L10 主成分分析 (2)

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論 L10(2021-12-07 Tue)

最終更新: Time-stamp: "2021-12-07 Tue 11:05 JST hig"

今日の目標

- 負荷 (loading), 得点 (score) の意味が説明できる
- 主成分分析の結果を解釈できる
- 全, 群内, 群間平方和を説明できる

L09-Q1

Quiz 解答:n 次元正規分布の等高面の主軸

1. 長軸の向きだから, 最大固有値の固有ベクトルの向きであり, $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.
2. $$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{3} \\ -1/\sqrt{2} & 2/\sqrt{6} & -1/\sqrt{3} \\ 0 & 1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} 1/4 & & \\ & 1/9 & \\ & & 1/25 \end{pmatrix}^{-1} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = C,$$

   すなわち,

$$\frac{(\frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2)^2}{2^2} + \frac{(\frac{1}{\sqrt{6}}x_1 + \frac{2}{\sqrt{6}}x_2 + \frac{1}{\sqrt{6}}x_3)^2}{3^2} + \frac{(\frac{1}{\sqrt{3}}x_1 - \frac{1}{\sqrt{3}}x_2 + \frac{1}{\sqrt{3}}x_3)^2}{5^2} = C.$$

ここまで来たよ

主成分分析 PCA=Principal Component Analysis

(母分布が多次元正規分布と限定せず) データ $\boldsymbol{x}_i \in \mathbb{R}^p$ $(i = 1, \ldots, n)$ が得られたとする (標本サイズ n, 列の個数 p).
x_{ik}: x 番目のデータ点の, 第 k 列の数値.
このデータが, (小さい量を無視すると) 実質的に p 次元空間の $0, 1, \ldots, p-1$ 次元部分空間に分布していることがありうる. このとき, 大事な (大きな), 少数の量だけで考えたい.
今回は, n 個のデータ点がいちばん広がって見える方向 (第 1 主成分の方向) を選びたい.

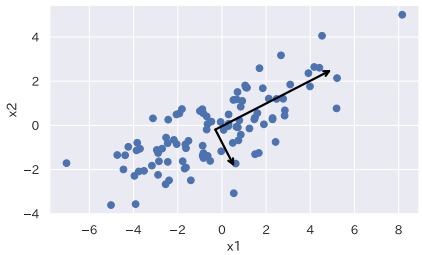
- この問題は教師なし学習
  - 対比:線形判別分析では, $y = 0, 1$ をいちばんよく分ける方向を考えた (教師あり)
- 注意: 単位や意味の違う量を比べてる可能性
- 次元圧縮 多次元のデータを, 情報をなるべく保って低次元のデータに変換する. 1,2,3 次元なら可視化容易. その後で他の手法を利用する.

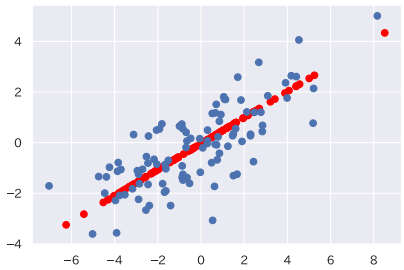
1 次元だけ選ぶとき

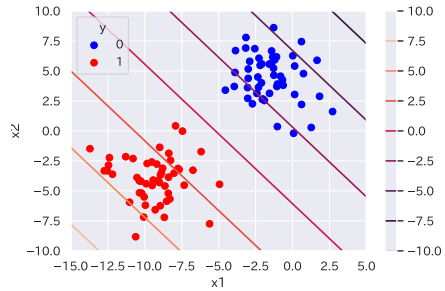
主成分分析では,

- 不偏標本共分散行列 S の, 最大 (第 1) 固有値 λ_1 の固有ベクトルの方向 $\boldsymbol{v}_1$ を選ぶ
- その方向へのデータのちらばりは $\simeq \sqrt{\lambda_1}$

なぜなら,
単位ベクトル $\boldsymbol{w} \in \mathbb{R}^p$, $|\boldsymbol{w}| = 1$ とするとき,
$z = {}^t\boldsymbol{w}\boldsymbol{x} = w_1x_1 + w_2x_2 + \cdots + w_px_p$ は, $\boldsymbol{w}$ 方向への射影.
$z = {}^t\boldsymbol{w}\boldsymbol{x}$ の不偏標本分散を最大にする (ただし $|\boldsymbol{w}| = 1$ という条件のもとで) ような $\boldsymbol{w}$ が単位固有ベクトル $\boldsymbol{v}_1/|\boldsymbol{v}_1|$. (要証明)







比較対象: 線形判別分析

- 第 1 主成分の方向 $\boldsymbol{v}_1$. 単位ベクトル $\boldsymbol{w}$. 広がりが最大. 最大固有値 λ_1 に対応.
- 第 1 主成分 $z = {}^{\mathrm{t}}\boldsymbol{w}\boldsymbol{x}$. $\boldsymbol{w}$ に平行な軸の座標.
- 第 1 主成分の主成分負荷量, 因子負荷量 (loading) $\frac{\sqrt{\lambda_1}w_1}{\sqrt{S_{11}}}, \frac{\sqrt{\lambda_1}w_2}{\sqrt{S_{22}}}, \ldots, \frac{\sqrt{\lambda_1}w_p}{\sqrt{S_{pp}}}$. 各 x_i と z との相関係数 (要証明).
- データ点 $\boldsymbol{x}_i$ の第 1 主成分得点 (score) ${}^{\mathrm{t}}\boldsymbol{w}\boldsymbol{x}_i$. データ点 x_i の, この軸での座標.

$\boldsymbol{v}_1$ を $\boldsymbol{v}_k$ にしたのが, 第 k 主成分… $z = z_1 =\rightsquigarrow z_k$
第 k 主成分の寄与率 $\frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j}$. その主成分が $\boldsymbol{x}$ のちらばりをどのくらい説明するか. z_k の分散/($\boldsymbol{x}$ の分散の和)
第 k 主成分までの累積寄与率 $\frac{\sum_{\ell=1}^{k}\lambda_\ell}{\sum_{j=1}^{p}\lambda_j}$. 第 $1, \ldots, k$ 主成分をあわせて $\boldsymbol{x}$ の分散をどのくらい説明するか.

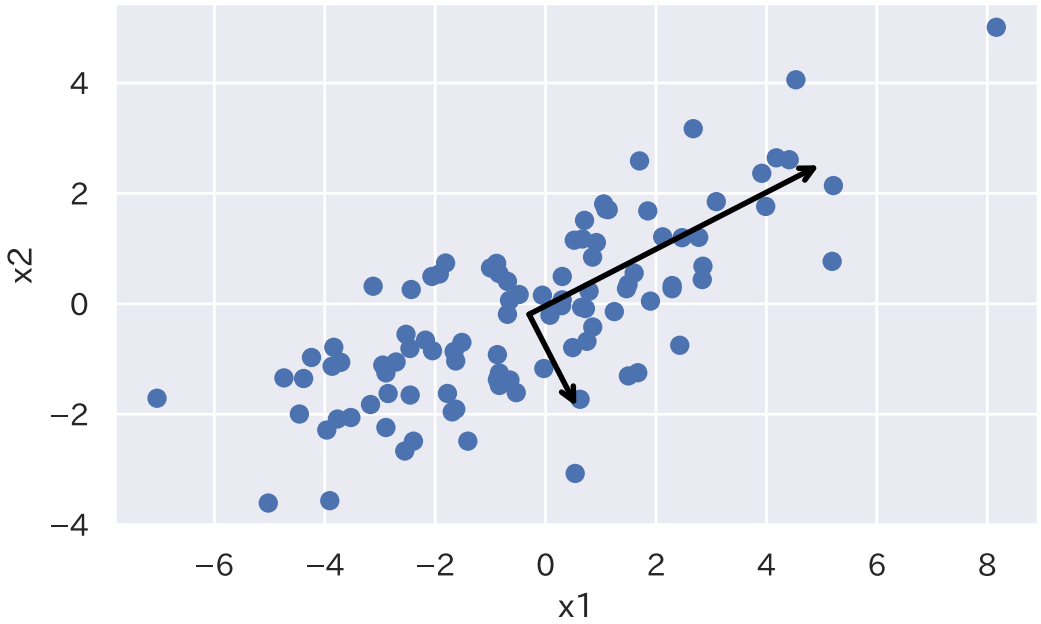
scikit-learn の主成分分析

```
1 from sklearn.decomposition import PCA
2
3 pca=PCA(n_components=2,random_state=seed) # インスタンス作成
4 pca.fit(df)                               # 学習
5
6 pca.components_ # 固有ベクトルのリスト=負荷
7 pca_x=pca.transform(df) # 各データ点の主成分得点
8 pca.explained_variance_ # 固有値=主成分の分散 のリスト
9 pca.explained_variance_ratio_ # 固有値/(すべての固有値の和)の
```

th-d09-pca.ipynb

numpy.pca もある

結果の例



L10-Q1

Quiz(主成分分析)

標準化された (平均 0, 分散 1 の)3 変量データについて, 共分散行列が次のように与えられる.

$$\begin{pmatrix} 1 & -\frac{4}{10} & \frac{3}{10} \\ -\frac{4}{10} & 1 & 0 \\ \frac{3}{10} & 0 & 1 \end{pmatrix}$$

1. 3 つの主成分を求めよう.
2. 第 1 主成分の因子負荷量を求めよう.
3. データ $(0.5, 0.3, -0.2)$ の第 1 主成分の主成分得点を求めよう.
4. 各主成分の寄与率と累積寄与率を求めよう.

固有値は $3/2, 1, 1/2$, 固有ベクトルは $\begin{pmatrix} 5 \\ -4 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix}$

なお, この行列の固有値は, $\lambda = 3/2, 1, 1/2$, 固有ベクトルは
$\begin{pmatrix} 5 \\ -4 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix}$

ここまで来たよ

現実の $p = 10$ 次元データ

ソウルオリンピック 10 種競技出場者の記録
`https://github.com/cran/ade4`

t100	(s) 100m 走
long	(m) 走り幅跳び
poid	(m) 砲丸投げ
haut	(m) 走り高跳び
t400	(s) 400m 走
t110	(s) 110m ハードル走
disq	(m) 円盤投げ
perc	(m) 棒高跳び
jave	(m) やり投げ
t1500	(s) 1500m 走

scikit-learn による標準化

- 大きい/小さい方がいいやつ混在 $\rightsquigarrow$ 困らない
- 単位が異なるやつ混在 $\rightsquigarrow$ 困る $\rightsquigarrow$ 標準化
- 100m の 1s と 1500m の 1s 混在 $\rightsquigarrow$ 困る $\rightsquigarrow$ 標準化

標準化 $x'_{ik} = \frac{x_{ik} - \overline{x}_{\cdot k}}{s_k}$.

```
1 from sklearn.Preprocessing import StandardScaler
2
3 scaler=StandardScaler() # インスタンス生成
4 scaler.fit(df)
5 df_std=scaler.transform(df)
6 # 2行は df_std=scaler.fit_transform(df) でまとめられる
```

主成分分析の結果

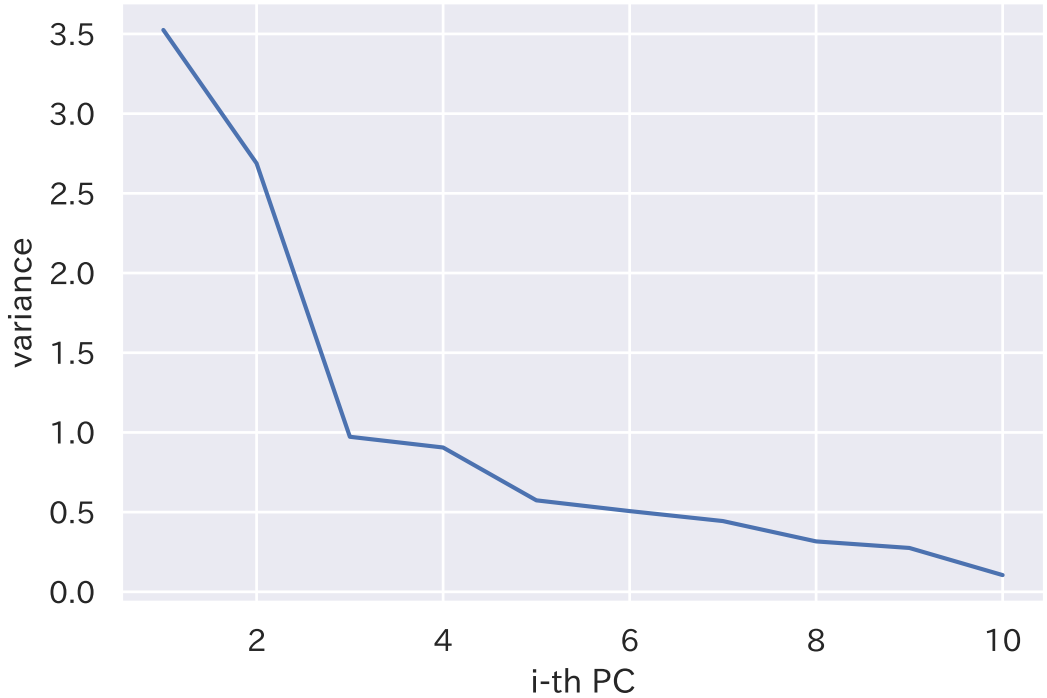
列名	(もとの単位) 種目	PC1	PC2
t100	(s) 100m 走	0.41588233	0.14880813
long	(m) 走り幅跳び	-0.39405149	-0.1520815
poid	(m) 砲丸投げ	-0.26910572	0.48353737
haut	(m) 走り高跳び	-0.21228177	0.0278985
t400	(s) 400m 走	0.35584739	0.35215981
t110	(s) 110m ハードル走	0.43348158	0.0695682
disq	(m) 円盤投げ	-0.17579228	0.50333471
perc	(m) 棒高跳び	-0.38408214	0.14958202
jave	(m) やり投げ	-0.17994361	0.371957
t1500	(s) 1500m 走	0.17014262	0.42096528

- 第 1 主成分 PC1= $(-1)\times$ 体力があるかないかの軸 =総合力=大きさ, 分析において興味ないことも多い
- 第 2 主成分 PC2 投擲力+持久力 があるかないかの軸, 個性のいちばん目立つ違い

主成分の個数の選択

経験的な方法.

- 方法 1 スクリープロット折れ曲がるところまでとる.



- 方法 2 (標準化されたとき) 固有値が 1, 累積寄与率が 0.8 になるところまでとる.

ここまで来たよ

平方和の分解とクラスター分析

本質は $p = 1$ 次元で見えるのでその表現で.
x_{ik}: k 番目のクラスターに属する i 番目のデータ点.
$i = 1, \ldots, n_k . k = 1, \ldots, C . \sum_{k=1}^{C} n_k = n.$
偏差平方和

$$
\begin{aligned}
S =& \sum_{i,k} [x_{ik} - \overline{x}_{..}]^2 \\
=& \sum_{i,k} [(x_{ik} - \overline{x}_{\cdot k}) + (\overline{x}_{\cdot k} - \overline{x}_{..})]^2 \\
\stackrel{!}{=}& \sum_{k} \sum_{i} (x_{ik} - \overline{x}_{\cdot k})^2 + \sum_{k} n_k (\overline{x}_{\cdot k} - \overline{x}_{..})^2 \\
=& S_{\mathrm{W}} + S_{\mathrm{B}}
\end{aligned}
$$

群=クラスター
$x_{..} = \frac{1}{n} \sum_{i,k} x_{ik}$ 全体平均
$x_{\cdot k} = \frac{1}{n_k} \sum_{i} x_{ik}$ 群内平均
S_{W}: 群内平方和 Within
S_{B}: 群間平方和 Between

L10-Q2

Quiz(平方和)

各組の学級で異なる教え方をした. テストの点数 x から学級 y をあてることはできそうだろうか.

y	学級	点数
0	A 組	78 79 79 80
1	B 組	78 86 81 83 82
2	C 組	86 85 87

群内平方和と群間平方和を求めよう.

平方和の分解と分類問題

線形判別分析実は, 線形判別分析の Fisher の線形判別関数は比 $S_{\mathrm{B}}/S_{\mathrm{W}}$ を最大化するものになっていた.
クラスター分析 CH 基準 (カリンスキ-ハラバシュ基準) クラスタの個数 C が異なる場合にも対応させたもの (回帰の自由度調整済決定係数みたいなアイデア).

$$\mathrm{CH}_C = \frac{(n-C)S_{\mathrm{B}}}{(C-1)S_{\mathrm{W}}}$$

Ward 法-トリッキーな距離の定義
クラスター間距離=(そのクラスターを合併したときの群内平方和 S_{W} の増分)

平方和の分解と n 群の差の検定

分散分析

3 個以上の群の差, F 分布 確率統計 II

2 群の t 検定

2 群の差 確率統計 II

平方和の分解と主成分分析

クラスターなし, 次元あり ($k = 1, \ldots, p$).
x_{ik}: i 番目のデータ点の, 第 k 次元の値. $i = 1, \ldots, n. k = 1, \ldots, p.$
標本平均値 $\overline{x}_{\cdot k} = 0.$
各座標の 分散の和

$$S = \sum_{i=1}^{n} \sum_{k=1}^{p} (x_{ik})^2 = \sum_{i=1^n} |\sum_{k=1}^{n} \boldsymbol{e}_k|^2 = \sum_{i=1^n} |\boldsymbol{x}_i|^2$$

別の正規直交基底 $\{\boldsymbol{v}_k\}$
$\boldsymbol{x}_i = \sum_k a_{ik} \boldsymbol{v}_k.$

$$S = \sum_{i=1}^{n} |\sum_{k} a_{ik} \boldsymbol{v}_k|^2 = \sum_{i} \sum_{k} a_{ik}^2 = \sum_{k} (\sum_{i} a_{ik}^2)$$

$(\sum_i a_{i1}^2)$ が最大になるように, 第 1 主成分の方向 $\boldsymbol{v}_1$ を選びたい ($\rightsquigarrow$ 固有ベクトル). それと直交する範囲で, 第 2 主成分…