主成分分析（2次元版）

Ver.2022.09.03

愛媛大学
データサイエンスセンター（CDSE）
理工学研究科／理学部

まつうら まさや
松浦　真也

2020年4月設立

CDSE
Center for Data Science, Ehime University

主成分分析（目的）

高次元だと取り扱いや直感的理解が困難　→　データの次元を落とす

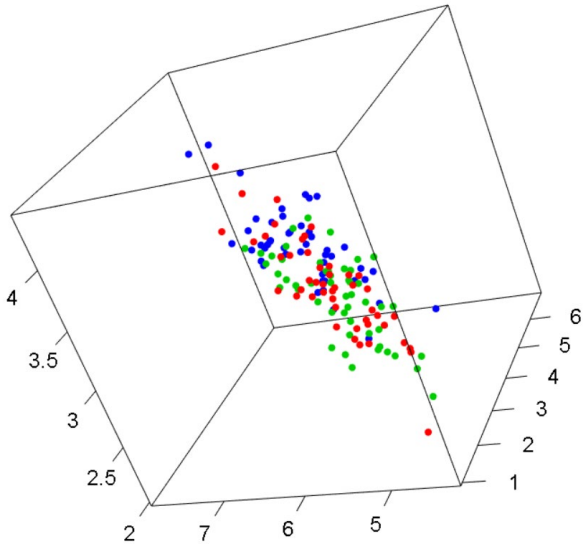
例：3次元データを2次元平面上に描くとき、どの角度から見た図がベスト？

Edgar Andersonによる3種類のアヤメの花のデータ（非常に有名）

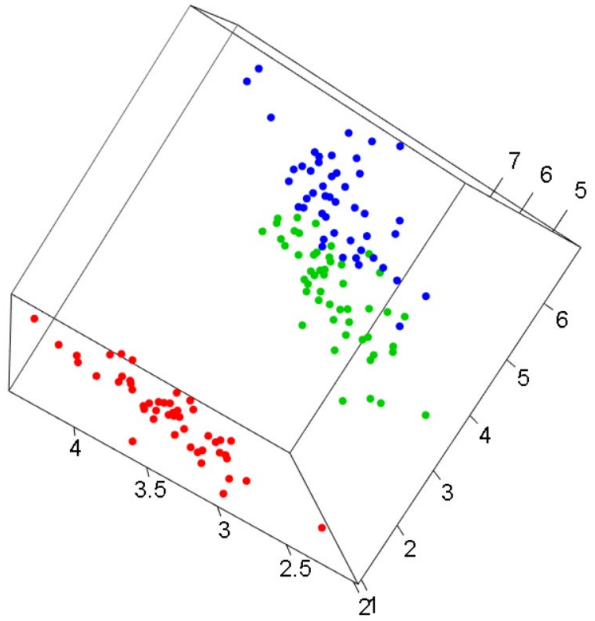
出典：R. A. Fisher, "The use of multiple measurements in taxonomic problems", Annals of Eugenics, Vol. 7, No. 2, 179–188, 1936.

※萼片（がくへん）の長さ・幅と花弁の長さ・幅の4次元データ

※ここでは、萼片の長さ・幅、花弁の幅の3次元データとしてプロット



点が重なって見にくい



点が分離して見やすい

主成分分析（目的）

【例】受講生の学力の分布を要約したい（2次元 → 1次元）

データ出典：The Data And Story LibraryのMidterms
https://dasl.datadescription.com/datafile/midterms/

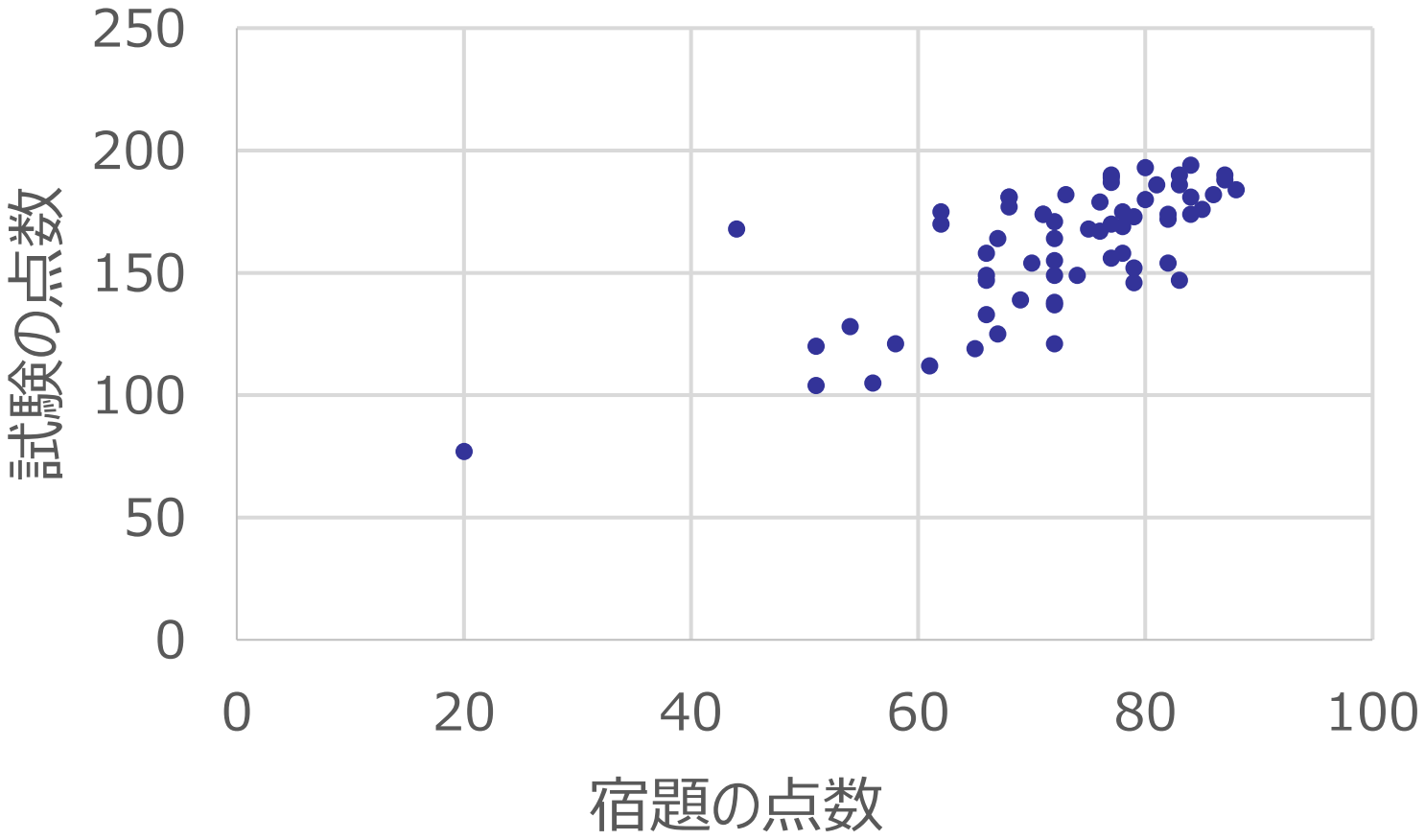
学生	宿題の点数	試験の点数
Adam	62	175
Alexandra	78	169
Alexis	68	181
Amandeep	54	128
Annie	68	181
Benjamin	72	149
Brian	76	167
Brian	82	172
⋮	⋮	⋮
Yvon	82	154

主成分分析（目的）

【例】受講生の学力の分布を要約したい（2次元 → 1次元）

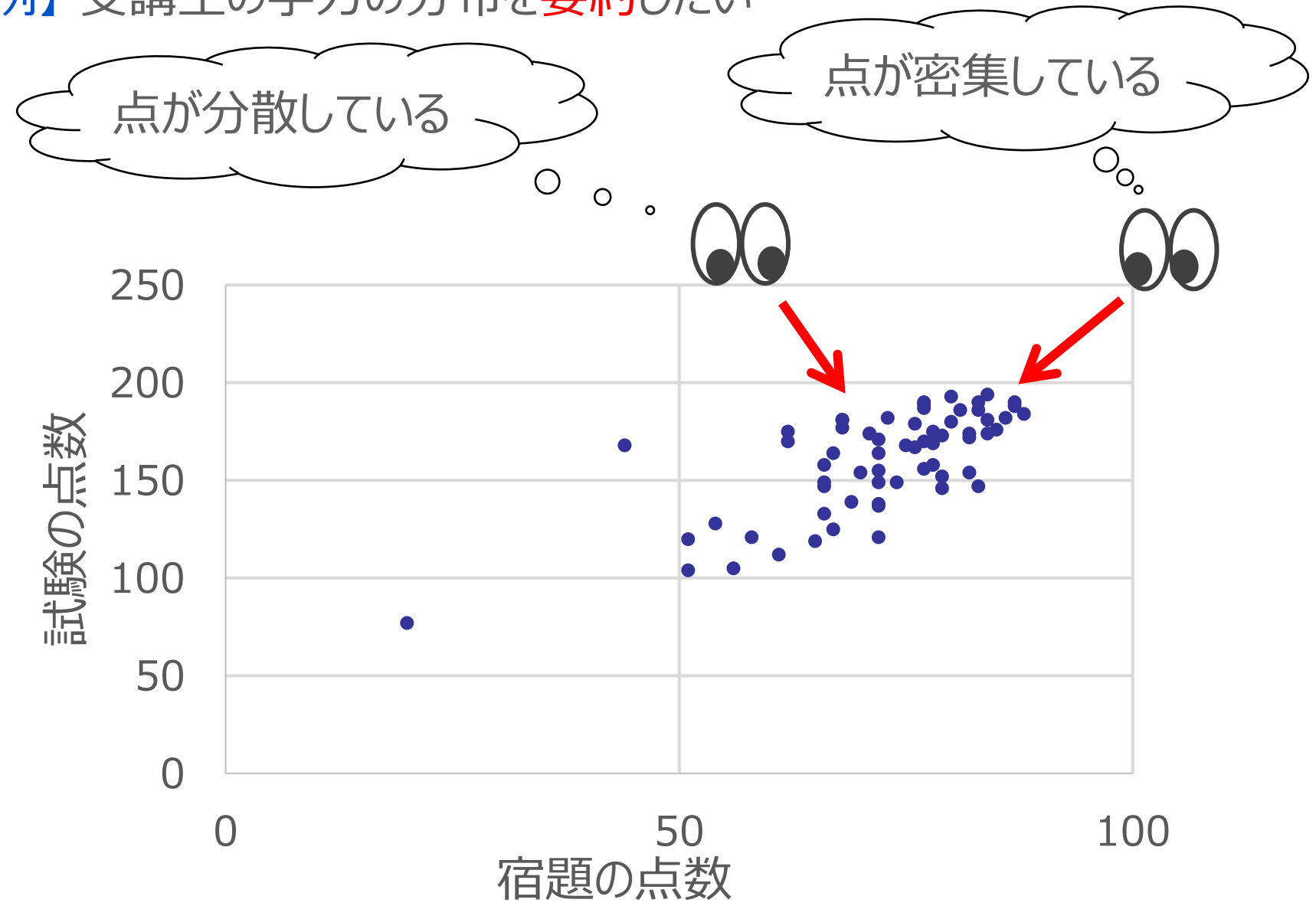
データ出典：The Data And Story LibraryのMidterms

https://dasl.datadescription.com/datafile/midterms/



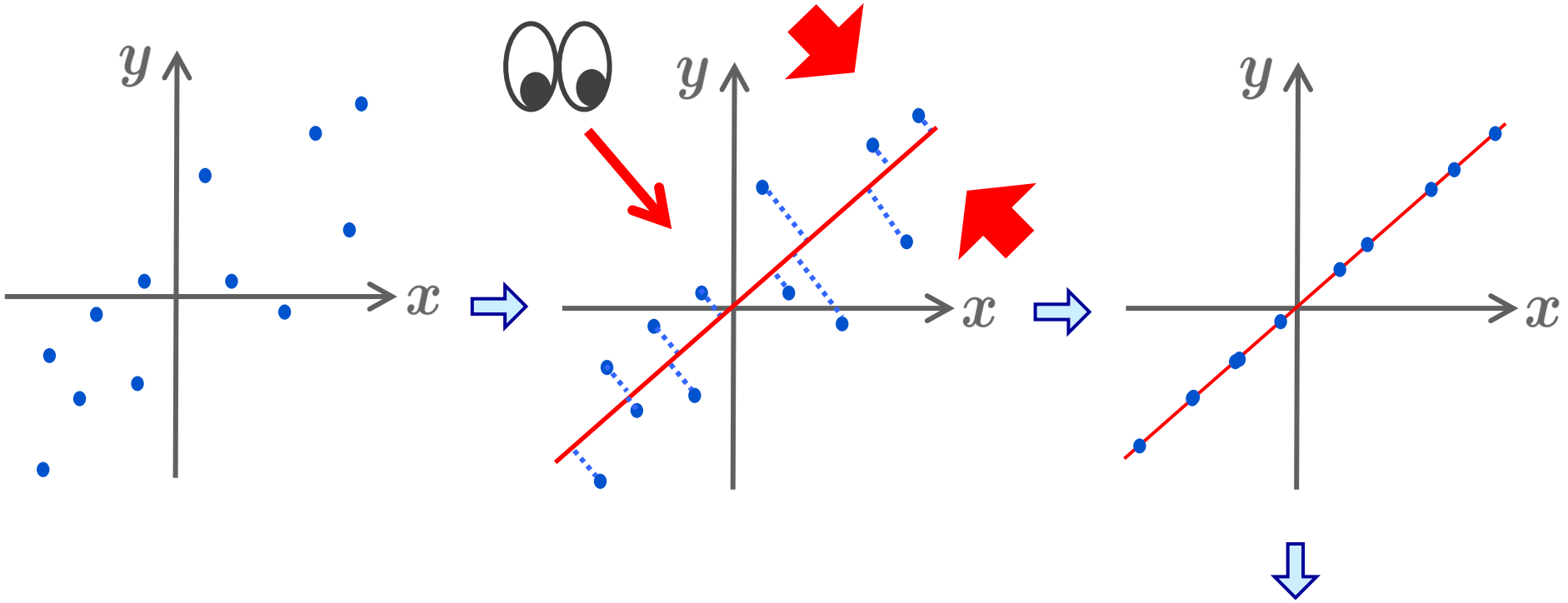
主成分分析（目的）

【例】受講生の学力の分布を要約したい



主成分分析（次元圧縮）

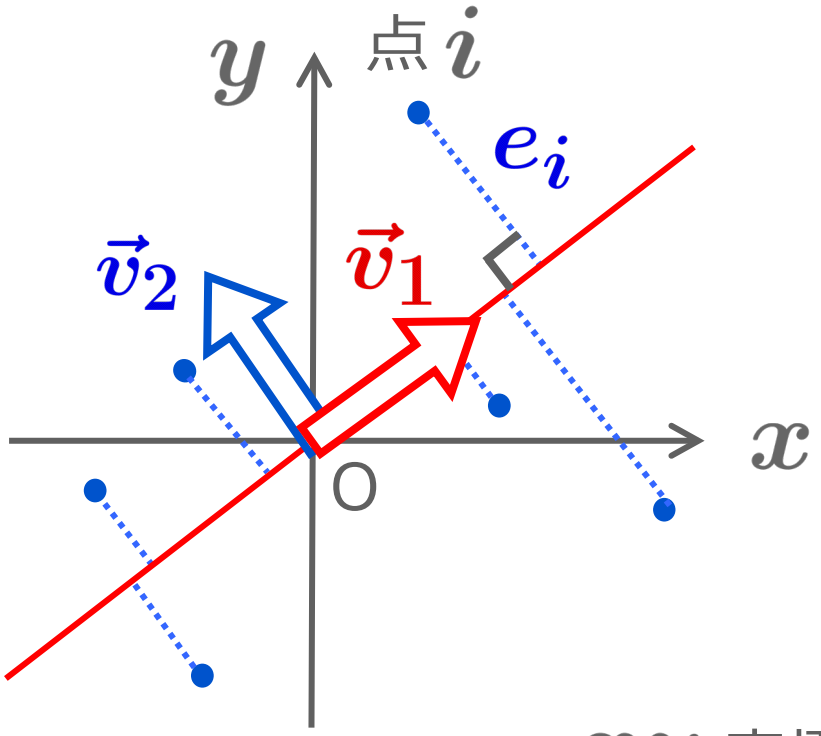
【例】受講生の学力の分布を要約したい（2次元 → 1次元）

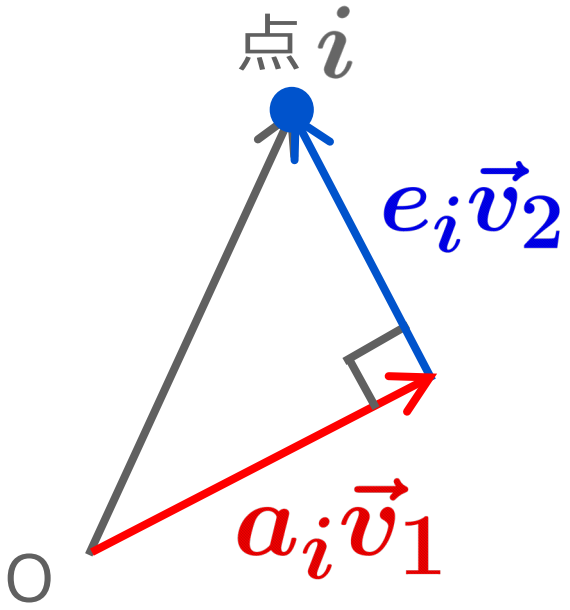


2次元データを1次元に圧縮

主成分分析（座標系の変換）

点と直線の距離の平方和 $\sum_{i=1}^{n} e_i^2$ を最小にする単位ベクトル $\vec{v}_1$ は？

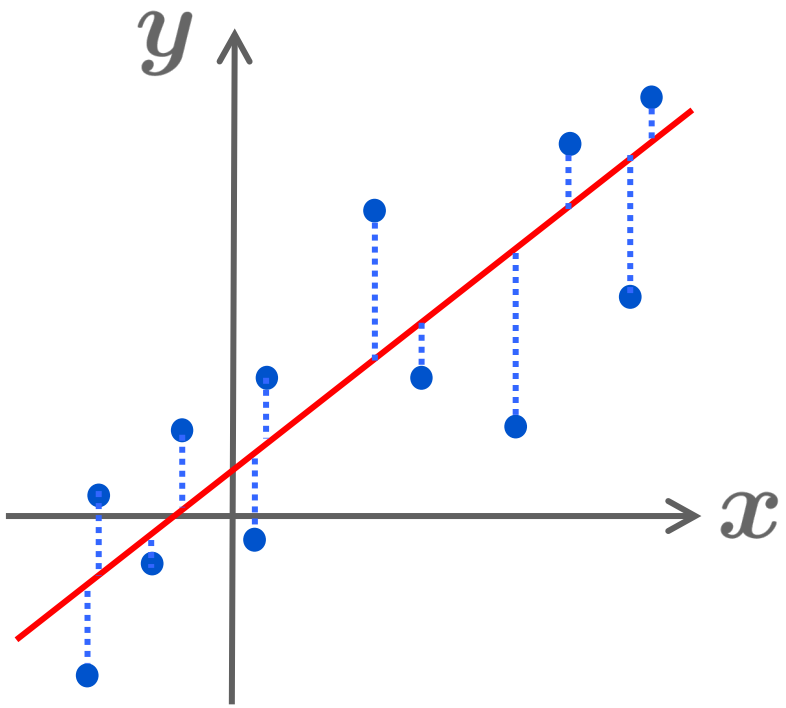




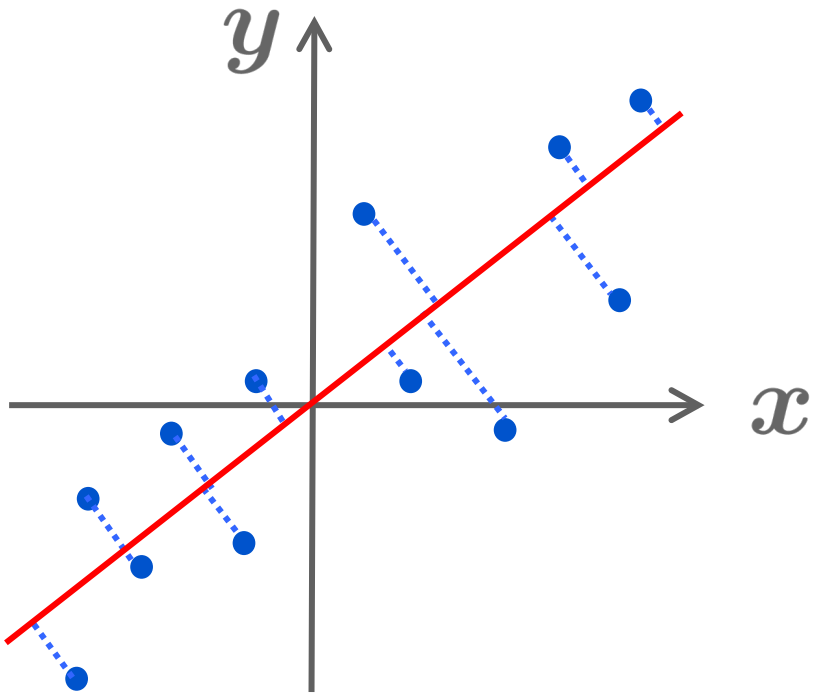
xy 座標系を $\vec{v}_1\vec{v}_2$ 座標系に変換

主成分分析（回帰分析との比較）

回帰分析



主成分分析



y 座標（目的変数）の値を近似

⇩

y 座標方向のズレを最小化

平面上の点の散らばりを近似

⇩

点と直線の距離を最小化

主成分分析（データの規格化）

名前 i	宿題の点数 x_i	試験の点数 y_i
1	62	175
2	78	169
⋮	⋮	⋮
n	68	181

標本平均 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$

標本標準偏差

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

規格化 $z_{i1} = \frac{x_i - \overline{x}}{s_x}, \quad z_{i2} = \frac{y_i - \overline{y}}{s_y}$ （平均 0、分散 1 ）

主成分分析（相関係数行列）

相関係数

$$r_{11} = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^2 = 1, \quad r_{12} = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}z_{i2},$$

$$r_{21} = \frac{1}{n-1}\sum_{i=1}^{n} z_{i2}z_{i1}, \quad r_{22} = \frac{1}{n-1}\sum_{i=1}^{n} z_{i2}^2 = 1$$

相関係数行列（対称行列）

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}$$

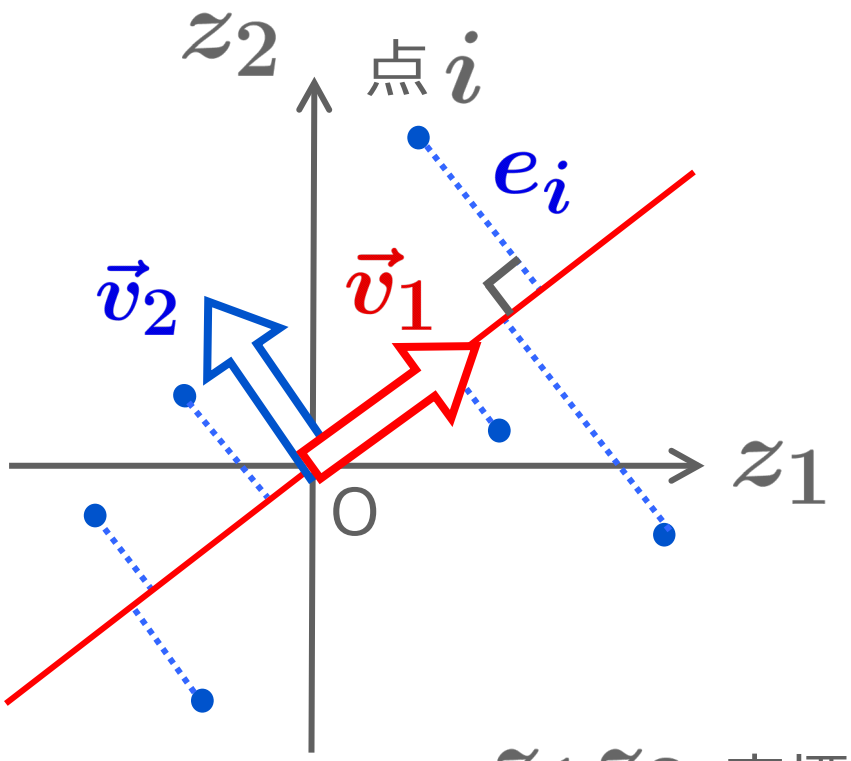
⇩

$$\vec{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix} \text{ とおくと } \quad R = \frac{1}{n-1}\begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix}\begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix}$$

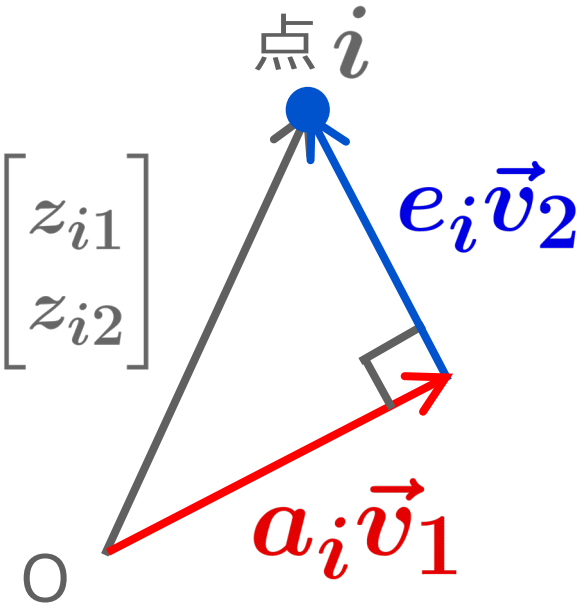
※ ${}^t\vec{z}_i$ は $\vec{z}_i$ の転置

主成分分析（座標系の変換）

点と直線の距離の平方和 $\sum_{i=1}^{n} e_i^2$ を最小にする単位ベクトル $\vec{v}_1$ は？



$$\vec{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}$$



$z_1 z_2$ 座標系を $\vec{v}_1 \vec{v}_2$ 座標系に変換

主成分分析（直交分解）

$$\vec{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}$$

$\vec{v}_1, \vec{v}_2$ を $\mathbb{R}^2$ の正規直交基底とすると

$\vec{z}_i = a_i\vec{v}_1 + b_i\vec{v}_2 \quad (a_i, b_i \in \mathbb{R})$ と書ける

$$\Downarrow$$

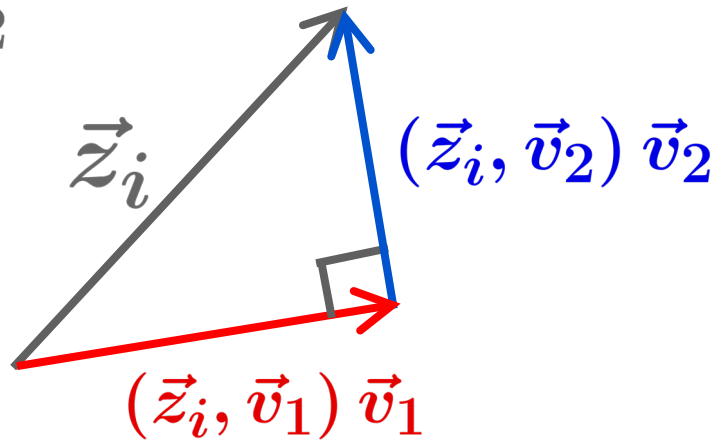
$$\begin{cases} (\vec{z}_i, \vec{v}_1) = a_i(\vec{v}_1, \vec{v}_1) + b_i(\vec{v}_2, \vec{v}_1) = a_i \\ (\vec{z}_i, \vec{v}_2) = a_i(\vec{v}_1, \vec{v}_2) + b_i(\vec{v}_2, \vec{v}_2) = b_i \end{cases}$$

$$\Downarrow$$

$$\vec{z}_i = (\vec{z}_i, \vec{v}_1)\vec{v}_1 + (\vec{z}_i, \vec{v}_2)\vec{v}_2$$

$$\Downarrow$$

$$|\vec{z}_i|^2 = (\vec{z}_i, \vec{v}_1)^2 + (\vec{z}_i, \vec{v}_2)^2$$

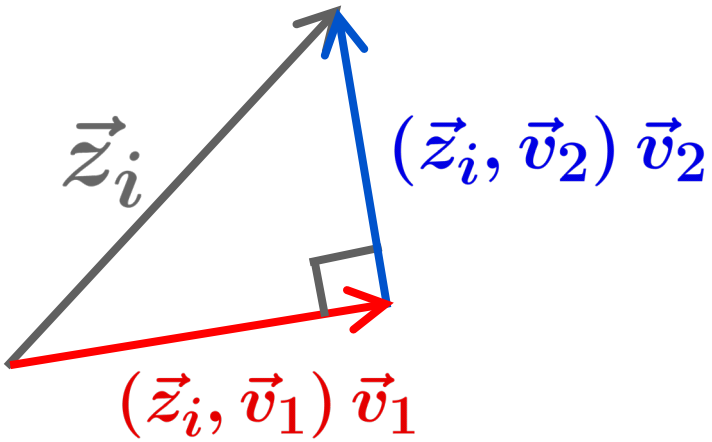


主成分分析（誤差最小化）

$$|\vec{z}_i|^2 = (\vec{z}_i, \vec{v}_1)^2 + (\vec{z}_i, \vec{v}_2)^2$$

⇩

$\sum_{i=1}^{n} (\vec{z}_i, \vec{v}_2)^2$ が最小になるように $\vec{v}_1$ を選ぶ



⇩

$\sum_{i=1}^{n} (\vec{z}_i, \vec{v}_1)^2$ が最大になるように $\vec{v}_1$ を選ぶ

$$\sum_{i=1}^{n} (\vec{z}_i, \vec{v}_1)^2 = \begin{bmatrix} (\vec{z}_1, \vec{v}_1) & \cdots & (\vec{z}_n, \vec{v}_1) \end{bmatrix} \begin{bmatrix} (\vec{z}_1, \vec{v}_1) \\ \vdots \\ (\vec{z}_n, \vec{v}_1) \end{bmatrix}$$

$$= {}^t\vec{v}_1 \begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix} \begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix} \vec{v}_1 = (n-1)\,{}^t\vec{v}_1 R \vec{v}_1$$

主成分分析（線形代数からの帰結）

定理（線形代数の一般論）

A を n 次実対称行列とする。

$|\vec{x}| = 1$ という条件のもと、${}^t\vec{x}A\vec{x}$ を最大にする $\vec{x} \in \mathbb{R}^n$ は、

A の最大固有値に対応する固有ベクトルとなる。

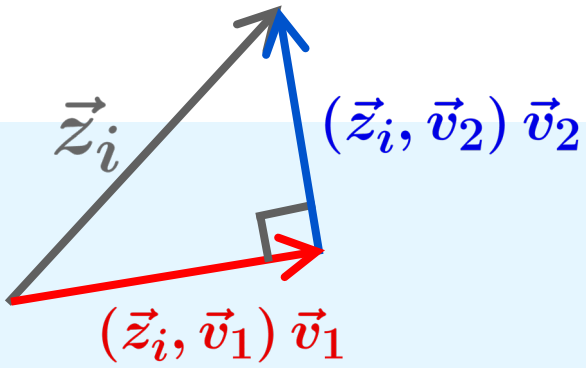
つまり、 $A\vec{x} = \lambda_{\max}\vec{x}$ （ $\lambda_{\max}$ は A の最大固有値）

$\sum_{i=1}^{n}(\vec{z}_i, \vec{v}_1)^2 = (n-1)\,{}^t\vec{v}_1 R\vec{v}_1$ を最大とする $\vec{v}_1$ は、

相関係数行列 R の最大固有値に対応する固有ベクトル

主成分分析（結論）

2次元データ (x_i, y_i) ⇒ 1次元データ $a_i = (\vec{z}_i, \vec{v}_1)$



標本平均 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$

標本標準偏差 $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

規格化 $z_{i1} = \frac{x_i - \bar{x}}{s_x}, \quad z_{i2} = \frac{y_i - \bar{y}}{s_y}, \quad \vec{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}$

相関係数行列

$$R = \frac{1}{n-1}\begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix}\begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix}$$

固有ベクトル

$\vec{v}_1$: R の最大固有値に対応する固有ベクトル

主成分分析（参考文献）

参考文献

・アヤメのデータ

R. A. Fisher, "The use of multiple measurements in taxonomic problems", Annals of Eugenics, Vol. 7, No. 2, 179–188, 1936.

・成績のデータ

The Data And Story Library, Midterms, https://dasl.datadescription.com/datafile/midterms/
（最終閲覧日 2022年9月3日）