L11 主成分分析 (2)

樋口さぶろお　https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習　L11(2021-12-09 Thu)

最終更新: Time-stamp: "2021-12-09 Thu 08:15 JST hig"

今日の目標

- 主成分分析のアルゴリズムが説明できる
- 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

L10-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,
   $$\frac{(\frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2)^2}{2^2} + \frac{(\frac{1}{\sqrt{6}}x_1 + \frac{2}{\sqrt{6}}x_2 + \frac{1}{\sqrt{6}}x_3)^2}{3^2} + \frac{(\frac{1}{\sqrt{3}}x_1 - \frac{1}{\sqrt{3}}x_2 + \frac{1}{\sqrt{3}}x_3)^2}{5^2} = C.$$

L10-Q2

Quiz 解答:主成分分析

固有値 $\lambda_1 = 12, \lambda_2 = 8$. 固有ベクトル $\boldsymbol{v}_1 = \frac{1}{2}\begin{pmatrix} 1 \\ -\sqrt{3} \end{pmatrix}, \boldsymbol{v}_2 = \frac{1}{2}\begin{pmatrix} \sqrt{3} \\ 1 \end{pmatrix}$

1. $z_1 = \frac{1}{2}x_1 - \frac{\sqrt{3}}{2}x_2$, $z_2 = \frac{\sqrt{3}}{2}x_1 + \frac{1}{2}x_2$.
2. $\frac{\sqrt{12}}{\sqrt{9}} \times \frac{1}{2}, \frac{\sqrt{8}}{\sqrt{11}} \times \frac{\sqrt{3}}{2}$,
3. $z_1 = \frac{1}{2} \cdot 0.5 - \frac{\sqrt{3}}{2} \cdot 0.3$.
4. 寄与率は, $\frac{12}{12+8}, \frac{8}{12+8}$. 累積寄与率は, $\frac{12}{12+8}, \frac{12+8}{12+8}$.

L10-Q3

Quiz 解答:主成分分析

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

1. $z_1 = \frac{1}{5\sqrt{2}}(5x_1 - 4x_2 + 3x_3)$, $z_2 = \frac{1}{5}(3x_2 + 4x_3)$, $z_3 = \frac{1}{5\sqrt{2}}(5x_1 + 4x_2 - 3x_3)$.
2. x_1, x_2, x_3 の分散が 1 なので, $\frac{5}{5\sqrt{2}}, \frac{-4}{5\sqrt{2}}, \frac{3}{5\sqrt{2}}$.
3. $z_1 = \frac{1}{50\sqrt{2}}(25 - 12 - 6) = \frac{7}{50\sqrt{2}}$.
4. 寄与率は, $\frac{3/2}{3/2+1+1/2} = \frac{1}{2}, \frac{1}{3}, \frac{1}{6}$. 累積寄与率は, $\frac{1}{2}, \frac{5}{6}, 1$.

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
7 pca_x=pca.transform(df) # 各データ点の主成分得点
8 pca.explained_variance_ # 主成分の分散のリスト
9 pca.explained_variance_ratio_ # 累積寄与率
```

mva-d10-0-pca.ipynb numpy.pca もある

ここまで来たよ

10 主成分分析

11 主成分分析 (2)
- 現実の $p = 10$ 次元データの主成分分析
- 平方和の分解

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 # 行は2 df_std=scaler.fit_transform(df) でまとめられる
```

mva-d10-0-pca.ipynb

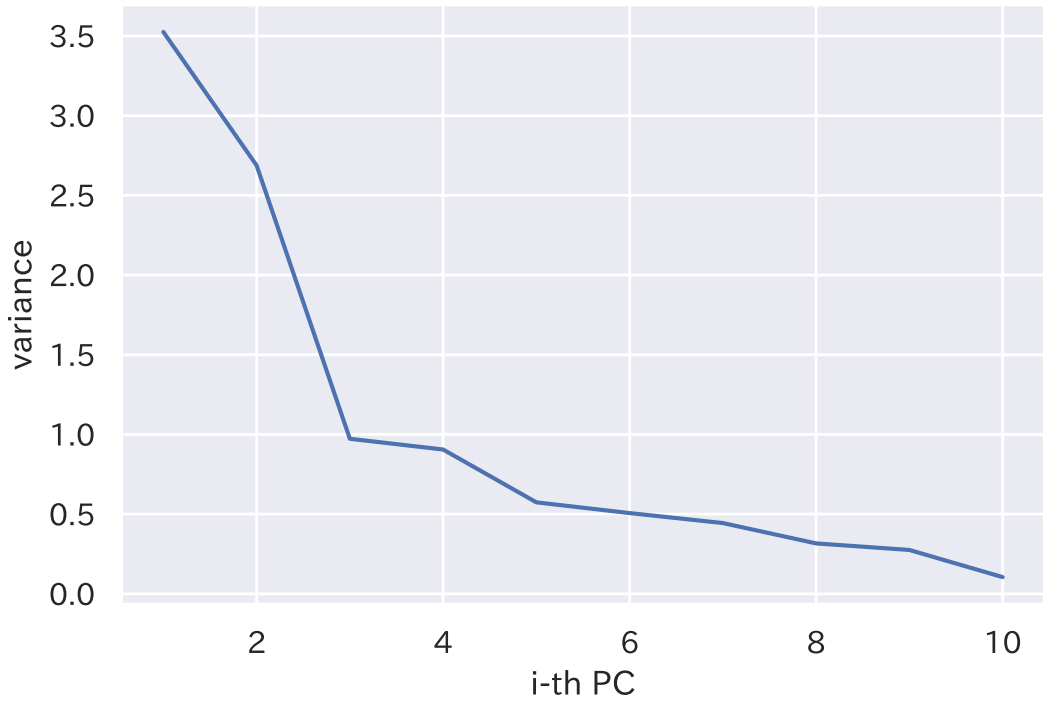
主成分分析の結果

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

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

10 主成分分析

11 主成分分析 (2)
- 現実の $p = 10$ 次元データの主成分分析
- 平方和の分解

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, しばらくその表現で.
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

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

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

L11-Q1

Quiz(平方和)

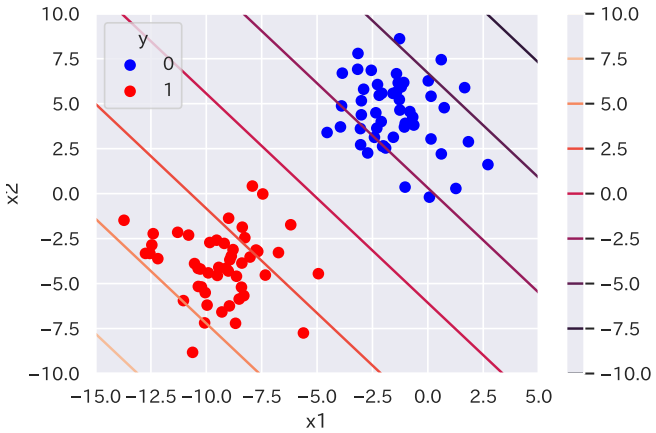
各組の学級で異なる教え方をした. テストの点数 x から学級 y をあてることはできそうだろうか.

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.

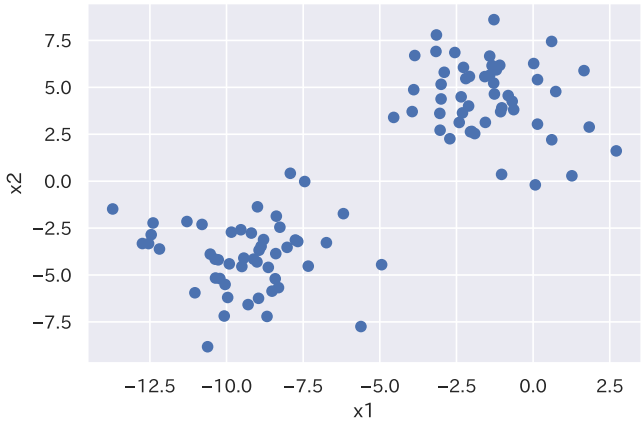
平方和の分解と分類問題

線形判別分析実は, 線形判別分析の Fisher の線形判別関数 z 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

$S_{\mathrm{B}}/S_{\mathrm{W}}$ を大きくなるようにラベル y を振りたい.



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).

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

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. $i = 1, \ldots, n. k = 1, \ldots, p.$
簡単のため, 標準化して標本平均値 $\overline{x}_{\cdot k} = 0$ とする.
平方和の, 各座標の和

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

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

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

$(\sum_i a_{i1})^2$ 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…