzR、Msnbase (BioConductor)
rmzTab-M、MRMConverteR、
Chromatogtams、Spectraなど

Peak Picking, Grouping and Alignment
LC-MS Focussed or General
xcms、IPO、Autotuner、yamss、
cosmiq (BioConductor)
xMSanalyzer、apLCMS、warpgroup、
AMDORAP、anviGCMS、enviPick、
massFlowR、KPIC2など

Statistical analysis
ChemoSpec、pcaMethods、
multtest、biosigner、
OmicsMarkeR、RankProd、
ropls、OmicsLonDA、impute、
MOFA(Bioconductor)
MetaboAnalystRなど

Stanstrup J. et al.,“The metaRbolomics Toolbox in Bioconductor and beyond”Metabolites 2019, 9(10), 200より

メタボロミクスにおけるデータ解析

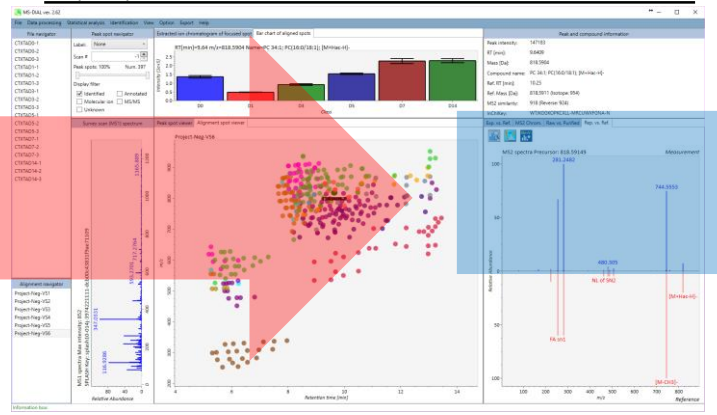
質量分析データ処理

統計解析

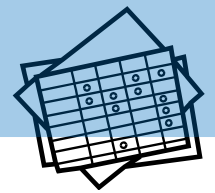
質量分析装置

質量分析装置から得られた信号を解析

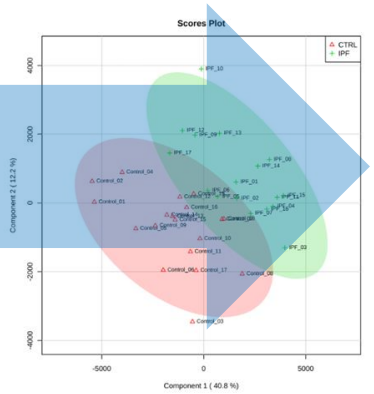
解析ソフトウェア　MS-DIAL



Excelデータ



MetaboAnalyst



MS Data Handling
mzR、Msnbase (BioConductor)
rmzTab-M、MRMConverteR、
Chromatogtams、Spectraなど

Peak Picking, Grouping and Alignment
LC-MS Focussed or General
xcms、IPO、Autotuner、yamss、
cosmiq (BioConductor)
xMSanalyzer、apLCMS、warpgroup、
AMDORAP、anviGCMS、enviPick、
massFlowR、KPIC2など

Statistical analysis
ChemoSpec、pcaMethods、
multtest、biosigner、
OmicsMarkeR、RankProd、
ropls、OmicsLonDA、impute、
MOFA(Bioconductor)
MetaboAnalystRなど

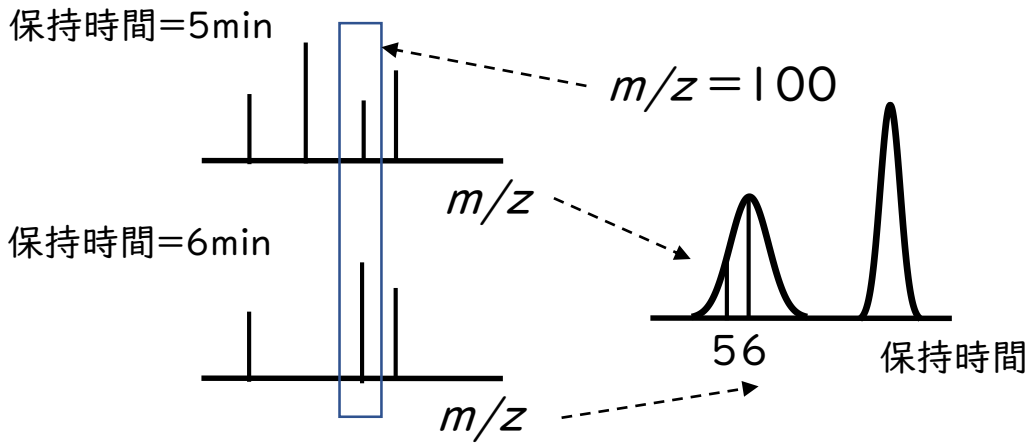
Stanstrup J. et al.,“The metaRbolomics Toolbox in Bioconductor and beyond” Metabolites 2019, 9(10), 200より

質量分析データ処理(1)

データの読み込み → ピークピッキング → クロマトアライメント → ピークアライメント → 欠損ピークの穴埋め → 行列データ

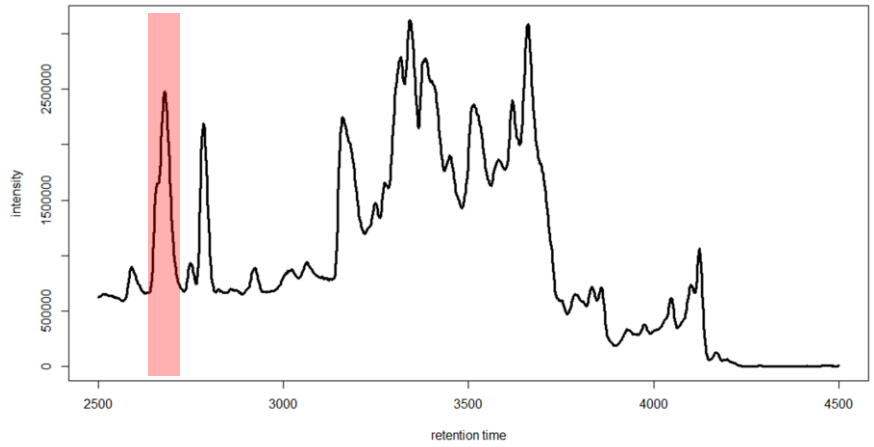
(時間補正)　(サンプル間のピークの対応付け)

データの読み込み



マススペクトル　　クロマトグラム

ピークピッキング

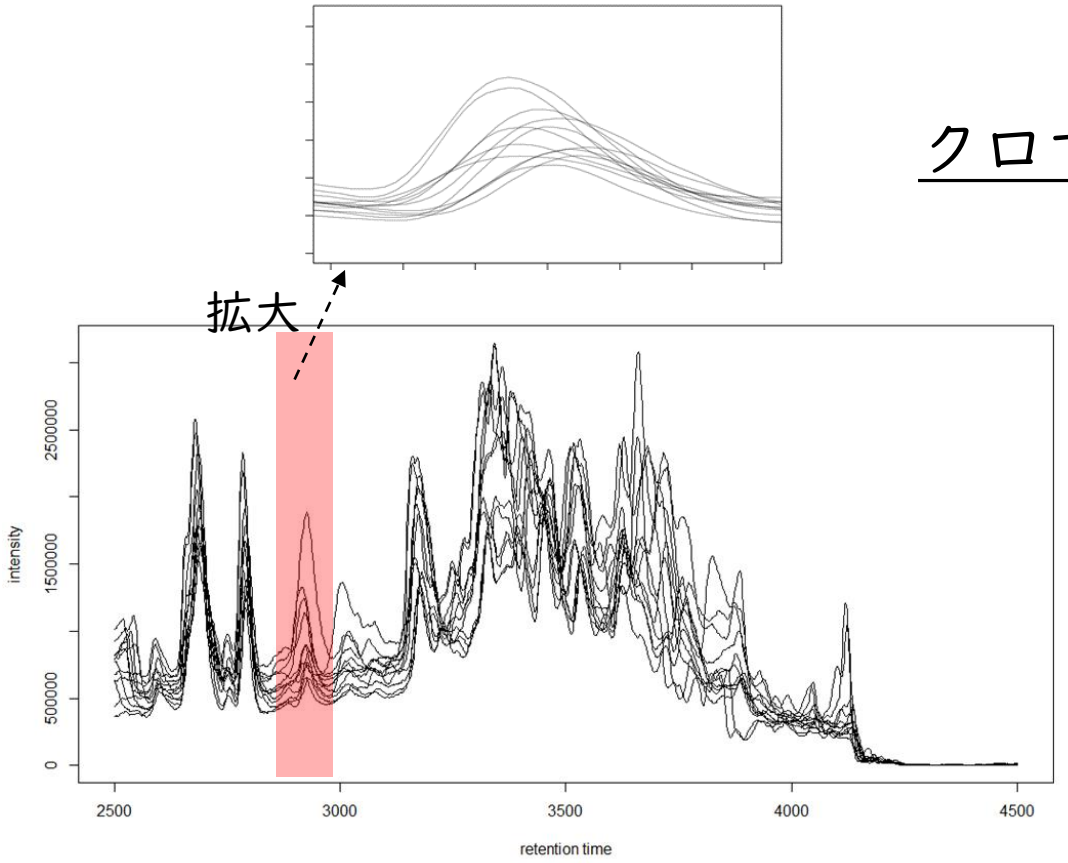


ガウス型関数に近い形状の領域を拾い上げる

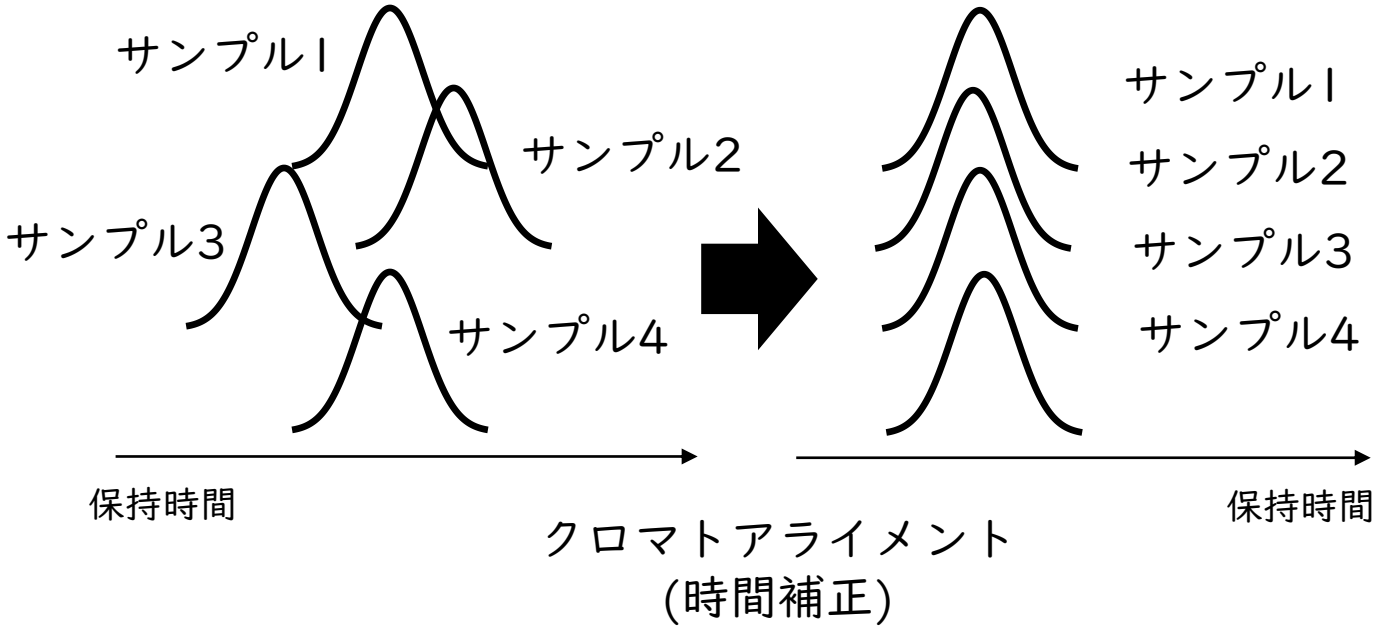
質量分析データ処理(2)

データの読み込み → ピークピッキング → クロマトアライメント → ピークアライメント → 欠損ピークの穴埋め → 行列データ

(時間補正)　(サンプル間のピークの対応付け)

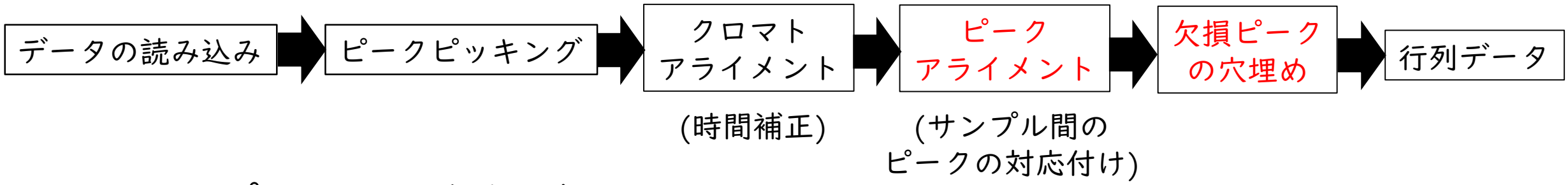


クロマトアライメント

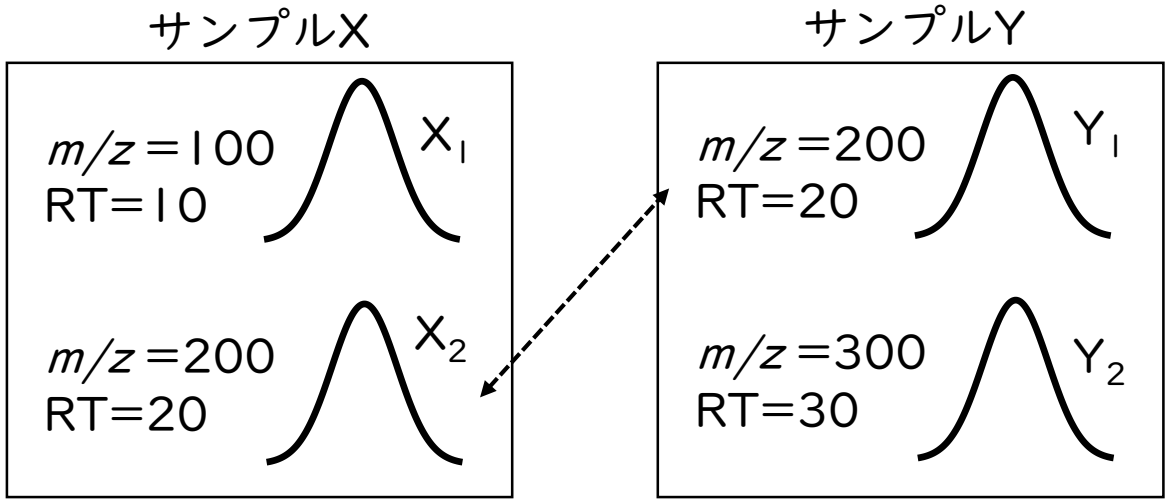


クロマトアライメント
(時間補正)

質量分析データ処理(3)

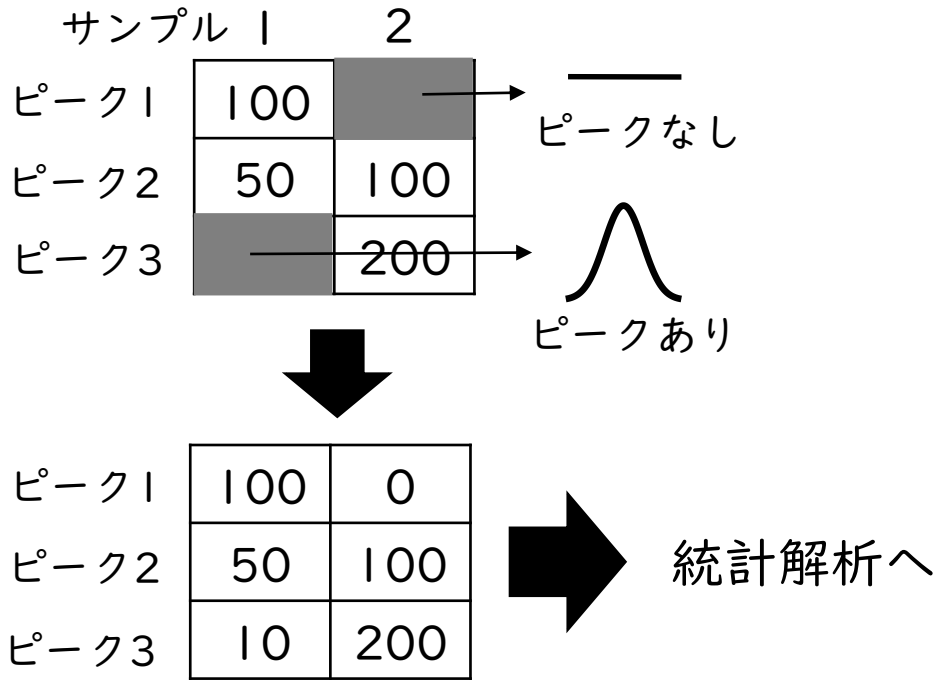


ピークアライメント



m/zとRTが同じなので、同一代謝物由来のピーク
(実際は測定の誤差があるので、ある範囲に入るピーク)

欠損ピークの穴埋め



メタボロミクスにおけるデータ解析

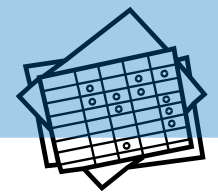
質量分析データ処理

統計解析

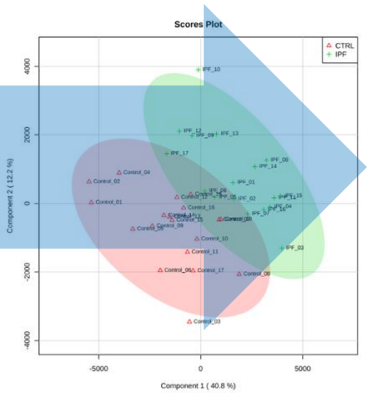
質量分析装置

解析ソフトウェア　MS-DIAL

Excelデータ



主成分分析



質量分析装置から得られた信号を解析

MS Data Handling

mzR、Msnbase (BioConductor)
rmzTab-M、MRMConverteR、
Chromatogtams、Spectraなど

Peak Picking, Grouping and Alignment

LC-MS Focussed or General

xcms、IPO、Autotuner、yamss、
cosmiq (BioConductor)
xMSanalyzer、apLCMS、warpgroup、
AMDORAP、anviGCMS、enviPick、
massFlowR、KPIC2など

Statistical analysis

ChemoSpec、pcaMethods、
multtest、biosigner、
OmicsMarkeR、RankProd、
ropls、OmicsLonDA、impute、
MOFA(Bioconductor)
MetaboAnalystRなど

loadings (2021.9リリース)

露崎さん、西田さん、久米さんにより毎月開催されている
BioPackathon (https://sites.google.com/view/biopackathon/)で開発

多変量解析によるメタボロームデータの解析の流れ

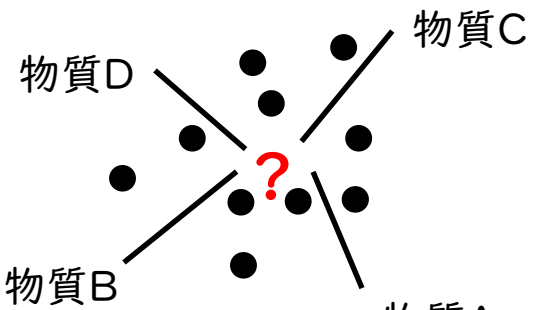
メタボロミクスでは、
低分子の代謝物を一斉に分析

Capillary electrophoresis(CE)

Mass spectrometer (MS)

質量分析
データ処理

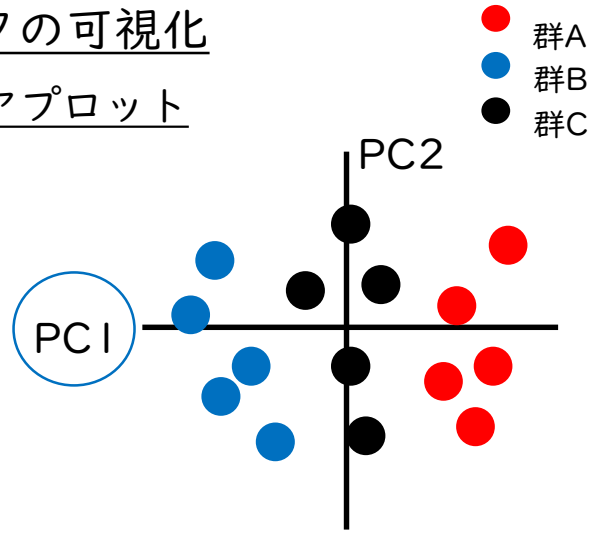
メタボロームデータ



主成分分析
PLS(群情報あり)

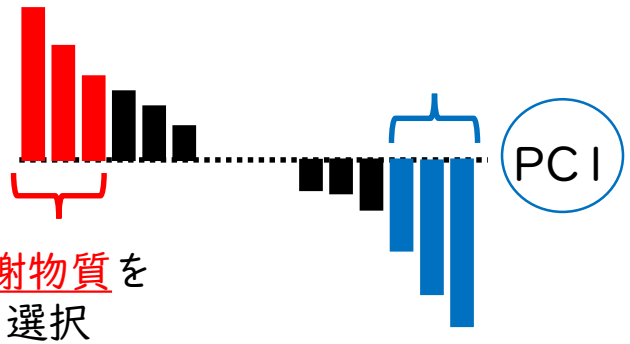
1. データの可視化

スコアプロット



2. 代謝物質の選択

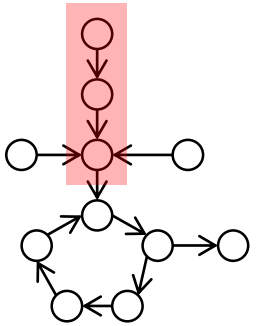
ローディングプロット



代謝物質を
選択

3. パスウェイとの関連

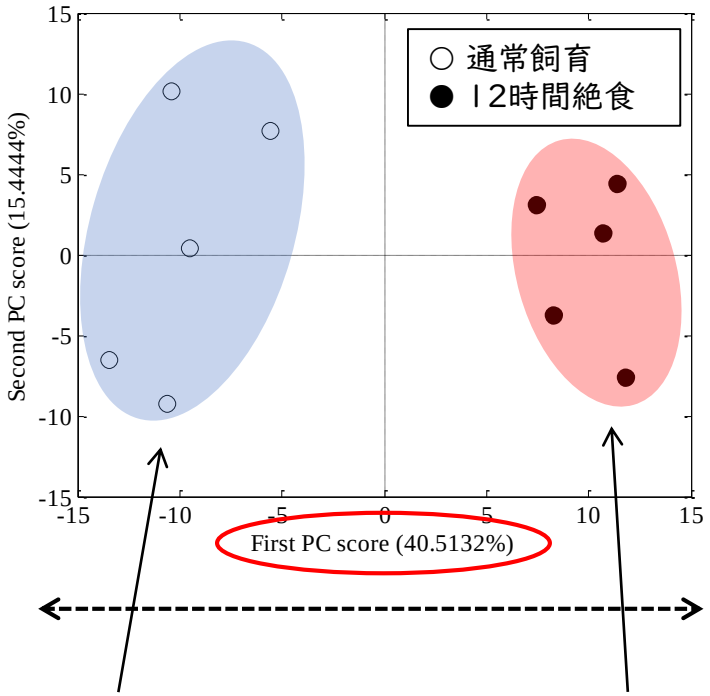
(データベースとの照合)



代謝経路を選択

絶食マウスでの主成分分析の解析例

データの可視化



通常飼育
マウスで低値(−)

絶食12時間
マウスで高値(+)

重み(主成分)係数を用いて変数を選ぶ

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

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

主成分スコアは、各変数のデータを重みwを係数として足し合わせたものであり、wの値が大きい変数が主成分スコアと関連が強く、wの値が小さい変数が主成分スコアと関連が弱い

主成分係数wの値が大きい上位10個
もしくは30個程度の代謝物を選ぶ

一般的なローディング(主成負荷量)の定義

杉山高一、多変量データ解析入門、朝倉書店(2001) 35ページ

『主成分と変数の相関係数を、主成分の因子負荷量(factor loading)という』

塩谷實、多変量解析概論、朝倉書店(2001) 114ページ

『共分散行列を用いた主成分分析の場合、第i主成分U_iとj番目の成分変数X_jの間の相関係数をX_jのU_iへの因子負荷量という』

(その後に、相関係数行列についても同じであることが書かれている)

中村永友、Rで学ぶデータサイエンス　多次元データ解析法、共立出版(2009) 102ページ

『得られた主成分と本来の変数の相関係数を主成分負荷量あるいは因子負荷量という』

主成分分析の主成分負荷量といえば、「主成分スコアと各変数との相関係数」として定義され、主成分負荷量を用いて重要な代謝物が選ばれる。

主成分係数(重み、固有ベクトル)と主成分負荷量の比較

正に値が大きな上位10個

ピーク	主成分係数	主成分負荷量
384.2/3993	0.1279	0.7489
491.2/3398	0.1260	0.7375
449.1/3290	0.1197	0.7007
322.1/3392	0.1195	0.6997
439.3/4056	0.1134	0.6638
438.3/4056	0.1127	0.6597
301.2/3389	0.1085	0.6352
354.2/3618	0.1076	0.6299
300.2/3392	0.1075	0.6293
410.3/3937	0.1064	0.6228

負に値が大きな上位10個

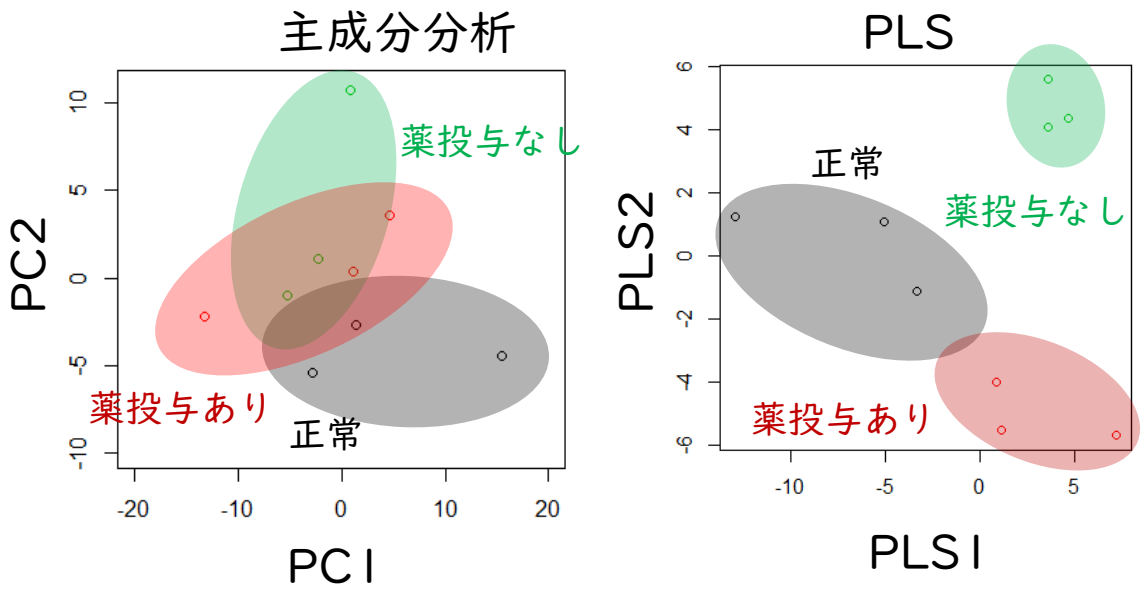
ピーク	主成分係数	主成分負荷量
306.1/2928	-0.1132	-0.6628
414.2/3060	-0.0943	-0.5523
246.1/2517	-0.0939	-0.5498
591.3/3003	-0.0926	-0.5419
288.1/2798	-0.0923	-0.5402
590.3/3003	-0.0916	-0.5365
546.3/3015	-0.0888	-0.5200
532.3/3729	-0.0884	-0.5177
547.2/3015	-0.0881	-0.5156
576.3/2860	-0.0844	-0.4944

主成分負荷量(＝主成分スコアと各代謝物の相関係数)の値が0.7以上が3物質であり、-0.7以下の代謝物は確認できなかった。

主成分係数と主成分負荷量は比例関係にあり、値が大きな代謝物の並び順は変わらないが、主成分負荷量を用いて重要な代謝物を選ぶことで、統計的な基準(相関係数の値)で重要な代謝物を選ぶことが出来る。

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

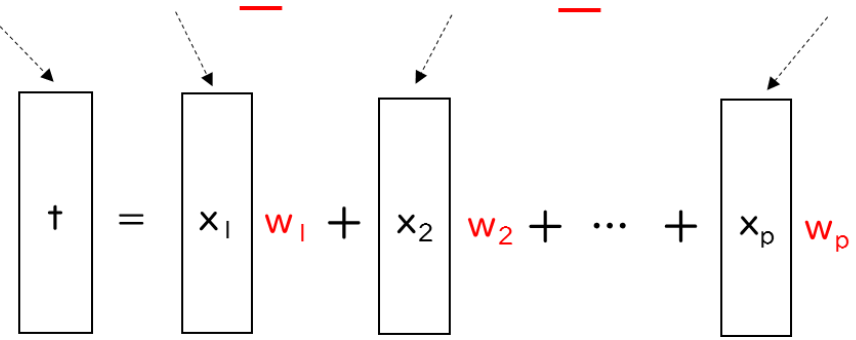


主成分分析の結果、主成分スコアで群間の差が表れなかったとき、PLSが用いられることが多い

Yamamoto H., "PLS-ROG: Partial least squares with rank order of groups.", Journal of Chemometrics, 31(3) (2017) e2883.

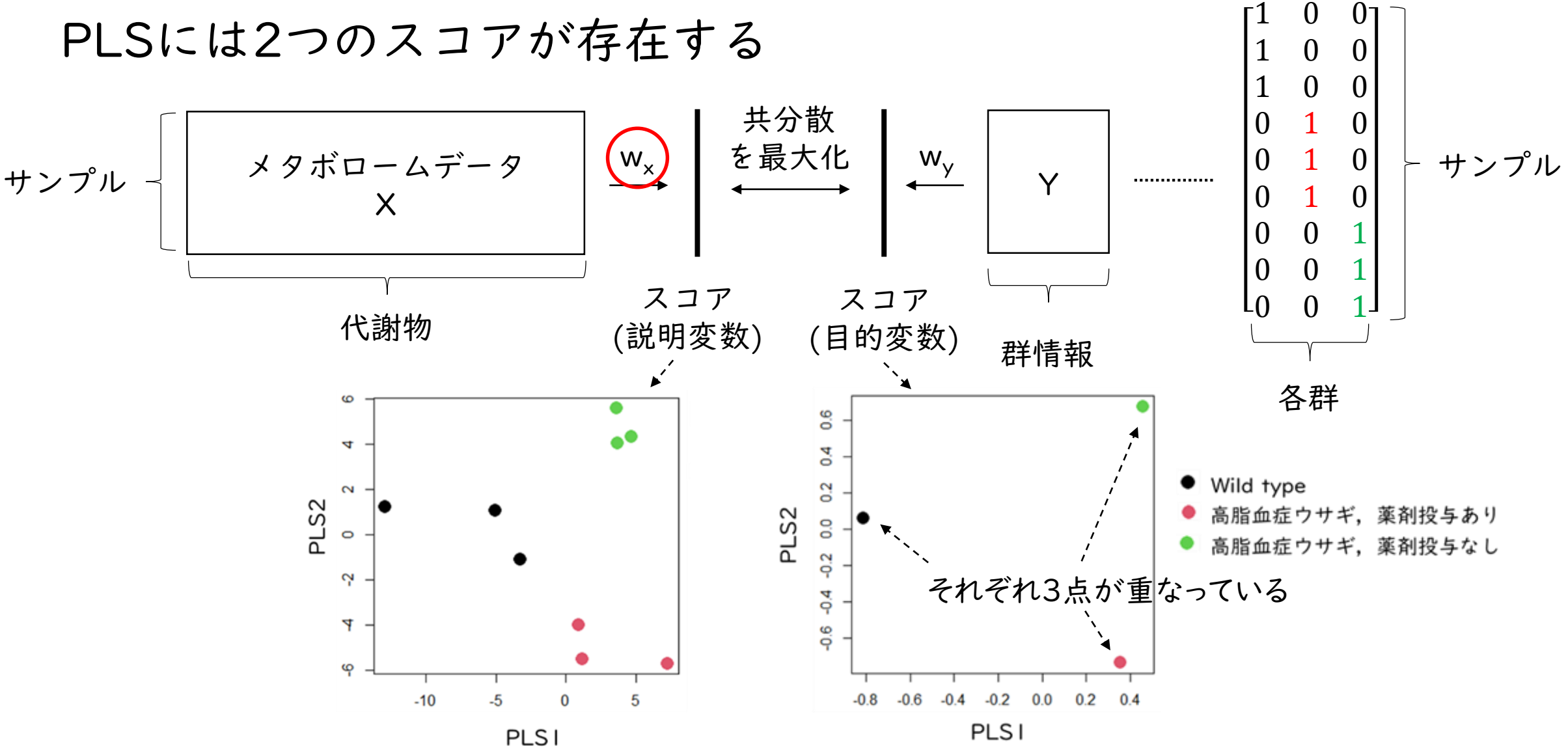
重み(PLS)係数を用いて変数を選ぶ

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



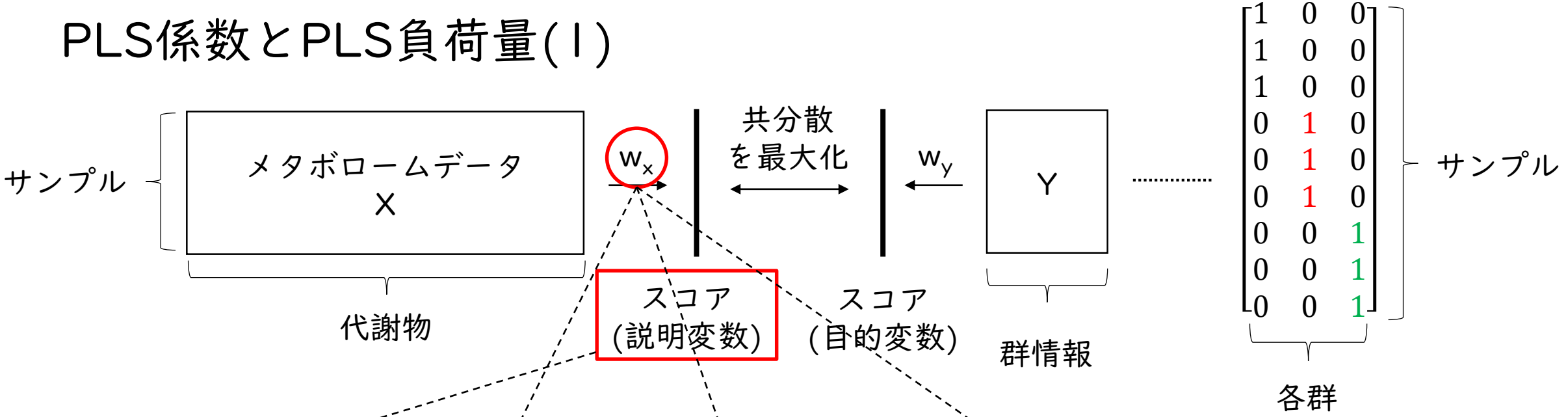
- PLS係数による変数の選び方
  - PLS係数**w**は、各代謝物に対する重要度を示す重みであり、PLS係数**w**が大きいものが重要な代謝物となる
  - PLS負荷量を定義し、統計的な基準で重要な代謝物を選ぶ

PLSには2つのスコアが存在する



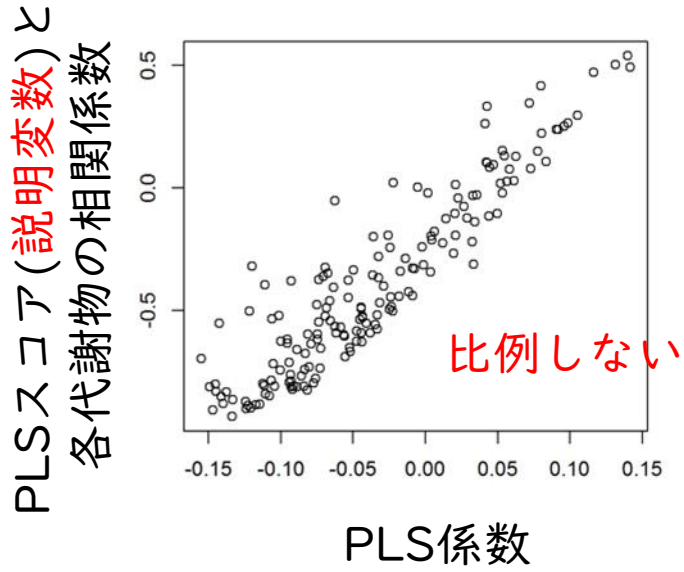
説明変数のスコア、目的変数のスコアいずれも同様の位置(左、右上、右下)に
配置されており、傾向が一致していることが確認できる。

PLS係数とPLS負荷量(1)

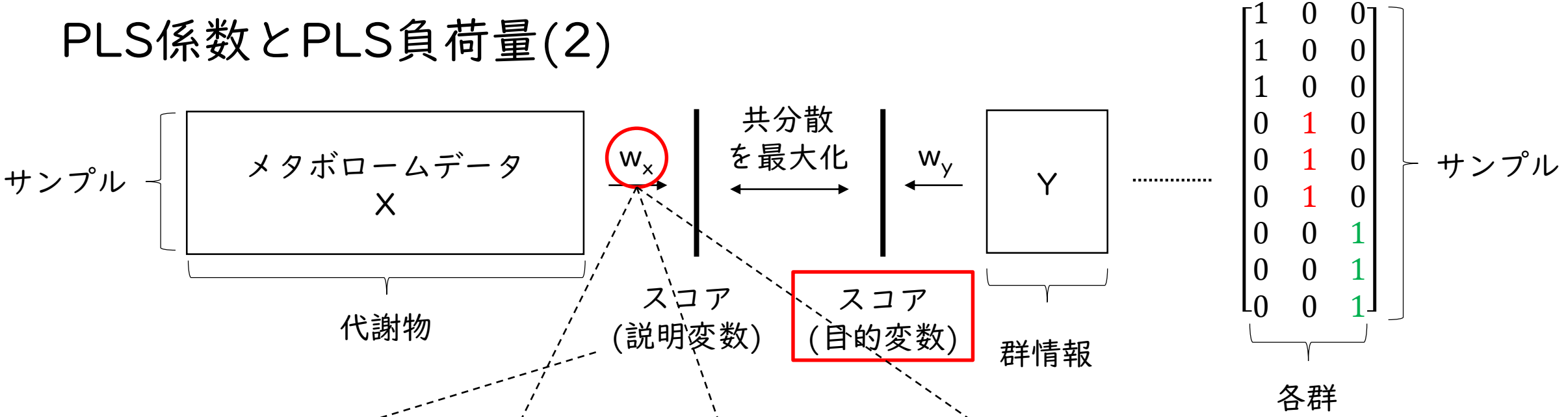


PLSスコア=(代謝物1)×w_1+(代謝物2)×w_2+・・・+(代謝物P)×w_p

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

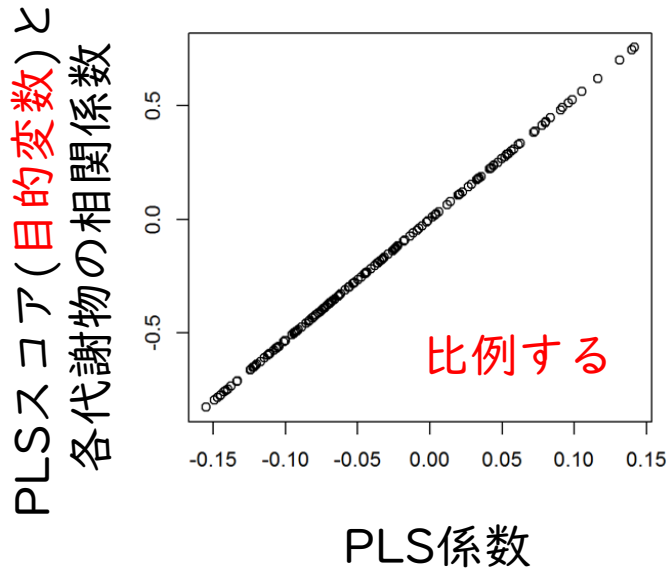


PLS係数とPLS負荷量(2)



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

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



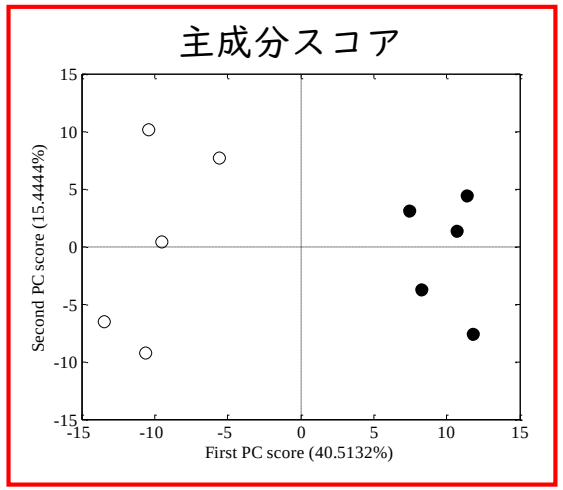
主成分係数、主成分負荷量、PLS係数、PLS負荷量

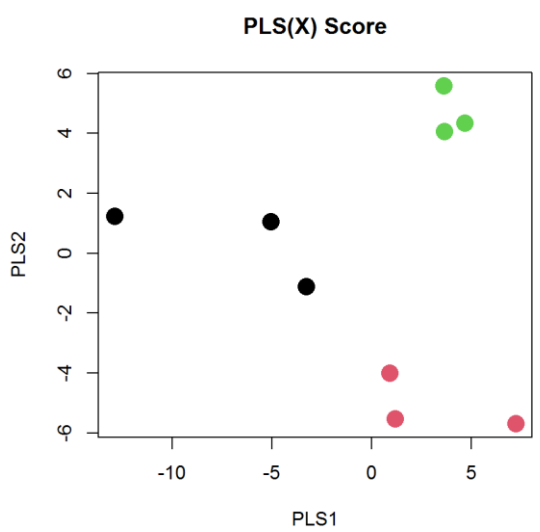
- 主成分分析の場合
  - 主成分係数は 「主成分スコアと各代謝物レベルの相関係数」に比例する

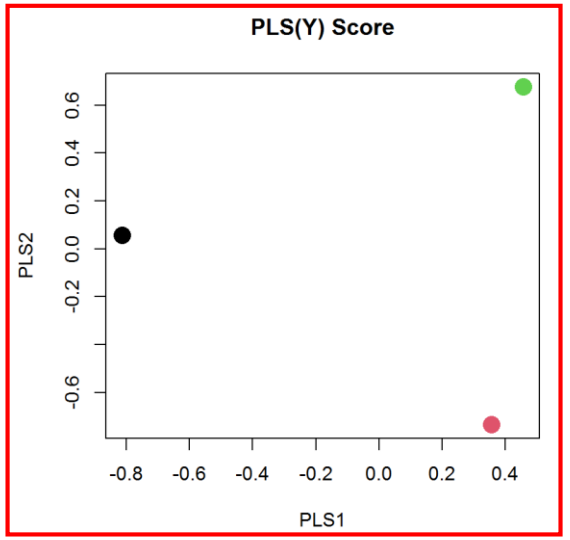
  → 主成分負荷量は「主成分スコアと各代謝物レベルの相関係数」と定義する

- PLSの場合
  - PLS係数は「 PLSスコア(説明変数)と各代謝物レベルの相関係数」に比例しない
  - PLS係数は「 PLSスコア(目的変数)と各代謝物レベルの相関係数」に比例する

  → PLS負荷量は「 PLSスコア(目的変数)と各代謝物レベルの相関係数」と定義する







loadingsパッケージを用いた主成分分析、PLS

- パッケージのインストール
  - install.packages(“loadings”)
- ライブラリの読み込み
  - library(loadings)

- 主成分分析
  - pca <- prcomp(X)
  - pca <- pca_loadings(pca)
  - pca$loading$R
  - pca$loading$p.value

(prcomp関数はstatsのものを利用)

- PLS
  - pls <- pls_svd(X,Y)
  - pls <- pls_loadings(pls)
  - pls$loading$R
  - pls$loading$p.value

(pls_svd関数の部分は、chemometrics
パッケージのpls_eigen関数も利用可能