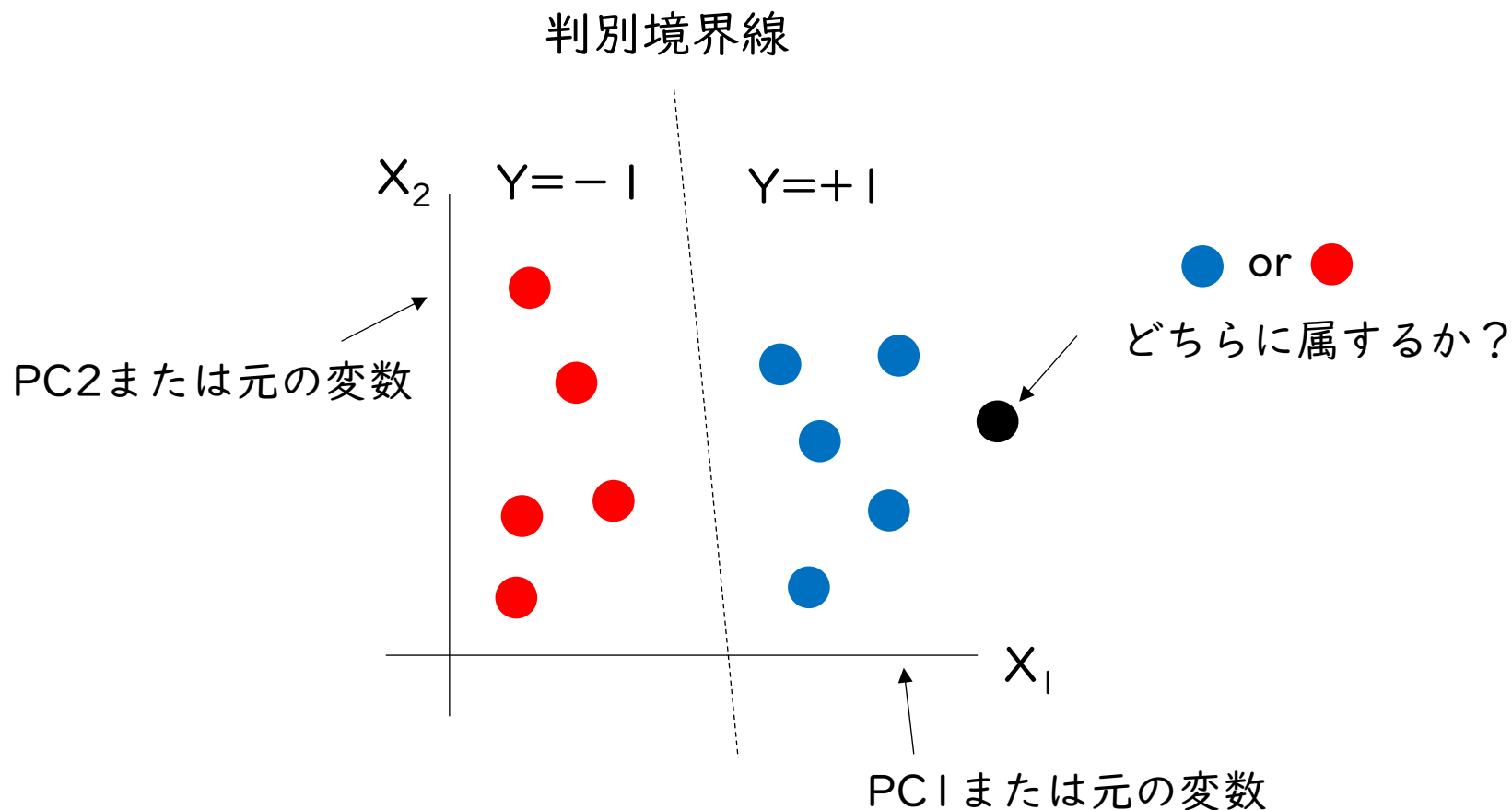


# 判別分析

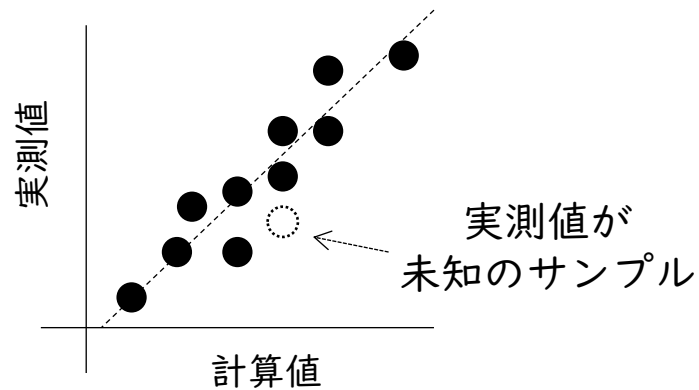
# (線形)判別分析のイメージ



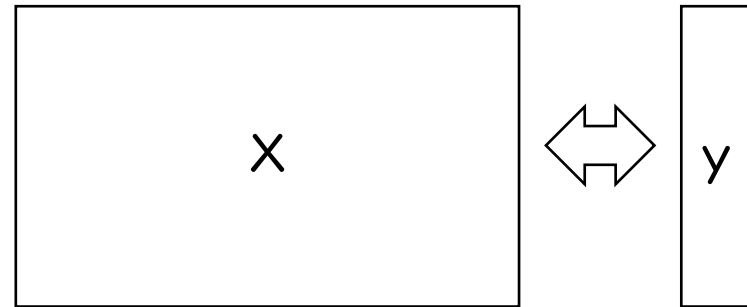
判別関数  $f(X_1, X_2) \geq 0 \longrightarrow \bullet$   $f(X_1, X_2) < 0 \longrightarrow \bullet$

# 回帰・判別モデルの共通点

- PLS回帰：回帰モデル

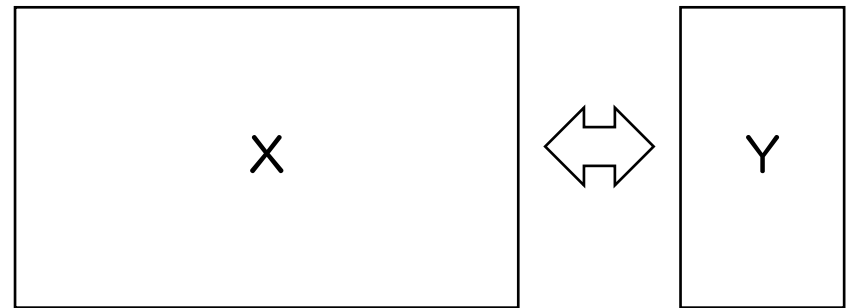
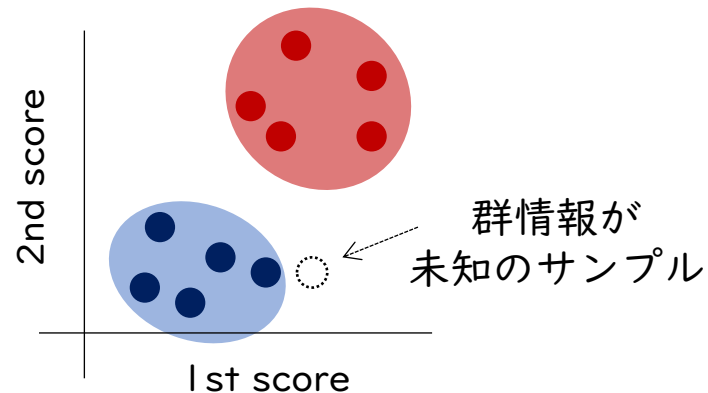


判別モデルも回帰モデルでも、  
PLSの理論としてはほとんど同じ



連続値のベクトル

- PLS-DA：判別モデル

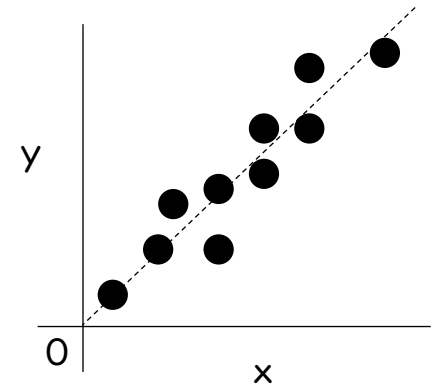


群の情報を表す  
0と1の行列

# [再掲] 回帰分析のモデル

- 単回帰分析

$$\begin{array}{ccccc} \text{目的変数} & & \text{1変数} & & \text{誤差} \\ \boxed{y} & = & \begin{array}{c} n \\ \text{(サンプル)} \end{array} \left\{ \boxed{x} \right\} \times b & + & \boxed{e} \end{array}$$



(原点を通るとき)

- 重回帰分析(2変数の場合)

$$\begin{array}{ccccccc} \text{目的変数} & & \text{変数1} & & \text{変数2} & & \text{誤差} \\ \boxed{y} & = & \begin{array}{c} n \\ \text{(サンプル)} \end{array} \left\{ \boxed{x_1} \right\} \times b_1 & + & \boxed{x_2} \times b_2 & + & \boxed{e} \end{array}$$

# 回帰分析と判別分析の違い

- 回帰分析(2変数の場合)

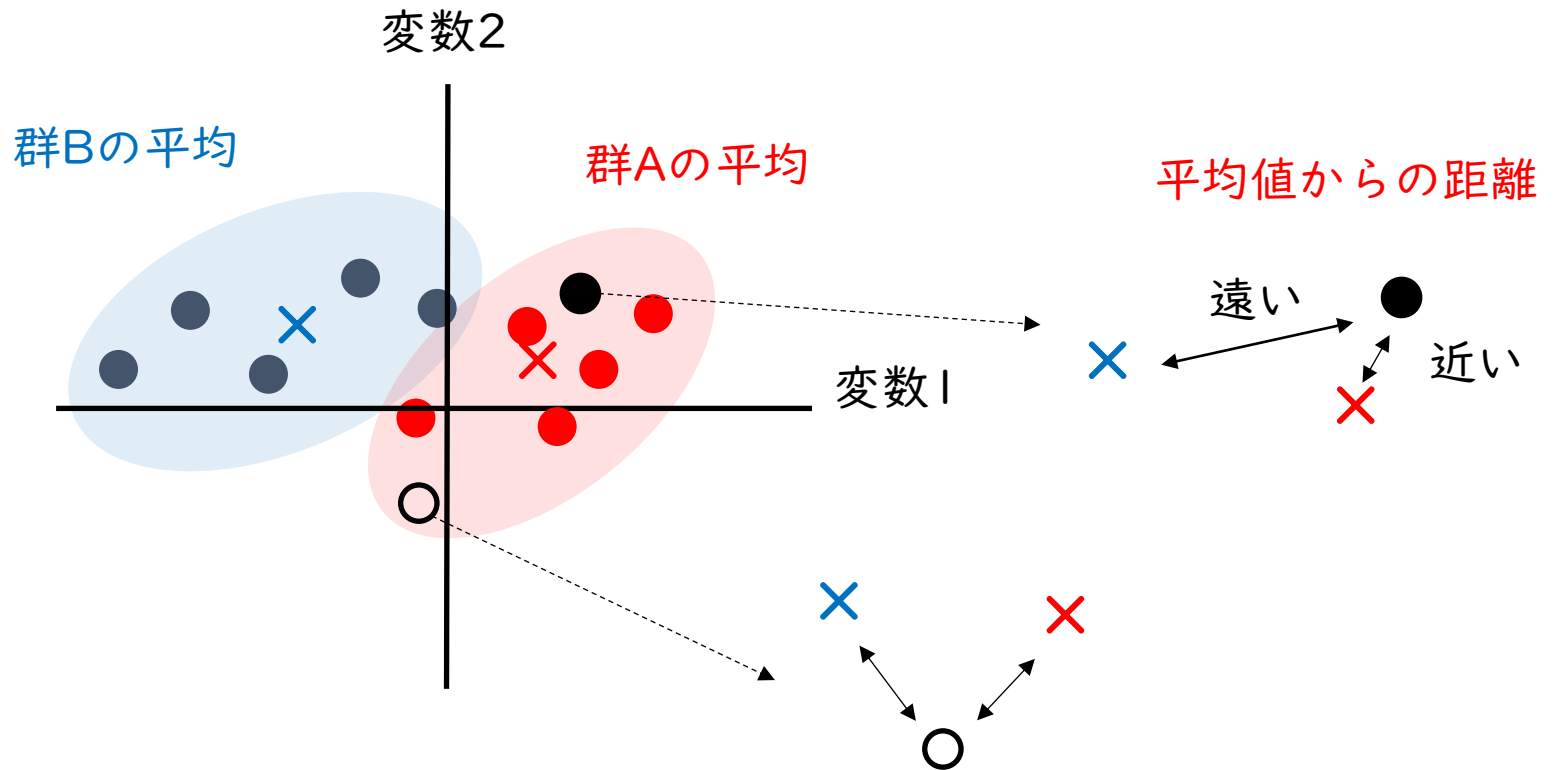
$$\begin{array}{ccccccc}
 \text{目的変数} & & & \text{変数1} & & \text{変数2} & \text{誤差} \\
 \boxed{y} & = & \begin{array}{c} n \\ \text{(サンプル)} \end{array} \left\{ \begin{array}{c} \boxed{x_1} \\ \boxed{x_2} \end{array} \right. & \times b_1 & + & \boxed{x_2} \times b_2 & + \boxed{e}
 \end{array}$$

- 判別分析(2変数の場合)

$$\begin{array}{ccccccc}
 \text{目的変数} & & & \text{変数1} & & \text{変数2} & \text{誤差} \\
 \boxed{y} & = & \begin{array}{c} n \\ \text{(サンプル)} \end{array} \left\{ \begin{array}{c} \boxed{x_1} \\ \boxed{x_2} \end{array} \right. & \times b_1 & + & \boxed{x_2} \times b_2 & + \boxed{e}
 \end{array}$$

$$\begin{array}{l}
 \text{サンプル1~3} \\
 \text{群A} \\
 \left\{ \begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right\} \\
 \left\{ \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right\} \\
 \left\{ \begin{array}{c} 0 \\ 1 \\ 1 \end{array} \right\} \\
 \left\{ \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right\}
 \end{array}
 \begin{array}{l}
 \text{サンプル4~6} \\
 \text{群B}
 \end{array}$$

# 判別分析のイメージ(2群の場合)

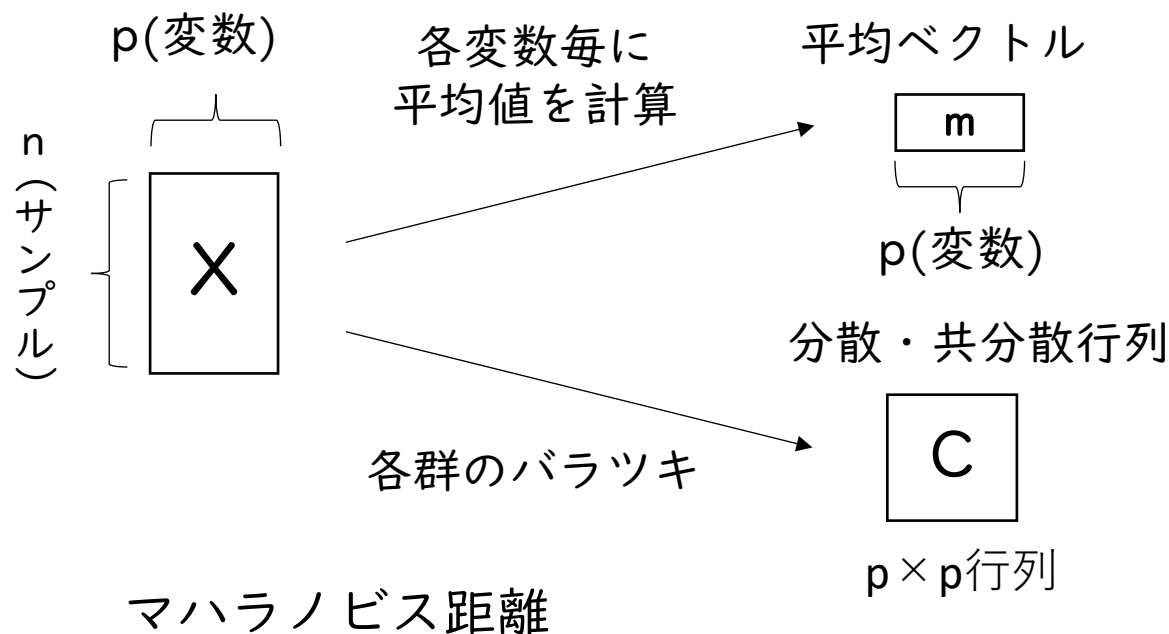


平均値からの距離は同程度だが、  
赤に近いように見える

各群のバラツキの様子を見ると、○は●に近いと判断できる

# 判別分析(2群の場合)

群A、群Bのそれぞれについて  
平均ベクトルと分散・共分散行列を計算



未知のサンプル

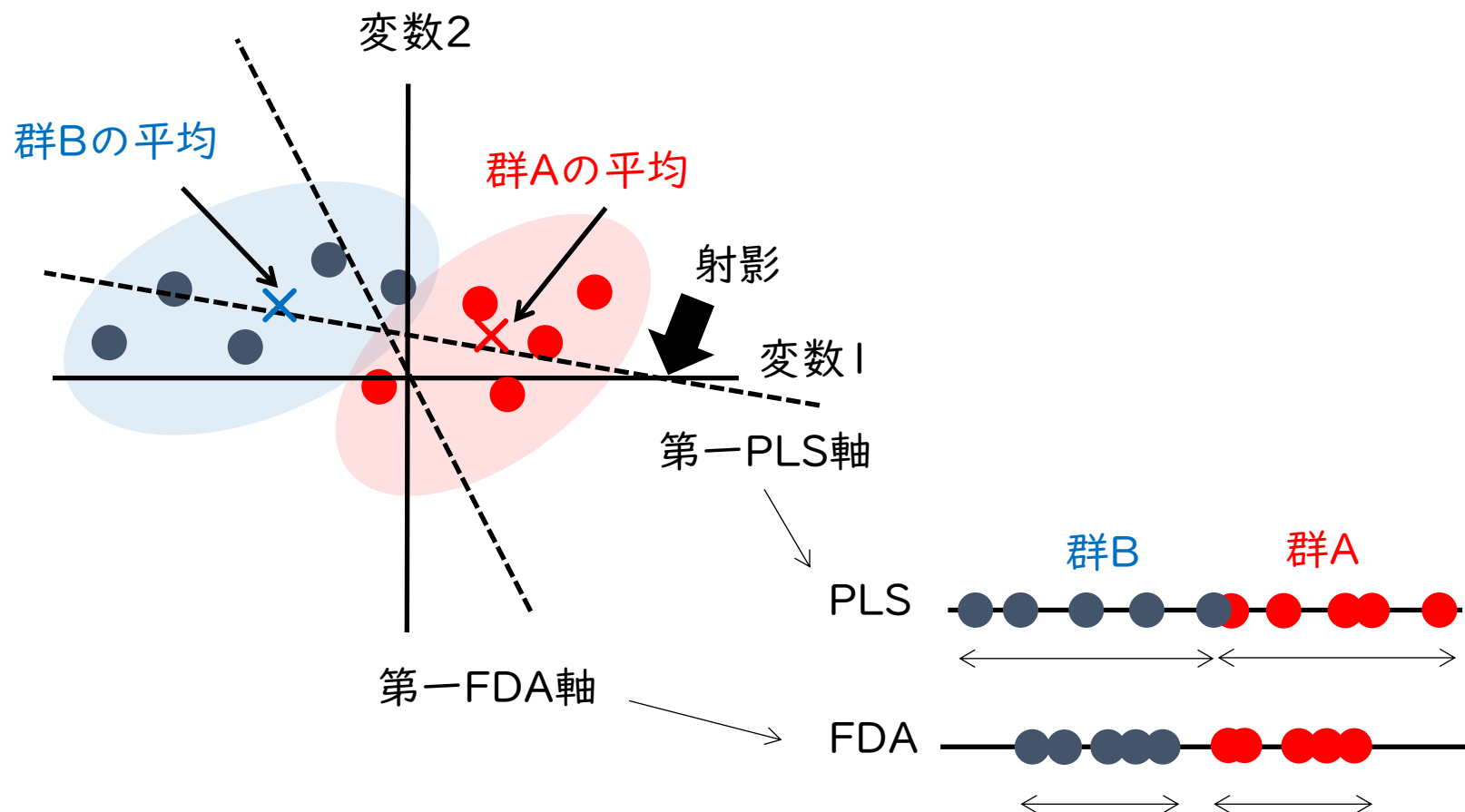
$x$

$$\text{群A } \sqrt{(\underline{x} - \underline{m}_1)^t \mathbf{C}_1^{-1} (\underline{x} - \underline{m}_1)}$$

$$\text{群B } \sqrt{(\underline{x} - \underline{m}_2)^t \mathbf{C}_2^{-1} (\underline{x} - \underline{m}_2)}$$

距離が近い方の  
群に割り当てる

# フィッシャーの線形判別分析(FDA)



データに依存するが、  
FDAの方が群内のバラツキが小さい



# FDAは、スコアの相関係数を最大化する方法

- FDA

$$\mathbf{t} = \mathbf{X}\mathbf{w}_x \quad \mathbf{s} = \mathbf{Y}\mathbf{w}_y$$

$$corr(\mathbf{t}, \mathbf{s}) = \frac{cov(\mathbf{t}, \mathbf{s})}{\sqrt{var(\mathbf{t})}\sqrt{var(\mathbf{s})}} \text{ を最大化}$$

(正準相関分析の定式化)

スコア $\mathbf{t}$ と $\mathbf{s}$ の共分散が大きい  
かつスコア $\mathbf{t}$ または $\mathbf{s}$ の分散が小さい  
→ 群内のバラツキが小さいのと同じ

3群 N=3のとき

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

- PLS

$$cov(\mathbf{t}, \mathbf{s}) \text{ を最大化}$$

スコア $\mathbf{t}$ と $\mathbf{s}$ の共分散が大きい

FDAとPLSはいずれも群間差が大きく、  
FDAはさらに群内のバラツキを小さくする方向が得られる

# 正則化フィッシャー線形判別分析

- FDA

$$\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x = \lambda \mathbf{X}'\mathbf{X}\mathbf{w}_x$$

$$\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{w}_y = \lambda \mathbf{Y}'\mathbf{Y}\mathbf{w}_y$$

$p > n$ だと逆行列の計算が出来ない

- 正則化FDA

$$\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x = \lambda \{(1-k)\mathbf{X}'\mathbf{X} + k\mathbf{I}\}\mathbf{w}_x$$

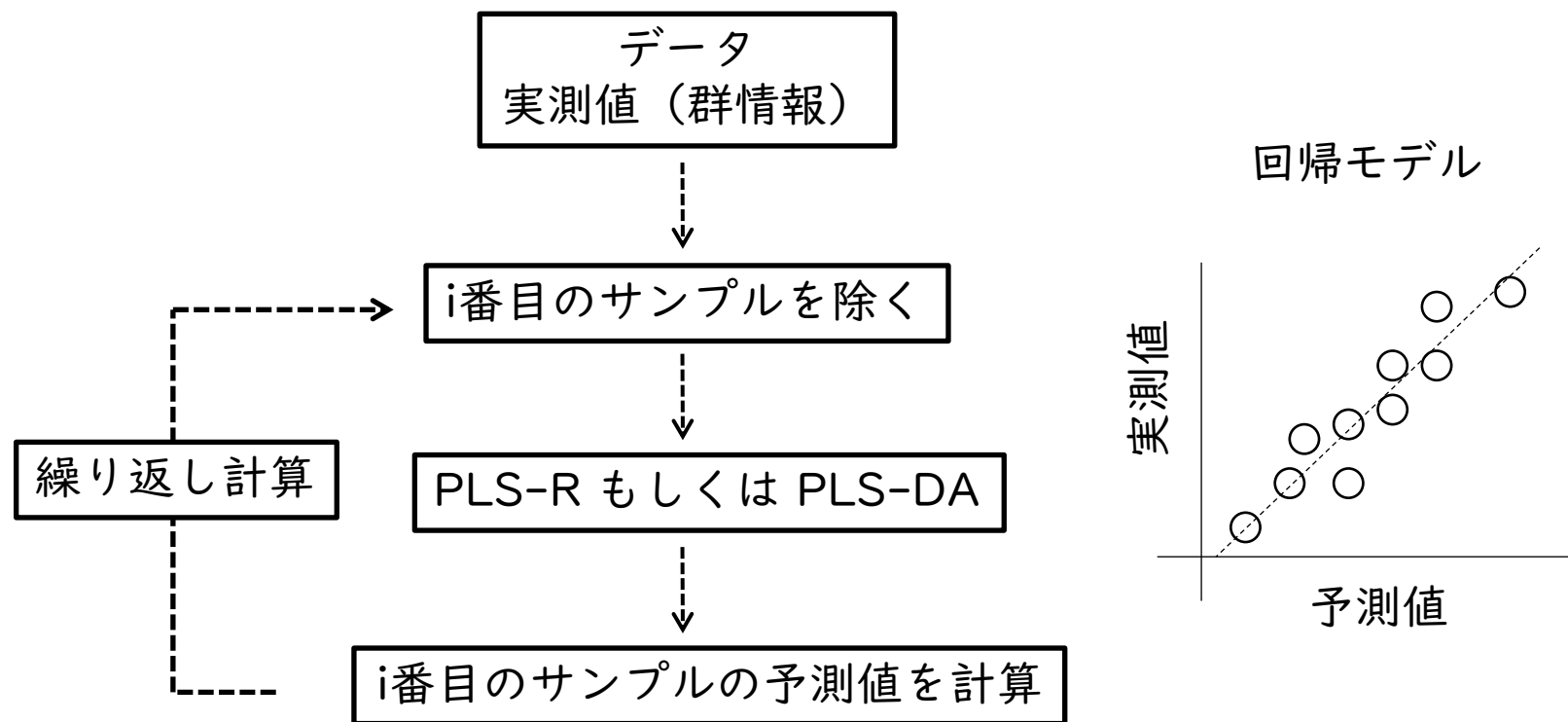
$$\mathbf{Y}'\mathbf{X}\{(1-k)\mathbf{X}'\mathbf{X} + k\mathbf{I}\}^{-1}\mathbf{X}'\mathbf{Y}\mathbf{w}_y = \lambda \mathbf{Y}'\mathbf{Y}\mathbf{w}_y$$

$p > n$ でも計算可能

重回帰分析とリッジ回帰の関係と同じように、  
逆行列の計算に単位行列を導入することで計算出来る

# [再掲] 判別モデルの性能評価 (回帰と同じ)

例. Leave-one-out cross validation (LOOCV)の場合



正しく性能評価を行うためには、さらに独立テストが必要

# 解析に用いるデータ

The screenshot shows the MetaboLights website interface. At the top, there is a navigation bar with links for EMBL-EBI, About us, Training, Research, and Services. The main header features the MetaboLights logo and a search bar with the text "Examples: Alanine, Homo sapiens, Urine, MTBLS1". Below the header is a menu bar with links for Home, Browse Studies, Browse Compounds, Browse Species, Download, Help, Give us feedback, About, Submit Study, and Login. The main content area displays the study MTBLS90: Large-scale non-targeted serum metabolomics in the Prospective Investigation of the Vasculature in Uppsala Seniors. It includes the status "Public", the release date "2014-08-30", and the authors "Andrea Ganna, Samira Salihovic, Erik Ingelsson, Lars Lind". A description of the study is provided, followed by a section titled "PUBLICATIONS" which lists two related publications with their titles, authors, and a Creative Commons license icon.

EMBL-EBI About us Training Research Services EMBL-EBI Hinxton

MetaboLights Search

Examples: Alanine, Homo sapiens, Urine, MTBLS1

Home Browse Studies Browse Compounds Browse Species Download Help Give us feedback About Submit Study Login

Status Public Release Date 2014-08-30

**MTBLS90: Large-scale non-targeted serum metabolomics in the Prospective Investigation of the Vasculature in Uppsala Seniors**

Andrea Ganna, Samira Salihovic, Erik Ingelsson, Lars Lind

The Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) is a community-based study where all men and women at age 70 living in Uppsala, Sweden were invited to participate in 2001. The 1,016 participants (50% women) have been extensively phenotyped. In March 2006 a reinvestigation of the cohort at the age of 75 was started. The major measurements performed at age 70 were also repeated at age 75. Here we make publically available all the metabolomics analysis on serum samples from the investigation at age 70. In the final analysis we included samples from 968 individuals.

PUBLICATIONS

**Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident Coronary Heart ...**

Ganna A, Salihovic S, Sundström J, Broeckling CD, Hedman AK...

**A workflow for UPLC-MS non-targeted metabolomic profiling in large human population-based ...**

Andrea Ganna, Tove Fall, Woojoo Lee, Corey D Broeckling, Jit...

968サンプル(女性:483名 男性:485名)  
189物質のメタボロームデータ

# PLS判別分析(Ⅰ)

- データの読み込み

```
X0 <- read.csv(file="C:/R/MTBLS90_train.csv")
```

- データの準備

```
X <- as.matrix(X0[,-c(1,2)]) # メタボロームデータ
```

```
Y0 <- X0[,2] # 群情報(性別)
```

```
Y <- model.matrix(~ Y0 + 0) # 目的変数
```

- PLS判別分析

```
library(chemometrics)
```

```
plsda <- cppls(Y~X, 30, data=data.frame(X=X, Y=Y),  
               scale=TRUE)
```

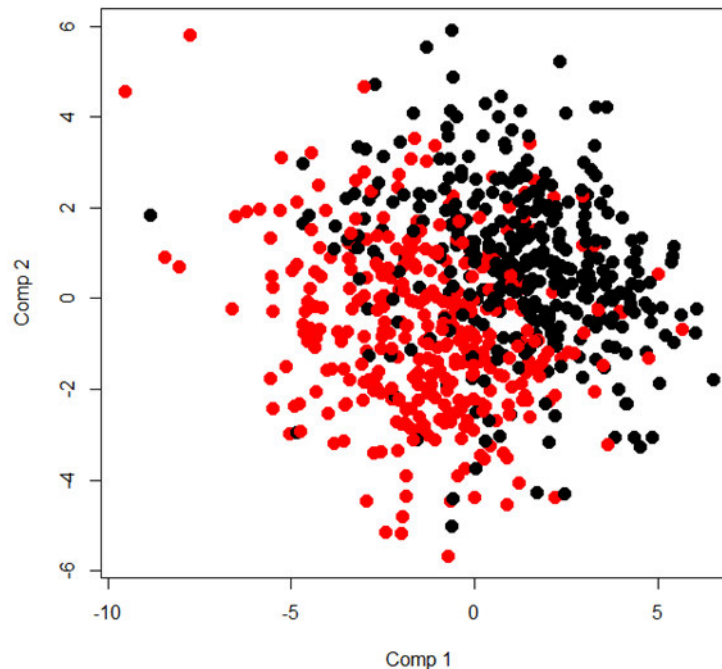
# PLS判別分析(2)

- PLS判別分析のスコア

```
col_class <- NULL;
```

```
col_class[Y0=="Male"] <- 1; col_class[Y0=="Female"] <- 2
```

```
plot(plsda$scores, col=col_class, pch=16, cex=1.5)
```



# PLS判別分析の比較

- chemometricsノパッケージ

```
library(chemometrics)
plsda1 <- cppls(Y~X, 2, data=data.frame(X=X, Y=Y),
               scale=TRUE)
```

- plsノパッケージ

```
library(pls)
plsda2 <- pls2_nipals(X, Y, a=2, scale=TRUE)
```

- mixOmicsノパッケージ

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

# PLS判別分析(3)

- クロスバリデーション(10 fold)

```
plsda <- cppls(Y~X, 30, data=data.frame(X=X, Y=Y),  
validation="CV", segment=10, segment.type="consecutive",  
scale=TRUE)
```

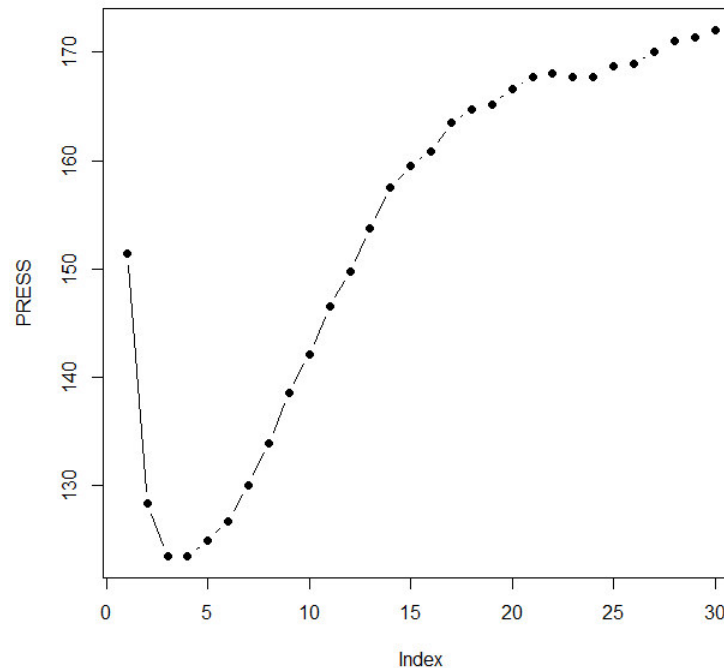
- 潜在変数の数の決定

```
PRESS <- plsda$validation$PRESS[1,]  
plot(PRESS, type="b")
```



# PLS判別分析(4)

- 潜在変数の数の決定



```
lvnum <- which.min(PRESS)  # 3がPRESS最小
```

# PLS判別分析(5)

- テストデータの読み込み

```
Z0 <- read.csv(file="C:/R/MTBLS90_test.csv")
```

- データの準備

```
Z <- as.matrix(Z0[,-c(1,2)]) # メタボロームデータ
```

```
Ytest <- Z0[,2] # テストデータの性別
```

- 予測

```
p0 <- predict(plsda, ncomp=lvnum, newdata=Z)
```

# PLS判別分析(6)

- 予測結果の整理

```
p1 <- max.col(p0[,1])
```

```
p <- NULL;
```

```
p[which(p1==1)] <- "Female"; p[which(p1==2)] <- "male"
```

- 集計結果

```
table(Ytest, p)
```

		予測結果		
		Female	male	合計
正解	Female	98	22	120
	Male	29	92	121

正解率は女性で81.7%、男性で76.0%



# A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification

Kevin M. Mendez<sup>1</sup> · Stacey N. Reinke<sup>1</sup> · David I. Broadhurst<sup>1</sup>

**Table 1** The ten data sets curated for this study

Study ID	Publication	Platform	Type	No. of samples (case/control)	No. of peaks	Case/control
MTBLS90 <sup>a</sup>	Ganna et al. (2014); Ganna et al. (2015)	LC–MS	Plasma	968 (485/483)	189	Sex (M/F)
MTBLS92 <sup>a</sup>	Hilvo et al. (2014)	LC–MS	Plasma	253 (142/111)	138	Breast cancer chemotherapy (before/after)
MTBLS136 <sup>a</sup>	Stevens et al. (2018)	LC–MS	Serum	668 (337/331)	689	Postmenopausal hormone (estrogen/estrogen + progesterone)
MTBLS161 <sup>a</sup>	Armstrong et al. (2015)	NMR	Serum	59 (34/25)	29	Chronic fatigue syndrome (case/control)
MTBLS404 <sup>a</sup>	Thévenot et al. (2015)	LC–MS	Urine	184 (101/83)	120	Sex (M/F)
MTBLS547 <sup>a</sup>	Zheng et al. (2017)	LC–MS	Caecal	97 (46/51)	42	High fat diet (case/control)
ST000369*	Fahrman et al. (2015)	GC–MS	Serum	80 (49/31)	181	Adenocarcinoma (case/control)
ST000496*	Sakanaka et al. (2017)	GC–MS	Saliva	100 (50/50)	69	Debridement (pre/post)
ST001000*	Franzosa et al. (2019)	LC–MS	Stool	121 (68/53)	747	Inflammatory bowel diseases (Crohn's disease/ulcerative colitis)
ST001047*	Chan et al. (2016)	NMR	Urine	83 (43/40)	149	Gastric cancer (gastric cancer/healthy)

\*Indicates data sourced from Metabolomics Workbench (<https://www.metabolomicsworkbench.org>)

<sup>a</sup>Indicates data sourced from Metabolights (<https://www.ebi.ac.uk/metabolights/>)

# 機械学習の計算の流れ

1. 訓練データとテストデータに分割する
2. 訓練データでモデル作成
3. クロスバリデーション
4. テストデータで予測

# 訓練データとテストデータに分割する

- パッケージの読み込み

```
library(caret)
```

- csvファイルの読み込み

```
X0 <- read.csv(file="C:/R/data/MTBLS90.csv")
```

- データの準備

```
X <- as.matrix(X0[, -c(1:4)])
```

```
y <- X0[, 4]
```

- データの分割(3/4を訓練データ、1/4をテストデータ)

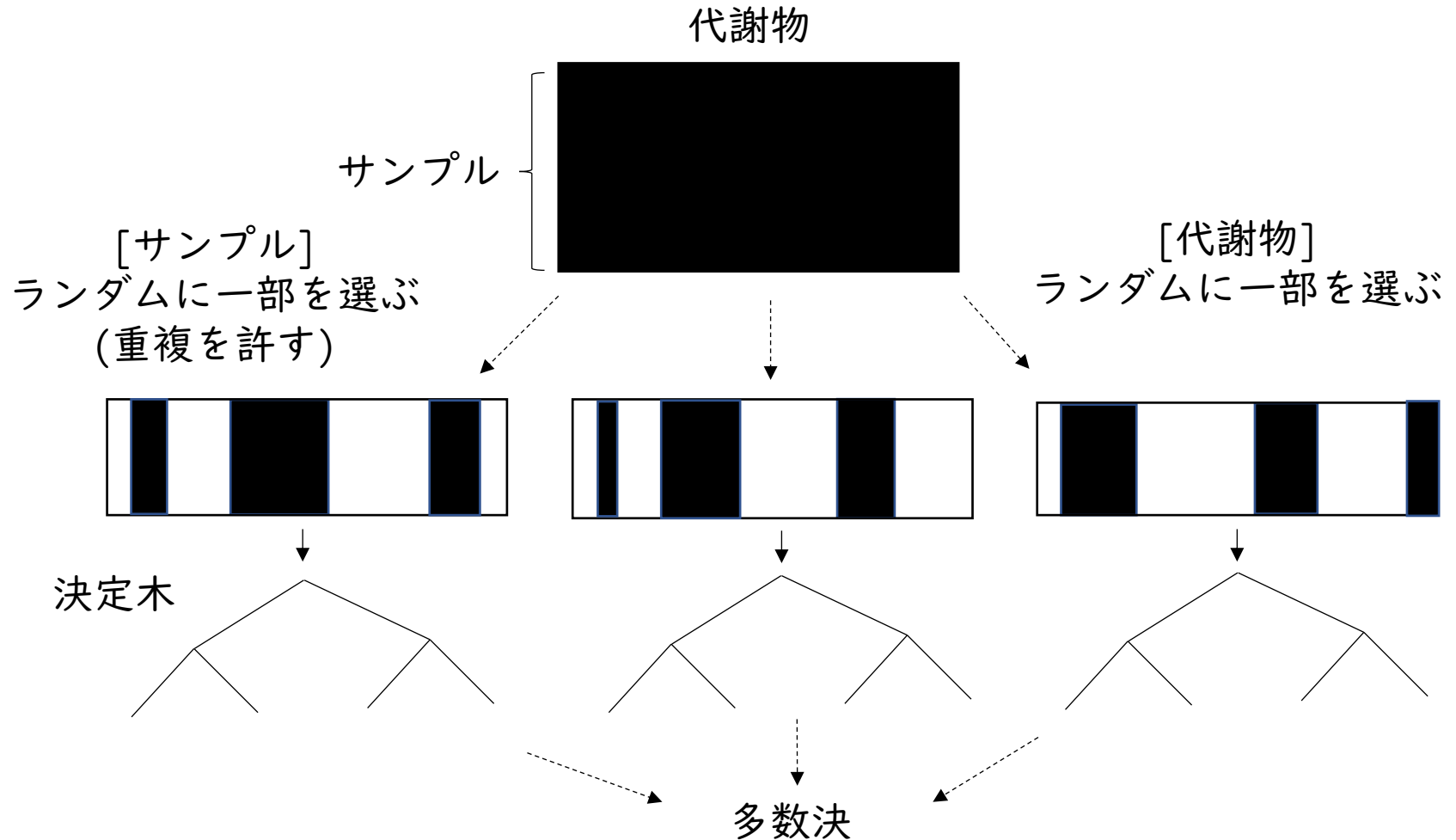
```
trainIndex <- createDataPartition(y, p = .75, list=FALSE)
```

```
train_data <- data.frame(X[trainIndex,],
```

```
                        y=factor(y[trainIndex]))
```

```
test_data <- data.frame(X[-trainIndex,], y=factor(y[-trainIndex]))
```

# ランダムフォレストの概要



- caretパッケージでのランダムフォレストの計算

```
cv_rf <- train(… (省略), method = "rf", … (省略))
```

# ランダムフォレスト

- モデル作成(ランダムフォレスト)

```
set.seed(1)
cv_rf <- train(
  target ~ .,
  data = data.frame(train_data[, -ncol(train_data)],
                    target=train_data$y),
  method = "rf",
  trControl = trainControl(method = "cv", number=10,
                           savePredictions = TRUE, classProbs=T)
)
```



# クロスバリデーションで性能評価

- クロスバリデーション

```
library(ROCR) # ROC curve
```

```
pred_cv <- cv_rf$pred
```

```
index_cv <- which(pred_cv$mtry == cv_rf$bestTune$mtry)
```

```
pred <- prediction(pred_cv[index_cv,]$Male, pred_cv[index_cv,]$obs)
```

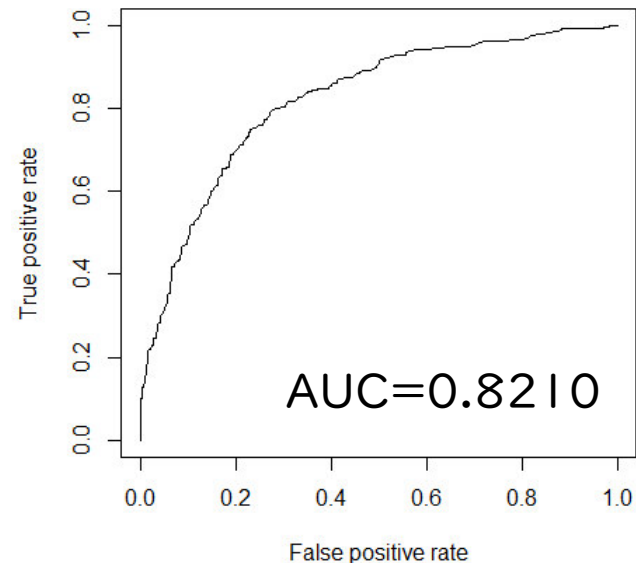
```
perf_cv <- performance(pred, "tpr", "fpr")
```

```
plot(perf_cv)
```

- AUC

```
auc.tmp <- performance(pred, "auc")
```

```
auc_cv <- auc.tmp@y.values[[1]]
```



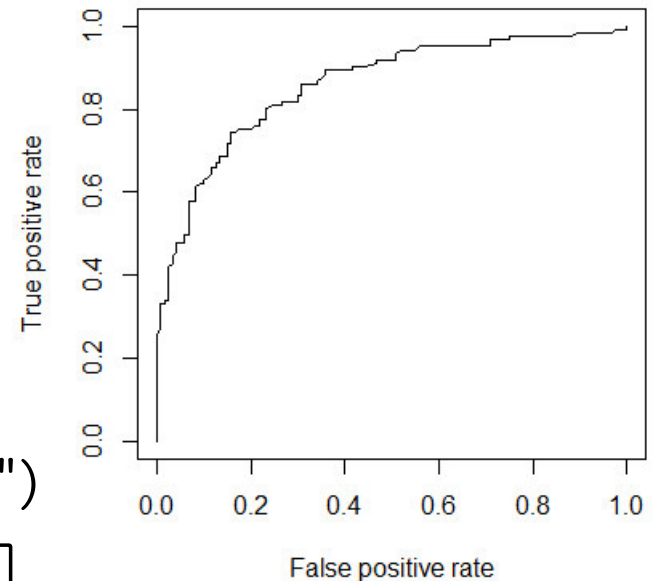
# テストデータに対する予測

- ROC曲線

```
p <- predict(cv_rf, newdata=data.frame(
  test_data[,-ncol(test_data)]), type="prob")
pred <- prediction(p[,2], test_data$y)
perf_test <- performance(pred, "tpr", "fpr")
plot(perf_test)
```

- AUC

```
auc.tmp <- performance(pred,"auc")
auc_test <- auc.tmp@y.values[[1]]
```



AUC=0.8579

# PLS判別分析 caretの場合

- クロスバリデーション

```
cv_plsda <- train(  
  target ~ .,  
  data = data.frame(  
    train_data[, -ncol(train_data)],  
    target=train_data$y),
```

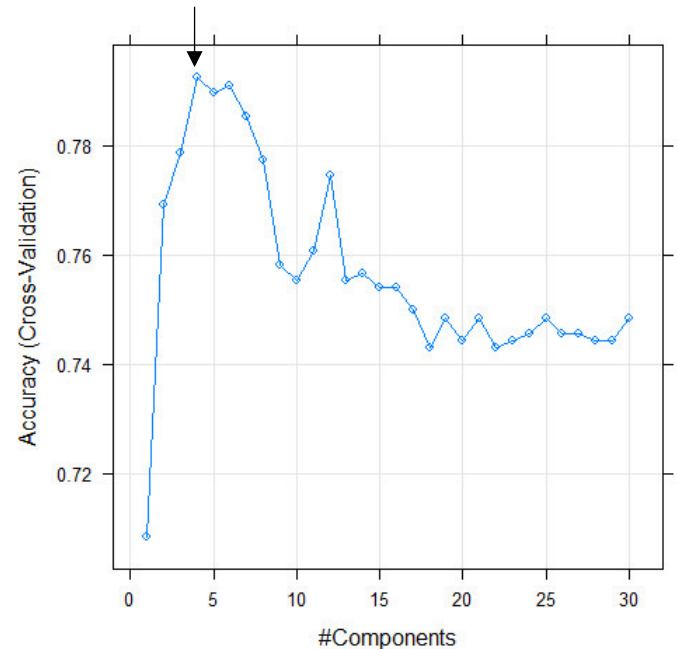
```
  method = "pls",
```

```
  trControl = trainControl(  
    method = "cv",  
    number=10,  
    savePredictions = TRUE,  
    classProbs=T),
```

```
  tuneLength = 30
```

```
)
```

4が最適



plot(cv\_plsda)

# Boostingアルゴリズムの概要

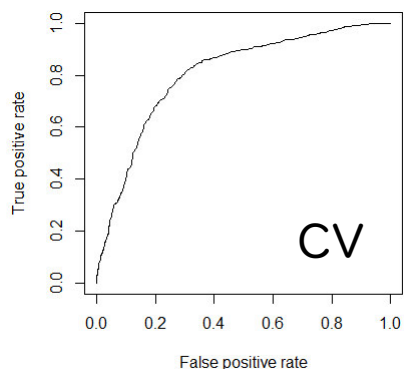
- 各サンプルについてデータと重みをセットで考える
- 以下の手順を繰り返す
  - 単純な方法(弱学習)で学習を行う
  - 性能評価を行って、重みを変更する
- それぞれに予測を行い、多数決で結果を統合する
- Gradient Boosting(勾配ブースティング)  
`cv_gbm <- train(... (省略), method = "gbm", ... (省略))`

# Elastic net (Lasso)の概要

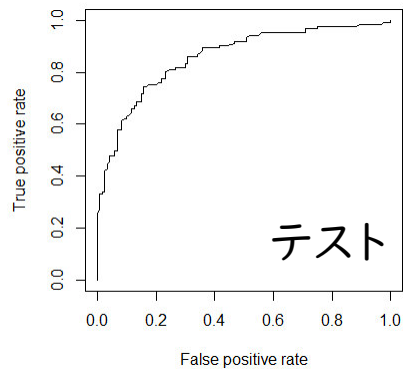
- 変数の数がサンプルの数よりも多い場合、回帰や判別が計算が出来ない(先述のリッジ回帰、FDAなど)
  - リッジ回帰では重み $w$ のL2ノルム正則化  $||w||$
  - Lassoは $w$ のL1ノルムの正則化  $|w|$
  - Elastic netは $w$ のL1とL2ノルム両方の正則化  
 $||w|| + |w|$
- Elastic net (Lasso)  
`cv_glmnet <- train(... (省略), method = "glmnet", ... (省略))`

# 予測結果の比較

## ランダムフォレスト

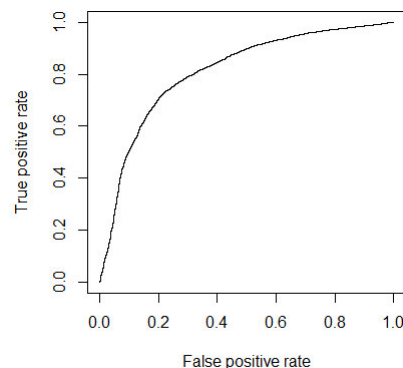


AUC: 0.8210

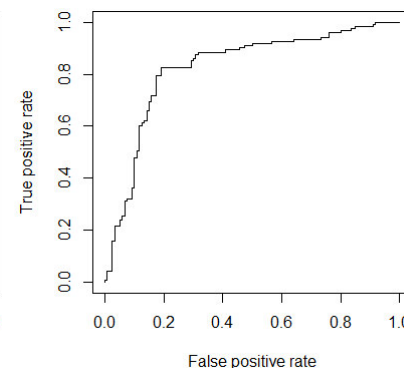


AUC: 0.8579

## PLS-DA

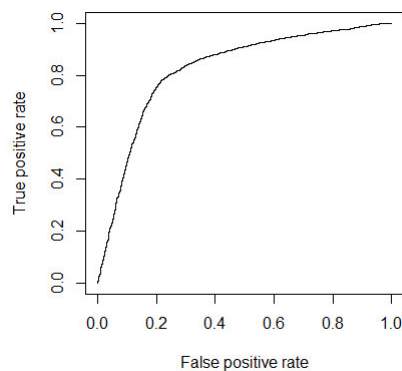


AUC: 0.8111

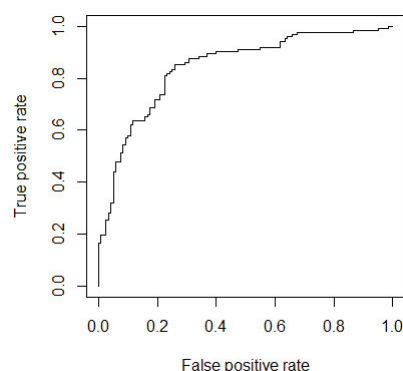


AUC: 0.8256

## Gradient Boosting

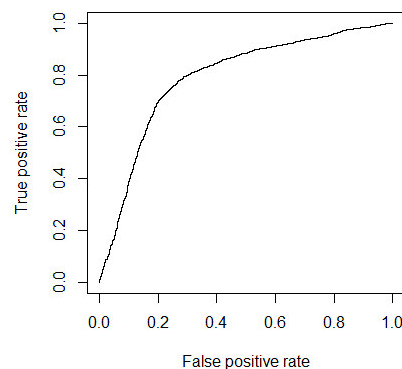


AUC: 0.8222

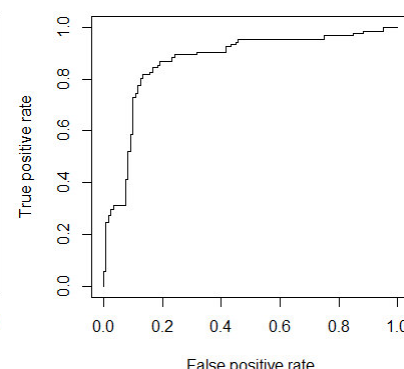


AUC: 0.8426

## Elastic net



AUC: 0.7913



AUC: 0.8683

どの方法でも同程度の予測精度が得られており、PLS-DAとも大差はない

# 各手法での重要な変数の比較

- 重要な変数を選ぶ

varImp(cv\_rf) # ランダムフォレストの場合

Rank	ランダムフォレスト	PLS-DA	Gradient boosting	Elastic net
1	Phosphatidylcholine (28:2)	Creatine	Phosphatidylcholine (28:2)	Creatinine
2	Ceramide phosphoethanolamine (35:2) Sphingomyelin (32:2)	1,3,7-Trimethyluric acid	Creatinine	Ceramide phosphoethanolamine (35:2) Sphingomyelin (32:2)
3	Creatinine	DL-2-Aminooctanoic acid	Creatine	Phosphatidylcholine (28:2)

上位3位以内の物質で比較すると、重複しているものが多い