主成分分析

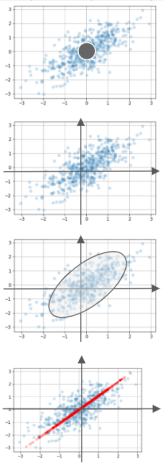
多変量データの持つ構造を**より少数個の指標に「圧縮」**

○変量の個数を減らすことに伴う、情報の損失はなるべく小さくしたい

○少数変数を利用した**分析や可視化**(2・3次元の場合)が実現可能

学習データ	$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{im}) \in \mathbb{R}^m$
平均 (ベクトル)	$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$
データ行列	$\bar{X} = (\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \cdots, \boldsymbol{x}_n - \bar{\boldsymbol{x}})^T \in \mathbb{R}^{n \times m}$
分散共分散行列 (復習用)	$\Sigma = Var(\bar{X}) = \frac{1}{n}\bar{X}^T\bar{X}$
線形変換後のベクトル	$\boldsymbol{s}_j = (s_{1j}, \cdots, s_{nj})^T = \bar{X}\boldsymbol{a}_j \qquad \boldsymbol{a}_j \in \mathbb{R}^m$

※jは射影軸のインデックス

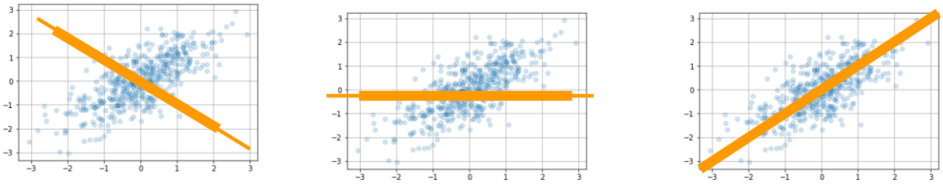


係数ベクトルが変われば線形変換後の値が変化

「情報量の分散の大きさ」ととらえる。

線形変換後の変数の分散が最大となる射影軸を探索

$$\boldsymbol{s}_j = (s_{1j}, \cdots, s_{nj})^T = \bar{X}\boldsymbol{a}_j \qquad \boldsymbol{a}_j \in \mathbb{R}^m$$



線形変換後の分散

$$Var(\boldsymbol{s}_j) = \frac{1}{n}\boldsymbol{s}_j^T\boldsymbol{s}_j = \frac{1}{n}(\bar{X}\boldsymbol{a}_j)^T(\bar{X}\boldsymbol{a}_j) = \frac{1}{n}\boldsymbol{a}_j^T\bar{X}\bar{X}\boldsymbol{a}_j = \boldsymbol{a}_j^T Var(\bar{X})\boldsymbol{a}_j$$

↑↑↑**一番右側のグラフが「分散最大」な射影軸がえられる。。**

よって、以下の制約付き最適化問題を解けばよい。

制約: ajT * aj = 1 ← ノルムが1.

このような制約を入れないと、例えば、(1, 1) (2, 2)など、大きさが違うベクトルが大量に出てきてしまい解けない。

目的関数 $\arg\max_{\boldsymbol{a}\in\mathbb{R}^m} \boldsymbol{a}_j^T Var(\bar{X})\boldsymbol{a}_j$ **制約条件** $\boldsymbol{a}_j^T\boldsymbol{a}_j = 1$

ラグランジュ関数 $E(\boldsymbol{a}_j) = \boldsymbol{a}_j^T Var(\bar{X})\boldsymbol{a}_j - \lambda(\boldsymbol{a}_j^T\boldsymbol{a}_j - 1)$

ラグランジュ乗数

目的関数　　制約条件

ラグランジュの未定乗数法

$g(x, y) = 0$のもとで $f(x, y)$を**最大化**したいという**等式制約つき**の問題において

$L(x, y, \lambda) = f(x, y) - \lambda g(x, y)$を作ると

(α, β)が極値を与える→(α, β)は$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} = \frac{\partial L}{\partial \lambda} = 0$ の解。

寄与率：第1-k主成分においていて保持している情報量の割合：λk / Σ(i=1~k) λi

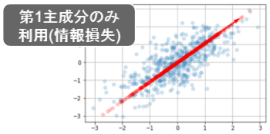
大量データの絞り込みなどに利用。

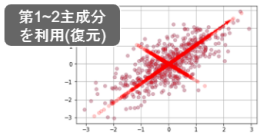
分散共分散行列を計算 → 固有値問題を解く → (最大)m個の固有値と固有ベクトルのペアが出現 → k番目の固有値(昇順)並べ、対応する固有ベクトルを第k主成分と呼ぶ

- 寄与率
  - 第1~元次元分の主成分の分散は、元のデータの分散と一致
    - 2次元のデータを2次元の主成分で表示した時、固有値の和と元のデータの分散が一致
    - 第k主成分の分散は主成分に対応する固有値

$$V_{total} = \sum_{i=1}^{m} \lambda_i$$

元データの総分散　主成分の総分散





  - 寄与率：第k主成分の分散の全分散に対する割合 (第k主成分が持つ情報量の割合)
  - 累積寄与率：第1-k主成分まで圧縮した際の情報損失量の割合

$$c_k = \frac{\lambda_k}{\sum_{i=1}^{m} \lambda_i}$$

第k主成分の分散 / 主成分の総分散

$$r_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{m} \lambda_i}$$

第1~k主成分の分散 / 主成分の総分散