Mastering PLS regression

YUGI, Katsuyuki
Kuroda Lab., The University of Tokyo

この講習会の目的

- PLS回帰の理解
- Janes et al. (2005) Science の追試に必要な知識の習得

Gradus ad Parnassum

PLS回帰	
主成分回帰(PCR)	
重回帰(MLR)	主成分分析(PCA)

- 特異値分解がわかるとこれらを統一的に理解できる

特異値分解と多変量解析の関係

- PLS回帰
  - 説明変数と目的変数の相関を最大化する条件
- 重回帰
  - 回帰係数行列を与えるMoore-Penrose一般逆行列の基礎理論
- 主成分分析
  - 主成分スコア行列、因子負荷量行列と特異値分解の間に対応関係

特異値分解(Singular Value Decomposition)

- 任意のm x n行列は3つの行列の積で書ける

$$\underset{m\times n}{\mathbf{M}} = \underset{\text{回転}}{\underset{m\times m}{\mathbf{U}}} \quad \underset{\text{伸展}}{\underset{m\times n}{\mathbf{\Sigma}}} \quad \underset{\text{回転}}{\underset{n\times n}{\mathbf{V}^T}}$$

$$\underset{m\times m}{\mathbf{U}} = \left(\underset{m\times r}{\text{Mの列空間の基底}} \;\middle|\; \underset{m\times(m-r)}{\text{Mの左零空間}} \right) \qquad \underset{n\times n}{\mathbf{V}^T} = \left(\begin{array}{c} \underset{r\times n}{\text{Mの行空間の基底}} \\ \hline \underset{(n-r)\times n}{\text{Mの零空間}} \end{array} \right)$$

$$\underset{m\times n}{\mathbf{\Sigma}} = \left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_r & \\ \hline & 0 & & 0 \end{array} \right)$$

各行列の性質

- U,Vは正規直交行列
  - 各列ベクトルは互いに直交
  - 各列ベクトルのノルムは 1

- σ_j を特異値(singular value)と呼ぶ
  - 特異値は M^TM, MM^T の固有値

$$\underset{m\times n}{\Sigma} = \left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right)$$

Moore-Penrose の一般逆行列

$$\underset{n\times m}{\mathbf{A}^{\#}} = \underset{n\times n}{\mathbf{V}}\ \underset{n\times m}{\boldsymbol{\Sigma}'^{-1}}\ \underset{m\times m}{\mathbf{U}^{T}}$$

$$\boldsymbol{\Sigma}'^{-1} = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & 0 & \\ & \ddots & & 0 \\ 0 & & \frac{1}{\sigma_r} & \\ \hline & 0 & & 0 \end{array}\right)$$

連立１次方程式 $\mathbf{Ax} = \mathbf{b}$

- Aが縦長の行列
  - $\hat{\mathbf{x}} = \mathbf{A}^{\#}\mathbf{b}$ は二乗誤差 $\left|\left|\mathbf{Ax} - \mathbf{b}\right|\right|^2$ が最小になる解
- Aが横長の行列
  - $\hat{\mathbf{x}} = \mathbf{A}^{\#}\mathbf{b}$ はノルム $\left|\left|\mathbf{x}\right|\right|$ が最小になる解

演習: 紙と鉛筆で特異値分解

- 行列 $\begin{pmatrix} 1 & 2 \\ -1 & -2 \end{pmatrix}$ を手計算で特異値分解する

重回帰(Multiple Linear Regression)

- 残差平方和を最小にする回帰係数aを求めることが重回帰分析の目的

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \varepsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

- 行数 > 列数 であれば最小二乗解は一般逆行列で与えられる

$$\hat{\mathbf{a}} = \mathbf{X}^{\#}\mathbf{y}$$

重回帰の問題点

- 変数が増えると、必要なデータ数も増える
  - 行数 < 列数・・・最小ノルム解。複数ある解の一つにすぎない。

- 多重共線性(multi-collinearity)
  - 説明変数の間に相関があると、回帰係数がおかしくなる

- 解決策: 相関のある説明変数をひとまとめに
  - 必要なデータが少なくて済み、多重共線性も回避できる

主成分分析

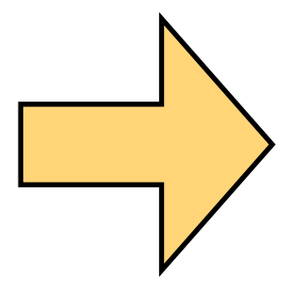
- 相関する変数を線形結合で1つにまとめる

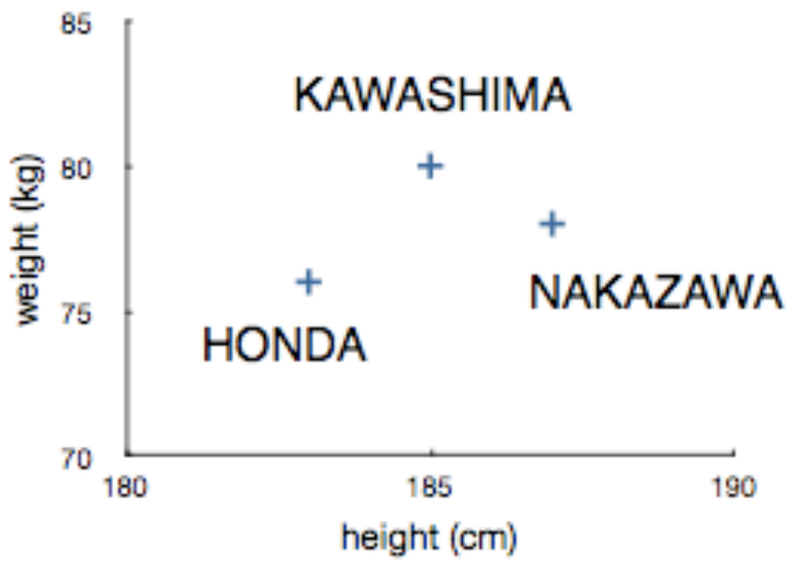
$$t = w_1 X_1 + w_2 X_2$$

  - 問題：「wをいかにして求めるか」
    - 方針: 「X_1 と X_2 との共分散(相関)が最大になるように w を決める」

- 定番の例: 「身長と体重」を「人間の大きさスコア」に集約

	身長(cm)	体重(kg)
川島	185	80
中澤	187	78
本田	183	76





主成分分析の手順

1. データ行列 $\mathbf{X}$ から平均値を差し引く (行列 $\hat{\mathbf{X}}$ とおく)
2. $\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ (共分散行列) を計算する
3. 共分散行列の固有ベクトルを求める (因子負荷量$\mathbf{w}$と定義)
4. 因子負荷量を右からデータ行列に掛けると主成分得点

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

- 重要： 共分散行列の固有ベクトルは因子負荷量(loading)

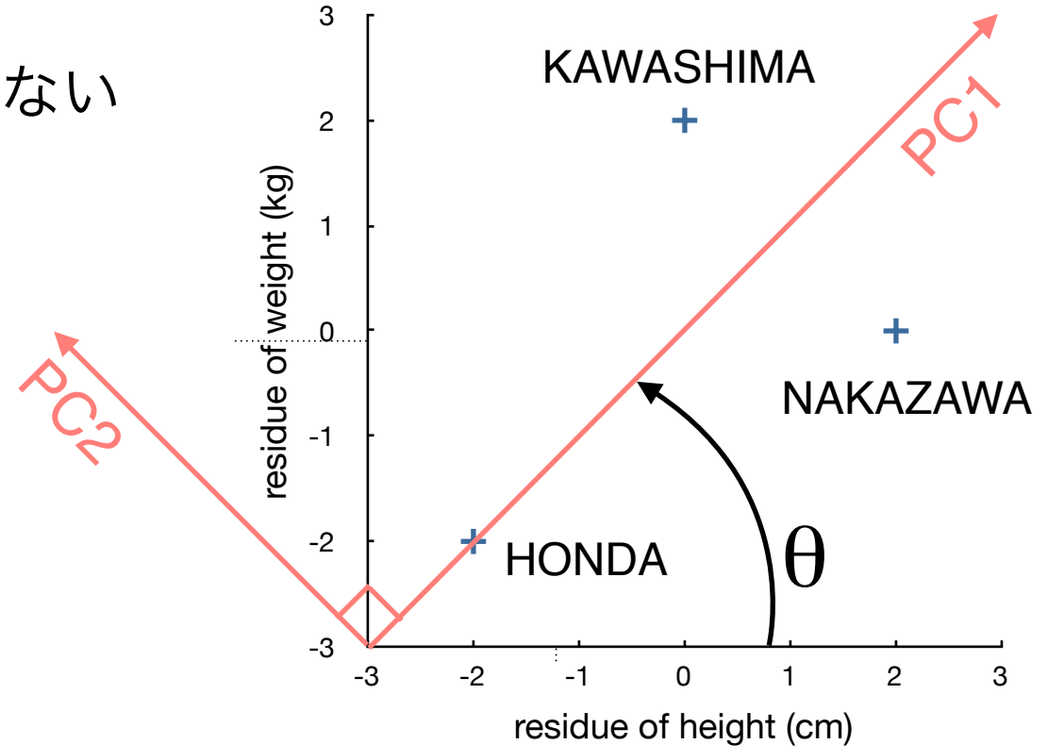
演習: 紙と鉛筆で主成分分析

- 以下3名の身体測定データから、「身体の大きさ」を表す主成分スコアを求める。

	身長(cm)	体重(kg)
川島	185	80
中澤	187	78
本田	183	76

主成分分析は座標の回転である

- 正規直交行列
  - ベクトルの長さを変えない
  - ベクトル同士の角度を変えない
  - ベクトルの回転



主成分分析の行列表記

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

第 i 主成分スコア (列ベクトル)

$$\mathbf{T} = \begin{pmatrix} \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_k \end{pmatrix}$$

列方向に並べる

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_k \end{pmatrix}$$

wの代わりにpで表記(慣習)

$$\mathbf{T} = \mathbf{X}\mathbf{P}$$

主成分得点 = データ x 共分散行列の固有ベクトル

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T$$

P は正規直交行列なので $PP^T=I$

主成分分析と特異値分解の関係

$$\begin{aligned} \mathbf{X} &= \mathbf{U}\ \ \mathbf{\Sigma}\ \ \mathbf{V}^T \\ &= \quad \mathbf{T} \quad \mathbf{P}^T \qquad\qquad (\mathbf{U\Sigma} = \mathbf{T},\quad \mathbf{V} = \mathbf{P}) \end{aligned}$$

- V(= P)は X^TX の固有ベクトルを並べた行列

復習: 重回帰の問題点

- 変数が増えると、必要なデータ数も増える
  - 行数 < 列数・・・最小ノルム解。複数ある解の一つにすぎない。
- 多重共線性(multi-collinearity)
  - 説明変数の間に相関があると、回帰係数がおかしくなる
- 解決策: 主成分分析で変数を減らす
  - 必要なデータが少なくて済み、多重共線性も回避できる

主成分回帰

- 重回帰の式 ＋ 主成分分析の行列表記 → 主成分回帰の式

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} \quad \mathbf{B} \\ &= \mathbf{T} \quad \mathbf{P}^T \quad \mathbf{B} \end{aligned}$$

$$(\mathbf{X} = \mathbf{T}\mathbf{P}^T)$$

PLS回帰	
主成分回帰	
重回帰	主成分分析

- 回帰係数

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}_{PCR}$$

$$\hat{\mathbf{B}}_{PCR} = \mathbf{P}\mathbf{T}^{\#}\mathbf{Y}$$

- 「Tの列数 < 行数」ならば B の最小二乗解が求まる

PLS回帰の基本アイディア

- 主成分回帰
  - 説明変数 X を主成分 T に変換
  - 目的変数 Y を T で回帰

PLS回帰	
主成分回帰	
重回帰	主成分分析

- PLS回帰
  - X , Y の両方をそれぞれの主成分(のようなもの) T , U に変換
  - 因子負荷量は、T と U の共分散を最大にするように決める
  - 目的変数 U を説明変数 T で回帰

因子負荷量をどのように決めるか？

- 方針: 「T と U の共分散を最大にするように決める」
  - t = Xw , u=Yc
  - Cov(Xw,Yc) が最大になるような w , c の条件を探す

- 特異値分解と密接な関係
  - w ・・・ $\mathbf{X}^T\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ としたとき、Uの1列目
  - c ・・・ $\mathbf{X}^T\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ としたとき、Vの1列目

PLS回帰係数を求める手順 (概略)

1. $\mathbf{X}^T\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ と特異値分解し、U、Vの1列目をw、cとする
2. $w^Tw=1$、$c^Tc=1$となるよう正規化
3. t=Xw、u=Yc よりt、uを求める。$t^Tt=1$、$u^Tu=1$に正規化。
4. u = bt の線形関係にあてはめて回帰。bを求める。
5. $p = X^Tt$、$q=Y^Tu$ よりp、qを求める
6. X から tp^T を、Yからuq^Tをそれぞれ差し引いて「1.」に戻る。主成分の数だけこれを繰り返す。

回帰係数行列の中身

- t, u, p, q, b をそれぞれ並べた行列を大文字で表す

- PLS回帰で求まる係数行列

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{U}\mathbf{Q}^T \\
&= \mathbf{T}\mathbf{B}\mathbf{Q}^T \\
&= \mathbf{X}(\mathbf{P}^T)^{\#}\mathbf{B}\mathbf{Q}^T \qquad (\because \hat{\mathbf{T}} = \mathbf{X}(\mathbf{P}^T)^{\#}) \\
&= \mathbf{X}\mathbf{B}_{\mathrm{PLS}}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{P} &= \mathbf{X}^T\mathbf{T} \\
\mathbf{P}^T &= \mathbf{T}^T\mathbf{X} \\
\mathbf{T}\mathbf{P}^T &= \mathbf{T}\mathbf{T}^T\mathbf{X} \\
\hat{\mathbf{T}} &= \mathbf{X}(\mathbf{P}^T)^{\#}
\end{aligned}
$$

PLSではPが直交しない。

そのため最小二乗推定値を用いる。

演習: IEG発現のPLS回帰モデル

- ERKの時系列を説明変数として、30分時点のc-Fos、c-Junの発現量を予測する

	ERK(t1)	ERK(t2)	ERK(t3)
EGF	0	4	2
NGF	2	4	4

	c-Fos30min	c-Jun30min
EGF	0	1
NGF	1	2

図解：各ベクトル・行列の関係