L09 主成分分析 (1)

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 L09(2021-11-30 Tue)

最終更新: Time-stamp: "2021-11-29 Mon 22:10 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

L08-Q1

Quiz 解答:クラスター分析

1. 距離 2 で, {{A,B},{C},{D}}. 距離 4 で, {{A,B},{C,D}}. 距離 5 で, {{A,B,C,D}}.
2. クラスター {A,B} の座標の平方偏差和は 2. クラスター {C,D} の座標の平方偏差和は 4. クラスター {A,B,C,D} の座標の平方偏差和は 40. よって, 距離は 34.
3. 2 個の (空でない) クラスターへの分割方法は 7 個あるが, その中で最小となるのは {A,B},{C,D}. 目的関数の値は $2 + 4 = 6$.
4. ステップ 0 でのクラスターは C_1 ={A,B,C},C_2 ={D}. それぞれの重心は $(\frac{8}{3}, \frac{7}{3}), (6, 7)$ ステップ 1 でのクラスターは C_1 ={A,B},C_2={C,D}.

L08-Q2

Quiz 解答:クラスター分析

1. 距離 2 で, {{A,B},{C},{D}}. 距離 4 で, {{A,B},{C,D}}. 距離 5 で, {{A,B,C,D}}.
2. クラスター {A,B} の座標の平方偏差和は 2. クラスター {C,D} の座標の平方偏差和は 4. クラスター {A,B,C,D} の座標の平方偏差和は 40. よって, 距離は 34.
3. 2 個の (空でない) クラスターへの分割方法は 7 個あるが, その中で最小となるのは {A,B},{C,D}. 目的関数の値は $2 + 4 = 6$.
4. ステップ 0 でのクラスターは C_1 ={A,B,C},C_2 ={D}. それぞれの重心は $(\frac{8}{3}, \frac{7}{3}), (6, 7)$ ステップ 1 でのクラスターは C_1 ={A,B},C_2={C,D}.

ここまで来たよ

n 次元正規分布

理論物理学特論 (2021)L02

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,

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

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

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

$\boldsymbol{\mu} = \boldsymbol{0}$ 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 $C'\lambda_1^{1/2}$.

⋮

L09-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).

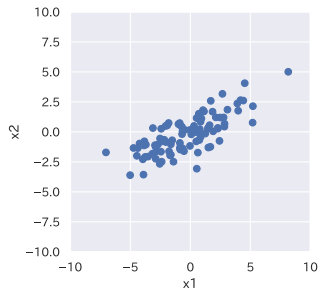
ここまで来たよ

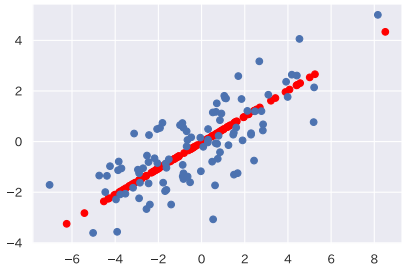
主成分分析 PCA=Principal Component Analysis

(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．その方向は，不偏標本共分散行列 $\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\,{}^{\mathrm{t}}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ の，最大 (第 1) 固有値の固有ベクトルの方向として求められる．その方向の広がり $\simeq \sqrt{\text{第 1 固有値}}$
- 2 次元だけ選ぶとき．その方向は，不偏標本共分散行列 $\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\,{}^{\mathrm{t}}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ の，最大 (第 1) 固有値，第 2 固有値の固有ベクトルの方向として求められる．その方向の広がり $\simeq \sqrt{\text{第 2 固有値}}$





主成分分析の原理

以下 $\overline{\boldsymbol{x}} = \boldsymbol{0}$ に調整済と仮定.
$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_i \boldsymbol{x}_i$: 標本平均値ベクトル ($p$ 次元).
S_{kk}: x_k の不偏標本分散
S_{jk}: x_j と x_k の不偏標本共分散
$p \times p$ 行列 $S = \frac{1}{n-1}\sum_i \boldsymbol{x}_i\, {}^{\mathrm{t}}\boldsymbol{x}_i$: 不偏標本共分散行列.

- 例 $p = 2$, $x_1 = x, x_2 = y$ のとき, $S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix}$.

単位係数ベクトル $\boldsymbol{w} = {}^{\mathrm{t}}\begin{pmatrix} w_1 & \cdots & w_p \end{pmatrix}$, $|\boldsymbol{w}| = 1$ に対して,
$z_i = {}^{\mathrm{t}}\boldsymbol{w}\boldsymbol{x}_i = \sum_k w_k x_{ik}$ を考える.
z の分散 $S_{zz} = \frac{1}{n-1}\sum_i \boldsymbol{w}\boldsymbol{x}_i\, {}^{\mathrm{t}}(\boldsymbol{w}\boldsymbol{x}_i) = {}^{\mathrm{t}}\boldsymbol{w} S \boldsymbol{w}$.
$\boldsymbol{w}$ を調節して, S_{zz} を最大にしたい.
$\rightsquigarrow$ 答: S_{zz} の最大値は共分散行列 S の最大固有値 λ_1, そのときの $\boldsymbol{w}$ は対応する固有ベクトル $\boldsymbol{v}_1$.

共分散行列 S の固有値を $\lambda_1 \geq \lambda_2 \geq \cdot \geq \lambda_p \geq 0$ とする. 対角行列 Λ の上から並べる.
対応する (p 次元) 固有ベクトルを $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ と書く.

気分: $S = P\Lambda\,{}^{\mathrm{t}}P$, ${}^{\mathrm{t}}PSP = \Lambda$ だから, $P \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \boldsymbol{v}_1$ が $\boldsymbol{w}$ としてよさそうじゃない?

第 k 主成分
$1, \ldots, k-1$ 番目の主成分の向きと直交し, その範囲で分散を最大にする係数でつくった z.
$\rightsquigarrow$ 答: S の第 k 固有値と, 対応する固有ベクトル.

第 k 主成分負荷量 (loading) $\lambda_k \boldsymbol{w}_k$ の各成分 (p 個, $1 \leq k \leq p$). ぜんぶ集めると規定変換行列になる.
第 k 主成分得点 (score) データ点 $\boldsymbol{x} = a_1 \boldsymbol{w}_1 + \cdots + a_p \boldsymbol{w}_p$ と書いた係数 a_k.
第 k 主成分の分散 λ_k.
分散の和は, 変数の分散の和, 主成分の分散の和, どちらで見ても同じ.
$\sum_k S_{kk} = \sum_k \lambda_k$. トレースは P, P^{-1} を掛けても同じだから.

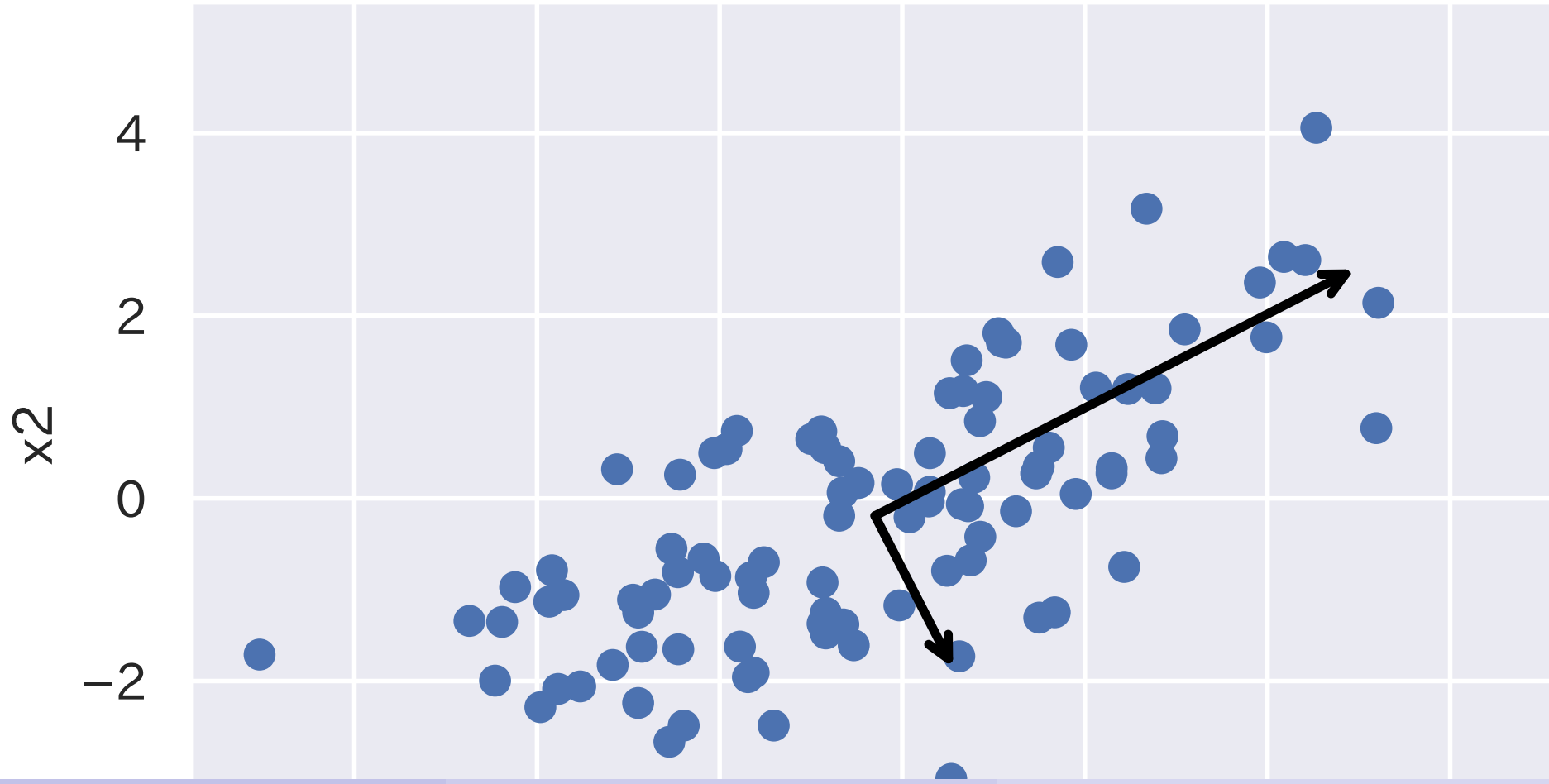
scikit-learn の主成分分析

```
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



L09-Q2

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
$\begin{pmatrix} 5 \\ -4 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix}$

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

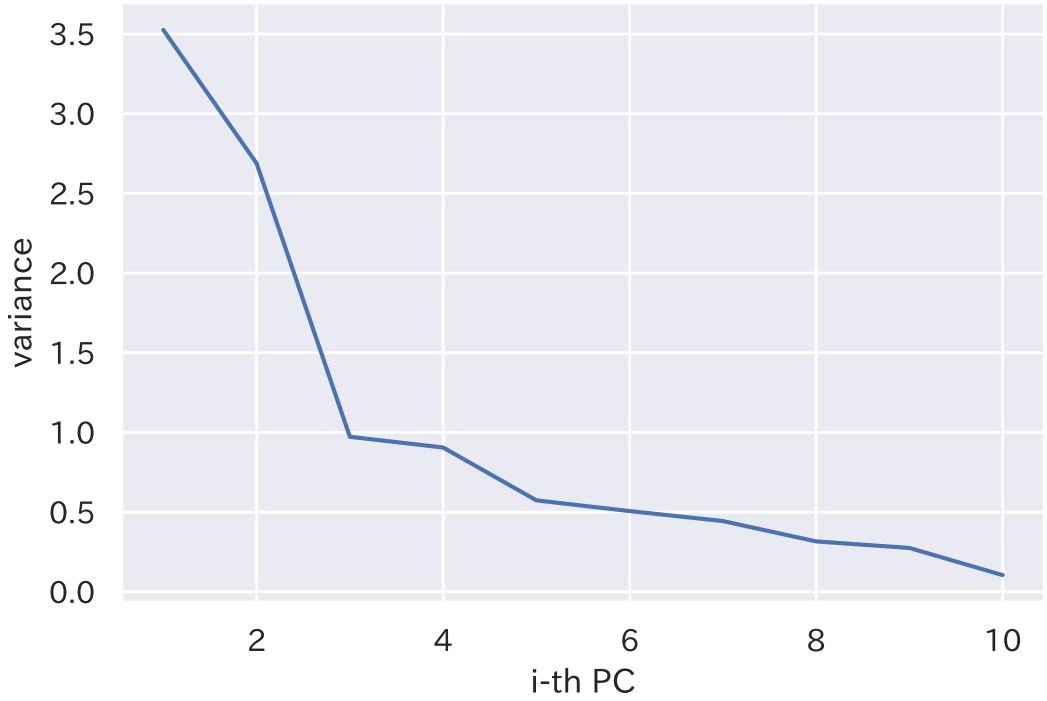
列名	(もとの単位) 種目	PC1	PC2
t100	(s) 100m 走	0.41588233	0.14880813
long	(m) 走り幅跳び	-0.39405149	-0.1520815
poid	(m) 砲丸投げ	-0.26910572	0.48353737
haut	(m) 走り高跳び	-0.21228177	0.0278985
t400	(s) 400m 走	0.35584739	0.35215981
t110	(s) 110m ハードル走	0.43348158	0.0695682
disq	(m) 円盤投げ	-0.17579228	0.50333471
perc	(m) 棒高跳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