「データ解析」＋「確率過程」＋「パターン認識」に関して共通に必要な基礎知識

【第1.6版(H30.7.4)】

工藤峰一、今井英幸、吉田哲也

平成24年10月1日

目 次

第1章　線形代数

執筆：吉田哲也（知識メディア研究室）
E-mail: yoshida@meme.hokudai.ac.jp

主な学習事項

1. 行列とベクトル
2. 固有値と固有ベクトル
3. ベクトルによる微分

以下に，線形代数の基本的な事項を述べる．主に定義や基本的な性質のまとめを示すにとどまっているため，さらに詳しくは良書 (たとえば [1]) を参照されたい．

1　行列とベクトル

1.1　定義と記法

以下では，ベクトルとは，列ベクトル（縦ベクトル）とする．行ベクトル（横ベクトル）を表す時には，転置の記号 ($'$) を用いる．

例

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \boldsymbol{x}' = (x_1, x_2, \cdots, x_n)$$

特に，$\mathbf{0}' = (0, 0, \cdots, 0)$ と表記する．

同じサイズのベクトルを並べると行列になる．

$$\mathbf{A} = (\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

行列 $\mathbf{A}$ は m 行 n 列で表現され，$m \times n$ 行列と呼ぶ．

以下では，行列 $\mathbf{A}$ の i 行 j 列の要素を a_{ij} あるいは $(\mathbf{A})_{ij}$ と表記する．

1.2 ベクトルの公理

ベクトルの和とスカラー積が定義され，以下を満たす.

$$
\begin{aligned}
\text{交換律:} &\quad \boldsymbol{x}+\boldsymbol{y}=\boldsymbol{y}+\boldsymbol{x} \\
\text{結合律:} &\quad (\boldsymbol{x}+\boldsymbol{y})+\boldsymbol{z}=\boldsymbol{x}+(\boldsymbol{y}+\boldsymbol{z}) \\
\text{零元} &\quad \boldsymbol{x}+\boldsymbol{0}=\boldsymbol{x} \\
\text{逆元} &\quad \boldsymbol{x}+(-\boldsymbol{x})=\boldsymbol{0} \\
\text{分配律 (1):} &\quad k(\boldsymbol{x}+\boldsymbol{y})=k\boldsymbol{x}+k\boldsymbol{y} \\
\text{分配律 (2):} &\quad (k_1+k_2)\boldsymbol{x}=k_1\boldsymbol{x}+k_2\boldsymbol{x} \\
\text{結合律:} &\quad (k_1k_2)\boldsymbol{x}=k_1(k_2\boldsymbol{x}) \\
\text{（スカラー積の）零元:} &\quad 0\boldsymbol{x}=\boldsymbol{0} \\
\text{（スカラー積の）単位元:} &\quad 1\boldsymbol{x}=\boldsymbol{x}
\end{aligned}
$$

1.3 行列の演算

行列の演算として，以下に示す行列の和，スカラー積，行列の積が定義される.

行列の和

$$
\mathbf{A}+\mathbf{B}=\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix} \tag{1.1}
$$

スカラー積

$$
k\mathbf{A}=\begin{pmatrix} ka_{11} & ka_{12} & \cdots & ka_{1n} \\ ka_{21} & ka_{22} & \cdots & ka_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ ka_{m1} & ka_{m2} & \cdots & ka_{mn} \end{pmatrix} \tag{1.2}
$$

で与えられる。

行列の積は、$\mathbf{A}$ の列数と $\mathbf{B}$ の行数が一致する場合のみ定義され、$\mathbf{C}=\mathbf{AB}$ の要素として、

$$
c_{ij}=(\mathbf{AB})_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj} \tag{1.3}
$$

と定義される。

上記の演算に対し，以下が成立する.

$$
\begin{aligned}
&\text{非交換律:} && \text{一般に} \quad \mathbf{AB} \neq \mathbf{BA} \\
&\text{結合律:} && \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} \\
&\text{分配律 (1):} && \mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \\
&\text{分配律 (2):} && (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC} \\
&\text{線型性 (1)} && \mathbf{A}(\boldsymbol{x}+\boldsymbol{y}) = \mathbf{A}\boldsymbol{x}+\mathbf{A}\boldsymbol{y} \\
&\text{線型性 (2)} && \mathbf{A}(k\boldsymbol{x}) = k(\mathbf{A}\boldsymbol{x})
\end{aligned}
$$

演習問題 1.1

非交換律が成り立つ例を示せ

演習問題 1.2

結合律，分配律，線形性が成立することを要素の演算として確認せよ．

1.4 転置とトレース

行数と列数が等しい行列を正方行列と呼ぶ．
$n \times n$ 正方行列 $\mathbf{A}$ に対して，転置行列 $\mathbf{A}'$ を以下で定義する．

$$(\mathbf{A})'_{ij} = (\mathbf{A})_{ji} \tag{1.4}$$

行列の転置と行列の積には以下が成立する．

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \tag{1.5}$$

演習問題 1.3

式 (1.5) が成立することを確認せよ．

$n \times n$ 正方行列 $\mathbf{A}$ に対し，トレース (tr) を以下で定義する．

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \tag{1.6}$$

式 (1.6) の定義より，トレースは正方行列 $\mathbf{A}$ の対角要素の和であるためスカラーである．
トレースに対しては以下が成り立つ．

$$\text{線型性 (1)} \quad \mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \tag{1.7}$$

$$\text{線型性 (2)} \quad \mathrm{tr}(k\mathbf{A}) = k\ \mathrm{tr}(\mathbf{A}) \tag{1.8}$$

$$\text{交換律} \quad \mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) \tag{1.9}$$

$$\text{スカラー} \quad \mathrm{tr}(c) = c \ \ (c\ \text{はスカラー}) \tag{1.10}$$

演習問題 1.4

トレースに対して線型性，交換律が成立することを確認せよ．

1.5 分割行列

$m \times n$ 行列 $\mathbf{A}$ を，$p \times q$ 行列 $\mathbf{A}_{11}$，$p \times (n-q)$ 行列 $\mathbf{A}_{12}$，$(m-p) \times q$ 行列 $\mathbf{A}_{21}$，$(m-p) \times (n-q)$ 行列 $\mathbf{A}_{22}$ の 4 つの行列に分割する．

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

同様に，$m \times n$ 行列 $\mathbf{A}$ を，$p \times q$ 行列 $\mathbf{B}_{11}$，$p \times (n-q)$ 行列 $\mathbf{B}_{12}$，$(m-p) \times q$ 行列 $\mathbf{B}_{21}$，$(m-p) \times (n-q)$ 行列 $\mathbf{B}_{22}$ の 4 つの行列に分割する．

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$$

このとき，式 (1.1), (1.1) と同様に，分割行列に対する和，スカラー積は以下のようになる．

分割行列の和

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{pmatrix} \tag{1.11}$$

分割行列のスカラー積

$$k\mathbf{A} = \begin{pmatrix} k\mathbf{A}_{11} & k\mathbf{A}_{12} \\ k\mathbf{A}_{21} & k\mathbf{A}_{22} \end{pmatrix} \tag{1.12}$$

そして、積は

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix} \tag{1.13}$$

となる。

つまり，あたかも分割行列がスカラーになったかのように行列の演算を適用することができる．

演習問題 1.5

式 (1.11), (1.12), (1.13) が成立することを確認せよ．

1.6 内積とノルム

2 つの $n \times 1$ ベクトル $\boldsymbol{x}'$=$(x_1, x_2, \cdots, x_n)$ と $\boldsymbol{y}'$=$(y_1, y_2, \cdots, y_n)$ の内積 $(\boldsymbol{x}, \boldsymbol{y})$ を以下で定義する．

$$(\boldsymbol{x}, \boldsymbol{y}) \overset{\text{def}}{\equiv} \boldsymbol{x}'\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i \tag{1.14}$$

$\boldsymbol{x}$ と $\boldsymbol{y}$ の内積がゼロとなる場合，その 2 つのベクトルは「直交する」という．

ベクトル $\boldsymbol{x}$ のノルムを，上記の内積を用いて以下で定義する．

$$||\boldsymbol{x}|| \overset{\text{def}}{\equiv} = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1.15}$$

1.7 逆行列

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

を n 次の単位行列と呼ぶ.

$n \times n$ 正方行列 $\mathbf{A}$ に対して,

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n \tag{1.16}$$

を満たす $n \times n$ 正方行列 $\mathbf{B}$ を逆行列と呼び，$\mathbf{A}^{-1}$ と表記する.

逆行列は常に存在するとは限らないが，存在する場合には一意である．逆行列を持つ行列を非特異行列 (または，正則行列) と呼び，逆行列を持たない行列を特異行列と呼ぶ.

逆行列が存在する 2 つの $n \times n$ 正方行列 $\mathbf{A}$，$\mathbf{B}$ に対して，以下が成り立つ.

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{1.17}$$

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})' \tag{1.18}$$

演習問題 1.6

式 (1.17), (1.18) が成立することを確認せよ.

1.8 直交行列

$n \times n$ 正方行列 $\mathbf{T}$ に対して

$$\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_n \tag{1.19}$$

が成立する場合，$\mathbf{T}$ を直交行列と呼ぶ.

直交行列の定義 (1.19) より，$(\mathbf{T}'\mathbf{T})_{ij} = \delta_{ij}$ [1] であるため，$\mathbf{T}$ の列ベクトルはノルムが 1 でお互いに直交する．また，直交行列 $\mathbf{T}$ の逆行列はその転置行列である.

$$\mathbf{T}^{-1} = \mathbf{T}' \tag{1.20}$$

演習問題 1.7

直交行列 $\mathbf{T}$ の列ベクトルはノルムが 1 でお互いに直交することを確認せよ.

演習問題 1.8

式 (1.20) が成立することを確認せよ.

[1] δ_{ij} は ”クロネッカーのデルタ ”と呼ばれ、$i = j$ の時に 1、それ以外の時に 0 をとる関数である。

1.9 二次形式

$n \times n$ 対称行列 $\mathbf{A}$ と n 次元ベクトル $\boldsymbol{x}$ に対し，以下で定義される量を二次形式と呼ぶ.

$$\begin{aligned}\boldsymbol{x}'\mathbf{A}\boldsymbol{x} &= (x_1, x_1, \ldots, x_n)\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \end{aligned} \tag{1.21}$$

式 (1.21) より，この量はスカラーとなる．このため，以下が成り立つ.

$$\boldsymbol{x}'\mathbf{A}\boldsymbol{x} = \mathrm{tr}\,(\mathbf{A}\boldsymbol{x}\boldsymbol{x}') \tag{1.22}$$

演習問題 1.9

式 (1.22) が成立することを確認せよ.

2 固有値と固有ベクトル

2.1 行列式

集合 $\{1, 2, \ldots, n\}$ からの $\{1, 2, \ldots, n\}$ への全単射を (n 次の）置換と呼び，σ で表現する.

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix}$$

特に，2 つの要素のみを置換したものを互換と呼ぶ．一般に，任意の置換は互換の積として表現できることが知られている．置換を互換の積として表現した際，互換の数が偶数の場合を偶置換，奇数の場合を奇置換と呼ぶ．置換の符号 $sgn(\sigma)$ を，偶置換の場合に $sgn(\sigma)$=1，奇置換の場合に $sgn(\sigma)$=-1 と定義する.

$n \times n$ 正方行列

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

に対して，

$$|\mathbf{A}| = \sum_{\sigma \in S_n} sgn(\sigma) a_{1\sigma(1)} \ldots a_{n\sigma(n)} \tag{1.23}$$

を $\mathbf{A}$ の行列式と呼ぶ．なお，行列式 $|\mathbf{A}|$ を $\det \mathbf{A}$, $\det(\mathbf{A})$ などとも表記する.

$n \times n$ 正方行列 $\mathbf{A}$, $\mathbf{B}$ に対して，以下が成り立つ.

$$|\mathbf{A}'| = |\mathbf{A}| \tag{1.24}$$

$$|\mathbf{A}\mathbf{A}^{-1}| = |\mathbf{I}_n| = 1 \tag{1.25}$$

$$|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}| \tag{1.26}$$

上記より，$n \times n$ 正方行列 $\mathbf{A}$ が正則行列の場合には，$|\mathbf{A}| \neq 0$ となる．

演習問題 1.10

式 (1.24), (1.25), (1.17) が成立することを確認せよ．

2.2 固有値と固有ベクトル

$n \times n$ 正方行列 $\mathbf{A}$ に対し，$\boldsymbol{x} \neq \mathbf{0}$ が

$$\mathbf{A}\boldsymbol{x} = \lambda \boldsymbol{x} \tag{1.27}$$

を満たすとき，スカラー λ を $\mathbf{A}$ の固有値，$\boldsymbol{x}$ を λ に対応する $\mathbf{A}$ の固有ベクトルと呼ぶ．$\mathbf{A}$ を n 次元ベクトル空間から n 次元ベクトル空間への線形写像と考えると，固有ベクトルはこの写像に関して方向は不変 (ただし，λ が負の場合には反対向きになる) であることを意味する．なお，固有ベクトル $\boldsymbol{x}$ に対して，任意のスカラー $\alpha \neq 0$ に対して $\alpha\boldsymbol{x}$ も固有ベクトルである．

固有値と固有ベクトルを得るには，

$$|\mathbf{A} - \lambda \mathbf{I}| = 0 \tag{1.28}$$

を解くことで求めることができる．式 (1.28) を固有方程式と呼ぶ．

演習問題 1.11

固有ベクトル $\boldsymbol{x}$ に対して，任意のスカラー $\alpha \neq 0$ に対して $\alpha\boldsymbol{x}$ も固有ベクトルであることを確認せよ．

2.3 対称行列のスペクトル分解

$n \times n$ 正方行列 $\mathbf{A}$ が対称（$\mathbf{A}'$=$\mathbf{A}$) であり，各要素 a_{ij} が実数とする．この場合，n 個の固有値は全て実数となり，対応する固有ベクトルも実数ベクトルとなる．また，異なる固有値 λ_i, λ_j $(\lambda_i \neq \lambda_j)$ に対応する固有ベクトル $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ は直交する

$n \times n$ 正方対称行列 $\mathbf{A}$ の固有値 $\lambda_1, \lambda_2, \ldots, \lambda_n$ が全て異なる場合を考える．固有ベクトルの長さを 1 に正規化した固有ベクトルを $\boldsymbol{t}_1$,$\boldsymbol{t}_2$,...,$\boldsymbol{t}_n$ とする．これらの固有ベクトルを並べた行列 $\mathbf{T}$=$(\boldsymbol{t}_1,\boldsymbol{t}_2,\ldots,\boldsymbol{t}_n)$ は直交行列になる．

演習問題 1.12

$n \times n$ 正方対称行列 $\mathbf{A}$ の固有値 $\lambda_1, \lambda_2, \ldots, \lambda_n$ が全て異なる場合，固有ベクトルの長さを 1 に正規化した固有ベクトル $\boldsymbol{t_i}$ を並べた行列 $\mathbf{T}$ は直交行列であることを確認せよ．

上記の固有値を対角要素に並べ，それ以外の要素をゼロとした対角行列を Λ とする．

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

この際，$\lambda_1, \lambda_2, \ldots, \lambda_n$，$\boldsymbol{t}_1$,$\boldsymbol{t}_2$,...,$\boldsymbol{t}_n$ はそれぞれ固有値，固有ベクトルであるため，それぞれ以下が成り立つ．

$$
\begin{aligned}
\mathbf{A}\boldsymbol{t}_1 &= \lambda_1 \boldsymbol{t}_1 \\
\mathbf{A}\boldsymbol{t}_2 &= \lambda_1 \boldsymbol{t}_2 \\
&\cdots \\
\mathbf{A}\boldsymbol{t}_n &= \lambda_1 \boldsymbol{t}_n
\end{aligned}
$$

上記をまとめて行列として表現する．

$$
\begin{aligned}
\mathbf{A}(\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n) &= (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n) \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \\
\mathbf{AT} &= \mathbf{T}\Lambda
\end{aligned}
$$

行列 $\mathbf{T}$ は直交行列であり，

$$\mathbf{T}'\mathbf{AT} = \Lambda \tag{1.29}$$

が成り立つ．式 (1.29) を対称行列 $\mathbf{A}$ の対角化と呼ぶ．

式 (1.29) に基づいて，対称行列 $\mathbf{A}$ を以下のように分解することができる．

$$\mathbf{A} = \sum_i \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i' \tag{1.30}$$

式 (1.30) を，対称行列 $\mathbf{A}$ のスペクトル分解と呼ぶ．

2.4 対称行列の固有値と固有ベクトル

$n \times n$ 正方行列 $\mathbf{A}$ に対しては，重解も含めると一般に固有値は n 個存在する．n 個の固有値を λ_1, λ_2, $\ldots$, λ_n, とすると，以下の関係式が成立する．

$$
\begin{aligned}
\mathrm{tr}\mathbf{A} &= \lambda_1 + \lambda_2 + \ldots + \lambda_n = \sum \lambda_i \qquad (1.31) \\
|\mathbf{A}| &= \lambda_1 \lambda_2 \ldots \lambda_n = \prod \lambda_i \qquad (1.32)
\end{aligned}
$$

$n \times n$ 正方行列 $\mathbf{A}$ が正則な場合，$|\mathbf{A}| \neq 0$ となるため，全ての固有値は 0 ではない．逆に，$n \times n$ 正方行列 $\mathbf{A}$ が正則でない場合，少なくとも一つはゼロとなる固有値が存在する．

3 ベクトルによる微分

n 次元ベクトル

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

を引数とする関数 $f(x_1, x_2, \ldots, x_n)$ を考える．ここで，関数の値はスカラーとする．関数 $f(x_1, x_2, \ldots, x_n)$ はベクトル $\boldsymbol{x}$ の各要素 x_i に依存するため，関数を n 変数の関数と考え，各 x_i で偏微分したものをベクトル $\boldsymbol{x}$ による微分と考えて以下で表現する．

$$\frac{\partial f}{\partial \boldsymbol{x}} \stackrel{\text{def}}{\equiv} \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \tag{1.33}$$

同様にして、転置ベクトル $\boldsymbol{x}'$ で微分や、ベクトル値関数 $\boldsymbol{f}$ の微分も

$$\frac{\partial f}{\partial \boldsymbol{x}'} \stackrel{\text{def}}{\equiv} \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right) \tag{1.34}$$

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \stackrel{\text{def}}{\equiv} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \tag{1.35}$$

と定義する（縦横の対応に注意）ことで標記が楽になる。

3.1 線形形式のベクトル微分

定数ベクトル $\boldsymbol{a}$ に対する内積 $\boldsymbol{x}'\boldsymbol{a}$ $(= \boldsymbol{a}'\boldsymbol{x})$ のベクトル微分

$$\begin{aligned} \frac{\partial \boldsymbol{x}'\boldsymbol{a}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} \\ &= \boldsymbol{a} \end{aligned} \tag{1.36}$$

演習問題 1.13

式 (1.36) が成立することを確認せよ．

3.2 二次形式のベクトル微分

一般の（対称行列ではない)$n \times n$ 正方行列 $\mathbf{A}$ (各 a_{ij} は定数) に対する二次形式 $\boldsymbol{x}'\mathbf{A}\boldsymbol{x}$ のベクトル微分については，以下が成立する．

$$\frac{\partial \boldsymbol{x}'\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\mathbf{A} + \mathbf{A}')\boldsymbol{x} \tag{1.37}$$

特に，$n \times n$ 正方行列 $\mathbf{A}$ が対称な場合には，二次形式 $\boldsymbol{x}'\mathbf{A}\boldsymbol{x}$ のベクトル微分は以下になる．

$$\frac{\partial \boldsymbol{x}'\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\mathbf{A}\boldsymbol{x} \tag{1.38}$$

さらに、この式を $\boldsymbol{x}'$ で微分することで、

$$\frac{\partial \boldsymbol{x}'\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}\partial \boldsymbol{x}'} = 2\mathbf{A} \tag{1.39}$$

を得る。

$\boxed{\text{演習問題 1.14}}$

式 (1.37)，(1.38)，(1.39) が成立することを確認せよ．

第2章　確率の基礎

執筆：今井英幸（情報解析学研究室）
E-mail: `imai@ist.hokudai.ac.jp`

主な学習事項

1. 確率空間 (Probability space)
2. 確率変数 (Random variable)
3. 累積分布関数 (Cumulative distribution function, cdf)
4. 確率密度関数 (Probability density function, pdf)
5. 期待値 (Expectation)

1　確率

1.1　標本空間

一つのサイコロを２回投げる。１回目に出た目を k, ２回目に出た目を l として、結果を (k,l) と書くことにすると、起こりうるすべての目の出かたは

$$\Omega_2 \overset{\text{def}}{\equiv} \{(1,1),(1,2),\ldots,(6,6)\}$$

である。Ω_2 を**標本空間**という。サイコロがゆがんでいなければ、Ω_2 のすべての要素の出現は同程度に確からしい、と考えていいので, (k,l) が起きる確率は

$$P((k,l)) = \frac{1}{36}, k,l \in \{1,2,3,4,5,6\}$$

としてよいであろう。例えば、

$$P((1,2)) = \frac{1}{36}$$

である。

同じように、一つのサイコロを n 回投げるとする。i 回目に出た目を x_i とし、結果を $(x_1,\ldots,x_n)$ と書くことにすると、起こりうるすべての目の出かたは、

$$\Omega_n \overset{\text{def}}{\equiv} \{(1,1,\ldots,1),(1,1,\ldots,2),\ldots,(6,6,\ldots,6)\}$$

である。この例では Ω_n が標本空間になる。前の例と同じ理由で、サイコロがゆがんでいなければ

$$P((x_1,\ldots,x_n)) = \frac{1}{6^n}, x_i \in \{1,2,3,4,5,6\}$$

であると考えてよい。

標本空間 Ω_n の中で $x_1 = 1, x_2 = 2$ という目の出かたをするものを A とすると、

$$A \stackrel{\text{def}}{\equiv} \{(x_1, x_2, \ldots, x_n) \in \Omega_n \mid x_1 = 1, x_2 = 2\}$$

であり、A に属する要素の個数は 6^{n-2} である。したがって、

$$P(A) = \frac{6^{n-2}}{6^n} = \frac{1}{36}$$

となる。これは Ω_2 の例で、$P((1,2))$ の確率と同じである。

以上、二つの例をまとめると、次のような手順で確率を決めていることがわかる。

1. 標本空間 Ω を決める。ただし、標本空間に属する要素はすべて同じ程度に確からしい。

2. 標本空間の部分集合 A の起きる確率は、集合 A に属する要素の数 [1] を $\#A$ として

$$P(A) = \frac{\#A}{\#\Omega}$$

と決める。ただし、$\#\Omega$ は標本空間の要素数である。

このような確率の決め方をここでは**頻度に基づく確率**ということにする。頻度に基づく確率は、次の性質 (P1'), (P2'), (P3'),(P4') を満たすことを示すことができる。

(P1') 標本空間 Ω のすべての部分集合 $A \subset \Omega$ に対して、$0 \leq P(A) \leq 1$.

(P2') $P(\Omega) = 1$.

(P3') $A, B \subset \Omega$ かつ $A \cap B = \emptyset$ であれば

$$P(A \cup B) = P(A) + P(B)$$

さらに、$A_1, A_2, \ldots, A_m \subset \Omega$ が、$i \neq j$ ならば $A_i \cap A_j = \emptyset$ をみたすとき [2]

$$P(A_1 \cup A_2 \cup \cdots \cup A_m) = P(A_1) + P(A_2) + \cdots + P(A_m)$$

あるいは

$$P(\bigcup_{i=1}^{m} A_i) = \sum_{i=1}^{m} P(A_i).$$

(P4') $P(A) + P(A^c) = 1$.

演習問題 2.1

頻度に基づく確率の定義にしたがって (P1'), (P2'), (P3'), (P4') を示せ。

[1] 集合の要素数のことを集合の濃度 (cardinality) という。有限集合の場合には濃度とは個数のことだが、無限集合の濃度にはいろいろと不思議なことが起こることが知られている。なお、集合 A の濃度を $|A|$ と表す本も多い。

[2] これを $A_1, \ldots, A_m$ は**互いに排反である**という。

1.2 確率測度

1.1 節ではサイコロを投げる回数をあらかじめ決めていた。では、次のような確率はどのように求めたらよいだろうか。

いつかは 1 の目が出る確率

1.1 節の頻度に基づく確率の定義にしたがえば、まず標本空間を決める必要がある。この例ではサイコロを投げる回数は決まっていないので、

$$\Omega_\infty \overset{\text{def}}{\equiv} \{(x_1, x_2, \ldots, x_n, \ldots) \mid x_i \in \{1, 2, 3, 4, 5, 6\}\}$$

という無限集合になり、「いつかは 1 の目が出る」に対応する部分集合 A もまた

$$\begin{aligned} A &\overset{\text{def}}{\equiv} \{(1, x_2, x_3, \ldots,)\} \ (1\text{ 回目に }1\text{ が出る}) \\ &\cup \{(x_1, 1, x_3, \ldots) \mid x_1 \neq 1\} \ (2\text{ 回目に }1\text{ が出る}) \\ &\cup \{(x_1, x_2, 1, x_4, \ldots) \mid x_1 \neq 1, x_2 \neq 1\} \ (3\text{ 回目に }1\text{ が出る}) \cdots \\ &\cup \{(x_1, \ldots, x_{n-1}, 1, x_{n+1}, \ldots) \mid x_i \neq 1, i = 1, \ldots, n-1\} \ (n\text{ 回目に }1\text{ が出る}) \cdots \end{aligned}$$

という無限集合になる。したがって

$$P(A) = \frac{\#A}{\#\Omega} = \frac{\infty}{\infty} = ??.$$

この例からわかるように、標本空間が無限集合の場合には頻度に基づく確率の定義はうまくいかない。そこで、これに代わってどのような標本空間に対しても定義できるように確率の考え方を変える必要がある。(そして、それは頻度に基づく確率の考え方と矛盾しないことが求められる。)

まず、標本空間 Ω の部分集合のなかで確率が定義されるものを**事象**という。そして、事象 A に対して条件 (P1), (P2), (P3) を満たす関数(写像)を**確率測度** (または、単に**確率**) という。

(P1) 任意の事象 $A \subset \Omega$ に対して、$0 \leq P(A) \leq 1$.

(P2) $P(\Omega) = 1$.

(P3) 無限個の事象 $A_1, A_2, \ldots \subset \Omega$ が互いに排反[3] であれば

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

これらの性質は、頻度に基づく確率の性質 (P1'), (P2'), (P3') に対応している。また、(P3) の性質を**σ-加法性**または**可算加法性**という。(P1), (P2), (P3) を満たすように定義されている関数 P を、ここでは**公理に基づく確率**ということにし、これからは、確率といえば、公理に基づく確率を指すことにする。

公理に基づく確率は次のような性質をもつことを示すことができる。

(P4) $P(\emptyset) = 0$.

(P5) (余事象の確率) $P(A^c) = 1 - P(A)$.

[3] 12 ページの脚注参照。

(P6) (有限加法性) m 個の事象 $A_1, A_2, \dots A_m \subset \Omega$ が互いに排反であれば

$$P(\bigcup_{i=1}^{m} A_i) = \sum_{i=1}^{m} P(A_i).$$

(P7) (単調性) 事象 $A, B \subset \Omega$ に対して

$$A \subset B \Longrightarrow P(A) \leq P(B).$$

(P8) 事象 $A, B \subset \Omega$ に対して

$$P(A \cup B) + P(A \cap B) = P(A) + P(B).$$

(P9) (劣加法性) 無限個の事象 $A_1, A_2, \dots \subset \Omega$ に対して

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i).$$

演習問題 2.2

公理に基づく確率が (P4)〜(P9) を満たすことを示せ。

公理に基づく確率が、頻度に基づく確率と矛盾しないことを示そう。そのためには、標本空間が有限集合で、かつその要素の起こる確率が同程度に確からしい場合に、部分集合 A の確率が $P(A) = \frac{\#A}{\#\Omega}$ になることを (P1), (P2), (P3) を用いて示せばよい。

標本空間 Ω の濃度（個数）を n とし、$A = \{a_1, a_2, \dots, a_m\}$ とすると、有限加法性 (P6) より

$$\begin{aligned} P(A) &= P(\{a_1, \dots, a_m\}) = P(\{a_1\} \cup \{a_2\} \cup \cdots \cup \{a_m\}) = P(\{a_1\}) + P(\{a_2\}) + \cdots P(\{a_m\}) \\ &= \frac{1}{n} + \cdots + \frac{1}{n} = \frac{m}{n} = \frac{\#A}{\#\Omega} \end{aligned}$$

であることがわかる。

1.3 事象と σ-集合体

公理に基づく確率では、標本空間 Ω の部分集合である事象 A に対して確率を定義した。事象についてもう少し詳しく考えてみよう。

「事象」とはどのように定義すればよいだろうか。「確率が定義される部分集合を事象という」では、定義にならない。まず、標本空間 Ω が有限集合である場合と同じように「標本空間 Ω のすべての部分集合の集まりを事象とする」ことが考えられる。このように事象を定義すれば「事象とは何か」について考えなくてもよいので都合がいいが、これから展開される確率論においては、「あらゆる部分集合」に対して確率が定義されている必要はない。むしろ、「すべての部分集合に対して確率が定義されていなければならない」のでは制約が強すぎて確率論が適用できる範囲が狭くなってしまう。

そこで、次のように考えることにする。標本空間 Ω の部分集合の中で、確率が定義されている部分集合だけを集めて、これを $\mathcal{B}$ としよう。集合族 $\mathcal{B}$ は次のような性質をもつことが望ましい。

(B1) $\Omega \in \mathcal{B}$.

(B2) $A \in \mathcal{B} \Longrightarrow A^c \in \mathcal{B}$.

(B3) 無限個の部分集合 $A_1, A_2, \ldots \subset \Omega$ に対して、

$$A_1, A_2, \ldots \in \mathcal{B} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}.$$

条件 (B1), (B2), (B3) をみたすような Ω の部分集合の族を Ω 上の σ-**集合体**という。そして、σ-集合体 $\mathcal{B}$ に属する部分集合に確率を定義する。

例 2.1. 標本空間 Ω のすべての部分集合からなる集合の族は Ω 上の σ-集合体である。

例 2.2. 標本空間 Ω 上の部分集合族 $\{\emptyset, \Omega\}$ は Ω 上の σ-集合体である。

例 2.3. $A \subset \Omega$ とすると $\sigma(A) \stackrel{\text{def}}{\equiv} \{\emptyset, A, A^c, \Omega\}$ は Ω 上の σ-集合体である。

演習問題 2.3

集合族 $\sigma(A)$ は σ-集合体であることを示せ。

定理 2.1. $\mathcal{B}$ を Ω 上の σ-集合体とする。$A_1, A_2, \ldots \in \mathcal{B}$ のとき、以下の性質が成り立つ。

(B4) $\displaystyle\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}$

(B5) $\displaystyle\limsup_{i\to\infty} A_i \stackrel{\text{def}}{\equiv} \bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} A_i \in \mathcal{B}$

(B3) $\displaystyle\liminf_{i\to\infty} A_i \stackrel{\text{def}}{\equiv} \bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} A_i \in \mathcal{B}$

演習問題 2.4

σ-集合体の定義を用いて、(B4), (B5), (B6) を示せ。

1.4 確率空間

これまでに説明してきたように、公理に基づく確率を定義するためには以下のものが必要であることがわかる。

(1) 標本空間 Ω.

(2) 標本空間 Ω の部分集合の族で、(B1), (B2), (B3) をみたす σ-集合体 $\mathcal{B}$.

(3) $\mathcal{B}$ に属する事象 A に対して定義された (P1), (P2), (P3) をみたす確率測度 P.

これらを組にした $(\Omega, \mathcal{B}, P)$ を**確率空間**という。

Table 2.1: 例 2.4 を使ったゲームの得点

事象	確率	得点
H	$\frac{1}{2}$	1
T	$\frac{1}{2}$	0

2 確率変数

2.1 確率変数

次の例を考えよう。

例 2.4. 1 つのコインを投げるとする。表が出ることを H (Head の略)、裏が出ることを T (Tail の略) であらわすことにすると、標本空間は

$$\Omega_1 = \{H, T\},$$

である。また、σ-集合体は

$$\mathcal{B}_1 = \{\emptyset, \{H\}, \{T\}, \{H, T\}\},$$

確率測度は

$$P(\{H\}) = \frac{1}{2},\ P(\{T\}) = \frac{1}{2},$$

である。

A 君が、このコイン投げで、次のようなゲームをする。

- 表が出たら A 君は 1 点もらう。

A 君の点数を表にすると表 2.1 のようになる。このゲームのように、標本空間の要素に対して、実数値を対応させる写像を**確率変数** (random variable) という [4]。

このゲームでは、A 君の得点を確率変数 X とすると

$$X(H) = 1,\ X(T) = 0$$

と定めることになる。このとき、表 2.1 から、この確率変数は

$$P(X = 1) = \frac{1}{2}, P(X = 0) = \frac{1}{2}$$

のような確率をもつこともわかる。

同じゲームを n 回繰り返したとしよう。この場合の確率空間は、標本空間を

$$\Omega_n = \{(H, H, \ldots, H), (T, H, \ldots, H), \ldots, (T, T, \ldots, T)\},$$

Ω_n のすべての部分集合からなる σ-集合体を $\mathcal{B}_n$, 確率測度を

$$P((\omega_1, \ldots, \omega_n)) = \frac{1}{2^n},\ \omega_i \in \{H, T\}$$

として、$(\Omega_n, \mathcal{B}_n, P)$ ができあがる。A 君の得点は表 2.2 のようになる。確率変数 X を A 君の得点とすると、今度は

$$X((H, \ldots, H)) = n, (T, H, \ldots, H)) = n - 1, \ldots, X((T, \ldots, T)) = 0$$

[4] 確率変数は大文字アルファベットで表されることが多い。

Table 2.2: 例 2.4 を使った n 回のゲームの得点

事象	確率	得点
$(H,\ldots,H)$	$\frac{1}{2^n}$	n
$\vdots$	$\vdots$	$\vdots$
$(T,\ldots,T)$	$\frac{1}{2^n}$	0

となる。また、確率は

$$P(X=k)=\binom{n}{k}\frac{1}{2^n},\ k=0,1,\ldots,n, \tag{2.1}$$

であることがわかる。

演習問題 2.5

コインを n 回投げるゲームについて、次の問いに答えよ。

(1) Ω_n の濃度を求めよ。

(2) $\mathcal{B}_n$ の濃度を求めよ。

(3) 確率が式 (2.1) であることを示せ。

この二つの例からわかるように

(1) 確率変数は標本空間の要素に実数を対応させる写像であり、

(2) 確率変数がとる値の確率は確率測度 P から

$$\begin{aligned}P(X=a)&=P(\{\omega\in\Omega \mid X(\omega)=a\})\\&=P(\{\omega\in\Omega \mid \omega\in X^{-1}(a)\})\end{aligned}$$

となる。

これが確率変数の確率の定義である。

これまで考察してきた標本空間 Ω_1,Ω_n はともに有限集合であった。したがって、確率変数の取る値も有限個である。このような確率変数を**離散型確率変数**という [5]。

次に、標本空間が無限集合である場合を考えよう。図 2.1 のルーレットを考える。盤面には 0 から 1 までの数が刻まれている。したがって標本空間は $\Omega=(0,1]$ となる。標本空間 Ω は (連続) 無限個の値からなる無限集合である。このとき、針の指し示す値を確率変数 X とすると、このような確率変数を**連続型確率変数** (continuous random variable) であるという。連続な標本空間においては，事象「X は値 x をとる」の確率はつねに 0, すなわち

$$P(X=x)=0$$

でなければならない。このとき，事象を、A:「針が a と b の間で静止する」とすると、事象 A の確率は

$$P(X\in A)=|a-b|,\quad a,b\in(0,1]$$

となる。

[5] 正しくは、取りうる値が有限個または可算無限個の場合

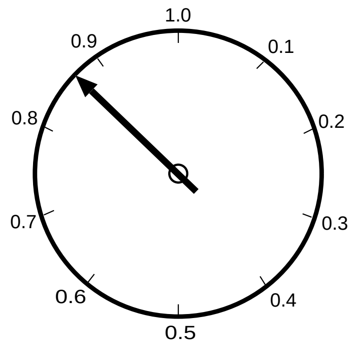

図 2.1: ルーレット盤

2.2 累積分布関数と確率密度関数

離散型確率変数は、「確率変数がある値をとる確率」が定義されるのに対し、連続型確率変数では、「確率変数の値がある範囲に入る確率」しか定義されない。これをまとめて扱うために**累積分布関数** (cumulative distribution function, cdf) を

$$F(a) \stackrel{\text{def}}{\equiv} P(X \leq a)$$

のように定義する。X が離散型確率変数の場合は、

$$F(a) = \sum_{x \leq a} P(X = x)$$

である。連続型確率変数の場合は、ある関数 $f(x)$ が存在して

$$F(a) = \int_{-\infty}^{a} f(x)dx \tag{2.2}$$

となることが証明できる[6]。この $f(x)$ を確率変数 X の**確率密度関数** (probability density function, pdf) という。確率の性質から、確率密度関数は

(PDF.1) $f(x) \geq 0$.

(PDF.2) $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1$.

をみたす。

演習問題 2.6

表 2.1 の確率変数の累積分布関数を求めよ。また、その概形をかけ。

[6] Lebesgue 積分論のラドン・ニコディムの定理による。

2.3 主な離散型確率変数

ベルヌイ (Bernoulli) 分布

確率変数 $X \in \{0,1\}$ の確率が

$$P(X=a)=\begin{cases} p, & a=1 \\ 1-p, & a=0 \\ 0, & \text{その他} \end{cases}$$

であるとき、X はパラメータ p の**ベルヌイ分布**にしたがうという。ただし、$0 \le p \le 1$ である。この確率変数は、表 2.1 の Ω_1 と $\mathcal{B}_1$ と確率

$$P(\{H\})=1,\ P(\{T\})=0$$

からなる確率空間 $(\Omega_1, \mathcal{B}_1, P)$ と確率変数

$$X(H)=1,\ X(T)=0$$

としたものである。つまり、表の出る確率が p のコインを振るとき、表が出たら 1、裏が出たら 0 という確率変数である。

二項分布

n を正の整数とするとき、確率変数 X の確率が

$$P(X=k)=\begin{cases} \binom{n}{k}p^k(1-p)^{n-k}, & k=0,1,\ldots,n \\ 0, & \text{その他.} \end{cases}$$

であるとき、X はパラメータ (n,p) の**二項分布** (binomial distribution) にしたがうといい、$X \sim Bin(n,p)$ とかく。ただし、$0 \le p \le 1$ である。この確率変数は、表 2.2 の Ω_n と $\mathcal{B}_n$ と確率

$$P((\omega_1,\ldots,\omega_n))=p^k(1-p)^{n-k},\ k=\#\{i \mid \omega_i=T\}$$

からなる確率空間 $(\Omega_n, \mathcal{B}_n, P)$ と確率変数

$$X((\omega_1,\ldots,\omega_n))=\#\{i \mid w_i=T\}$$

としたものである。つまり、表の出る確率が p のコインを振るとき、表が出る回数を表す確率変数である。

2.4 主な連続型確率変数

連続型確率変数を定義するためには、累積分布関数か確率密度関数のどちらかがわかればよい。確率密度関数が与えられた場合には、累積分布関数は式 (2.2) から求めることができる。

一様分布

実数 $a, b(a < b)$ とするとき、確率変数 X の確率密度関数が

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b] \\ 0, & \text{その他} \end{cases}$$

であるとき、X はパラメータ (a, b) の一様分布 (Uniform distribution) にしたがうといい、$X \sim U(a, b)$ とかく。

演習問題 2.7
一様分布 $U(a, b)$, $(a < b)$ にしたがう確率変数 X の累積分布関数を求めよ。

指数分布

確率変数 X の確率密度関数が

$$f(x) = \begin{cases} \lambda \exp(-\lambda x), & x \geq 0 \\ 0, & \text{その他} \end{cases}$$

であるとき、X はパラメータ λ の指数分布 (Exponential distribution) にしたがうといい、$X \sim \mathrm{Exp}(\lambda)$ とかく[7]。ただし、$\lambda > 0$ である。

演習問題 2.8
指数分布 $\mathrm{Exp}(\lambda)$, $(\lambda > 0)$ にしたがう確率変数 X の累積分布関数を求めよ。また、その概形をかけ。

正規分布

確率変数 X の確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

であるとき、X はパラメータ (μ, σ^2) の正規分布 (Normal distribution) にしたがうといい、$X \sim \mathrm{N}(\mu, \sigma^2)$ とかく。ただし、$\sigma^2 > 0$ である[8]。

特に、$\mu = 0$, $\sigma^2 = 1$ の場合、この分布を標準正規分布 (Standard normal deistribution) といい、その確率密度関数を $\phi(x)$、累積分布関数を $\Phi(x)$ であらわすことが多い。ただし、正規分布の累積分布関数は初等関数で表すことはできない。つまり、$X \sim \mathrm{N}(0, 1)$ のとき、

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$$
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2) dt$$

である。

演習問題 2.9
正規分布にしたがう確率変数 X の確率密度関数の極値および変曲点を求め、その概形をかけ。

---

[7] $\exp(x)$ は e^x のこと。
[8] μ と σ^2 はそれぞれ X の平均と分散をあらわす。3 節参照。

3 期待値とモーメント

3.1 期待値

確率変数 X の**期待値** (Expectation) (または平均 (Mean)) を

$$\mathbb{E}(X) = \begin{cases} \displaystyle\sum_x xP(X = x), & X \text{ が離散型確率変数} \\ \displaystyle\int_{-\infty}^{\infty} xf(x)dx, & X \text{ が連続型確率変数} \end{cases}$$

と定義する。確率変数 X が小さい値をとる確率が大きければ期待値の値は小さくなり、大きい値をとる確率が大きければ期待値の値も大きくなる。つまり、期待値は確率変数がとる値の大体の場所を表す数値であると考えることができる。期待値のような確率変数の場所をあらわす値のことを位置パラメータ（または、位置母数）という。

演習問題 2.10

離散型確率変数 X と、実数 a, b に対して

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

が成り立つことを示せ。

X を確率変数、$g(x)$ を関数とするとき、$g(X)$ の期待値も同じように

$$\mathbb{E}(g(X)) = \begin{cases} \displaystyle\sum_x g(x)P(X = x), & X \text{ が離散型確率変数} \\ \displaystyle\int_{-\infty}^{\infty} g(x)f(x)dx, & X \text{ が連続型確率変数} \end{cases}$$

と定義する。たとえば、$g(x) = x^2$ のときは、

$$\mathbb{E}(X^2) = \begin{cases} \displaystyle\sum_x x^2P(X = x), & X \text{ が離散型確率変数} \\ \displaystyle\int_{-\infty}^{\infty} x^2f(x)dx, & X \text{ が連続型確率変数} \end{cases}$$

となる。特に、実数 c と正の整数 m に対して $g(x) = (x - a)^m$ の期待値

$$\mathbb{E}((X - a)^m) = \begin{cases} \displaystyle\sum_x (x - a)^mP(X = x), & X \text{ が離散型確率変数} \\ \displaystyle\int_{-\infty}^{\infty} (x - a)^mf(x)dx, & X \text{ が連続型確率変数} \end{cases}$$

を a **の回りの** m **次モーメント**とよぶ。また、期待値の定義から $g(x), h(x)$ を関数、α, β を実数とするとき、

$$\mathbb{E}(\alpha) = \alpha$$
$$\mathbb{E}(\alpha g(X) + \beta h(X)) = \alpha\mathbb{E}(g(X)) + \beta\mathbb{E}(h(X)) \quad \text{(期待値の線形性)}$$

であることがわかる。

演習問題 2.11

確率変数 X_1 と X_2 の期待値を求めよ。

(1) $P(X_1 = 100) = \frac{1}{2},\ P(X_1 = 0) = \frac{1}{2}.$

(2) $P(X_2 = 40) = \frac{1}{4},\ P(X_2 = 50) = \frac{1}{2},\ P(X_2 = 60) = \frac{1}{4}.$

3.2 分散

確率変数 X の**分散** (Variance) を

$$\mathbb{V}(X) \overset{\text{def}}{\equiv} \begin{cases} \displaystyle\sum_x (x-\mathbb{E}(X))^2 P(X=x), & X \text{ が離散型確率変数} \\ \displaystyle\int_{-\infty}^{\infty} (x-\mathbb{E}(X))^2 f(x)dx, & X \text{ が連続型確率変数} \end{cases}$$

と定義する [9]。分散は平均の回りの 2 次のモーメントである。

演習問題 2.12

演習問題 2.11 の確率変数 X_1, X_2 の分散を求めよ。

演習問題 2.13

$\mathbb{V}(X)=\mathbb{E}(X^2)-(\mathbb{E}(X))^2$ が成り立つことを示せ。(分散公式)

演習問題 2.12 からわかるように、確率変数 X が期待値に近い値をとる確率が大きければ分散の値は小さくなり、期待値から離れた値をとる確率が大きければ分散の値も大きくなる。つまり、分散は確率変数がとる値の期待値からの散らばりの程度を表す数値であると考えることができる。分散のような確率変数の散らばり具合をあわらす値のことを尺度パラメータ（または、尺度母数）という。

分散の定義からわかるように、$\mathbb{V}(X) \geq 0$ である。分散の正の平方根を**標準偏差** (standard deviation) といい、$\sigma \overset{\text{def}}{\equiv} \sqrt{\mathbb{V}(X)}$ で表す。

演習問題 2.14

確率変数 X と実数 a に対して、

$$f(a)=\mathbb{E}((X-a)^2)$$

とするとき、$f(a)$ の最小値を求めよ。

[9] 分散を表す記号として、$\mathbb{V}(X)$ のほかに σ^2、あるいは確率変数を明示して σ_X^2 が使われる場合もある。

第3章　多変量解析の基礎

執筆：工藤峰一（情報認識学研究室）
E-mail: `mine@main.ist.hokudai.ac.jp`

主な学習事項

1. 確率ベクトル (Random vector)
2. 同時分布 (Joint distribution)、周辺分布 (Marginal distribution)
3. ベイズの定理 (Bayes theorem)
4. 多次元正規分布 (Muliti-dimensional normal distribution)

1　確率ベクトル

複数の確率変数を一緒に考えることにしよう。p 個の確率変数 $X_1, X_2, \ldots, X_p$ をベクトルとしてまとめて

$$\mathbf{X} = (X_1, X_2, \ldots, X_p)' \qquad (' \text{ は転置}) \tag{3.1}$$

と書こう。この時、一変数の時と同様に、分布関数 F は一点 $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ に対して、

$$F(\mathbf{x}) \stackrel{\mathrm{def}}{\equiv} P\{\mathbf{X} \le \mathbf{x}\} = P\{X_1 \le x_1, X_2 \le x_2, \ldots, X_p \le x_p\} \tag{3.2}$$

つまり、x_i を第 i 軸 $(i = 1, 2, \ldots, p)$ の最大値とする超区間を事象とした場合のその事象の確率、として定義される。勿論、$F(-\infty, -\infty, \ldots, -\infty) = 0$ である。さらに、$F(\mathbf{x})$ が $\mathbf{x}$ で微分可能な場合、密度関数を

$$f(x_1, x_2, \ldots, x_p) = \frac{\partial F(x_1, x_2, \ldots, x_p)}{\partial x_1 \partial x_2 \cdots \partial x_p} \tag{3.3}$$

と定義する。微積分学の基本定理により、

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_p} f(x_1, x_2, \ldots, x_p) dx_1 dx_2 \cdots dx_p \tag{3.4}$$

の関係が成り立つ。

そうすると、一変数のときと同様に、確率ベクトル $\mathbf{X}$ の関数 $g(\mathbf{X})$ [1] の期待値は

$$\mathbb{E}g = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(\mathbf{x}) f(x_1, x_2, \ldots, x_p) dx_1 dx_2 \cdots dx_p \tag{3.5}$$

と定義される。特に、第 i 軸への**射影関数** $g_i(\mathbf{X}) = X_i$ を考えると、各変量の期待値（平均）が

$$\mathbb{E}X_i = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i f(x_1, x_2, \ldots, x_p) dx_1 dx_2 \cdots dx_p \tag{3.6}$$

[1] 確率変数（ベクトル）の関数は確率変数であることに注意。

と定義される。これを用いて、平均ベクトルを

$$\mathbb{E}\mathbf{X} = \begin{pmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \vdots \\ \mathbb{E}X_p \end{pmatrix} \tag{3.7}$$

と定義する。以降では、記法の便利さとして、

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)' = \mathbb{E}\mathbf{X} \tag{3.8}$$

を平均ベクトルとして使うことにする。

分散に関しては、一変数と異なり、第 i 軸における平均回りの二次モーメントだけでなく、第 i 軸と第 j 軸での平均回りの関係を考慮するために、分散 $\mathbb{V}$ に加え**共分散** $\mathbb{C}$ を

$$\mathbb{V}X_i = \mathbb{E}(X_i - \mu_i)^2 \tag{3.9}$$

$$\mathbb{C}(X_i, X_j) = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j) \tag{3.10}$$

として考える。また、それらをまとめて**分散（共分散）行列**を

$$\begin{aligned}
\Sigma &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} && (3.11) \\
&= \begin{pmatrix} \mathbb{V}X_1 & \mathbb{C}(X_1, X_2) & \cdots & \mathbb{C}(X_1, X_p) \\ \mathbb{C}(X_2, X_1) & \mathbb{V}X_2 & \cdots & \mathbb{C}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}(X_p, X_1) & \mathbb{C}(X_p, X_2) & \cdots & \mathbb{V}X_p \end{pmatrix} && (3.12) \\
&= \begin{pmatrix} \mathbb{E}(X_1 - \mu_1)^2 & \mathbb{E}(X_1 - \mu_1)(X_2 - \mu_2) & \cdots & \mathbb{E}(X_1 - \mu_1)(X_p - \mu_p) \\ \mathbb{E}(X_2 - \mu_2)(X_1 - \mu_1) & \mathbb{E}(X_2 - \mu_2)^2 & \cdots & \mathbb{E}(X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(X_p - \mu_p)(X_1 - \mu_1) & \mathbb{E}(X_p - \mu_p)(X_2 - \mu_2) & \cdots & \mathbb{E}(X_p - \mu_p)^2 \end{pmatrix} && (3.13)
\end{aligned}$$

と定義する。分散行列に関してはよりすっきりと

$$\Sigma = \mathbb{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \tag{3.14}$$

とも表現できる（演習問題 3.1，3.2）。

分散行列は軸のスケーリングに関して不変ではない。そこで分散行列を相対化するために、**相関係数**

行列を

$$\begin{pmatrix} 1.0 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1.0 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1.0 \end{pmatrix} \stackrel{\text{def}}{\equiv} \begin{pmatrix} 1.0 & \sigma_{12}/(\sigma_1\sigma_2) & \cdots & \sigma_{1p}/(\sigma_1\sigma_p) \\ \sigma_{21}/(\sigma_2\sigma_1) & 1.0 & \cdots & \sigma_{2p}/(\sigma_2\sigma_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}/(\sigma_p\sigma_1) & \sigma_{p2}/(\sigma_p\sigma_2) & \cdots & 1.0 \end{pmatrix} \tag{3.15}$$

$$= \begin{pmatrix} 1.0 & \mathbb{C}(X_1,X_2)/\sqrt{\mathbb{V}X_1}\sqrt{\mathbb{V}X_2} & \cdots & \mathbb{C}(X_1,X_p)/\sqrt{\mathbb{V}X_1}\sqrt{\mathbb{V}X_2} \\ \mathbb{C}(X_2,X_1)/\sqrt{\mathbb{V}X_2}\sqrt{\mathbb{V}X_1} & 1.0 & \cdots & \mathbb{C}(X_2,X_p)/\sqrt{\mathbb{V}X_2}\sqrt{\mathbb{V}X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}(X_p,X_1)/\sqrt{\mathbb{V}X_p}\sqrt{\mathbb{V}X_1} & \mathbb{C}(X_p,X_2)/\sqrt{\mathbb{V}X_p}\sqrt{\mathbb{V}X_2} & \cdots & 1.0 \end{pmatrix} \tag{3.16}$$

$$= \left(\frac{\mathbb{E}(X_i-\mu_i)(X_j-\mu_j)}{\sqrt{\mathbb{E}(X_i-\mu_i)^2}\sqrt{\mathbb{E}(X_j-\mu_j)^2}} \right) \quad \text{第 } (i,j) \text{ 成分} \tag{3.17}$$

を考えることも直感的な議論には向いている。この**相関係数** $\rho_{ij} = \frac{\mathbb{E}(X_i-\mu_i)(X_j-\mu_j)}{\sqrt{\mathbb{E}(X_i-\mu_i)}\sqrt{\mathbb{E}(X_j-\mu_j)}}$ に対して、

$$-1 \leq \rho_{ij} \leq 1 \tag{3.18}$$

を示すこともたやすい（演習問題 3.3, 3.4）。

さて、準備ができたところで、ベクトルと行列を使ってこれらをエレガントに演算できように準備しよう。

始めに、線形性を確かめる。色々な線形性はあるものの基本は積分が線形、つまり、スカラー a, b と関数 $f(\cdot), g(\cdot)$ に対して

$$\int (af(x)+bg(x))dx = a\int f(x)dx + b\int g(x)dx \tag{3.19}$$

が成立することによる。これは、積分作用素を I としたとき、

$$I(af(\cdot)+bg(\cdot)) = aI(f(\cdot)) + bI(g(\cdot)) \tag{3.20}$$

と書き直すと見やすい。積分作用素 $\mathbb{E}$ に関しても、一変数の時には、第 3.1 章の演習問題 (2.10) で既に示している。二変数の時に一例を示す。（ただし、後述の周辺分布の概念を使っている。）確率密度関数を $p(x,y)$ とすると、次が示される。

$$\begin{aligned} \mathbb{E}(af(\mathbf{X})+bg(\mathbf{Y})) &= \int\int (af(x)+bg(y))p(x,y)dxdy & (3.21) \\ &= \int \left(\left(\int p(x,y)dy \right) af(x) \right) dx + \int \left(\left(\int p(x,y)dx \right) bg(y) \right) dy & (3.22) \\ &= a\int f(x)p(x)dx + b\int g(y)p(y)dy & (3.23) \\ &= a\mathbb{E}(f(\mathbf{X})) + b\mathbb{E}(g(\mathbf{Y})) & (3.24) \end{aligned}$$

比較的わかりやすいのは、確率ベクトル $\mathbf{X}$（次元 p）を行列 A（サイズ $q \times p$）およびベクトル b（次元 q）で変換した確率ベクトル $\mathbf{Y} = A\mathbf{X} + b$ に関して、

$$\mathbb{E}\mathbf{Y} = \mathbb{E}(A\mathbf{X}+b) = A(\mathbb{E}\mathbf{X}) + b \tag{3.25}$$

が成立することである。どうようにして、$\mathbb{E}(\mathbf{X}A+b) = (\mathbb{E}\mathbf{X})A + b$（行列演算できるようにサイズが合っているとする）も成立する。

また、分散行列 Σ と自己相関行列 $S = \mathbb{E}\mathbf{X}\mathbf{X}'$ との関係は

$$\begin{aligned}
\Sigma &= \mathbb{E}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' \\
&= \mathbb{E}(\mathbf{X}\mathbf{X}' - \mathbf{X}\boldsymbol{\mu}' - \boldsymbol{\mu}\mathbf{X}' + \boldsymbol{\mu}\boldsymbol{\mu}') \\
&= \mathbb{E}\mathbf{X}\mathbf{X}' - \mathbb{E}\mathbf{X}\boldsymbol{\mu}' - \mathbb{E}\boldsymbol{\mu}\mathbf{X}' + \mathbb{E}\boldsymbol{\mu}\boldsymbol{\mu}' \quad \text{（線形性より）} \\
&= S - \boldsymbol{\mu}\boldsymbol{\mu}' - \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}' \\
&= S - \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned}$$

となり、どちらかからもう一方への変換は容易であることがわかる。

次に、$\mathbf{Y} = A\mathbf{X} + b$ に関して平均と分散がどう変化するか見てみよう。すぐに、

$$\begin{aligned}
\boldsymbol{\mu}_Y &= \mathbb{E}(A\mathbf{X}+b) = A\mathbb{E}\mathbf{X} + b = A\boldsymbol{\mu}_X + b \\
\Sigma_Y &= \mathbb{E}(\mathbf{Y}-\boldsymbol{\mu}_Y)(\mathbf{Y}-\boldsymbol{\mu}_Y)' \\
&= \mathbb{E}(A\mathbf{X}-A\boldsymbol{\mu}_X)(A\mathbf{X}-A\boldsymbol{\mu}_X)' \\
&= \mathbb{E}A(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{X}-\boldsymbol{\mu}_X)'A' \\
&= A\mathbb{E}(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{X}-\boldsymbol{\mu}_X)'A' \\
&= A\Sigma_X A'
\end{aligned}$$

を得る。特殊な場合として、要素が全て 1 であるようなベクトル $\mathbf{1}$ と要素が全て 0 であるようなベクトル $\mathbf{0}$ を準備して、$A = \mathbf{1}', b = \mathbf{0}$ を当てはめると、$Y = \sum_{i=1}^{p} X_i$（確率変数の和）に対して、

$$\mathbb{E}(\sum_{i=1}^{p} X_i) = \sum_{i=1}^{p} \mu_i$$

や

$$\mathbb{V}(\sum_{i=1}^{p} X_i) = \sum_{i=1}^{p} \sigma_i^2 + 2\sum_{i=1, j>i}^{p} \sigma_{ij} \tag{3.26}$$

を得る。さらに、全ての p 変量が互いに独立であれば（後述）、

$$\mathbb{V}(\sum_{i=1}^{p} X_i) = \sum_{i=1}^{p} \sigma_i^2 \tag{3.27}$$

と、確率変数の和の平均も分散も各要素のそれらの線形和で求まることが確認される。

それでは、このような変換において密度関数はどう変わるであろう。今、p 個の確率変数を持つ確率ベクトル $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ を別の p 個の確率変数を持つ確率ベクトル $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)'$ に変換することを考える。ただし、この変換は正則、つまり、逆変換を持つものとする。ここで変換の関数を、$Y_i = y_i(X_1, X_2, \ldots, X_p)$ とする。同様にして、逆変換 $X_i = x_i(Y_1, Y_2, \ldots, Y_p)$ も考える。この時、$\mathbf{X}$ が密度関数 $f(x_1, x_2, \ldots, x_p)$ を持つならば、$\mathbf{Y}$ の密度関数 $g(y_1, y_2, \ldots, y_p)$ は、

$$g(y_1, y_2, \ldots, y_p) = f(x_1(y_1, \ldots, y_p), \ldots, x_p(y_1, \ldots, y_p))J(y_1, \ldots, y_p) \tag{3.28}$$

となる。ここで、$J(y_1, \ldots, y_p)$ は

$$J(y_1, \ldots, y_p) = \left| \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{pmatrix} \right| \tag{3.29}$$

と定義さる。また、行列は「ヤコビアン」と呼ばれる。ここで、$|\cdot|$ は絶対値を取ることを意味する（行列式は正とは限らない）。直感的には、対応する微小区間で確率が同じくなるためのう要請である（多重積分の変換公式）。

2 同時分布、周辺分布

これまで、多変量の確率変数を一緒に扱う扱う方式を述べてきたが、そもそも何故、一緒に扱う必要があるのだろう？答えは、「変量間に関連があるのであれば、それを明示的に扱うためには、ベクトルとして関連づけるのがわかりやいからである。さらに、、平均ベクトルや分散行列はユークリッド空間上でのどのような分布に従って確率ベクトルが生成されるかを理解しやすい。

肝心なのが、「すべてを考える。すべて知る。」ことと「一部しか考えない。一部しか知らない。」こととの区別である。「すべてを考える。すべて知る。」から始めよう。「同時分布」とは先に述べた分布関数 $F(\mathbf{x})$ のことであり、密度関数が存在する場合には、密度関数 $f(\mathbf{x})$ であると言っても差し支えない。本節で覚えておいて欲しいことは、「**同時分布を知ることはすべてを知ることであり、確率的な振舞いはするものの、その振舞を含んで、目の前の情報をすべて知っている。**」ということである。逆に言えば、「**周辺分布（後述）を知るだけでは、全体を知ることにはならない。**」と言える。

ここでは、直感を養うために、離散確率の場合を扱う。例として、一束のトランプを考えよう。種類は $\spadesuit, \heartsuit, \clubsuit, \diamondsuit$ の四種類で、それぞれ、$A, 2-10, J, Q, K$ の 13 枚、合計 52 枚あるものとする。ゲームの内容は「ディラーが持っている 52 枚の十分よくシャッフルされた伏せたカードから一枚を選ぶ。」ことである（なんらギャンブル性はない）。さて、$\heartsuit A$ を選ぶ確率は、$P(X=\heartsuit, Y=A)=1/52$ と書くことができる。この時、確率変数は二つで、それぞれ、$X=$suit、$Y=$number　である。取り得る値はそれぞれ、4 種類と 13 種類である。

このとき、同時確率は $P(X,Y)$ であり、全ての可能な (x,y) のペアに関して、その確率 $P(X=x, Y=y)$ がわかるということが、すなわち、「全てを知る」を意味する。この時、事象は、全体集合 $\Omega=\{(x,y)|x=1,2,3,4,\quad y=1,2,\ldots,13\}$[2] の任意の部分集合であり、$A=$「偶数の赤いスートのカードの集合」など、自由なカードの組み合わせが相当する。

「周辺分布」は一部の変量だけに関する確率情報であり、$P(X=x)$ などと書かれる。正式には、

$$P(X=x) \overset{\text{def}}{\equiv} \sum_y P(X=x, Y=y)$$

と 定義する[3]。わかりやすさのため、ここでは、全範囲に渡って総和を取ることを $P(X=x)=P(X=x,*)$ と書くことにする。連続変数の場合も同様であり、よく「積分消去して周辺化する」などと言われる。Y に関しても同様に、$P(Y=y)=P(*,Y=y)$ と書ける。これらは定義ではあるものの直感と整合している。つまり、注目している変量だけの確率を求めるにはその他の変量の変化をすべて吸収すればよいことを定義としている。ただし、多変量の場合には、あくまで「周辺分布は同時分布から定義されたもの」であることを認識しておく必要がある。具体例では、$P(X=\spadesuit)=\sum_{y=1}^{13} P(X=\spadesuit, y)=1/4$ や $P(Y=Q)=\sum_{x=1}^{4} P(x, Y=Q)=1/13$ となる。

すべての周辺分布を知ることが同時分布を知ることにならないことを確認するために特殊なカードを使おう。$\spadesuit$ は $A, 2, 3$ の 3 枚,$\heartsuit$ は $4, 5, 6$ の 3 枚、$\clubsuit$ は $8, 9, 10$ の 3 枚,$\diamondsuit$ は J, Q, K の 3 枚、合計で 12 枚のカードのみを使おう（ラッキー 7 を使わない）。その時、周辺分布は、$P(X=\spadesuit)=1/4$（他のスート

[2]確率変数のありがた味がここでよくわかる。X は取り得る（値）の集合から実数値（今は整数値）への変換で、Y も同様である。この例では、それぞれを $X: \spadesuit \to 1, \heartsuit \to 2, \clubsuit \to 3, \diamondsuit \to 4$ および $Y: A \to 1, J \to 11, Q \to 12, K \to 13, i \to i (i=2,\ldots,10)$ と定めることで、扱いが容易になっている。

[3]どれが定義（与えたもの）でどれが定理（導かれたもの）であるかを常に意識することはいつでも重要である。

も同様）と $P(Y=i)=1/12\ (i \neq 7)$ である。しかし、3 つの数字の割り当てを自由に変更しても（例えば、♠ と ♡ の数字を取り替える）同じ周辺分布を得る。よって、「すべての周辺分布を知っても同時分布を知ることにはならない」。

最後は「条件つき分布」である。これは、二つ以上の確率変数がある場合に、どちらを知ることで、もう一方の確率情報が変わることがあるために、とても重要な役割を果たす。$X=x$ が既知である場合の Y の条件つき分布は

$$P(Y=y|X=x) \overset{\text{def}}{\equiv} \frac{P(X=x,Y=y)}{P(X=x)}$$

と 定義される。この式は、周辺分布の定義を使うと、

$$P(Y=y|X=x)=\frac{P(X=x,Y=y)}{P(X=x)}=\frac{P(X=x,Y=y)}{\sum_{y'} P(X=x,Y=y')}$$

と、同時分布のみを使っても定義される。トランプの例では、

$$P(X=\clubsuit|Y=K)=\frac{P(X=\clubsuit,Y=K)}{P(Y=K)}==\frac{1/52}{1/13}=1/4$$

と周辺分布 $P(X=\clubsuit)=1/4$ と変わらない。これは、数はスートに何ら影響を与えないという事実を反映しており、式で書くと、

$$P(X=x|Y=y)=P(X=x)$$

が成立していることを示している。この背景には、より強く、

$$P(X=x,Y=y)=P(X=x)P(Y=y)$$

ということ、つまり、X と Y が統計的に独立であるという事実が隠れている。一般的には、（任意の順位づけにおける）チェインルール [4]

$$P(X=x,Y=y)=P(X=x)P(Y=y|X=x)=P(Y=y)P(X=x|Y=y)$$

は常に成り立ち、「今回は偶然 $P(X=x|Y=y)=P(X=x)$ であった。」つまり、「X の発生は Y とは無関係に一定であった。」とも言える。例えば、スペードの絵札だけを除いた 49 枚のカードを考えた場合、条件つき確率

$$P(X=\clubsuit|Y=K)=\frac{P(X=\clubsuit,Y=K)}{P(Y=K)}==\frac{1/49}{3/49}=1/3$$

は周辺確率 $P(X=\clubsuit)=13/49$ より大きな値となる。また、$P(X=\spadesuit|Y=K)=0$ であり、$P(X=\spadesuit)=10/49$ とまったく異なる。このように二つの確率変数間に相互作用がある場合、片方の値を知ることは、もう一方の値を推測することの確度を高めることにつながる。この辺りの事情は、相互情報量を考えることでより明確に捉えることができる。

連続分布に関しても、これらの概念はそのまま定義される。二変量 X,Y の場合を考えよう。X の周辺（累積）分布は

$$F(x) \overset{\text{def}}{\equiv} P(X \leq x)=P(X \leq x, Y \leq \infty)=F(x,\infty) \tag{3.30}$$

明らかに

$$F(x)=\int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)dudv \tag{3.31}$$

[4] 一般に $P(X_1,X_2,\ldots,X_p)=P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\cdots P(X_p|X_1,\ldots,X_{p-1})$ が成立し、変数の順番は自由に変えてよい。この式は条件つき確率の定義を繰り返し適用することで得られる。

が成立するので、

$$f(x) = \int_{-\infty}^{\infty} f(x, v) dv \tag{3.32}$$

を X の周辺密度関数 $f(x)$ と呼ぶ。この時、

$$F(x) = \int_{-\infty}^{x} f(u) du \tag{3.33}$$

となる。条件つきの場合も同様にして、

$$F(x|y) = F(x, y)/F(y) \tag{3.34}$$

$$f(x|y) = f(x, y)/f(y) \tag{3.35}$$

と定義される。p 変量のときも同様である。

3 独立性

独立性は、一つの変量の発生確率に他方の変量がまったく影響したいことを意味する。例えば、二つの確率変数 X,Y が独立であるとは、

$$F(x, y) = F(x)F(y) \quad \text{あるいは}\ f(x, y) = f(x)f(y) \tag{3.36}$$

であることをいう。p 変量の場合も同様に、それらが **互いに独立である**とは

$$F(x_1, x_2, \ldots, x_p) = F_1(x_1)F_2(x_2)\cdots F_p(x_p) \tag{3.37}$$

の等式が成立することを意味する。

独立な場合には種々の計算が簡略される。例えば、複雑な原点回りのモーメントでさえも、

$$\begin{aligned}
\mathbb{E}X_1^{h_1}X_2^{h_2}\cdots X_p^{h_p} &= \int_{-\infty}^{\infty} x_1^{h_1}x_2^{h_2}\cdots x_p^{h_p} f_1(x_1)f_2(x_2)\cdots f_p(x_p)dx_1dx_2\cdots dx_p \\
&= \prod_{i=1}^{p}\int_{-\infty}^{\infty} x_i^{h_i} f_i(x_i)dx_i \\
&= \prod_{i=1}^{p}\mathbb{E}X_i^{h_i}.
\end{aligned}$$

と変量毎のモーメントの積で書ける。

4 ベイズの定理

ベイズの定理は単純であるものの、確率推論に欠かせない重要な定理である。離散の場合には、二つの事象 A,B の同時確率は、チェインルールにより、

$$\begin{aligned}
P(A, B) &= P(A)P(B|A) \qquad (3.38)\\
&= P(B)P(A|B) \qquad (3.39)
\end{aligned}$$

と二通りに書ける。解釈は、「同時に二つの事象が観測される場合、先ず A が起き、その後に B が起きたと見なすことも、その逆でも良く、結果的に、二つとも生じれば良い。」ということになる。右辺の上下の式からベイズの定理を得る：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{3.40}$$

連続量の場合も密度関数を使い、

$$f(\mathbf{y}|\mathbf{x}) = \frac{f(\mathbf{x}|\mathbf{y})f(\mathbf{y})}{f(\mathbf{x})} \tag{3.41}$$

となる。

このベイズ定理は単純であるが、時間順序を入れ、事象を「因」（要因）と「果」（結果）に分けると、その破壊力がわかる。

$$P(\text{要因} \mid \text{結果}) = \frac{P(\text{結果} \mid \text{要因})P(\text{要因})}{P(\text{結果})} \tag{3.42}$$

つまり、ある結果からその要因を探る場合、各要因の起こりやすさ $P(\text{要因})$ と、各要因の下で起こり得る結果の確率 $P(\text{結果} \mid \text{要因})$ がわかれば、結果から元の要因を推測できる、ということである。ここで、分母の $P(\text{結果})$ は気にしなくてよい。なぜなら、$P(\text{結果}) = P(\text{結果}, *) = \sum_{\text{要因}} P(\text{結果} \mid \text{要因})P(\text{要因})$ であるので、個々の $P(\text{結果} \mid \text{要因})P(\text{要因})$ がわかればその和を取ることで容易に求まるからである。

一方、このままでは、敢えて順序をいれた効果は見えない。それは、今後検討する「推論」のし易さに関係している。通常、要因はコントロール可能であるか頻度を容易に推測しく、同様に、要因を固定した上で結果を観測する「頻度分析」も多くの場合可能であるからである。

ベイズ推論（ベイズの定理を利用した推論）は、ベイズの定理を連続的に施すことでその強力さがわかる。今度は、証拠が積み重なって、要因の推論が徐々に確からしくなることを概念的に示そう。

$$\begin{aligned} P(\text{要因} \mid \text{証拠１、証拠２}) &\propto P(\text{証拠１、証拠２} \mid \text{要因})P(\text{要因}) & (3.43)\\ &= P(\text{証拠２} \mid \text{証拠１}, \text{要因})P(\text{証拠１} \mid \text{要因})P(\text{要因}) & (3.44) \end{aligned}$$

これを、証拠が一つしかない場合

$$P(\text{要因} \mid \text{証拠１}) \propto P(\text{証拠１} \mid \text{要因})P(\text{要因}) \tag{3.45}$$

と比べると違いが明確になる。増える証拠とともに推論が強まる様子が見える。

5 多次元正規分布

各種の検討でよく使われる分布は「多次元正規分布」であり、次式で定義される：

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \tag{3.46}$$

ここで、

1. $\boldsymbol{\mu}$ は平均ベクトルで $\boldsymbol{\mu} = \mathbb{E}X$ で定義され、
2. Σ は分散（共分散）行列で $\Sigma = \mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})'$ で定義される。

ここで、本質的な部分は、$\mathbf{x}$ を含む指数部であり（それ以外は積分値が 1 となるような正規化定数と見なすことも多い）、マハラノビス（2 乗）距離

$$D_M^2(\mathbf{x}, \boldsymbol{\mu}; \Sigma) = (\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{3.47}$$

が大きい程密度関数の値が小さく、$\mathbf{x} = \boldsymbol{\mu}$ で最大値を取るようになっている [5]。

マハラノビス距離の解釈は様々な知見を得るために重要である：

1. Σ は正値対称行列であるため、距離の公理を満足する。

2. $\Sigma = I$(単位行列) のときは、ユークリッド二乗距離に他ならない。

3. Σ^{-1} は（後述のように）方向に応じて距離を調節する役割を担っている。

もう少し詳しく説明するには、第 1 章の線形代数のところで扱った対称行列のスペクトル分解式（1.30）が役に立つ。分散行列は、固有値 λ の大きい順に、対応する固有ベクトル $\boldsymbol{\phi}$ を使って、

$$\begin{aligned} \Sigma &= \lambda_1 \boldsymbol{\phi}_1 \boldsymbol{\phi}_1' + \lambda_2 \boldsymbol{\phi}_2 \boldsymbol{\phi}_2' + \cdots + \lambda_p \boldsymbol{\phi}_p \boldsymbol{\phi}_p' & (3.48) \\ &= \Phi \Lambda \Phi' & (3.49) \end{aligned}$$

と展開される。ここで、

$$\Phi = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_p)$$

は固有ベクトル（縦）を横に順に並べてできる $p \times p$ 行列で、

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

は固有値を対角に並べたの対角行列である（固有ベクトルと対応が取れていることに注意）。重要なことは、Σ は正値対称行列なので、

1. 固有値 λ_i は全て実数かつ非負

2. 異なる固有値に対応する固有ベクトルは直交、つまり、$\phi_i'\phi_j = \delta_{ij}$ （クロネッカーのデルタ）。ただし、各固有ベクトルはノルム 1 に正規化するとする。

となることである（演習 3.2）。このことは、

$$\Phi'\Phi = I \quad \text{かつ} \quad \lambda_i \geq 0 \quad (i = 1, 2, \ldots, p) \tag{3.50}$$

と書ける。

その場合、Σ の逆行列もすぐに求められ、

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\lambda_1}\boldsymbol{\phi}_1 \boldsymbol{\phi}_1' + \frac{1}{\lambda_2}\boldsymbol{\phi}_2 \boldsymbol{\phi}_2' + \cdots + \frac{1}{\lambda_p}\boldsymbol{\phi}_p \boldsymbol{\phi}_p' & (3.51) \\ &= \Phi \Lambda^{-1} \Phi' & (3.52) \end{aligned}$$

となる。

[5]ただし、最大値は 1 を超すこともあることに注意する。

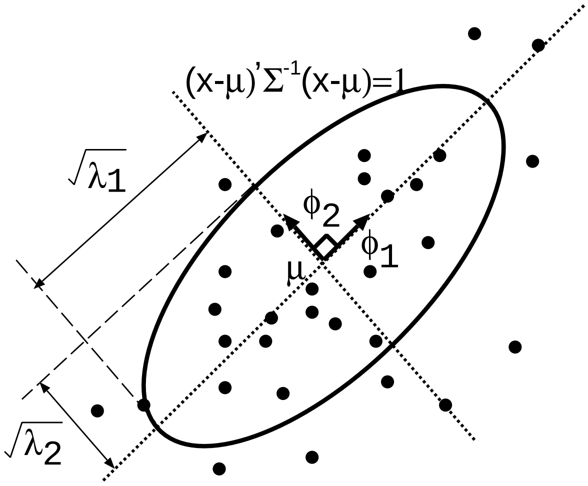

図 3.1: データの分布と当てはめた正規分布におけるマハラノビス距離 = 1 の楕円。

さて、それではマハラノビス距離をこの式を使って展開すると、

$$\begin{aligned} D_M^2(\mathbf{x}, \boldsymbol{\mu}; \Sigma) &= (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) &(3.53)\\ &= (\mathbf{x}-\boldsymbol{\mu})'\sum_{i=1}^{p}\frac{1}{\lambda_i}\boldsymbol{\phi}_i\boldsymbol{\phi}_i'(\mathbf{x}-\boldsymbol{\mu}) &(3.54)\\ &= \sum_{i=1}^{p}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\phi}_i\boldsymbol{\phi}_i'(\mathbf{x}-\boldsymbol{\mu}) &(3.55)\\ &= \sum_{i=1}^{p}\frac{1}{\lambda_i}||(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\phi}_i||^2 &(3.56)\\ &= \sum_{i=1}^{p}||\frac{1}{\sqrt{\lambda_i}}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\phi}_i||^2 &(3.57)\end{aligned}$$

つまり、$\mathbf{x}-\boldsymbol{\mu}$ を ϕ_i $(i=1,2,\ldots,p)$ へ射影 [6] した後、射影長を二乗した後に $\frac{1}{\lambda_i}$ $(i=1,2,\ldots,p)$ 倍してから和をとったものであることがわかる。ここで、各固有ベクトルは直交しているので、要は、原点を $\boldsymbol{\mu}$ に取り、$\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_p$ を正規直交基底とする座標系に変換した後に、各軸を $\frac{1}{\sqrt{\lambda_i}}$ で伸縮し、最後に、ユークリッド（二乗）距離を計算することと等しい（演習 3.5）。意味的には、「分散が大きい（λ が大きい）程、そちら方向へデータが大きく広がる傾向があり、分散を考慮する場合、それらで正規化して置くことが望ましい。」ことを実現している。

具体的な様子を図 3.1 に示す。これまでの議論がこの図から明瞭に見ることができる。ここでは、データ点から平均ベクトルおよび分散行列を推定したとして描写している。

さらに、行列式やトレースの意味も、正規分布の場合は式（3.48）より明瞭である。実際、

$$\Sigma = \Phi\Lambda\Phi', \quad \Phi'\Phi = I$$

[6] $||\mathbf{a}|| = 1$ のとき、$\mathbf{x}'\mathbf{a}$ は $\mathbf{x}$ を $\mathbf{a}$ へ射影した像の長さに対応する。

に気づくだけで、行列式の分解定理 (1.26) およびトレースの交換律 (1.9) より、

$$|\Sigma| = \prod_{i=1}^{p} \lambda_i \tag{3.58}$$

$$\mathrm{tr}\Sigma = \sum_{i=1}^{p} \lambda_i \tag{3.59}$$

であることがわかる (演習 3.6)。つまり、図 3.1 において $\pm\sqrt{\lambda_i}\phi_i$ $(i = 1, 2, \ldots, p)$ で囲まれる超長方形を考えたとき、行列式 $|\Sigma|$ はその体積の二乗、トレースは、その対角線の長さの二乗であることがわかる。大雑把に言うと、この二つは分散の程度を計る値である。

6 関連文献

これまでの議論は、データ解析、確率過程、パターン認識のいずれにおいても必要な基礎知識を必要最小限で用意したものである。記法は文献 [2] の第 2 章と付録に基づく。また、内容はパターン認識の標準的な教科書である [3, 4] の付録に整理されているものとほぼ同じである。

7 演習問題

いずれも、曖昧な理解を明解化するのに助けになるので、是非、解いて欲しい。

演習問題 3.1

式 (3.14) を証明せよ。

演習問題 3.2

分散共分散行列 $\Sigma = \mathbb{E}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})'$ において、次を示せ。

1. Σ は対称である。つまり、$\Sigma' = \Sigma$
2. Σ は非負定値である。つまり、任意のベクトル $\mathbf{x}$ に対して、常に、$\mathbf{x}'\Sigma\mathbf{x} \geq 0$.

さらに、対称非負定値行列に関して、

1. 固有値はすべて実数である。
2. 固有値はすべて非負である。
3. 異なる固有値に対する固有ベクトルは直交する。
4. 同じ固有値が二つある場合、対応する固有ベクトルは無限に存在するものの、それらの張る空間は一意である。

(ヒント) 複素数ベクトルの内積の定義 $(\mathbf{x}, \mathbf{y}) \stackrel{\mathrm{def}}{\equiv} \mathbf{x}'\bar{\mathbf{y}}$ ($^-$ は共役複素数) を使う。この場合、$(c\mathbf{x}, \mathbf{y}) = c(\mathbf{x}, \mathbf{y})$ や $(\mathbf{x}, c\mathbf{y}) = \bar{c}(\mathbf{x}, \mathbf{y})$、さらに $(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A^*\mathbf{y})$ (A^* は A の随伴行列 $A^* = \bar{A}'$) が得られる。

演習問題 3.3

二つの確率変数に関する Cauchy-Schwarz の不等式

$$(\mathbb{E}g(X)h(X))^2 \leq \left(\mathbb{E}g^2(X)\right)\left(\mathbb{E}h^2(X)\right) \tag{3.60}$$

を証明せよ。（ヒント：t を任意定数として $\mathbb{E}(g(X)t + h(X))^2$ を評価せよ。）この式は、両辺の平方根をとり、

$$|\mathbb{E}g(X)h(X)| \leq \sqrt{\mathbb{E}g^2(X)}\sqrt{\mathbb{E}h^2(X)} \tag{3.61}$$

と表されることも多い。

演習問題 3.4

Cauchy-Schwarz の不等式 (3.61) を用いて、相関係数 ρ_{ij} の絶対値が 1 を越えないこと（式 (3.18)）を証明せよ。

演習問題 3.5

平均ベクトル $\boldsymbol{\mu}_X$ で分散行列 Σ_X である p 次元正規分布に従う確率ベクトル $\mathbf{X}$ を考える。また、$\Sigma_X = \Phi\Lambda\Phi'$ と分解できるとする。このとき、変換 $\mathbf{Y} = \Lambda^{-\frac{1}{2}}\Phi' X$ を施すことで、

$$\Sigma_Y = I \tag{3.62}$$

となること、それ故に、

$$(\mathbf{x} - \boldsymbol{\mu}_X)'\Sigma_X^{-1}(\mathbf{x} - \boldsymbol{\mu}_X) = ||\mathbf{y} - \boldsymbol{\mu}_Y||^2 \tag{3.63}$$

となることの両方を示せ。ここで、$\Lambda^{-\frac{1}{2}} = \mathrm{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \ldots, 1/\sqrt{\lambda_p})$ とする。

演習問題 3.6

式 (3.58) を証明せよ。

参考文献

[1] 新井仁之. 線形代数 – 基礎と応用 –. 日本評論社, 2006.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc., 2nd edition, 1984.

[3] P. E. Hart R. O. Duda and D. G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., 2nd edition, 2001.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 2nd edition, 1990.