L09 共分散行列と固有値固有ベクトル

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　L09(2021-11-18 Thu)

最終更新: Time-stamp: "2021-11-26 Fri 08:35 JST hig"

今日の目標

- 2 次形式から楕円の長軸短軸を求められる
- n 次元正規分布の確率密度関数の等高面の主軸を求められる
- 平方和を群内と群間に分解できる

L07-Q1

Quiz 解答:混合ガウス分布のナイーブベイズ

$$P(Y=0|X=1)=\frac{\frac{1}{(2\pi 2^2)^{1/2}}\mathrm{e}^{-\frac{(1-2)^2}{2\cdot 2^2}}\cdot\frac{3}{10}}{\frac{1}{(2\pi 2^2)^{1/2}}\mathrm{e}^{-\frac{(1-2)^2}{2\cdot 2^2}}\cdot\frac{3}{10}+\frac{1}{(2\pi(1/2)^2)^{1/2}}\mathrm{e}^{-\frac{(1-6)^2}{2\cdot(1/2)^2}}\cdot\frac{7}{10}}$$

L07-Q2

Quiz 解答:混合ガウス分布の線形判別関数

$z=\frac{2-6}{2}\cdot\frac{x-4}{2}$.

L07-Q3

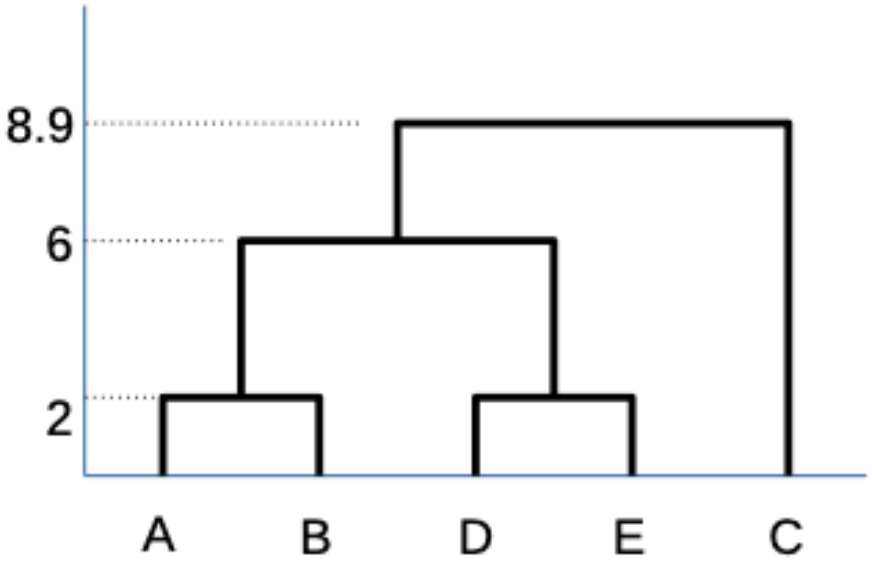
Quiz 解答:2 次元混合ガウス分布の線形判別関数

$$z={}^{\mathrm{t}}\left(\Sigma^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\right)\boldsymbol{x}={}^{\mathrm{t}}\left(\begin{pmatrix}4^{-1}&0\\0&6^{-1}\end{pmatrix}\begin{pmatrix}-2\\-6\end{pmatrix}\right)\begin{pmatrix}x_1\\x_2\end{pmatrix}=-\frac{1}{2}x_1-x_2.$$

L08-Q1

Quiz 解答:階層的クラスター分析

1. 距離 2 で, A,B が合併. 距離 2 で D,E が合併. . これら 2 つのクラスターの代表点は $(10, 3)$ と $(10, 9)$.
2. 距離 6 で, $\{$A,B$\}$,$\{$D,E$\}$ が合併. このクラスターの代表点は $(10, 6)$.
3. 距離 $((10-2)^2+(6-10)^2)^{1/2} = 4\sqrt{5}$ で, $\{$A,B,D,E$\}$,C が合併. このクラスターの代表点は $(\frac{42}{5}, \frac{34}{5})$.



L08-Q2

Quiz 解答:非階層的クラスター分析 (k-means)

- $(1,0)$ を代表点とするクラスター 1: $\{$A$\}$, $(3,3)$ を代表位置とするクラスター 2: $\{$B,C,D$\}$.
- 更新後の代表点 クラスター 1: $(1,1)$, クラスター 2: $(\frac{1+5+5}{3}, \frac{3+1+1}{3})$
- クラスター 1: $\{$A,B$\}$ クラスター 2: $\{$C,D$\}$. (B とクラスター 1 の距離 2, B とクラスター 2 との距離 2.75 なので.

ここまで来たよ

7. ナイーブベイズ・線形判別分析
8. クラスター分析
9. 共分散行列と固有値固有ベクトル
   - 固有値固有ベクトルと楕円の主軸
   - 対称行列の直交行列による対角化
   - n 次元正規分布とその等高面
   - 平方和の分解

一般の 2 次元正規分布

多変量解析☆演習 (2021)L2

一般の 2 次元正規分布の同時確率密度関数

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

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{2/2}\sqrt{\det \Sigma}} \mathrm{e}^{-\frac{1}{2}{}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$\text{母平均値 (ベクトル)}\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \end{pmatrix}$$

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

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

L09-Q1

Quiz(楕円の主軸)

2 次元正規分布 $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ の確率密度関数の等高線の方程式を求めよう.
ただし,

$$\boldsymbol{\mu} = \begin{pmatrix} 4 \\ -3 \end{pmatrix}, \Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}.$$

L09-Q2

Quiz(楕円の主軸)

$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ の方程式

$$ {}^{\mathrm{t}}\boldsymbol{x}\Sigma^{-1}\boldsymbol{x} = C, \quad \Sigma = \begin{pmatrix} 14 & -2 \\ -2 & 11 \end{pmatrix} $$

(C は正の定数) に対して, 長軸, 短軸に平行なベクトルをそれぞれ求めよう. 長半径, 短半径をそれぞれ求めよう.

数式処理ソフトウェア Maple

数値ばかりでなく，変数を用いた形式的計算 (Python でいえば sympy 相当) に優れている
学内ネットワーク (ru-wifi) 接続時のみ，学習目的のみ利用可能．Windows or Mac.
Web サイト (Teams で URL を共有) からダウンロードしてインストール．

Quiz 解答:楕円の主軸

Σ は固有値 $\lambda_1 = 15$ と対応する固有ベクトル $\boldsymbol{v}_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} t$, 固有値 $\lambda_2 = 10$ と対応する固有ベクトル $\boldsymbol{v}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} t$ を持つ $(t \in \mathbb{R}, t \neq 0)$.
よって,

$$\Sigma = \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} 15 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix},$$

$$\Sigma^{-1} = \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} \frac{1}{15} & 0 \\ 0 & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix}.$$

これは, x_1, x_2 軸を $\begin{pmatrix} 2/\sqrt{5} & -1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix}$ だけ回転すると, 主軸が x_1, x_2 軸に平行になることを意味する.
よって式は,

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} \frac{1}{15} & 0 \\ 0 & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = C$$

$$\frac{(\frac{2}{\sqrt{5}}x_1 - \frac{1}{\sqrt{5}}x_2)^2}{15} + \frac{(\frac{1}{\sqrt{5}}x_1 + \frac{2}{\sqrt{5}}x_2)^2}{10} = C$$

よって, 長軸は $\boldsymbol{v}_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$ に平行で, 長さは $2\sqrt{15C}$, 短軸は $\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ に平行, 長さは $2\sqrt{10C}$. 中心は $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

ここまで来たよ

7. ナイーブベイズ・線形判別分析
8. クラスター分析
9. 共分散行列と固有値固有ベクトル
   - 固有値固有ベクトルと楕円の主軸
   - 対称行列の直交行列による対角化
   - n 次元正規分布とその等高面
   - 平方和の分解

実対称行列 (線形代数の復習)

固有値固有ベクトルの定義

n 次正方行列 A に対して,

$$A\boldsymbol{x} = \lambda\boldsymbol{x}$$

を満たす数 λ を A の固有値 eigenvalue , ベクトル $\boldsymbol{x}$ を固有ベクトル eigenvector という (ただし $\boldsymbol{x} \neq \boldsymbol{0}$ の時だけ考える).

実対称行列の定義

n 次正方行列 A が実対称行列とは, 成分がすべて実で, 次が成立することをいう.

$$^{\mathrm{t}}A = A$$

n 次実対称行列の性質

- n 次実対称行列の固有値は実数.
- n 次実対称行列には, 互いに直交する n 個の実の固有ベクトルがある.

$\rightsquigarrow$ 実対称行列の固有ベクトルから, 正規直交基底を作れる.

直交行列 (線形代数の復習)

直交行列の定義

n 次正方行列 P が直交行列であるとは, 成分がすべて実で, 次が成立すること.

$$ {}^{\mathrm{t}}PP = E $$

2 次の直交行列の性質

2 つの 2 次元列ベクトルを並べた 2 次正方行列 $P = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2]$ について, 次の 3 つは同じこと (必要十分条件).

1. P は直交行列.
2. $\boldsymbol{x}_1, \boldsymbol{x}_2$ は, 互いに直交する単位ベクトル,
3. $\boldsymbol{x}_1 \cdot \boldsymbol{x}_2 = \delta_{ij}$

$\rightsquigarrow$ 実対称行列を対角化する基底変換行列は, 直交行列にとれる.

L09-Q3

Quiz(実対称行列の直交行列による対角化)

実対称行列

$$A = \begin{bmatrix} \frac{23}{5} & \frac{36}{5} \\ \frac{36}{5} & \frac{2}{5} \end{bmatrix}$$

を, 直交行列で対角化しよう.

2 次元正規分布の確率密度関数の等高線

2 次元正規分布の確率密度関数 f の等高線

2 次元正規分布 $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ の確率密度関数 $f(x_1, x_2)$ を考える.
対称行列 Σ の固有値 $\lambda_1 \geq \lambda_2 > 0$, 対応する固有ベクトル $\boldsymbol{v}_1, \boldsymbol{v}_2$ とするとき, 等高線 $f(x_1, x_2) = C$ は,
中心が $\boldsymbol{\mu}$,
長軸が $\boldsymbol{v}_1$ に平行, 短軸が $\boldsymbol{v}_2$ に平行,
長半径:短半径$=\sqrt{\lambda_1} : \sqrt{\lambda_2}$
の楕円である.

ここまで来たよ

7. ナイーブベイズ・線形判別分析

8. クラスター分析

9. 共分散行列と固有値固有ベクトル
   - 固有値固有ベクトルと楕円の主軸
   - 対称行列の直交行列による対角化
   - n 次元正規分布とその等高面
   - 平方和の分解

n 次元正規分布 多変量解析☆演習 (2021)L6

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)

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

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

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. $|\boldsymbol{x}|^2 = \sum_{i=1}^n a_i^2$.

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

$a_1 = (C'\lambda_1)^{1/2}$, 他の $a_i = 0$ とすると, $|\boldsymbol{x}|^2$ が最大になる.

理由 (続き)

$a_1 = 0$ という制約下では, $a_2 = (C'\lambda_2)^{1/2}$, 他の $a_i = 0$ とすると, $|\boldsymbol{x}|^2$ が最大になる.

L09-Q4

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.

ここまで来たよ

7. ナイーブベイズ・線形判別分析

8. クラスター分析

9. 共分散行列と固有値固有ベクトル
   - 固有値固有ベクトルと楕円の主軸
   - 対称行列の直交行列による対角化
   - n 次元正規分布とその等高面
   - 平方和の分解

平方和の分解とクラスタの良さの評価と Ward の距離

本質は $p = 1$ 次元で見えるのでその表現で.
x_{ik}: k 番目のクラスターに属する i 番目のデータ点.
$i = 1, \ldots, n_k . k = 1, \ldots, C . \sum_{k=1}^{C} n_k = n.$
偏差平方和

$$
\begin{aligned}
S =& \sum_{i,k} [x_{ik} - \overline{x}_{..}]^2 \\
=& \sum_{i,k} [(x_{ik} - \overline{x}_{.k}) + (\overline{x}_{.k} - \overline{x}_{..})]^2 \\
\stackrel{!}{=}& \sum_{k} \sum_{i} (x_{ik} - \overline{x}_{.k})^2 + \sum_{k} n_k (\overline{x}_{.k} - \overline{x}_{..})^2 \\
=& S_{\mathrm{W}} + S_{\mathrm{B}}
\end{aligned}
$$

群=クラスター
$x_{..} = \frac{1}{n} \sum_{i,k} x_{ik}$ 全体平均
$x_{.k} = \frac{1}{n_k} \sum_{i} x_{ik}$ 群内平均
S_{W}: 群内平方和 Within
S_{B}: 群間平方和 Between

L09-Q5

Quiz(平方和)

各組の学級で異なる教え方をした. テストの点数 x から学級 y をあてることはできそうだろうか.

y	学級	点数
0	A 組	78 79 79 80
1	B 組	78 86 81 83 82
2	C 組	86 85 87

群内平方和と群間平方和を求めよう.

クラスタリングへの応用

‘よい’ クラスタリングとは, S_{B} に対して S_{W} が小さいことでは?

Ward 法=トリッキーな距離の定義での階層的クラスター分析

クラスター間距離=(そのクラスターを合併したときの群内平方和 S_{W} の増分)

判別分析への応用

実は, 線形判別分析の Fisher の線形判別関数は比 $S_{\mathrm{B}}/S_{\mathrm{W}}$ を最大化するものになっていた.

CH 基準 (カリンスキ-ハラバシュ基準)

クラスタの個数 C が異なる場合にも対応させたもの (回帰の自由度調整済決定係数みたいなアイデア).

$$\mathrm{CH}_C = \frac{(n-C)S_{\mathrm{B}}}{(C-1)S_{\mathrm{W}}}$$

分散分析

確率統計 II,III で学ぶ分散分析では, 多群の間に差があるかどうかを判定する検定を, 群内平方和, 群間平方和を利用して行う.

主成分分析

来週学ぶ主成分分析には ‘群’ がない,