主成分分析
Principal Component Analysis
PCA

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

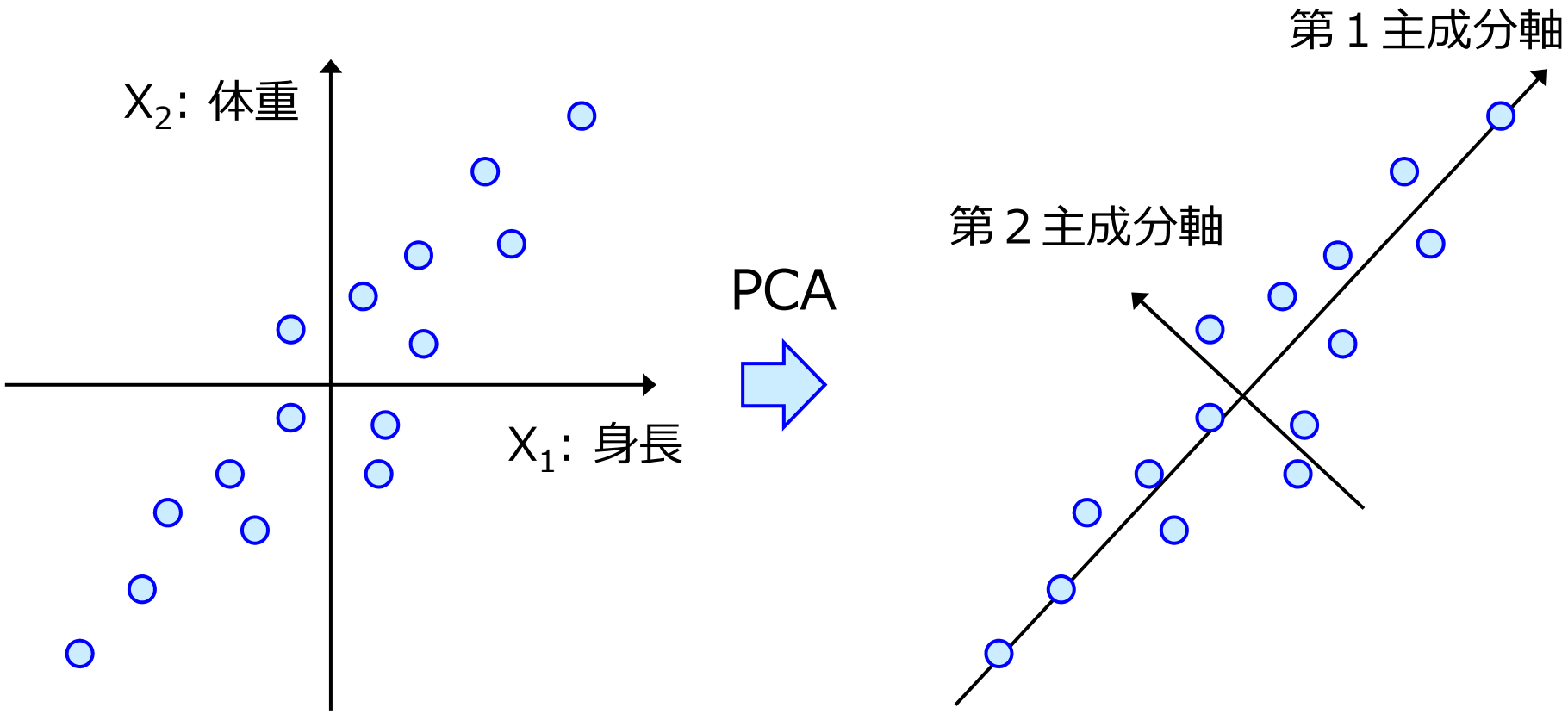
主成分分析 (PCA) とは？

✓主成分分析 (Principal Component Analysis, PCA)

- 見える化 (可視化) する手法
- 多変量 (多次元) のデータセットを低次元化する方法
- データセットのもつ情報量をなるべく失わないように
  元の次元から より低い次元でデータセットを表現
  - “より低い次元” を 2 次元にすれば可視化を達成
- 軸を回転 (+ 反転) させる

PCAの図解

例) １５人の身長･体重データ (多次元のデータ)



第１主成分だけでも、１５人のだいたいの情報はおさえられる

PCAで できること

✓データセットのだいたいの様子を見る
- いろいろな主成分同士のプロットを見る
- それぞれの主成分の角度を見ることで、データセットがどんな方向に分布しているか分かる

✓ノイズを除く
- 第4成分以降をノイズとみなして、第1,2,3主成分のみ使う、とか

✓データセットの中で外れているサンプルを探す
- PCAをした後に主成分のプロットを見たとき、他のサンプルと離れているサンプルは、PCA前のサンプル同士も必ず離れている

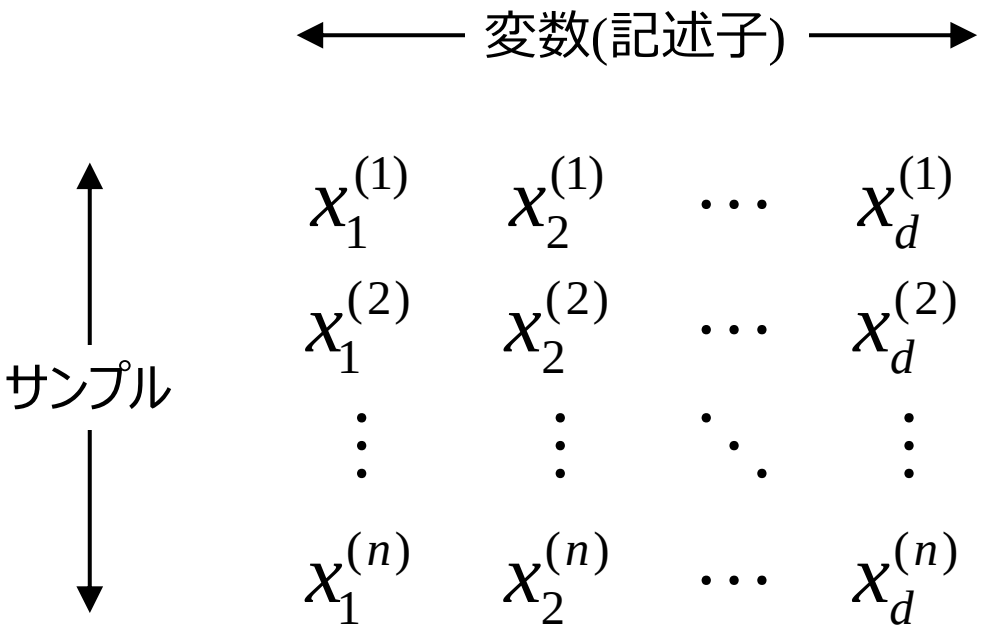
✓変数 (PCA後は成分) の間の相関を 0 にする
- 回帰分析をしたときの回帰係数の値が安定になる

データセットの表し方

$x_i^{(k)}$: k 個目のサンプルにおける、i 番目の変数(記述子) の値

d : 変数(記述子) の数

n : サンプルの数



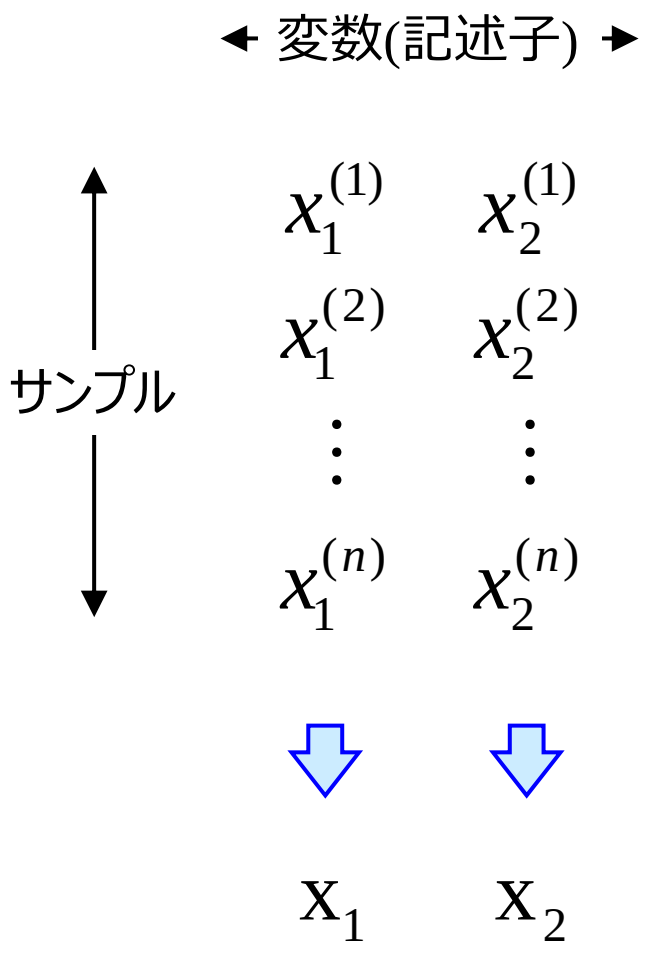
PCAの前に

✓PCAの前に、必ず前処理を行いましょう

- 分散が 0、もしくは同じ値を多くもつ変数の削除
- オートスケーリング

✓詳しくは こちら

2変数のときのPCA (3変数以上への拡張も簡単)



主成分とローディング

$$t_1 = x_1 p_1^{(1)} + x_2 p_1^{(2)}$$

$$t_2 = x_1 p_2^{(1)} + x_2 p_2^{(2)}$$

t_i : 第 i 主成分

$p_i^{(j)}$: 第 i 主成分に対応する、j 番目の変数(記述子) の重み (ローディング)

行列で表すと・・・

$$\begin{bmatrix} t_1^{(1)} & t_2^{(1)} \\ t_1^{(2)} & t_2^{(2)} \\ \vdots & \vdots \\ t_1^{(n)} & t_2^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} \end{bmatrix} \begin{bmatrix} p_1{}^{(1)} & p_2{}^{(1)} \\ p_1{}^{(2)} & p_2{}^{(2)} \end{bmatrix}$$

$x_i^{(k)}$: k 個目のサンプルにおける、i 番目の変数(記述子) の値

$t_i^{(k)}$: k 個目のサンプルにおける、第 i 主成分の値

$p_i{}^{(j)}$: 第 i 主成分に対応する、j 番目の変数(記述子) の重み (ローディング)

第１主成分を考える

$$t_1 = x_1 p_1^{(1)} + x_2 p_1^{(2)}$$

ローディングの制約条件

✓ローディング(重み)を定数倍することで、主成分が変わってしまうため
ローディングの二乗和は 1 とする

$$\left(p_1^{(1)}\right)^2+\left(p_1^{(2)}\right)^2=1$$

主成分の分散を最大化

✓データセットのばらつき (分散) が最大の方向を第一主成分軸とする

✓元のデータセットはオートスケーリングしてあり、各変数の平均は 0

✓p.7 のように変数の線形結合で表される主成分の平均も 0

✓分散を最大化させることは、主成分の値の二乗和を最大化させることに対応する

✓ $S = \sum_{i=1}^{n} \left(t_1^{(i)} \right)^2$ を最大化させる！

$t_1^{(i)}$: i 個目のサンプルにおける、第 1 主成分の値

Sを最大化するローディングを求める

$$S = \sum_{i=1}^{n} \left(t_1^{(i)} \right)^2$$

$$= \sum_{i=1}^{n} \left(x_1^{(i)} p_1^{(1)} + x_2^{(i)} p_1^{(2)} \right)^2$$

$$= \left(p_1^{(1)} \right)^2 \sum_{i=1}^{n} \left(x_1^{(i)} \right)^2 + 2 p_1^{(1)} p_1^{(2)} \sum_{i=1}^{n} x_1^{(i)} x_2^{(i)} + \left(p_1^{(2)} \right)^2 \sum_{i=1}^{n} \left(x_2^{(i)} \right)^2$$

$p_1^{(1)}$, $p_1^{(2)}$ が規格化条件 $\left(p_1^{(1)} \right)^2 + \left(p_1^{(2)} \right)^2 = 1$ を満たしながら

S を最大化する ⇨ Lagrange の未定乗数法

Lagrangeの未定乗数法

λ を未知の定数として下の G が最大となる λ , $p_1^{(1)}$, $p_1^{(2)}$ を求める

$$G = S - \lambda\left(\left(p_1^{(1)}\right)^2 + \left(p_1^{(2)}\right)^2 - 1\right)$$

$$= \left(p_1^{(1)}\right)^2 \sum_{i=1}^{n}\left(x_1^{(i)}\right)^2 + 2p_1^{(1)}p_1^{(2)}\sum_{i=1}^{n}x_1^{(i)}x_2^{(i)} + \left(p_1^{(2)}\right)^2 \sum_{i=1}^{n}\left(x_2^{(i)}\right)^2 - \lambda\left(\left(p_1^{(1)}\right)^2 + \left(p_1^{(2)}\right)^2 - 1\right)$$

G が最大 ⇨ G が極大 ⇨ G を λ , $p_1^{(1)}$, $p_1^{(2)}$ で偏微分して 0

Gを偏微分して 0

G が最大 ⇒ G が極大 ⇒ G をλ , $p_1^{(1)}$, $p_1^{(2)}$ で偏微分して 0

$$\left(\sum_{i=1}^{n}\left(x_1^{(i)}\right)^2-\lambda\right)p_1^{(1)}+\left(\sum_{i=1}^{n}x_1^{(i)}x_2^{(i)}\right)p_1^{(2)}=0$$

$$\left(\sum_{i=1}^{n}x_1^{(i)}x_2^{(i)}\right)p_1^{(1)}+\left(\sum_{i=1}^{n}\left(x_2^{(i)}\right)^2-\lambda\right)p_1^{(2)}=0$$

行列で表現 ⇒

$$\begin{bmatrix}\sum_{i=1}^{n}\left(x_1^{(i)}\right)^2-\lambda & \sum_{i=1}^{n}x_1^{(i)}x_2^{(i)}\\ \sum_{i=1}^{n}x_1^{(i)}x_2^{(i)} & \sum_{i=1}^{n}\left(x_2^{(i)}\right)^2-\lambda\end{bmatrix}\begin{bmatrix}p_1^{(1)}\\ p_1^{(2)}\end{bmatrix}=\mathbf{0}$$

行列で表す

$$\begin{bmatrix} \sum_{i=1}^{n}\left(x_1^{(i)}\right)^2-\lambda & \sum_{i=1}^{n}x_1^{(i)}x_2^{(i)} \\ \sum_{i=1}^{n}x_1^{(i)}x_2^{(i)} & \sum_{i=1}^{n}\left(x_2^{(i)}\right)^2-\lambda \end{bmatrix}\begin{bmatrix} p_1^{(1)} \\ p_1^{(2)} \end{bmatrix}=\mathbf{0}$$

$$\Rightarrow \quad \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}-\lambda\mathbf{E}\right)\mathbf{p}_1=\mathbf{0}$$

ただし、

$$\mathbf{X}=\begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} \end{bmatrix},\ \mathbf{E}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ \mathbf{p}_1=\begin{bmatrix} p_1^{(1)} \\ p_1^{(2)} \end{bmatrix}$$

固有値問題へ

$$\begin{bmatrix} \sum_{i=1}^{n}\left(x_1^{(i)}\right)^2 - \lambda & \sum_{i=1}^{n} x_1^{(i)} x_2^{(i)} \\ \sum_{i=1}^{n} x_1^{(i)} x_2^{(i)} & \sum_{i=1}^{n}\left(x_2^{(i)}\right)^2 - \lambda \end{bmatrix} \begin{bmatrix} p_1^{(1)} \\ p_1^{(2)} \end{bmatrix} = \mathbf{0}$$

$$\Rightarrow \quad \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda\mathbf{E}\right)\mathbf{p}_1 = \mathbf{0}$$

$p_1^{(1)} = p_1^{(2)} = 0$ 以外の解をもつためには、

$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda\mathbf{E}\right)$ の行列式が 0 である必要がある

$\Rightarrow$ λ を固有値、$\mathbf{p}_1$ に加えて $\mathbf{p}_2$ を固有ベクトルとする固有値問題

これによって $\mathbf{p}_1, \mathbf{p}_2$ を求め、対応する主成分を計算する

寄与率

第 i 主成分に対応する固有値 λ_i は、その主成分の二乗和に等しい

つまり、 $$\lambda_i = \sum_{j=1}^{n} \left(t_i^{(j)} \right)^2$$

固有値 λ_i を第 i 主成分のもつ情報量と仮定する

全固有値の中の λ_i の割合を寄与率 c_i として、
第 i 主成分のもつ情報量の割合として用いる

$$c_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j}$$

m : すべての主成分の数

累積寄与率

✓第 i 主成分までの寄与率の和を、第 i 主成分までの累積寄与率とする

✓たとえば、

- 可視化した第 2 主成分までの累積寄与率は 0.75 であった
- 累積寄与率が 0.9 を超えた最初の主成分までを用いる

といったように用いられる

逆写像

✓PCAにより、あるサンプルを低次元空間に写像できる

✓低次元空間に写像された点を、元の空間に戻すことを逆写像という

✓元のサンプル点と逆写像された点との距離を見ることで、
サンプル点が写像先とどれくらい近いかが分かる

✓離れているサンプルは、適切に写像されていない、外れ値である、
などの議論ができる

逆写像のしかた

- ✓第 i 主成分までのローディング $\mathbf{P}$ を用いる
- ✓あるサンプル $\mathbf{x}$ に対して、$\mathbf{T} = \mathbf{xP}$ で第 i 主成分までのスコア $\mathbf{T}$ を計算する
- ✓$\mathbf{TP}^{\mathrm{T}}$ が逆写像されたサンプルである
- ✓つまり、$\mathbf{xPP}^{\mathrm{T}}$ で逆写像されたサンプルを計算できるあ