進捗報告

2022/05/11

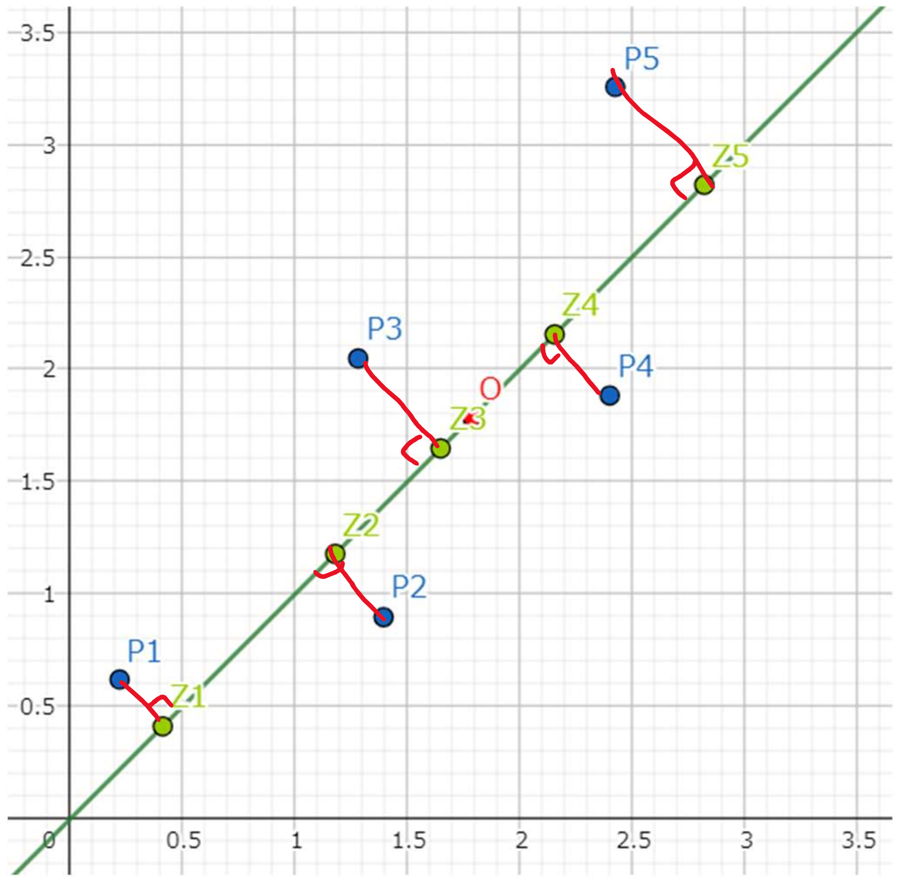
今回やったこと

- 主成分分析について

主成分分析：イメージ

- 点$P_i (i = 1 \ldots n)$の平均点Oに対して$OZ_1^2 + OZ_2^2 + \cdots + OZ_n^2$が最大となる新しい軸（緑色）を探す

→**新しい軸の分散を最大にする**



主成分分析：データの準備

- データ$\{\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}\}$に対してその平均を求める

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n}$$

- また、平均が原点に来るように平行移動する

$$\mathbf{x_n'} = \mathbf{x_n} - \bar{\mathbf{x}}$$

※以降、$\mathbf{x}'$は$\mathbf{x}$と表記する

主成分分析：射影されたデータを求める

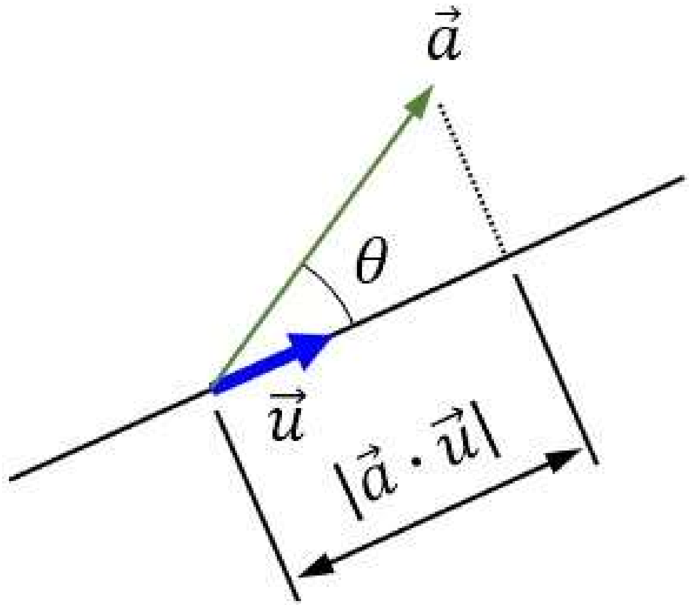
- データ$\{\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}\}$に対して、単位ベクトル$\mathbf{u}$を使って直交射影し、次の$\{z_1, z_2, \ldots z_n\}$のスカラー値を得る

$$z_i = \mathbf{u}^{\mathbf{T}}\mathbf{x_i}$$

- 単位ベクトルと任意ベクトルの内積を計算すると、単位ベクトル方向の成分（大きさ）が取得できる

主成分分析：射影されたデータを求める

- 例えば、$\vec{a}$の単位ベクトル$\vec{u}$の方向の成分は、 $\vec{a}$と$\vec{u}$の内積の絶対値で求まる



主成分分析：射影されたデータを求める

- $\vec{a}$の単位ベクトル$\vec{u}$の方向の成分は、 $\vec{a}$と$\vec{u}$の内積の絶対値で求まる
  - $\vec{a}$と$\vec{u}$のなす角をθとすると

$$L = |\vec{a}|cos\theta$$

ここで、$\vec{a}$と$\vec{u}$の内積は

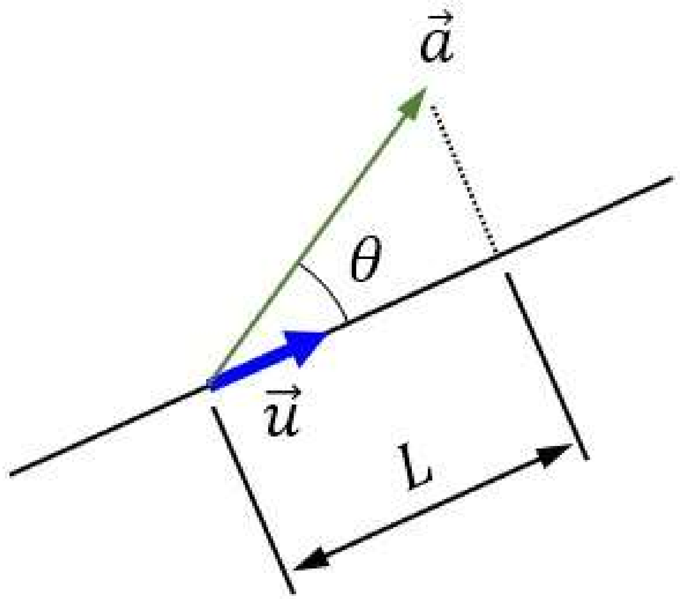
$$\vec{a} \cdot \vec{u} = |\vec{a}||\vec{u}|cos\theta$$

単位ベクトルの大きさは1なので

$$\vec{a} \cdot \vec{u} = |\vec{a}|cos\theta$$

$$L = \vec{a} \cdot \vec{u}$$

$$\boxed{z_i = \mathbf{u}^{\mathbf{T}}\mathbf{x_i}}$$



主成分分析：問題の定式化

- **射影されたデータの集合$\{z_n\}$の分散が最大化されるようにuの値を決定したい。つまり、分散をV_zとすると、**

$$V_z = \frac{1}{N}\sum_{n=1}^{N} z_n^2$$

を最大化したいという問題になる

分散$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

主成分分析：問題の定式化

$$V_z = \frac{1}{N}\sum_{n=1}^{N} z_n^2$$

$$\therefore z_n = \mathbf{u}^{\mathbf{T}}\mathbf{x}_{\mathbf{n}}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{u}^{\mathbf{T}}\mathbf{x}_{\mathbf{n}}\right)^2$$

$$= \mathbf{u}^{\mathbf{T}}\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_{\mathbf{n}}\mathbf{x}_{\mathbf{n}}^{\mathbf{T}}\right)\mathbf{u}$$

$$= \mathbf{u}^{\mathbf{T}}\mathbf{V}_{\mathbf{x}}\mathbf{u}$$

主成分分析：問題の定式化

$$= \mathbf{u}^{\mathrm{T}}\mathbf{V}_{\mathbf{x}}\mathbf{u}$$

$$\mathbf{V}_{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_{\mathbf{n}}\mathbf{x}_{\mathbf{n}}^{\mathrm{T}}$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

は分散共分散行列

分散…データのばらつきを表す $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$

共分散…二組の対応するデータの間の関係を表す $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$

分散共分散行列…n変量のデータから得られるn個の分散とn(n-1)個の共分散をn次の正方行列にまとめたもの

主成分分析：問題を解く

- 式$V_z = \mathbf{u}^{\mathbf{T}}\mathbf{V_x}\mathbf{u}$を$\mathbf{u}$に関して最大化する
  - $\mathbf{u}$は単位ベクトルなので$\mathbf{u}^{\mathbf{T}}\mathbf{u} = \mathbf{1}$
  - ラグランジュの未定乗数λを導入して

$$L(\lambda, \mathbf{u}) = \mathbf{u}^{\mathbf{T}}\mathbf{V_x}\mathbf{u} + \lambda(1 - \mathbf{u}^{\mathbf{T}}\mathbf{u})$$

$\mathbf{u}$に関して微分して0と置くことで

$$\mathbf{2}\mathbf{V_x}\mathbf{u} - 2\lambda\mathbf{u} = \mathbf{0}$$

$$\mathbf{V_x}\mathbf{u} = \lambda\mathbf{u}$$

→分散共分散行列$\mathbf{V_x}$の固有値問題に帰着する

主成分分析：問題を解く

- 式$V_z = \mathbf{u}^{\mathrm{T}}\mathbf{V_x}\mathbf{u}$を$\mathbf{u}$に関して最大化する

$$\mathbf{V_x}\mathbf{u} = \lambda\mathbf{u}$$

→分散共分散行列$\mathbf{V_x}$の固有値問題に帰着する

両辺に$\mathbf{u}^{\mathrm{T}}$をかける。また、$\mathbf{u}^{\mathrm{T}}\mathbf{u} = 1$より、

$$\lambda = \mathbf{u}^{\mathrm{T}}\mathbf{V_x}\mathbf{u}$$

これはもともと最大化したいと考えていたV_zにほかならないので、求めている問題は固有値問題を解いたときの最大固有値が最大値となり、その最大値を満たす$\mathbf{u}$は最大固有値に属する固有ベクトルである