主成分分析
Principal Component Analysis
(PCA)

pp.35~49

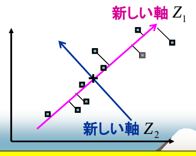
回帰分析と主成分分析

◎回帰分析は条件の付かない最小問題
◎主成分分析は条件付の最大最小問題
⇒手順は殆ど同じ

◎回帰分析は
多変量解析の手法の中で最もポピュラーな手法

主成分分析(今回のポイント)

◎主成分分析は条件付の最大最小問題
問題:
観測データにおける分散が最大の軸を求める
条件:
軸同士が直交
p次元の軸の長さ＝1



主成分分析の問題 ➡ 固有値問題

主成分分析(PCA)

データ解析の数学的考え方の基本形は主成分分析にある
主成分分析が理解できれば
データ解析は半分以上が理解できたと言える
観測データの主要な変動を要約し、特徴を把握する
ための統計的手法を主成分分析と呼ぶ

主成分分析の目的

・情報の要約(縮約)
・新しい尺度の構築
・構造の探索

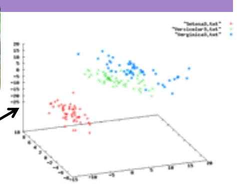
次元縮小・情報要約・新変数

- 質量÷体積　＝ 密度
- 体重・身長・胸囲・座高 → 体格
  - 具体的・現象的(計測できる) → 抽象的・概念的
    - 次元縮小 → 抽象度上昇
- 科目テストの成績 → 能力
  - 観測変数 → 合成変数
    - 単純化 → 理解・解釈が容易に

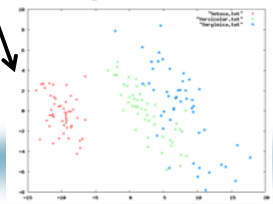
不可視 → 可視化

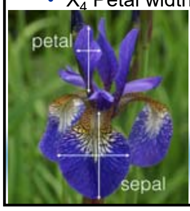
- X_1 Sepal length(がく長)
- X_2 Sepal width(がく幅)
- X_3 Petal length(花びら長)
- X_4 Petal width(花びら幅)

3D



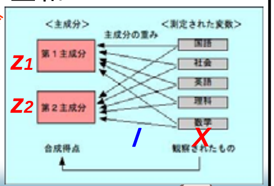
2D





データの要約(縮約)とは

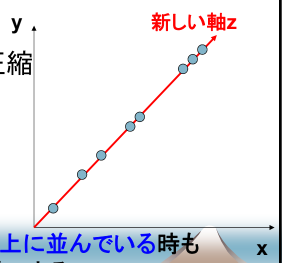
- 1変量の要約→　平均・分散・標準偏差
- 2変量の要約→　相関係数(分散、共分散)
- 多変量の要約→　総和・加重和
  - 主成分分析、回帰分析など
    $z = Xl$



- 多変量 X → 1変量の z に ⇒ 次元縮小

データの要約(縮約)の特例

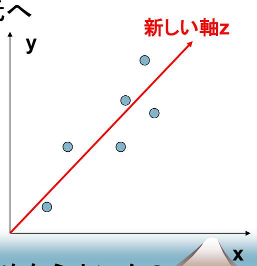
- 2変量(●)は一直線上に並んでいる場合
  → 新しいz軸で表現
  →2次元表現が1次元に圧縮
  ⇒ 次元縮小



拡張：N次元の変量は一直線上に並んでいる時も同じように1次元に圧縮できる

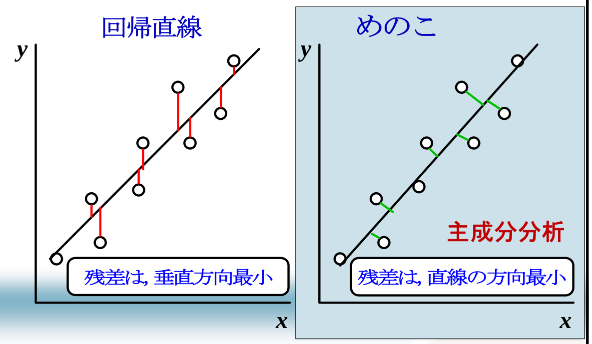
データの要約(縮約)の一般例

- N変量は直線に近い状態で並んでいる場合
  →N(=2)次元から1次元へ
  ⇒次元縮小



問題：Z軸をどのように定めたらよいか？

回帰分析と主成分分析における誤差の考え方

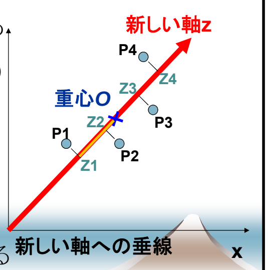


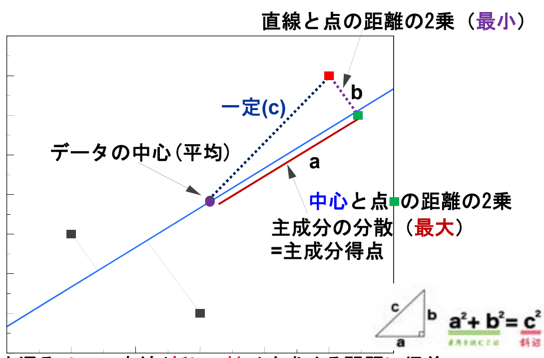
主成分分析のイメージ

点$P_i(i=1...4)$の平均点Oに対して　$OZ_1+OZ_2+OZ_3+OZ_4$
が最大にするz軸を探す
負の値を取ることが避けるため

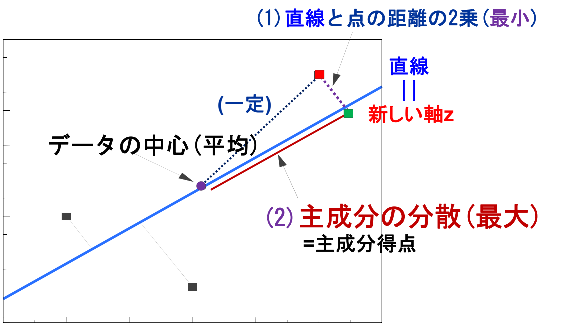
$$\boxed{OZ_1^2+OZ_2^2+OZ_3^2+OZ_4^2} \quad (1)$$

考え方を用いる
主成分分析の基本は式(1)が
最大となる新しい軸を求める
問題に帰着する
=> 新しい軸の分散を最大にする





中心を通る以下の直線(新しい軸Z)を求める問題に帰着：
(1)各点■からこの直線におろした垂線の長さの2乗和の値が最小にする直線
直角三角形の有名なピタゴラスの定理より　||　(内容的にはまったく同じことを意味)
(2)各点からこの直線におろした点■と中心●の距離2乗総和を最大にする直線



これから(2)の言い方で統一

⇒ 全てのデータに対して、新しい軸Zの分散を最大にする

POINT!

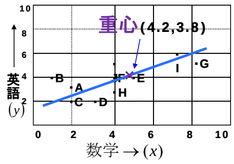
主成分分析の例

	数学	英語	合計	順位
A	2	3	5	7
B	1	4	5	7
C	2	2	4	10
D	3	2	5	7
E	5	4	9	3
F	4	4	8	6
G	8	5	13	1
H	6	3	9	3
I	7	6	13	1
J	4	5	9	3

左の順位は単純合計値によるが、

主成分分析を行い、主成分得点で順位を出す

1)点数グラフを作成
2)英語と数学のそれぞれの平均点を重心とし
3)重心を通る線を引く



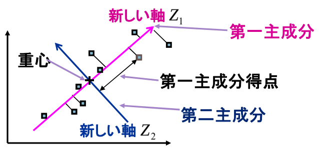
主成分分析で求めた主成分(主軸)は

1)重心(全データの平均点)を通る直線である

2)各点からこの直線におろした点と重心の距離2乗総和を最大にする直線(軸)(=分散最大)

主成分分析の例(続)

	数学	英語	合計	順位
A	2	3	5	7
B	1	4	5	7
C	2	2	4	10
D	3	2	5	7
E	5	4	9	3
F	4	4	8	6
G	8	5	13	1
H	6	3	9	3
I	7	6	13	1
J	4	5	9	3



Z_1 軸:第1主成分で総合能力

Z_2 軸:第2主成分で系別能力

Z_1と Z_2 垂直

POINT!

2次元のデータ形式と主成分(主軸)

2個の変数

	x	y
1	x_1	y_1
2	x_2	y_2
:	⋮	⋮
n	x_n	y_n

n個のケース

各変数の観測データ

得られる主成分(主軸):

$$z_1 = l_{1x}x + l_{1y}y$$

$$z_2 = l_{2x}x + l_{2y}y$$

2個の変量の線形(一次)結合
但し、以下の条件を満足:

単位ベクトル
軸同士直交

分散が最大の軸を探す

2次元のデータ形式と主成分(主軸)

得られる主成分(主軸):

$$z_1 = l_{1x}x + l_{1y}y$$

$$z_2 = l_{2x}x + l_{2y}y$$

2個の変量の線形結合であり、以下の条件を満足:

単位ベクトル:$\mathbf{l}_1 = \begin{bmatrix} l_{1x} \\ l_{1y} \end{bmatrix}, \mathbf{l}_2 = \begin{bmatrix} l_{2x} \\ l_{2y} \end{bmatrix},$

$$\mathbf{l}_1 \bullet \mathbf{l}_1 = \mathbf{l}_2 \bullet \mathbf{l}_2 = \|\mathbf{l}_1\| = \|\mathbf{l}_2\| = 1$$

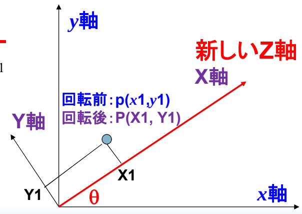
軸同士直交: $\mathbf{l}_1 \bullet \mathbf{l}_2 = 0$

主成分分析
(2次元の場合、数式で表す)

*x-y*座標系上の点$p(x1, y1)$は、直交軸の回転をした場合の新しい*X-Y*座標系では点$P(X1,Y1)$となる。回転角度をθとすると、回転後の点$P(X1, Y1)$と回転前の点$p(x1, y1)$の関係:

$$\begin{cases} X_1 = (\cos\theta)x_1 + (\sin\theta)y_1 \\ Y_1 = (-\sin\theta)x_1 + (\cos\theta)y_1 \end{cases}$$

$$\boxed{Z_1 = (\cos\theta)x_1 + (\sin\theta)y_1}$$



この基礎知識から、(x_i, y_i)が新しいZ軸(第1主軸)上の点Z_iを次式で表現する:

$$Z_i = (\cos\theta)x_i + (\sin\theta)y_i \quad (i = 1,2,\ldots n)$$

主成分分析
（2次元の場合、数式で表す）

$Z_i = (\cos\theta)x_i + (\sin\theta)y_i \quad (i=1,2,...,n)$より

$Z_i = l_1 x_i + l_2 y_i \quad (i=1,2,...,n)$

ここで、$l_1=\cos\theta; \quad l_2=\sin\theta$

$Z_i = x_i l_1 + y_i l_2 \quad (i=1,2,...,n)$

ベクトルZとl、行列Xで表現: $\mathbf{Z} = \begin{bmatrix}Z_1\\Z_2\\\vdots\\Z_n\end{bmatrix} = \begin{bmatrix} x_1 & y_1\\ x_2 & y_2\\ \vdots & \vdots\\ x_n & y_n\end{bmatrix}\begin{bmatrix}l_1\\l_2\end{bmatrix} = \mathbf{Xl}$

ベクトルZの分散 $S_{ZZ}=\frac{1}{n}\mathbf{Z}'\mathbf{Z}$ を最大にする問題

主成分分析
（2次元の場合、数式で表す）

$S_{ZZ}=\frac{1}{n}\mathbf{Z}'\mathbf{Z}$の最大となる時のベクトルlの値を求めれば、新しいZ 軸（第1主軸）が求められる

ただし、下の制約条件が常に付いている:

$l_1=\cos\theta; \quad l_2=\sin\theta$ から $l_1^2+l_2^2=1$

ベクトル$\mathbf{l}=\begin{bmatrix}l_1\\l_2\end{bmatrix}$で表現すると $\mathbf{l}'\mathbf{l}=1$となる

ラグランジュの未定乗数法を使って新しい関数を定義

その問題を解く式は？（各自で書いてみてください）

タイトル「演習レポート」、日付、学生番号、氏名を書く

主成分分析（数式で表す）

$\boxed{v=\frac{1}{n}\mathbf{Z}'\mathbf{Z} - \lambda(\mathbf{l}'\mathbf{l}-1)}$ にZ=Xl を代入して

$v = \frac{1}{n}(\mathbf{Xl})'(\mathbf{Xl}) - \lambda(\mathbf{l}'\mathbf{l}-1) = \frac{1}{n}\mathbf{l}'\mathbf{X}'\mathbf{Xl} - \lambda(\mathbf{l}'\mathbf{l}-1) = \mathbf{l}'\boldsymbol{\Sigma}\mathbf{l} - \lambda\mathbf{l}'\mathbf{l}+\lambda$

ここで、$\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ （分散共分散行列） →対称行列

ベクトルlについて偏微分して、0とおくと

どんな式が得られるか？

タイトル「演習レポート」、日付、学生番号、氏名を書く

「数式で表す」のまとめ

●主成分分析はとても簡単な計算で求めることができる

●今まで復習した内容を綺麗にかつ簡潔に使った

●データから分散共分散行列Σさえ計算されれば、

$$(\boldsymbol{\Sigma}-\lambda\mathbf{E})\mathbf{l}=\mathbf{0}$$

簡単にヤコビ法で解くことができる

主成分分析の計算を行ってみよう

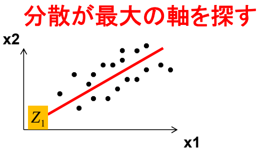
2 個の変数

$$\begin{array}{c|cc} & \mathbf{x} & \mathbf{y} \\ \hline 1 & x_1 & y_1 \\ 2 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ n & x_n & y_n \end{array}$$

n 個のケース

各変数の観測データ

分散が最大の軸を探す



$$\mathbf{Z}=\mathbf{Xl}=\begin{bmatrix}50 & 85\\70 & 75\\65 & 80\\60 & 70\\75 & 90\end{bmatrix}\begin{bmatrix}l_1\\l_2\end{bmatrix}$$

ベクトルで分散・共分散を書くと（復習）

•変数が2個、個体がn個の場合

$\mathbf{Z}=\mathbf{Xl}=[\mathbf{x}\quad\mathbf{y}]\begin{bmatrix}l_1\\l_2\end{bmatrix}$ $\quad \mathbf{x}=\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}, \mathbf{y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}$

•変数xとyの平均をそれぞれ求めてから

•以下の式で分散、共分散をそれぞれ計算できる

分散(Sxx)は: $Sxx=\frac{1}{n}(\mathbf{x}-\bar{\mathbf{x}})'(\mathbf{x}-\bar{\mathbf{x}})$

分散(Syy)は: $Syy=\frac{1}{n}(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}})$

共分散(Sxy)は: $Sxy=\frac{1}{n}(\mathbf{x}-\bar{\mathbf{x}})'(\mathbf{y}-\bar{\mathbf{y}})$

$$\boldsymbol{\Sigma}=\begin{bmatrix}S_{\mathbf{xx}} & S_{\mathbf{xy}}\\ S_{\mathbf{xy}} & S_{\mathbf{yy}}\end{bmatrix}$$

主成分分析の計算

2次分散共分散行列Σを求める

$$\mathbf{Z}=\mathbf{Xl}=\begin{bmatrix}50 & 85\\ 70 & 75\\ 65 & 80\\ 60 & 70\\ 75 & 90\end{bmatrix}\begin{bmatrix}l_1\\ l_2\end{bmatrix} \Rightarrow [\mathbf{x}\quad \mathbf{y}]=\begin{bmatrix}50 & 85\\ 70 & 75\\ 65 & 80\\ 60 & 70\\ 75 & 90\end{bmatrix}$$

平均値は: 64、80

ベクトル**x,y**の分散共分散行列

$$\mathbf{\Sigma}=\begin{bmatrix}Sxx & Sxy\\ Sxy & Syy\end{bmatrix}=\begin{bmatrix}74 & 10\\ 10 & 50\end{bmatrix}$$

主成分分析（数式で表す）

$$\mathbf{\Sigma l}=\lambda\mathbf{l}\Rightarrow(\mathbf{\Sigma}-\lambda\mathbf{E})\mathbf{l}=\mathbf{0}$$

新しい軸の分散＝固有値

分散共分散行列Σの固有値を求めればOK

$$\mathbf{\Sigma}=\begin{bmatrix}S_{xx} & S_{xy}\\ S_{xy} & S_{yy}\end{bmatrix}\text{の固有値を求める}|\mathbf{\Sigma}-\lambda\mathbf{E}|=0$$

$$\rightarrow\begin{vmatrix}S_{xx}-\lambda & S_{xy}\\ S_{xy} & S_{yy}-\lambda\end{vmatrix}=0$$

2次分散共分散行列Σの固有値と固有ベクトルを求めよう

$$\mathbf{\Sigma}=\begin{bmatrix}S_{xx} & S_{xy}\\ S_{xy} & S_{yy}\end{bmatrix}\qquad \mathbf{\Sigma l}=\lambda\mathbf{l}\text{より}\quad(\mathbf{\Sigma}-\lambda\mathbf{E})\mathbf{l}=\mathbf{0}$$

$$\mathbf{\Sigma}\text{の固有値を求める}|\mathbf{\Sigma}-\lambda\mathbf{E}|=0\rightarrow\begin{vmatrix}S_{xx}-\lambda & S_{xy}\\ S_{xy} & S_{yy}-\lambda\end{vmatrix}=0$$

$$\lambda_k\text{の固有ベクトルを求める}(\mathbf{\Sigma}-\lambda_k\mathbf{E})\mathbf{l}_k=\mathbf{0}$$

$$\rightarrow\begin{cases}(S_{xx}-\lambda_k)l_{kx}+S_{xy}l_{ky}=0\\ S_{xy}l_{kx}+(S_{yy}-\lambda_k)l_{ky}=0\\ \sum_{i=x,y}l_{ki}^2=1\end{cases}$$

←**制約条件**

主成分分析の計算問題

データの分散共分散行列

$$\mathbf{\Sigma}=\begin{bmatrix}Sxx & Sxy\\ Sxy & Syy\end{bmatrix}=\begin{bmatrix}74 & 10\\ 10 & 50\end{bmatrix}$$

Σの固有値と固有ベクトルを求めて下さい

タイトル「演習レポート」、日付、学生番号、氏名を書く

主成分分析の計算問題 （固有ベクトルを求める）

$$\text{データの分散共分散行列}\mathbf{\Sigma}=\begin{bmatrix}Sxx & Sxy\\ Sxy & Syy\end{bmatrix}=\begin{bmatrix}74 & 10\\ 10 & 50\end{bmatrix}$$

$(\mathbf{\Sigma}-\lambda\mathbf{E})\mathbf{l}=\mathbf{0}$より,**Σ**の固有ベクトルを求める

$$\begin{pmatrix}Sxx-\lambda_i & Sxy\\ Sxy & Syy-\lambda_i\end{pmatrix}\begin{pmatrix}l_{i1}\\ l_{i2}\end{pmatrix}=\begin{pmatrix}74-\lambda_i & 10\\ 10 & 50-\lambda_i\end{pmatrix}\begin{pmatrix}l_{i1}\\ l_{i2}\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

求めた固有値：$\lambda_1=77.62$、 $\lambda_2=46.38$より

固有ベクトル：$\mathbf{l}_1=\begin{bmatrix}?\\ ?\end{bmatrix}$

タイトル「演習レポート」、日付、学生番号、氏名を書く

主成分分析の宿題：

下記の表に基づいて、以下の問題を答えなさい。

No	標本	英語（x1）	数学（x2）
1	A	5	8
2	B	5	5
3	C	7	4
4	D	8	5

（1）x1とx2の平均値を求めなさい。
（2）x1とx2の分散と共分散を求める定義式を書き、分散共分散行列Sを求めなさい。
（3）分散共分散行列Sの固有値を求める定義式を書き、固有値を求めなさい。
（4）分散共分散行列Sのランクを答え、その理由を述べなさい。
（5）分散共分散行列Sの固有ベクトルを求める定義式を書き、固有ベクトルを求めなさい。