部分的最小二乗回帰
Partial Least Squares Regression
PLS

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

部分的最小二乗回帰 (PLS) とは？

✓部分的最小二乗回帰
(Partial Least Squares Regression, PLS)

- 線形の回帰分析手法の１つ
- 説明変数(記述子)の数がサンプルの数より多くても計算可能
- 回帰式を作るときにノイズの影響を受けにくい
- 説明変数の間の相関が高くても対応可能
- 主成分分析をしたあとの主成分と目的変数との間で最小二乗法を行うのは主成分回帰 (PCR) であり、PLSとは異なるので注意
- PLS回帰とかPLSRとも呼ばれているが、ここでは PLS

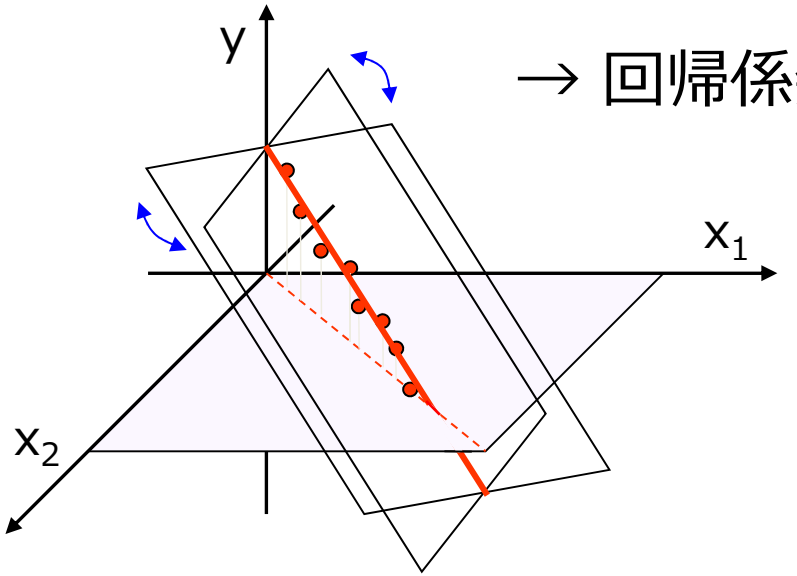
どうして PLS を使うの？～多重共線性～

✓多重共線性の問題

- 説明変数の間に強い相関がある場合、回帰係数が不安定になる
- わずかなデータの変化（追加、削除）で回帰係数が大きく変わってしまう

赤い線を中心に回帰平面が回りやすい

→ 回帰係数が変わりやすい



多重共線性への対策

✓事前に共線性のある変数(記述子)を削除 → 変数選択
- 必要な変数(記述子)を取り除いてしまう危険もある

✓Xを無相関化 (相関係数=0 に) してから重回帰分析

✓Xの情報の一部のみを使用して重回帰分析

主成分分析 (Principal Component Analysis, PCA)
+
重回帰分析

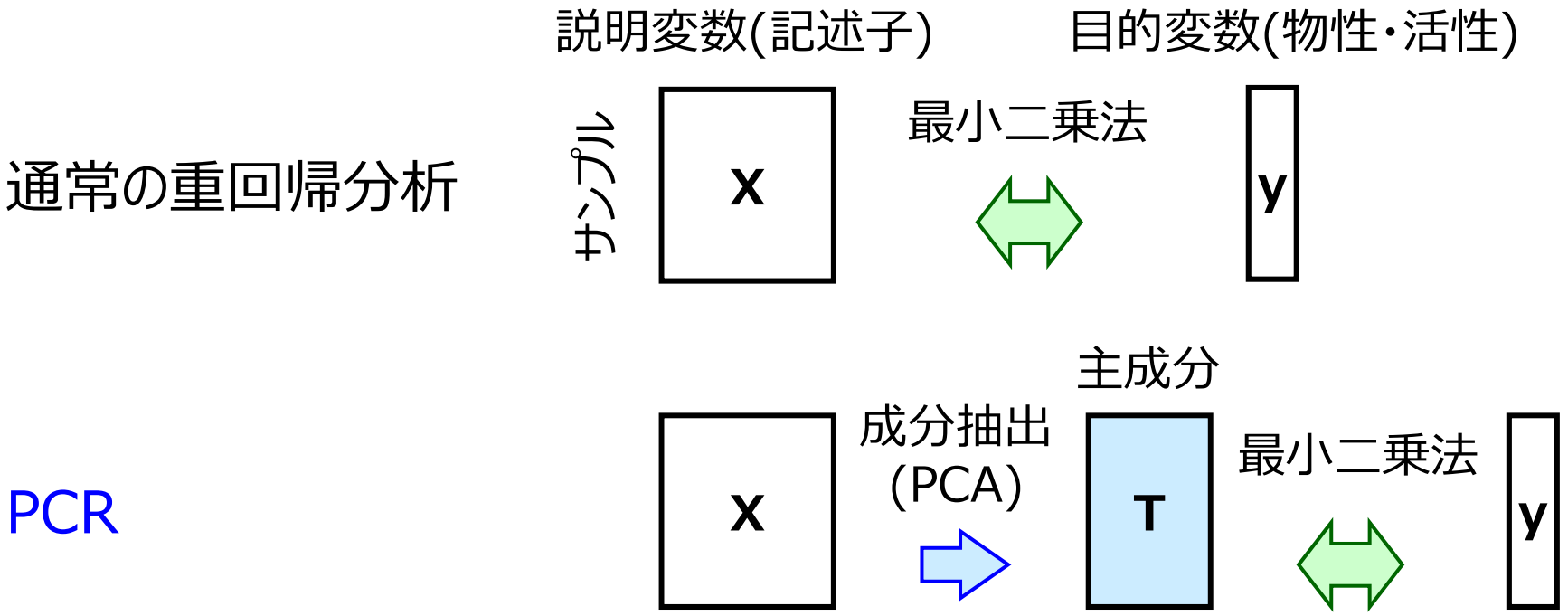
主成分回帰 (Principal Component Regression, PCR)

重回帰分析については こちら、PCAについては こちら

主成分回帰 (PCR)

✓主成分回帰 (Principal Component Regression, PCR)

- 説明変数のデータ **X** のみを用いて主成分分析を行い
  主成分 **T** を得る
  - **T** の成分(変数)の間は無相関
- **T** と目的変数 **y** との間で最小二乗法による重回帰分析

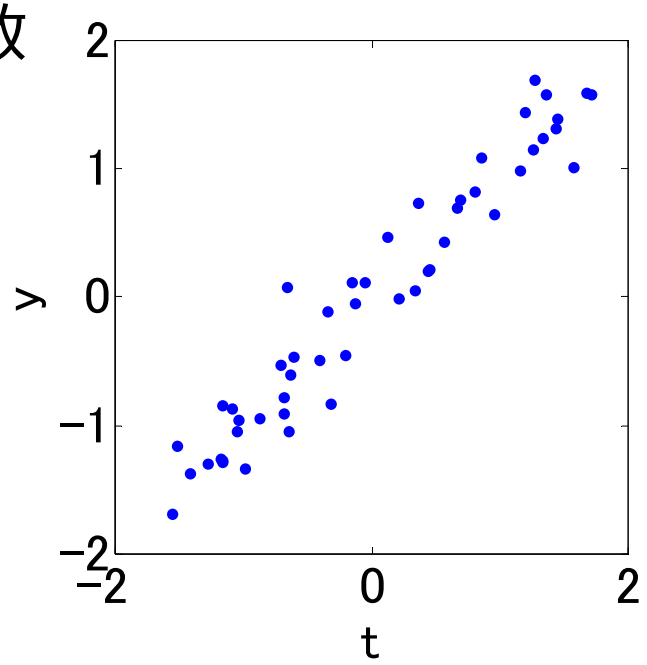


PCR と PLS との違い

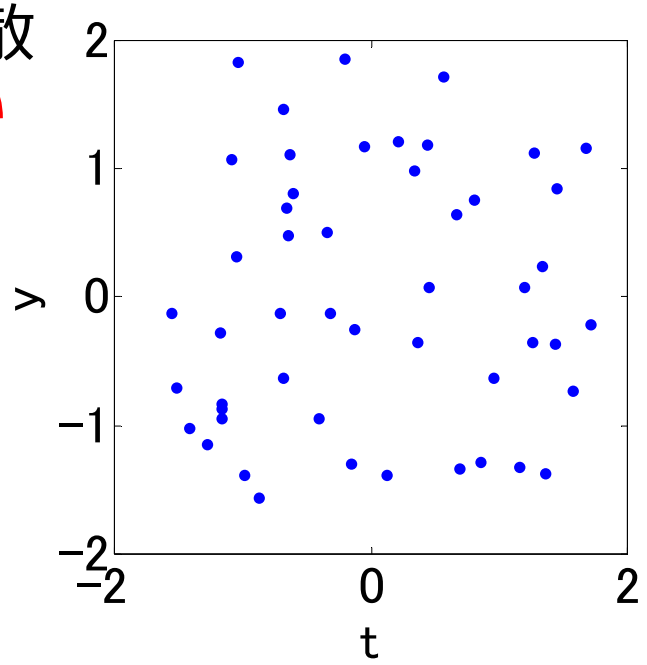
✓PCA
- 主成分 **t** の分散（ $\mathbf{t}^T\mathbf{t}$ ）が最大になるように主成分を抽出

✓PLS
- 主成分 **t** と目的変数 **y** との共分散（ $\mathbf{t}^T\mathbf{y}$ ）が最大になるように主成分を抽出

共分散
大きい



共分散
小さい



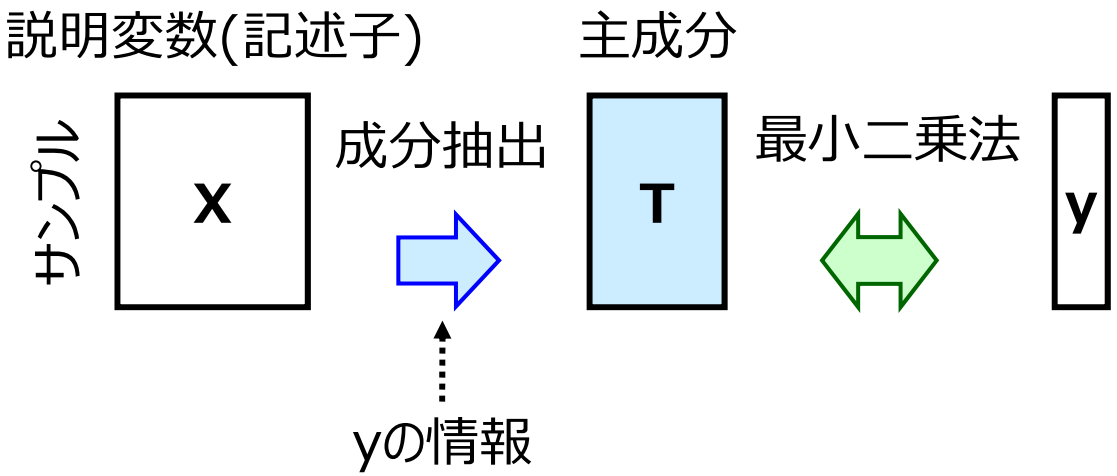
PLS の概要

✓PCA

- 主成分 **t** の分散（ $\mathbf{t}^T\mathbf{t}$ ）が最大になるように主成分を抽出

✓PLS

- 主成分 **t** と目的変数 **y** との共分散（ $\mathbf{t}^T\mathbf{y}$ ）が最大になるように主成分を抽出



PLSの基本式 (yは１変数)

X、**y** はオートスケーリング後 (平均0、標準偏差1)

オートスケーリングについては こちら

$$\mathbf{X} = \sum_{a}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \mathbf{y} = \sum_{a}^{A} \mathbf{t}_a q_a + \mathbf{f} = \mathbf{T}\mathbf{q} + \mathbf{f}$$

- ✓ A : PLS の成分数
- ✓ $\mathbf{t}_a$: a 番目の主成分
- ✓ $\mathbf{p}_a$: a 番目のローディング
- ✓ $\mathbf{E}$: $\mathbf{X}$ の残差
- ✓ q_a : a 番目の係数
- ✓ $\mathbf{f}$: $\mathbf{y}$ の残差

行列の表し方やローディングについては こちら

1成分のPLSモデル

PLSモデル式

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^{\mathrm{T}} + \mathbf{E} \qquad \mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{f}$$

$\mathbf{t}_1$ は $\mathbf{X}$ の線形結合で表わされると仮定

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$$

✓$\mathbf{w}_a$：a番目の重みベクトル
大きさ(ノルム)は1とする

$$\|\mathbf{w}_1\| = 1$$

$\mathbf{t}_1$の計算 yとの共分散の最大化

y との関連性が大きい $\mathbf{t}_1$ を抽出したい

y と $\mathbf{t}_1$ の共分散 $\mathbf{y}^T\mathbf{t}_1$ を最大化するよう $\mathbf{t}_1$ を求める

- ✓ オートスケーリングしているため **X** と **y** は平均0

ただし、 $\|\mathbf{w}_1\| = 1$ (制約条件)

$\mathbf{t}_1$の計算 Lagrangeの未定乗数法

制約条件がある中での最大化

⇨ Lagrangeの未定乗数法

μ を未知の定数として、下の G を最大化

$$\begin{aligned} \mathrm{G} &= \mathbf{y}^{\mathrm{T}}\mathbf{t}_1 - \mu\left(\|\mathbf{w}_1\|^2 - 1 \right) \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1 - \mu\left(\|\mathbf{w}_1\|^2 - 1 \right) \end{aligned}$$

$\mathbf{t}_1$の計算 Gの最大化

G は $\mathbf{w}_1$ の関数

G が最大値のとき、G を $\mathbf{w}_1$ の要素ごとに偏微分した値は 0

$$\mathrm{G} = \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1 - \mu\left(\left\|\mathbf{w}_1\right\|^2 - 1 \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{d} y_i x_{i,k} w_{1,k} - \mu\left(\sum_{k=1}^{d} w_{1,k}^2 - 1 \right)$$

- ✓n : データ数
- ✓d : 説明変数の数

$$\frac{\partial \mathrm{G}}{\partial w_{1,k}} = \sum_{i=1}^{n} y_i x_{i,k} - 2\mu w_{1,k} = 0$$

- ✓ k : 変数番号

$\mathbf{t}_1$の計算 式変形

$$\sum_{i=1}^{n} y_i x_{i,k} - 2\mu w_{1,k} = 0$$ より、 $$\sum_{i=1}^{n} y_i x_{i,k} = 2\mu w_{1,k}$$

$w_{1,\mathrm{k}}$を両辺に掛けると、 $$\sum_{i=1}^{n} y_i x_{i,k} w_{1,k} = 2\mu w_{1,k}^2$$

kについて1からdまで和を取る
(制約条件を使ってwが消える) $$\sum_{i=1}^{n} \sum_{k=1}^{d} y_i x_{i,k} w_{1,k} = 2\mu$$

よって、 $$\mathbf{y}^{\mathrm{T}} \mathbf{t}_1 = 2\mu$$

$\mathbf{t}_1$の計算 $\mathbf{w}_1$の計算

$$\sum_{i=1}^{n} y_i x_{i,k} = 2\mu w_{1,k}$$ より、

$$w_{1,k} = \frac{\sum_{i=1}^{n} y_i x_{i,k}}{2\mu}$$

μは $\mathbf{y}^{\mathrm{T}}\mathbf{t}_1$ の値、$\mathbf{w}_1$の大きさ(ノルム)は1より、

$$\mathbf{w}_1 = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{y}}{\left\|\mathbf{X}^{\mathrm{T}}\mathbf{y}\right\|}$$

$\mathbf{w}_1$ が得られた後、$\mathbf{t}_1$ も計算

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$$

$\mathbf{p}_1$とq_1の計算

$\mathbf{p}_1$ は **X** の残差 **E** の要素の二乗和が最小になるように求める
(最小二乗法)

$$\mathbf{p}_1 = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{t}_1}{\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1}$$

q_1 は **y** の残差 **f** の要素の二乗和が最小になるように求める
(最小二乗法)

$$q_1 = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{t}_1}{\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1}$$

2成分のPLSモデル

PLSモデル式

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \mathbf{E}_2 \qquad \mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{t}_2 q_2 + \mathbf{f}_2$$

$$\mathbf{X}_2 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} \qquad \mathbf{y}_2 = \mathbf{y} - \mathbf{t}_1 q_1$$

✓$\mathbf{X}_2$：$\mathbf{X}$ の中で1成分のPLSモデルでは説明できない部分

✓$\mathbf{y}_2$：$\mathbf{y}$ の中で1成分のPLSモデルでは説明できない部分

$\mathbf{t}_2$ は $\mathbf{X}_2$ の線形結合　　$\mathbf{t}_2 = \mathbf{X}_2\mathbf{w}_2$

ただし、$\mathbf{w}_2$ の大きさ(ノルム)は1　　$\|\mathbf{w}_2\| = 1$

$\mathbf{w}_2$、$\mathbf{t}_2$、$\mathbf{p}_2$、$\mathbf{q}_2$の計算

$\mathbf{y}_2$ との関連性が大きい $\mathbf{t}_2$ を抽出したい

$\mathbf{y}_2$ と $\mathbf{t}_2$ の共分散 $\mathbf{y}_2^{\mathrm{T}}\mathbf{t}_2$ を最大化するよう $\mathbf{t}_2$ を計算する

1成分の時と同様にして、

$$\mathbf{w}_2 = \frac{\mathbf{X}_2^{\mathrm{T}}\mathbf{y}_2}{\left\|\mathbf{X}_2^{\mathrm{T}}\mathbf{y}_2\right\|} \qquad \mathbf{t}_2 = \mathbf{X}\mathbf{w}_2$$

$$\mathbf{p}_2 = \frac{\mathbf{X}_2^{\mathrm{T}}\mathbf{t}_2}{\mathbf{t}_2^{\mathrm{T}}\mathbf{t}_2} \qquad q_2 = \frac{\mathbf{y}_2^{\mathrm{T}}\mathbf{t}_2}{\mathbf{t}_2^{\mathrm{T}}\mathbf{t}_2}$$

3成分以降も同様に計算する

何成分まで用いるか？

✓多くの成分を用いるとモデルの自由度が大きく(モデルが複雑に)なり、過学習の恐れがある

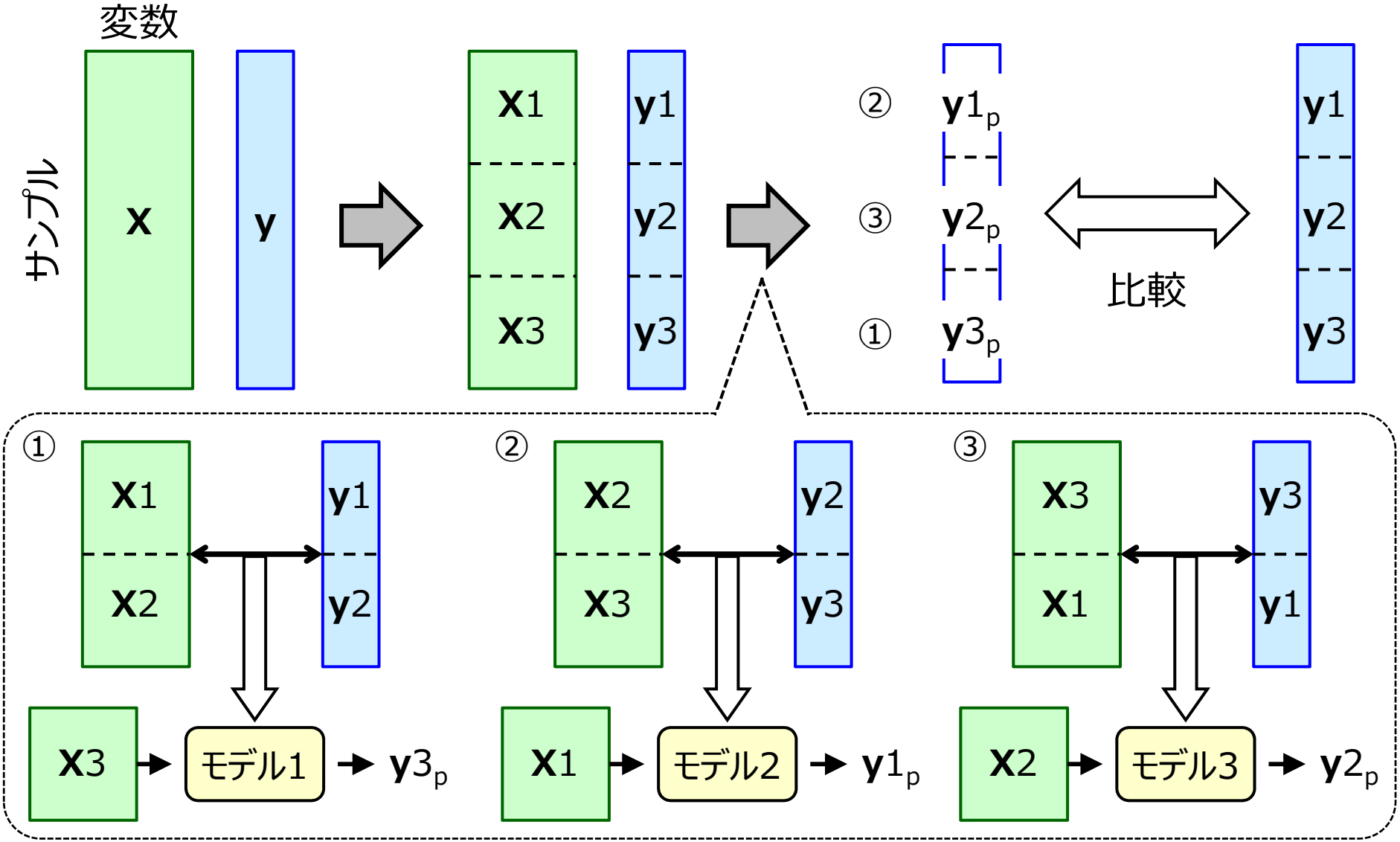
- 過学習：モデル構築用データには回帰式(回帰モデル)がよく当てはまるが、新しいデータに対する予測誤差が大きくなってしまうこと

- 予測性の高いモデルが得られる適切な成分数を選択

⇨ クロスバリデーション

クロスバリデーション

✓例) 3-fold クロスバリデーション



r^2_{CV} (予測的説明分散)

- ✓クロスバリデーションによる予測値を用いた説明分散 r^2
  - Leave-one-out クロスバリデーション
  - N-fold クロスバリデーション
  など
- ✓モデルの予測性を表す指標
- ✓1に近いほど良い

$$r^2_{CV} = 1 - \frac{\sum_{i=1}^{n} \left(y^{(i)} - y_{CV}{}^{(i)} \right)^2}{\sum_{i=1}^{n} \left(y^{(i)} - y_A \right)^2}$$

$y^{(i)}$：i 番目のサンプルにおける目的変数の値

$y_{CV}{}^{(i)}$：i 番目のサンプルにおけるクロスバリデーションによる目的変数の推定値

y_A：目的変数の平均値

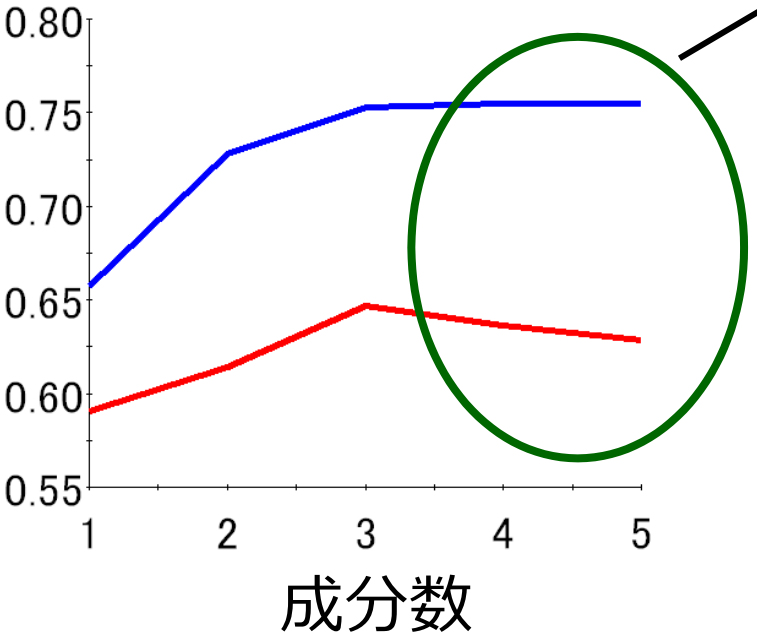
n：サンプル数

成分数の決め方

✓例) r^2_{CV}値を指標にして判断

- r^2_{CV}値が最大値を取る成分数
- r^2_{CV}値が最初の極大値を取る成分数
- r^2_{CV}値の上昇が最初に0.03以下となる成分数

r^2(青)
r^2_{CV}(赤)



モデル構築用データに対する性能は高くなっているが、予測性能は低下

→ 過学習が起きている

Root Mean Squared Error (*RMSE*) : 誤差の指標

RMSE$_C$ (RMSE of Calibration)

yの計算値

$$RMSE_C = \sqrt{\frac{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{n}} \quad \Leftrightarrow \quad r^2 = 1 - \frac{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{n}\left(y^{(i)} - \bar{y}\right)^2}$$

RMSE$_{CV}$ (RMSE with Cross-Validation)

クロスバリデーションによるyの予測値

$$RMSE_{CV} = \sqrt{\frac{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}_{CV}^{(i)}\right)^2}{n}} \quad \Leftrightarrow \quad r^2{}_{CV} = 1 - \frac{\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}_{CV}^{(i)}\right)^2}{\sum_{i=1}^{n}\left(y^{(i)} - \bar{y}\right)^2}$$

データが同じであれば、 r^2, $r^2{}_{CV}$ が大きい ⇔ *RMSE*$_C$, *RMSE*$_{CV}$ が小さい