主成分分析 (PCA)

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(2026-11-26 Thu)

最終更新: Time-stamp: "2026-11-27 Fri 11:29 JST hig"

今日の目標

- 特徴量の個数が小さい時に主成分分析をて計算で実行できる
- 主成分分析の様々な量の定義と意味が説明できる

多変量解析の手法の分類

教師あり

タイプ	教師 Y	手法・モデル
回帰	連続値	線形単回帰 L02 線形重回帰 L03,L04,L05
	非負整数	ポアソン回帰
分類	2 値（ラベル）	ロジスティック回帰 L06
	多値（ラベル・カテゴリ）	ベイズ分類 L07 QDA,LDA L07

教師なし

タイプ	手法・モデル
クラスタリング	モデルベース L08 k-means 法 L09 階層的クラスタリング L09
次元圧縮	主成分分析 (PCA) L10,L11
潜在構造の推定	因子分析 L11

時系列解析 L12 or L13–

とりあえず別世界

ここまで来たよ

10. 主成分分析 (PCA)
    - 2 次元正規分布とその等高面
    - 主成分分析

一般の 2 次元正規分布 多変量解析☆演習 ()L2026 03

$$f(x,y) = C \times \mathrm{e}^{-ax^2-by^2+cxy+px+qy}$$

定義 (一般の 2 次元正規分布の同時確率密度関数)

$\boldsymbol{Y} = \binom{X}{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ とは，確率密度関数が次のようになること．

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}$$

$$\text{母平均値 (ベクトル)}\boldsymbol{\mu} = \begin{pmatrix} \mu_{\mathrm{X}} \\ \mu_{\mathrm{Y}} \end{pmatrix} \overset{\text{要証明}}{=} \begin{pmatrix} \mathrm{E}[X] \\ \mathrm{E}[Y] \end{pmatrix}$$

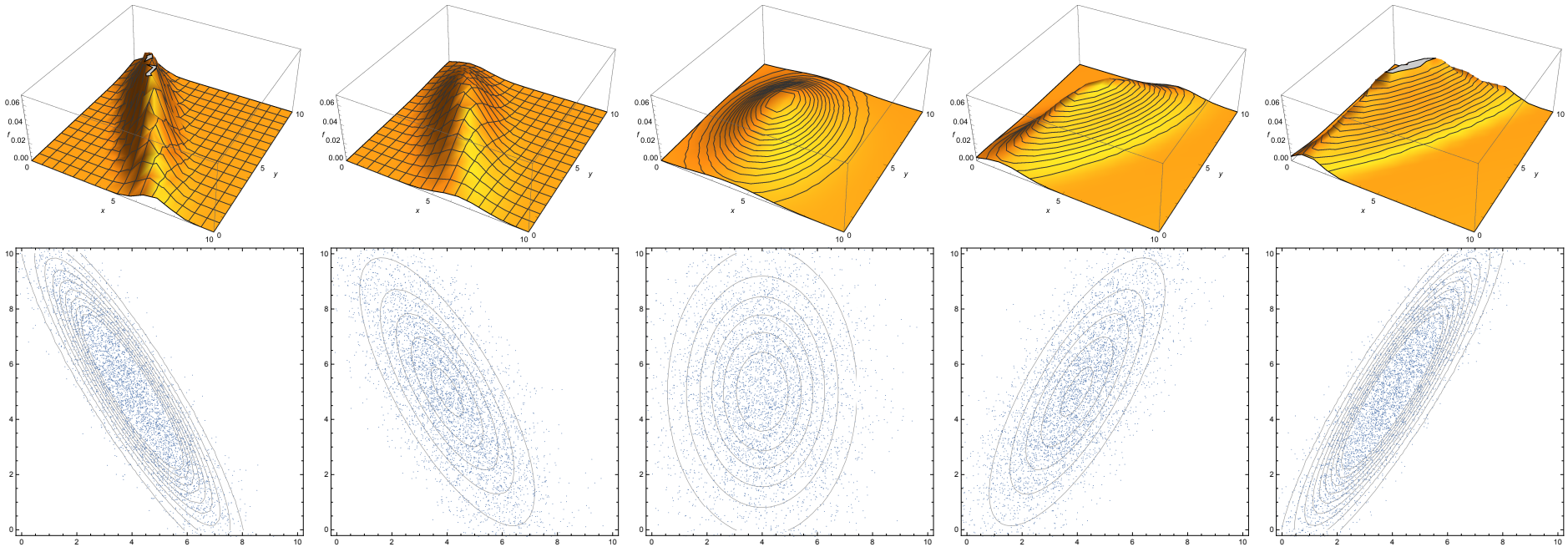
$$\text{母共分散行列}\Sigma = \begin{pmatrix} \sigma_{\mathrm{X}}^2 & \sigma_{\mathrm{XY}} \\ \sigma_{\mathrm{XY}} & \sigma_{\mathrm{Y}}^2 \end{pmatrix} \overset{\text{要証明}}{=} \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix} = \mathrm{Cov}(\boldsymbol{Y})$$

- たしかに正規分布
- $\mathrm{Cov}(X,Y) = 0$ のとき独立なケースに帰着．
- 平面上の重積分が収束する必要 $\rightsquigarrow$ Σ^{-1} の固有値 石村畑 §6.3 はすべて正である必要．

母相関係数 r, 共分散行列 Σ. $c = \frac{9\sqrt{21}}{10}$

$r = -0.90 \qquad -0.76 \qquad 0 \qquad +0.76 \qquad +0.90$

$$\Sigma = \begin{pmatrix} 3 & -c \\ -c & 7 \end{pmatrix}, \begin{pmatrix} 3 & -2\sqrt{3} \\ -2\sqrt{3} & 7 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix}, \begin{pmatrix} 3 & +2\sqrt{3} \\ +2\sqrt{3} & 7 \end{pmatrix}, \begin{pmatrix} 3 & +c \\ +c & 7 \end{pmatrix}$$



2 次元正規分布の確率密度関数の等高線

簡単のため $\boldsymbol{\mu} = 0$ とおくと，

$$\boldsymbol{y}^\top \Sigma^{-1} \boldsymbol{y} = C$$
$$ax^2 + by^2 + 2cxy = C$$

実対称行列 Σ の固有値を $\lambda_1 > \lambda_2 > 0$, 対応する規格化された固有ベクトルを $\boldsymbol{w}_1, \boldsymbol{w}_2$ とすると，直交行列 $P = (\boldsymbol{w}_1\ \boldsymbol{w}_2)$ で $P^\top \Sigma P = D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ と対角化される．

$$(P^{-1}\boldsymbol{y})^\top D^{-1} (P^{-1}\boldsymbol{y}) = C$$

$P^{-1}\boldsymbol{y}$ の長さが最大になるのは，$P^{-1}\boldsymbol{y} = \sqrt{C}\sqrt{\lambda_1}\binom{1}{0}$. すなわち，
$\boldsymbol{y} = \sqrt{C}\sqrt{\lambda_1} P \binom{1}{0} = \sqrt{C}\sqrt{\lambda_1}\boldsymbol{w}_1$
等高線が中心からいちばん遠い方向（長軸）第 1 固有ベクトル $\boldsymbol{w}_1$ の方向．長さは $\sqrt{\lambda_1}$ に比例
等高線が中心にいちばん近い方向（短軸）第 2 固有ベクトル $\boldsymbol{w}_2$ の方向．長さは $\sqrt{\lambda_2}$ に比例

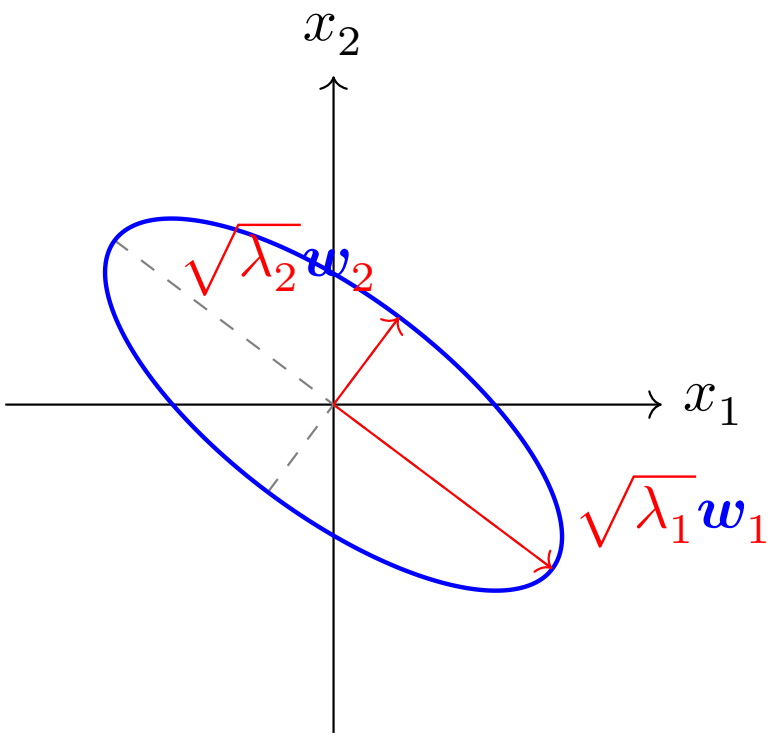
直交射影（正射影 石村畑 p.206）

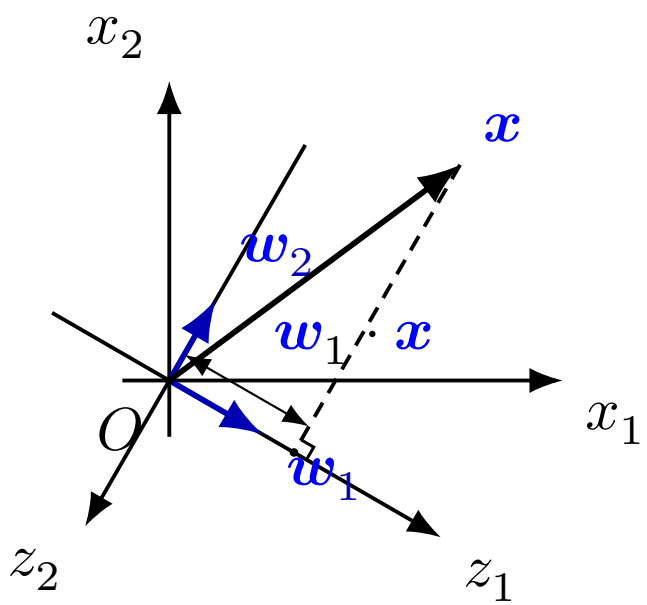
ベクトル $\boldsymbol{x}$ の，z_1, z_2 軸での座標を求めたい．垂線の足までの長さを読み取ればよい．

$z_1 = \boldsymbol{x} \cdot \boldsymbol{w}_1 = |\boldsymbol{x}| 1 \cos\theta.$

$z_2 = \boldsymbol{x} \cdot \boldsymbol{w}_2 = |\boldsymbol{x}| 1 \cos(\frac{1}{2}\pi - \theta).$

ベクトル $(\boldsymbol{x} \cdot \boldsymbol{w}_1)\boldsymbol{w}_1$ のことを，$\boldsymbol{x}$ の $\boldsymbol{w}_1$ への直交射影という．

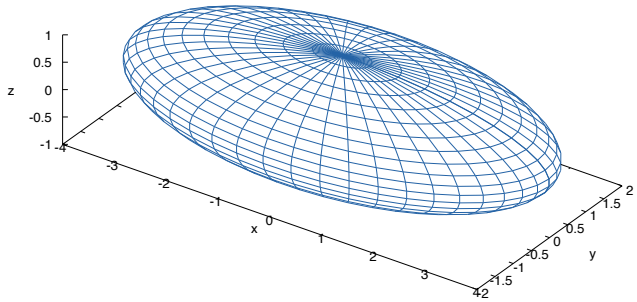




楕円面 (ellipsoid)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

で定義される曲面は，単位球面を 3 方向に，a, b, c 倍に膨らませたもの（ティッシュペーパーの箱の角をとったもの）．内部を楕円体という．



$\boldsymbol{y} \in \mathbb{R}^3$ として，等高面

$$\boldsymbol{y}^\top \Sigma^{-1} \boldsymbol{y} = C$$

は楕円面になる．
直交する 3 軸（長さ $a > b > c > 0$) の方向は，Σ の 3 個の固有ベクトルに対応する．
1 番目 3 番目は最長最短の方向
2 番目の軸の意味は？ → 長軸に直交する平面で切った断面として得られる楕円の長軸．「2 番目に長い方向」

n 次元正規分布とその等高面

一般に,

命題 (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.

- 最長軸の向きは？ 長半径は？ $\rightsquigarrow \Sigma$ の最大固有値 λ_1 の固有ベクトル $\boldsymbol{w}_1$ の向き．半径 $\sqrt{C'\lambda_1}$.
- 最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 λ_2 の固有ベクトル $\boldsymbol{w}_2$ の向き．半径 $\sqrt{C'\lambda_2}$.
- 最も長軸とも次の軸とも直交する軸のうち… $\boldsymbol{w}_3$

L10-Q1

Quiz(n 次元正規分布の等高面の主軸)

3 次元正規分布 $\boldsymbol{x} = (X_1, X_2, X_3)^\top \sim \mathrm{N}(0, \Sigma)$ 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
$\boldsymbol{v}_i = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} t, \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} t, \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} t,$ 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).

略解

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

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \\ -1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} 1/4 & & \\ & 1/9 & \\ & & 1/25 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & 2/\sqrt{6} \\ -1/\sqrt{3} & -1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

すなわち,

$$\frac{(\frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2)^2}{2^2} + \frac{(\frac{1}{\sqrt{6}}x_1 + \frac{1}{\sqrt{6}}x_2 + \frac{2}{\sqrt{6}}x_3)^2}{3^2} + \frac{(-\frac{1}{\sqrt{3}}x_1 - \frac{1}{\sqrt{3}}x_2 + \frac{1}{\sqrt{3}}x_3)^2}{5^2}$$

ここまで来たよ

10. 主成分分析 (PCA)
    - 2 次元正規分布とその等高面
    - 主成分分析

主成分分析 PCA=Principal Component Analysis

- 前提 確率モデルや分布は仮定しない
- 入力 $(x_{i1}, x_{i2}, \dots, x_{ip})$ $(i = 1, \dots, n)$ 重回帰や判別分析に似てるけど正解 y はない．教師なし学習
- 出力 変数変換の式 $z_1 = \boldsymbol{w}_1 \cdot \boldsymbol{x}, z_2 = \boldsymbol{w}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{w}_k \cdot \boldsymbol{x}$. z は，入力の特徴を残しつつ，個数 $m < p$. 次元圧縮 $k = 2, 3$ なら目で見える．
- この式を使って，個々のデータ点に対する $(z_{i1}, \dots, z_{im})$ も出力できる

特徴を残すとは？

x_{ij} のうち，あまり変化のない方向の $x_{\bullet j}$ を捨てれば？

あまり変化のない方向=楕円面の薄い方向は傾いてるかもしれない

(小さい誤差を無視すると) 実質的に p 次元空間の $0, 1, \dots, p-1$ 次元部分空間に分布していることがありうる．このとき，部分空間の座標だけで考えたい．

特徴を残すとは楕円面の長い方向だけ残すことなのでは？

$\overline{x}_{\bullet j} = 0$ とする（あらかじめ中心化されているとする）.

標準化が必要な局面もある

命題 (分散の最大化)

データ点 $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ に対して，$|\boldsymbol{w}| = 1$ であるようなベクトル $\boldsymbol{w}$ で，n 個の実数データ $z_i = \boldsymbol{w}_1 \cdot \boldsymbol{x}_i$ の不偏標本分散が最大になるようなベクトル $\boldsymbol{w}_1$ は，不偏標本共分散行列 S の，最大固有値に対応する固有ベクトルである.

そのような z_1 を第 1 主成分 principal component という.

命題 (第 1 主成分と直交する範囲での分散の最大化)

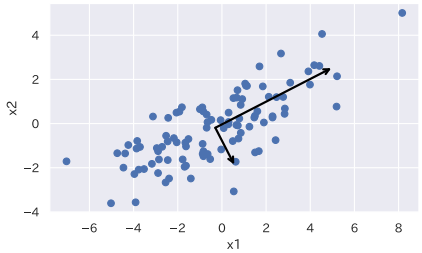
データ点 $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ で，$|\boldsymbol{w}| = 1$, $\boldsymbol{w}_1 \cdot \boldsymbol{w} = 0$ であるようなベクトル $\boldsymbol{w}$ で，n 個の実数データ $z_i = \boldsymbol{w}_2 \cdot \boldsymbol{x}_i$ の不偏標本分散が最大になるようなベクトル $\boldsymbol{w}_2$ は，不偏標本共分散行列 S の，第 2 固有値に対応する固有ベクトルである.

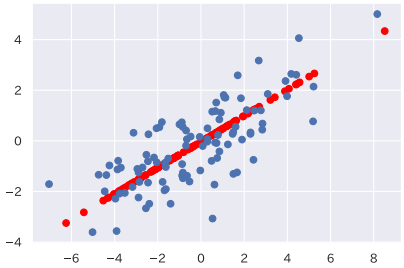
そのような z_2 を第 2 主成分という.

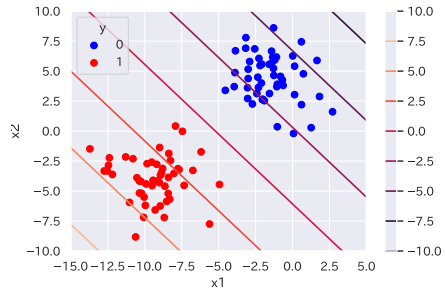
同様に，第 $1, \ldots, p$ 主成分を考えることができる

[

主成分分析] 第 $1, \ldots, m (< p)$ 主成分までを残し，残りを無視して次元圧縮するのが主成分分析 PCA である．







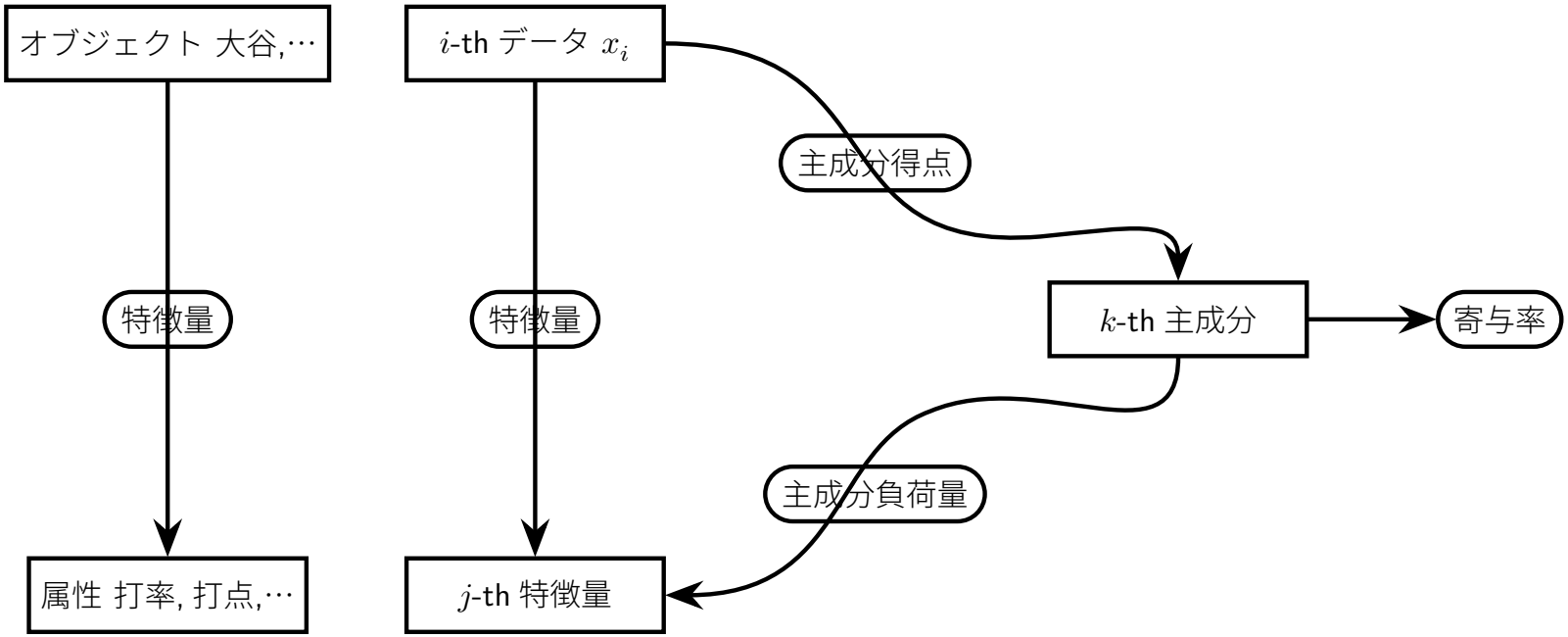
主成分分析の性質

- 第 k 主成分の方向は，不偏標本共分散行列の第 k 固有値（大きい方から k 番目の固有値）に対応する単位固有ベクトル $\boldsymbol{w}_k$ の方向.
- 第 k 主成分 $z = \boldsymbol{w}_k{}^\top \boldsymbol{x}$. $\boldsymbol{w}_k$ に平行な軸の座標. その分散は λ_k くらい.
- データ点 $\boldsymbol{x}_i$ の，第 k 主成分得点 (score) $\boldsymbol{w}_k{}^\top \boldsymbol{x}_i$. データ点 $\boldsymbol{x}_i$ の, k 番目の座標.
- 第 k 主成分の，特徴量 j の主成分負荷量 (loading) ($\boldsymbol{w}_k$ の第 j 成分)$\times \sqrt{\lambda_k}/\sqrt{S_{jj}}$.
  - (x_{ij}, z_{ik}) の相関係数 $(i = 1, \dots, n)$ に等しい (要証明).
- 第 k 主成分の寄与率 $\frac{\lambda_k}{\sum_{k=1}^{p} \lambda_k}$. 第 k 主成分が $\boldsymbol{x}_i$ 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 1 から m 主成分までの累積寄与率 $\frac{\sum_{k=1}^{m} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$. 第 $1, \dots, m$ 主成分の分散の和は，$\boldsymbol{x}_i$ の分散のどのくらいの部分を説明するか.

用語の整理

主語 — 関係を表す量 → 相手の量

主語の量のに属する，相手の量との，関係を表す量



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 主成分のすべての主成分負荷量を求めよう.
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.

略解

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. $\frac{\sqrt{12}}{\sqrt{9}} \times \frac{1}{2}, \frac{\sqrt{12}}{\sqrt{11}} \times \frac{-\sqrt{3}}{2}$,
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(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 主成分のすべての主成分負荷量を求めよう.
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.

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

略解

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, $\sqrt{3/2} \cdot \frac{5}{5\sqrt{2}}, \sqrt{3/2} \cdot \frac{-4}{5\sqrt{2}}, \sqrt{3/2} \cdot \frac{3}{5\sqrt{2}}$.
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, $\frac{1}{2}, \frac{5}{6}, 1$.