クレジット:
UTokyo Online Educationデータマイニング入門 2018 森　純一郎

東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

データマイニング入門 第8回

2018年度

学習目標

次元削減（縮約）について理解する

共分散行列とその固有値・固有ベクトルについて理解する

主成分分析について理解する

- 射影と分散の最大化について

主成分分析の実装について理解する

次元削減（次元縮約）

一般にデータの特徴量は高次元であり、中には互いに相関がある特徴量やノイズとなる特徴量が存在することもある

データをより低次の特徴量で表現したい

教師なし学習とも考えられる

次元削減

- データの圧縮
- 可視化（高次元データを2次元や3次元に削減）
- 特徴量抽出 など

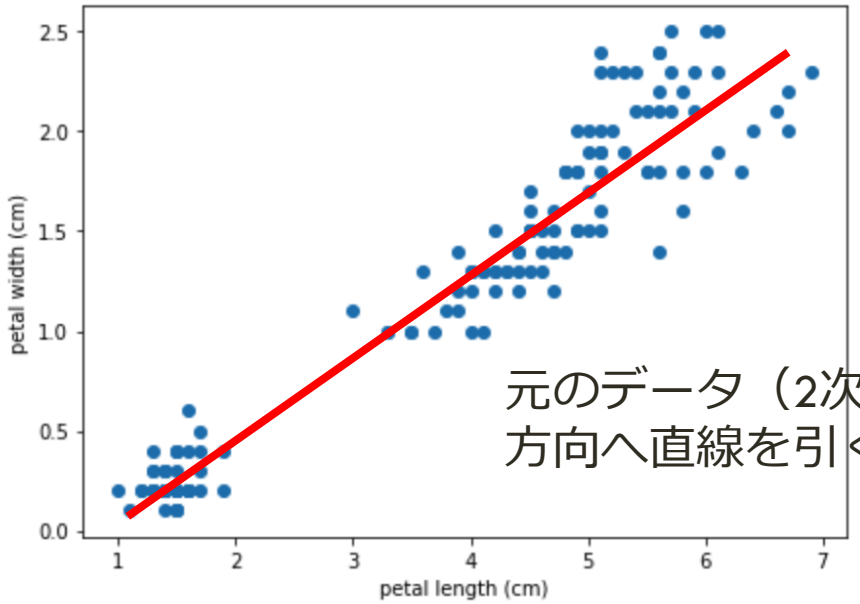
主成分分析

線形な次元削減方法

- データのn個の特徴量について、それらの特徴量をよく表す低次のk個（k<n）の特徴量を見つける

2次元から1次元（直線）への次元削減を考えてみる

- 直感的に次元削減をすると以下の図ではデータの広がっている方向に直線を引くことで2次元から1次元にデータに削減できそう



元のデータ（2次元）が分散している
方向へ直線を引く

復習：共分散

共分散

- 2つの変数の関係性を見るとき使われ以下の特徴を持つ
  - 共分散が正
    - 片方の変数が大きい値をとれば、もう片方も大きい値をとる
  - 共分散が負
    - 片方の変数が大きい値をとれば、もう片方は小さい値をとる
  - 共分散が0
    - データ間に関係性がない
- 具体的には各変数の平均からの差の積を取る

$$(x \text{ と } y \text{ の共分散}) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$(x \text{ の分散}) = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

復習：共分散

分散共分散行列

- 複数の変数において、分散と共分散の一覧を行列の形でまとめたもの
- 2変数においては以下のようにかける

	x	y
x	[xの分散	xとyの共分散]
y	[xとyの共分散	yの共分散]

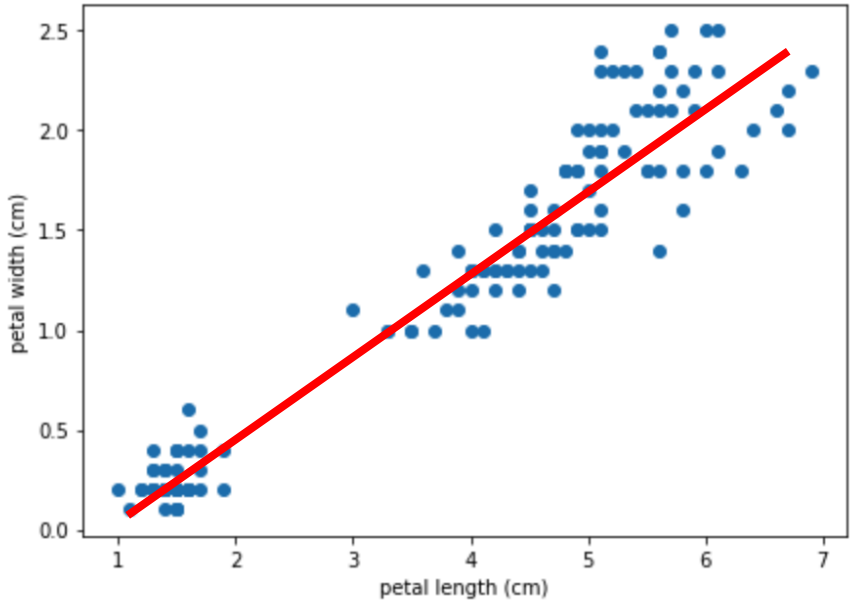
相関行列

	x	y
x	[1	xとyの相関係数]
y	[xとyの相関係数	1]

主成分分析

データの特徴量間の共分散行列は特徴空間におけるデータの分散を表している

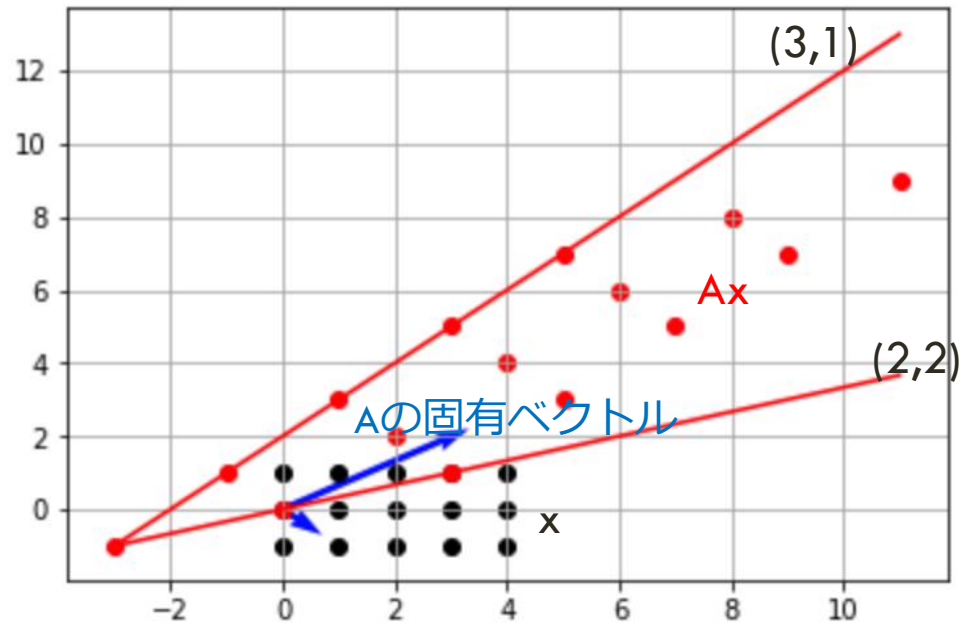
2つの特徴量の間の共分散行列を考えると、データ（2次元）が分散している方向へ引いた直線は、この共分散行列の主固有ベクトルが示す方向へ引いた直線となっている



線形変換と固有値・固有ベクトル

- ベクトルxに行列Aをかけることはxを線形変換（回転や折り返しなど）している
- 標準基底による座標であらわされる点xを、行列の各列を新たな基底とした時の座標点Axに変換している
- 行列Aの固有値λと固有ベクトルvはそれぞれ変換が働く力と向きを表しているとイメージできる

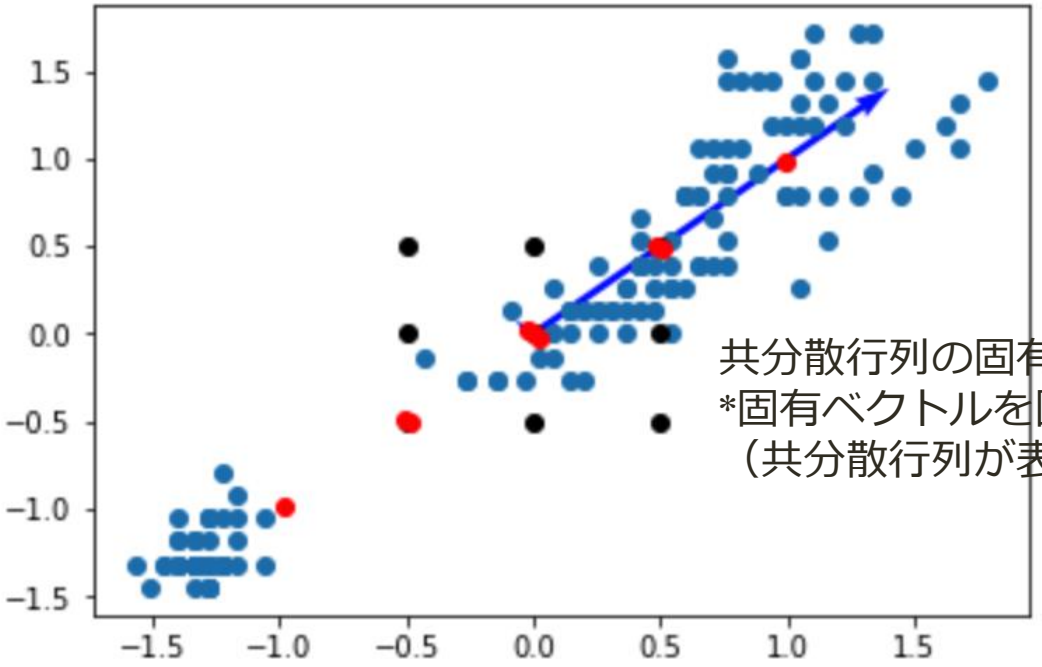
$$A = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$$



共分散行列と固有値・固有ベクトル

任意のベクトルを共分散行列で線形変換することを考えると、この線形変換は共分散行列が表すデータの分散の方向にベクトルを変換していることになる

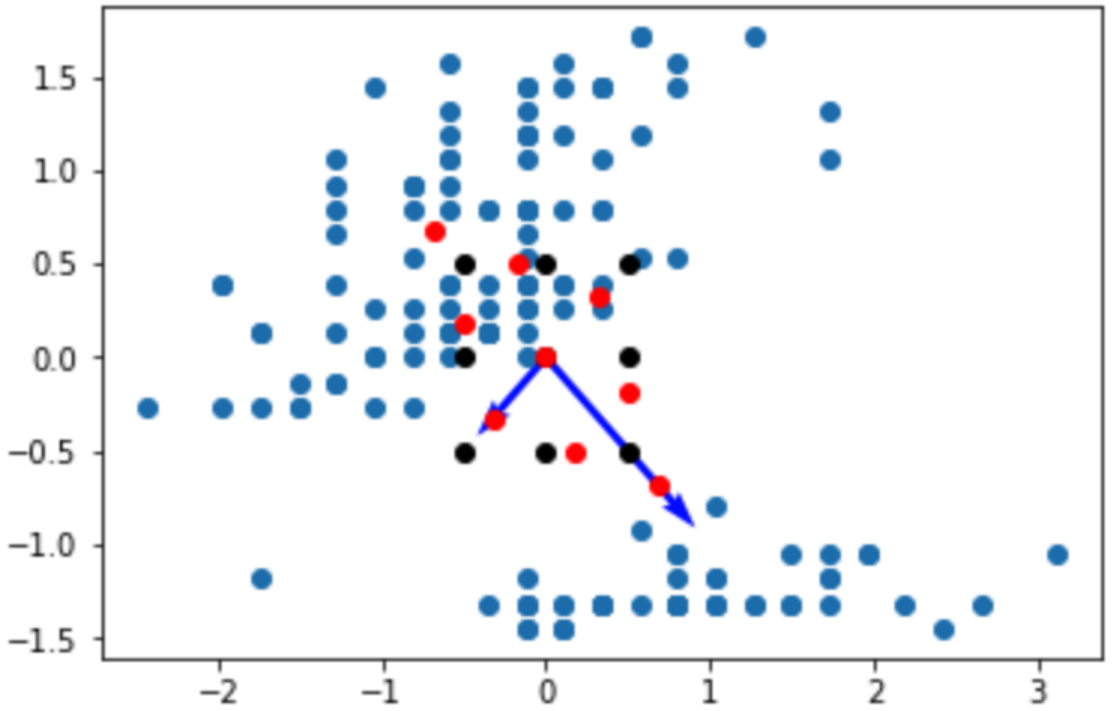
共分散行列の固有ベクトルと固有値はデータの分散方向とその共分散を表している



共分散行列の固有値と固有ベクトル
*固有ベクトルを固有値倍したもの
（共分散行列が表すデータの分散と方向）

共分散行列と固有値・固有ベクトル

データの分散が小さい（特徴量間の相関が小さい）時の共分散行列の
固有値と固有ベクトル（固有ベクトルを固有値倍したもの）

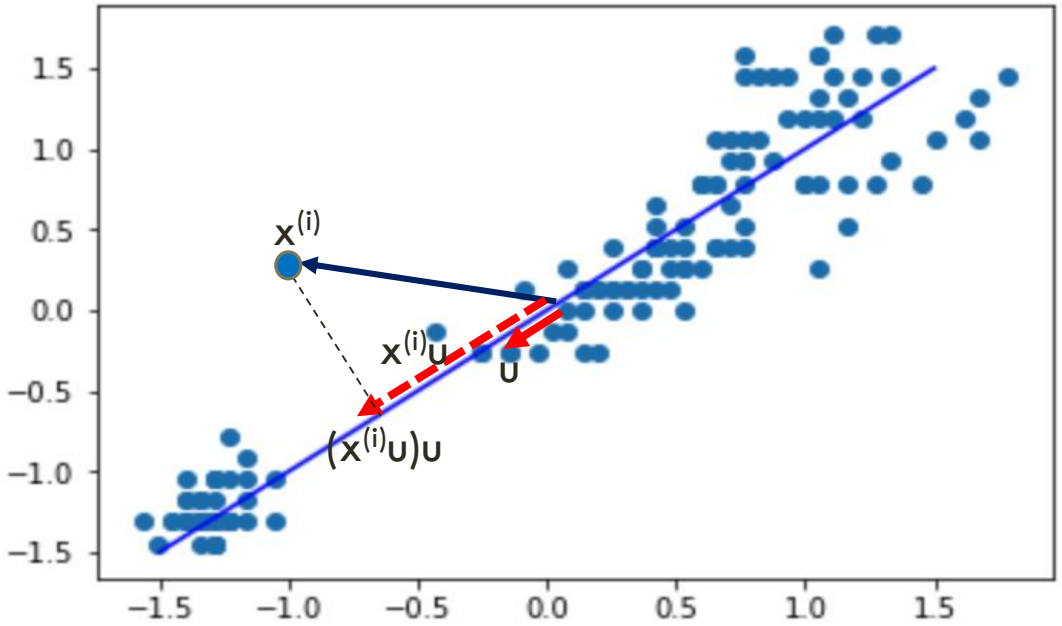


主成分分析

射影と分散の最大化

2次元から1次元への次元削減で考える

- データは各次元（特徴量）ごとに平均0, 分散1に標準化されているとする
  - データ　$x^{(i)} = (x_1^{(i)}, x_2^{(i)})^T$　　直線の単位ベクトル $u = (u_1, u_2)^T$
  - 各データを直線に射影したベクトルの大きさ $x^{(i)T}u$
- 射影後のデータの分散が最大化される（元のデータの分散がなるべく保たれる）ような直線を選びたい
  - $|x^{(i)}-(x^{(i)T}u)u|$ を最小にするような直線を選ぶとも考えられる



主成分分析

分散の最大化

射影後のデータの分散を最大化する（データを射影したベクトルの大きさを最大化する）ような単位ベクトルを求めたい

$$\Sigma_{i=1}^{m}(x^{(i)^T}u)^2/m = \Sigma_{i=1}^{m}(u^T x^{(i)} x^{(i)^T} u)/m$$

$$= u^T(\Sigma_{i=1}^{m} x^{(i)} x^{(i)^T}/m)u = u^T \Sigma_{i=1}^{m} \begin{pmatrix} x_1^{(i)} x_1^{(i)}/m & x_1^{(i)} x_2^{(i)}/m \\ x_2^{(i)} x_1^{(i)}/m & x_2^{(i)} x_2^{(i)}/m \end{pmatrix} u$$

$$= u^T \Sigma u$$

データの次元（特徴量）間の共分散行列

* 共分散行列はΣで表現するが、和を表すΣではないことに注意

主成分分析

分散の最大化

$u^T\Sigma u$を最大化するような単位ベクトルuを求めるため、uの大きさがである1であるという制約より、ラグランジュの未定乗数λを導入する

$$u^T \Sigma u + \lambda(1 - u^T u)$$

uについて微分して勾配が0となるようなuを求めると

$$\Sigma u = \lambda u$$

射影後のデータの分散を最大化する単位ベクトルは共分散行列Σの固有ベクトルとなる

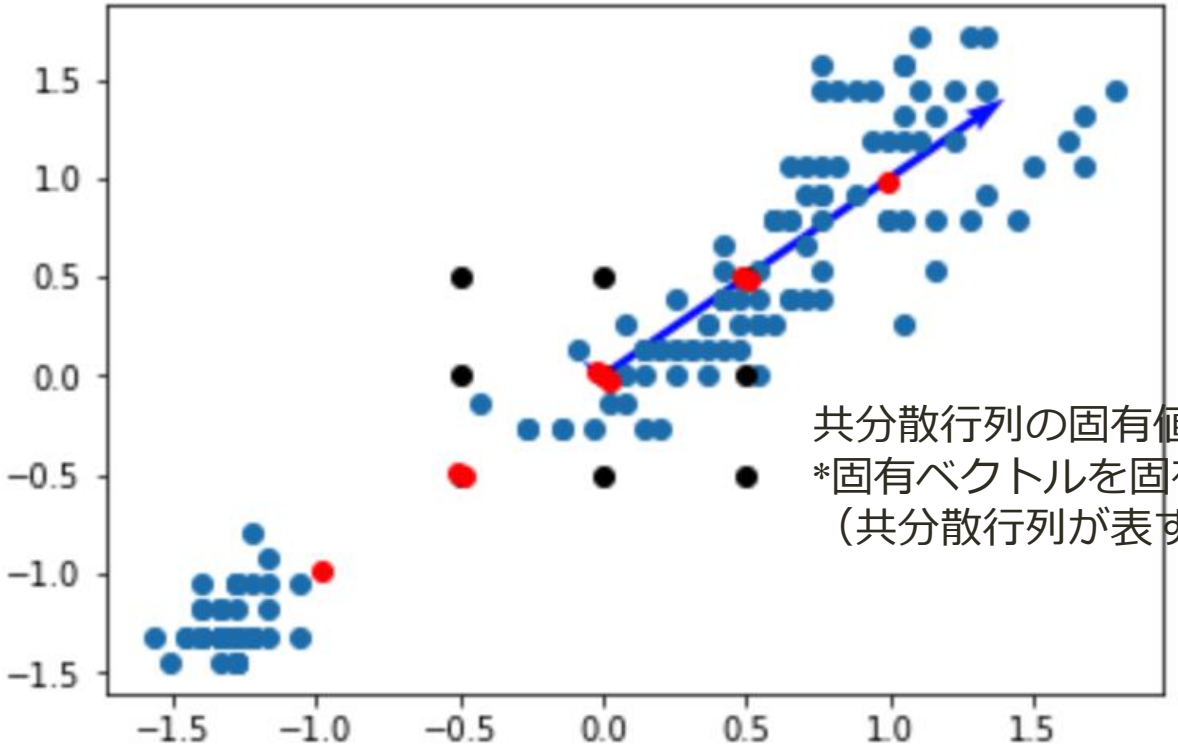
また、 $u^T \Sigma u = \lambda$ より、共分散行列Σの固有値λは射影後のデータの分散となる

射影後のデータの分散を最大化するのは最大固有値に対応する固有ベクトルでありこの固有ベクトルを第1主成分と呼ぶ

2次元から1次元への次元削減では第1主成分を単位ベクトルとする直線を選べばよい

共分散行列と固有値・固有ベクトル

共分散行列の固有ベクトルと固有値はデータの分散方向とその共分散を表している



共分散行列の固有値と固有ベクトル
*固有ベクトルを固有値倍したもの
（共分散行列が表すデータの分散と方向）

主成分分析

正規直交基底

第1主成分vを新たな基底（座標軸）として元のデータを直線に射影したデータは

$$x_{new}^{(i)} = (v^T x^{(i)})v$$

以上は2次元から1次元への次元削減だが、n次元からk次元への次元削減でも同様

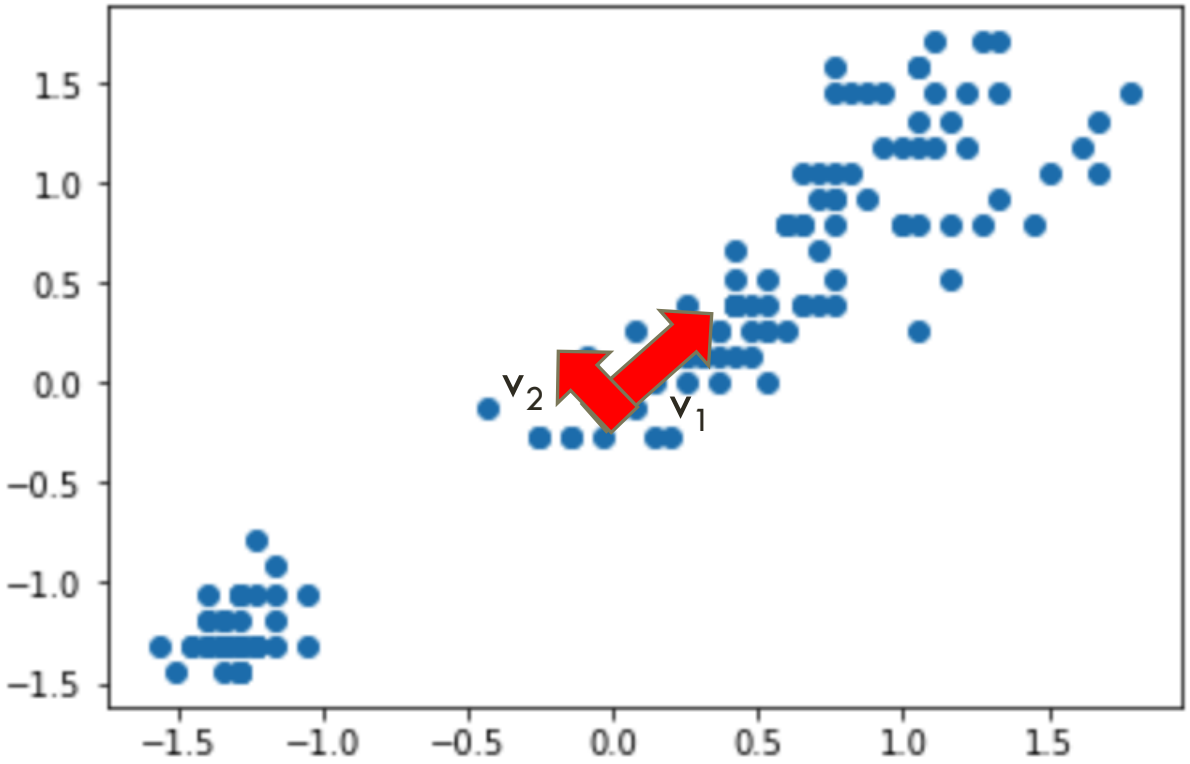
- 元のデータのk次元空間への射影の分散を最大化するには、共分散行列の固有値の大きさに従ってk個の固有ベクトルを選び、それらを新たな基底する
- 固有ベクトル$v_1, v_2, .., v_k$を新たな基底として元のデータをk次元空間で表したデータの座標点は

$$x_{new}^{(i)} = (v_1^T x^{(i)}, v_2^T x^{(i)}, \ldots, v_k^T x^{(i)})^T$$

  - 共分散行列の固有ベクトルは互いに直交するので正規直交基底

正規直交基底

固有ベクトルを基底とする正規直交基底



主成分分析の手順

1. データの各特徴量を標準化する（各特徴量のスケールを合わせる）
   各特徴量の各値についてその特徴量の平均を引き、その特徴量の標準偏差で割る
2. 特徴量間の共分散行列を作成する
3. 共分散行列の固有値と固有ベクトルを計算する
4. k次元に削減したい場合は、固有値が大きい順に対応する固有ベクトルをk個（k個の主成分）選び、それらを新たな直交基底とする
5. これらの直交基底を用いて元のデータを次元削減された低次の次元におけるデータに変換する

削減する次元数

累積寄与率

kの固有ベクトル（主成分）までの固有値（分散）の和が全部の固有値（分散）の総和に占める割合

$$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n} \lambda_i$$

共分散行列の固有値λ（各主成分方向のデータの分散）を用いて、各主成分において元のデータの分散がどれぐらい反映されているかを評価する

累積寄与率が閾値（例えば90%など）より大きくなるような次元数kを選ぶとよい

