線形代数学 I: 第6回講義

データサイエンスに必要なベクトルと行列

中村 知繁

T. Nakamura | Juntendo Univ.

2025/02/08

1. 講義情報と予習ガイド

講義回: 第8回

テーマ: 2次元データと行列の積

関連項目: データの共分散、相関係数、行列表現

予習内容: 第7回「1次元データとベクトルの和と積」の復習、特に平均・分散の計算方法

1. 2次元データの共分散と相関係数の概念を理解する
2. 行列とベクトルを用いた共分散の計算方法を習得する
3. 行列とベクトルを用いた相関係数の計算方法を習得する

2. 基本概念

2.1 2次元データとは

2次元データとは、各観測対象に対して2つの変数の値が記録されているデータのことです。例えば：

- 学生の身長と体重
- 都市の気温と降水量
- 商品の価格と販売数

2次元データは、$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ のような形式で表されます。ここで:

- x_i: i番目のデータの第1変数の値
- y_i: i番目のデータの第2変数の値
- n: データ数

2.2 共分散の定義

“ **定義**: 2つの変数 X と Y の共分散 $\mathrm{Cov}(X, Y)$ は、各変数の平均からの偏差の積の平均として定義されます。

$$\mathrm{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

ここで、$\bar{x}$ と $\bar{y}$ はそれぞれ変数 X と Y の平均値です。 ”

共分散は2つの変数の関連性を測る指標であり、以下のような性質を持ちます:

- $\mathrm{Cov}(X, Y) > 0$: X が大きくなると Y も大きくなる傾向（正の相関）
- $\mathrm{Cov}(X, Y) < 0$: X が大きくなると Y は小さくなる傾向（負の相関）
- $\mathrm{Cov}(X, Y) \approx 0$: X と Y に明確な関連がない（無相関）
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$: 変数自身との共分散は分散に等しい

2.3 相関係数の定義

共分散は変数のスケールに依存するため、変数間の関連性を標準化した指標として相関係数があります。

“ **定義**: 2つの変数 X と Y の相関係数 ρ_{XY} は次のように定義されます。

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

ここで、σ_X と σ_Y はそれぞれ変数 X と Y の標準偏差です。 ”

相関係数の性質:

- $-1 \leq \rho_{XY} \leq 1$
- $\rho_{XY} = 1$: 完全な正の相関（直線的な比例関係）
- $\rho_{XY} = -1$: 完全な負の相関（直線的な反比例関係）
- $\rho_{XY} = 0$: 無相関（線形の関連性がない）

3. 共分散と相関行列の行列表示

3.1 行列とベクトルによるデータ表現

2次元データを行列とベクトルで表現しましょう。

まず、データを以下のようにベクトルで表します:

- $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$: 第1変数の値を集めたベクトル
- $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$: 第2変数の値を集めたベクトル

または、データ行列 $\mathbf{X}$ として以下のように表すこともできます:

$$\mathbf{X} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

3.2 行列とベクトルを用いた共分散の計算

ベクトル表記を用いた共分散の計算方法を示します。

まず、各変数の平均ベクトルを求めます:

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$: 第1変数の平均
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: 第2変数の平均

偏差ベクトルを定義します:

- $\mathbf{x}_{dev} = \mathbf{x} - \bar{x}\mathbf{1}$: 第1変数の偏差ベクトル
- $\mathbf{y}_{dev} = \mathbf{y} - \bar{y}\mathbf{1}$: 第2変数の偏差ベクトル

ここで、$\mathbf{1}$ は全ての要素が1である長さ n のベクトルです。

このとき、共分散は次のように計算されます:

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\mathbf{x}_{dev}^{T}\mathbf{y}_{dev}$$

3.3 行列とベクトルを用いた相関係数の計算

相関係数は、共分散を各変数の標準偏差で除することで計算できます。

各変数の分散を計算します:

- $\mathrm{Var}(X) = \frac{1}{n}\mathbf{x}_{dev}^T\mathbf{x}_{dev}$: 第1変数の分散
- $\mathrm{Var}(Y) = \frac{1}{n}\mathbf{y}_{dev}^T\mathbf{y}_{dev}$: 第2変数の分散

標準偏差を計算します:

- $\sigma_X = \sqrt{\mathrm{Var}(X)}$: 第1変数の標準偏差
- $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$: 第2変数の標準偏差

相関係数は次の式で計算されます:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathbf{x}_{dev}^T\mathbf{y}_{dev}}{\sqrt{(\mathbf{x}_{dev}^T\mathbf{x}_{dev})(\mathbf{y}_{dev}^T\mathbf{y}_{dev})}}$$

3.4 共分散行列

多変量データを扱う際には、全ての変数のペアに対する共分散を含む共分散行列が重要になります。2変数 X と Y の場合、共分散行列 $\mathbf{\Sigma}$ は次のようになります:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{bmatrix}$$

一般に、データ行列 $\mathbf{X}$ の各列が変数を表す場合、共分散行列は以下のように計算できます:

$$\mathbf{\Sigma} = \frac{1}{n}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T)^T(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T)$$

ここで、$\boldsymbol{\mu}$ は各変数の平均値を含むベクトルです。

4. 計算例と実装

4.1 数値例: 共分散と相関係数の計算 (1/3)

以下の5つのデータポイントに対して共分散と相関係数を計算してみましょう:

データ番号	X (身長 cm)	Y (体重 kg)
1	160	55
2	170	65
3	180	75
4	165	60
5	175	70

ステップ1: 各変数の平均を計算します。

- $\bar{x} = \frac{160+170+180+165+175}{5} = 170$
- $\bar{y} = \frac{55+65+75+60+70}{5} = 65$

4.1 数値例: 共分散と相関係数の計算 (2/3)

ステップ2: 各データポイントの偏差を計算します。

データ番号	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-10	-10	100
2	0	0	0
3	10	10	100
4	-5	-5	25
5	5	5	25

ステップ3: 共分散を計算します。

$$\mathrm{Cov}(X, Y) = \frac{100 + 0 + 100 + 25 + 25}{5} = \frac{250}{5} = 50$$

4.1 数値例: 共分散と相関係数の計算 (3/3)

ステップ4: 各変数の分散と標準偏差を計算します。

$$\mathrm{Var}(X) = \frac{(-10)^2 + 0^2 + 10^2 + (-5)^2 + 5^2}{5} = \frac{250}{5} = 50$$

$$\mathrm{Var}(Y) = \frac{(-10)^2 + 0^2 + 10^2 + (-5)^2 + 5^2}{5} = \frac{250}{5} = 50$$

$$\sigma_X = \sqrt{50} \approx 7.07$$

$$\sigma_Y = \sqrt{50} \approx 7.07$$

ステップ5: 相関係数を計算します。

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{50}{7.07 \times 7.07} \approx \frac{50}{50} = 1$$

この例では、相関係数が1であり、身長と体重の間に完全な正の線形相関があることを示しています。

4.2 行列を用いた計算例 (1/2)

同じデータを行列形式で表現し、計算してみましょう。

$$x = [160, 170, 180, 165, 175]^T$$

$$y = [55, 65, 75, 60, 70]^T$$

ステップ1: 各変数の平均値ベクトルを計算します。

- $\bar{x} = 170$
- $\bar{y} = 65$

ステップ2: 偏差ベクトルを計算します。

$$\mathbf{x}_{dev} = [160 - 170, 170 - 170, 180 - 170, 165 - 170, 175 - 170]^T = [-10, 0, 10, -5, 5]^T$$

$$\mathbf{y}_{dev} = [55 - 65, 65 - 65, 75 - 65, 60 - 65, 70 - 65]^T = [-10, 0, 10, -5, 5]^T$$

4.2 行列を用いた計算例 (2/2)

ステップ3: 共分散を計算します。

$$\mathrm{Cov}(X,Y) = \frac{1}{5}\mathbf{x}_{dev}^T\mathbf{y}_{dev} = \frac{1}{5}((-10)\times(-10) + 0\times 0 + 10\times 10 + (-5)\times(-5) + 5\times 5) = \frac{250}{5} = 50$$

ステップ4: 分散を計算します。

$$\mathrm{Var}(X) = \frac{1}{5}\mathbf{x}_{dev}^T\mathbf{x}_{dev} = \frac{1}{5}((-10)^2 + 0^2 + 10^2 + (-5)^2 + 5^2) = \frac{250}{5} = 50$$

$$\mathrm{Var}(Y) = \frac{1}{5}\mathbf{y}_{dev}^T\mathbf{y}_{dev} = \frac{1}{5}((-10)^2 + 0^2 + 10^2 + (-5)^2 + 5^2) = \frac{250}{5} = 50$$

ステップ5: 相関係数を計算します。

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\times\mathrm{Var}(Y)}} = \frac{50}{\sqrt{50\times 50}} = \frac{50}{50} = 1$$

5. 演習問題

5.1 基本問題

問題1: 以下のデータについて、共分散と相関係数を手計算で求めなさい。

学生	テスト点数 (X)	勉強時間 (Y) (時間)
A	85	10
B	70	7
C	90	12
D	65	6
E	80	9

問題2: 以下の2セットのデータの相関係数を比較しなさい。また、その結果からどのような解釈ができるか説明しなさい。

データセット1:

```
X: [1, 2, 3, 4, 5]
Y: [2, 4, 6, 8, 10]
```

データセット2:

```
X: [1, 2, 3, 4, 5]
Y: [5, 4, 3, 2, 1]
```

問題3: 共分散行列が以下のように与えられている場合、相関行列を求めなさい。

$$\boldsymbol{\Sigma} = \begin{bmatrix} 16 & 12 \\ 12 & 25 \end{bmatrix}$$

問題5: あるクラスの10人の学生について、数学のテスト点数 (X)、英語のテスト点数 (Y)、勉強時間 (Z)（時間/週）のデータが以下のように得られた。

```
数学: [85, 70, 90, 65, 80, 75, 95, 60, 85, 75]
英語: [80, 75, 85, 70, 75, 80, 90, 65, 80, 70]
勉強時間: [10, 7, 12, 6, 9, 8, 14, 5, 11, 8]
```

(1) 3つの変数の共分散行列を計算しなさい。

(2) 3つの変数の相関行列を計算しなさい。

(3) 数学の点数と英語の点数の関係、数学の点数と勉強時間と勉強量の関係、英語の点数と勉強時間の関係を比較し、それぞれの相関の強さについて考察しなさい。