L10 主成分分析

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　L10(2021-12-02 Thu)

最終更新: Time-stamp: "2021-12-02 Thu 07:20 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

L09-Q1

Quiz 解答:楕円の主軸

等高線の方程式は,

$$\begin{pmatrix} x_1 - 4 & x_2 + 3 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - 4 \\ x_2 + 3 \end{pmatrix} = C$$

$$\frac{(x_1 - 4)^2}{2^2} + \frac{(x_2 + 3)^2}{3^2} = C$$

これは 中心 $(4, -3)$, x_1 軸に平行な短軸 (短半径 2), x_2 軸に平行な長軸 (長半径 3) の楕円を表す.

L09-Q2

L09-Q3

Quiz 解答:実対称行列の直交行列による対角化

固有値 $\lambda_1 = 10$ に対応する規格化された固有ベクトルのひとつは $\boldsymbol{v}_1 = \frac{1}{5}\begin{pmatrix} 4 \\ 3 \end{pmatrix}$.

固有値 $\lambda_2 = -5$ に対応する規格化された固有ベクトルのひとつは
$\boldsymbol{v}_2 = \frac{1}{5}\begin{pmatrix} 3 \\ -4 \end{pmatrix}$.
よって, 直交行列 $P = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{pmatrix}$ で,

$$ {}^{\mathrm{t}}PAP = \Lambda $$

と対角化される. ただし, 対角行列 $\Lambda = \begin{pmatrix} 10 & 0 \\ 0 & -5 \end{pmatrix}$.

ここまで来たよ

9 共分散行列と固有値固有ベクトル

10 主成分分析
- n 次元正規分布とその等高面
- 主成分分析
- 現実の $p = 10$ 次元データの主成分分析

n 次元正規分布

永田棟方 多変量解析法入門 §2.2(5)

3 次元正規分布の確率密度関数

3 次元正規分布 $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ の確率密度関数は, 確率変数を $\boldsymbol{X} = {}^{\mathrm{t}}(X_1, X_2, X_3)$ とするとき,

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{3/2}\sqrt{\det \Sigma}} \mathrm{e}^{-\frac{1}{2}\,{}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\text{パラメタは, 母平均値 (ベクトル)} \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \mathrm{E}[X_3] \end{pmatrix},$$

$$\text{母共分散行列} \Sigma = \begin{pmatrix} \mathrm{V}[X_1] & \mathrm{Cov}[X_1, X_2] & \mathrm{Cov}[X_1, X_3] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{V}[X_2] & \mathrm{Cov}[X_2, X_3] \\ \mathrm{Cov}[X_3, X_1] & \mathrm{Cov}[X_3, X_2] & \mathrm{V}[X_3] \end{pmatrix}.$$

等高面

等高面 $f(x_1, x_2, x_3) = C$ は, X_1, X_2, X_3 が独立なら ($\Leftrightarrow \Sigma$ が対角行列 $\lambda = \sigma_1^2, \sigma_2^2, \sigma_3^2$ なら)

$$ {}^{\mathrm{t}}(\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = C' $$

$$ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \frac{(x_3 - \mu_3)^2}{\sigma_3^2} = C'. $$

一般に,

n 次元正規分布の確率密度関数 f の等高面

n 次元正規分布の確率密度関数 f の等高面は, n 次元楕円体の表面.

- 長軸の向きは? $\rightsquigarrow \Sigma$ の最大固有値の固有ベクトルの向き
- 長軸と直交する軸のうち, いちばん長い軸の向きは? $\rightsquigarrow \Sigma$ の 2 番目の固有値の固有ベクトルの向き
- 長軸とも次の軸とも直交する軸のうち…

理由

$\boldsymbol{\mu} = \mathbf{0}$ とする.

Σ の固有値を, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$, 対応する単位固有ベクトルを $\boldsymbol{v}_i$ $(i = 1, \ldots, n)$ とする.

Σ^{-1} の固有値は, $0 < \lambda_1^{-1} \leq \lambda_2^{-1} \leq \cdots \leq \lambda_n^{-1}$, 対応する単位固有ベクトルは $\boldsymbol{v}_i$ $(i = 1, \ldots, n)$.

ある等高面上の点を $\boldsymbol{x} = \sum_{i=1}^n a_i \boldsymbol{v}_i$ と書く. $|\boldsymbol{x}|^2 = \sum_{i=1}^n a_i^2$.

$$
\begin{aligned}
\text{等高面 } {}^{\mathrm{t}}\left(\sum_{i=1}^n a_i \boldsymbol{v}_i\right) \Sigma^{-1} \left(\sum_{j=1}^n a_j \boldsymbol{v}_j\right) &= C' \\
\sum_{i=1}^n a_i \, {}^{\mathrm{t}}\boldsymbol{v}_i \sum_{j=1}^n \lambda_j^{-1} a_j \boldsymbol{v}_j &= C' \\
\sum_{i=1}^n \lambda_i^{-1} a_i^2 &= C'
\end{aligned}
$$

つまり, $\boldsymbol{v}_i$ による直交座標 a_i で考えると,

$$
\sum_{i=1}^n \frac{a_i^2}{((\lambda_i)^{1/2})^2} = C'
$$

これは, n 次元楕円体. いちばん長い軸は $\boldsymbol{v}_1$ に平行, 半径 $C'\lambda_1^{1/2}$

平面 $a_1 = 0$ の切り口は $n-1$ 次元楕円体. いちばん長い軸は $\boldsymbol{v}_2$ に平行, 半径 $C'\lambda_1^{1/2}$.

$\vdots$

L10-Q1

Quiz(n 次元正規分布の等高面の主軸)

3 次元正規分布 $\boldsymbol{x} = {}^{\mathrm{t}}(X_1, X_2, X_3) \sim \mathrm{N}(\mathbf{0}, \Sigma)$ を考える.
共分散行列 Σ は, 固有値 $\lambda_i = 4, 9, 25$, 対応する固有ベクトル
$\boldsymbol{v}_i = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} t, \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} t, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} t,$ を持つ $(t \in \mathbb{R}, t \neq 0)$.

1. 確率密度関数のひとつの等高面を考えたとき, 原点からもっとも遠い 2 点を結ぶ向きを求めよう.
2. 確率密度関数のひとつの等高面の式を書こう (行列やベクトルを使った未整理の式でよい).

ここまで来たよ

9 共分散行列と固有値固有ベクトル

10 主成分分析
- n 次元正規分布とその等高面
- 主成分分析
- 現実の $p = 10$ 次元データの主成分分析

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.
まずは, 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.

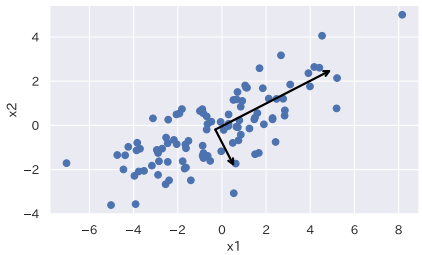
- この問題は教師なし学習
  - 対比:線形判別分析では, $y = 0, 1$ をいちばんよく分ける方向を考えた (教師あり)
- 注意: 単位や意味の違う量を比べてる可能性
- 次元圧縮 多次元のデータを, 情報をなるべく保って低次元のデータに変換する. 1,2,3 次元なら可視化容易. その後で他の手法を利用する.

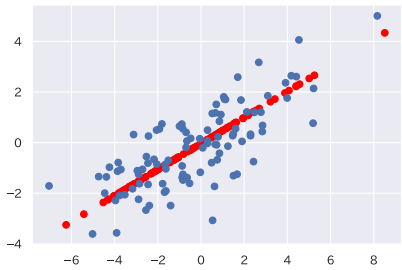
1 次元だけ選ぶとき

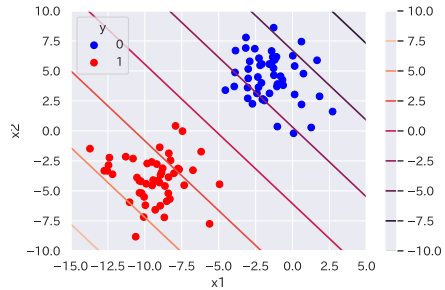
主成分分析では,

- 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 $\boldsymbol{w}$.
- 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$\cdots$ $z = z_1 =\rightsquigarrow z_k$
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.

第 k 主成分

第 k 固有値, 固有ベクトルで, 同じことをやったもの.

$1, \ldots, k-1$ 番目の主成分の向きと直交する範囲で, 不偏標本分散を最大にする $\boldsymbol{w}$.

L10-Q2

Quiz(主成分分析)

2 変量データについて, 不偏標本共分散行列が次のように与えられる.

$$\begin{pmatrix} 9 & -\sqrt{3} \\ -\sqrt{3} & 11 \end{pmatrix}$$

1. 2 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 $(0.5, 0.3)$ 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.

L10-Q3

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.

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

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

9 共分散行列と固有値固有ベクトル

10 主成分分析
- n 次元正規分布とその等高面
- 主成分分析
- 現実の $p = 10$ 次元データの主成分分析

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
```

主成分分析の結果

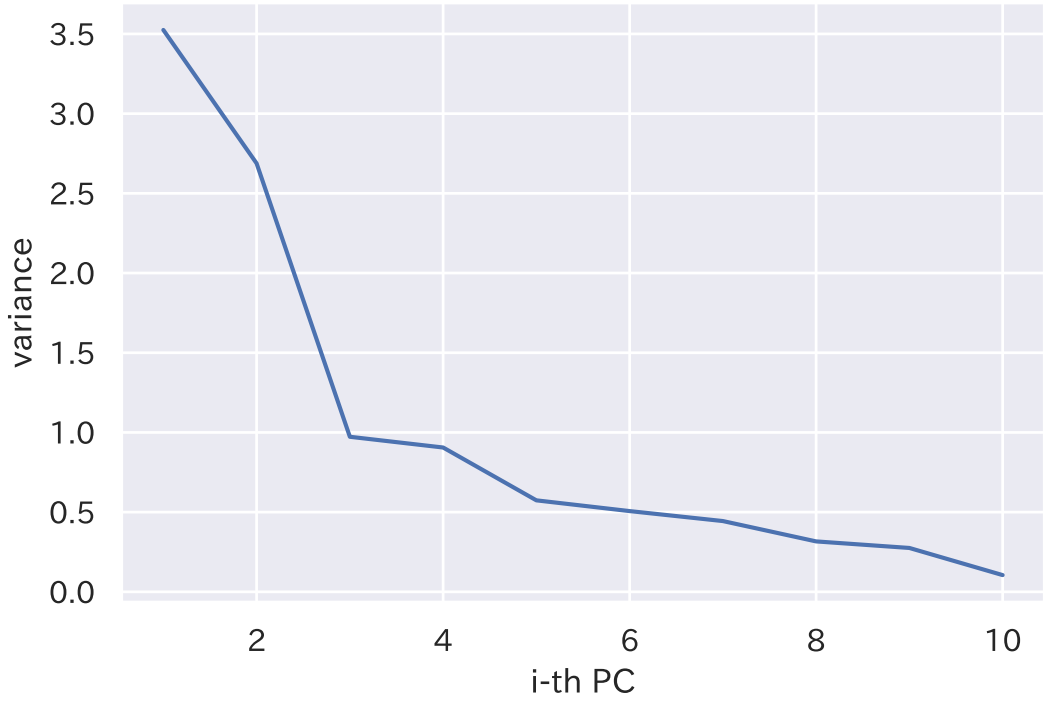
列名	(もとの単位) 種目	PC1	PC2
t100	(s) 100m 走	0.41588233	0.14880813
long	(m) 走り幅跳び	-0.39405149	-0.1520815
poid	(m) 砲丸投げ	-0.26910572	0.48353737
haut	(m) 走り高跳び	-0.21228177	0.0278985
t400	(s) 400m 走	0.35584739	0.35215981
t110	(s) 110m ハードル走	0.43348158	0.0695682
disq	(m) 円盤投げ	-0.17579228	0.50333471
perc	(m) 棒高跳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