主成分分析（一般の次元）

Ver.2022.09.04

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（2次元への圧縮）

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

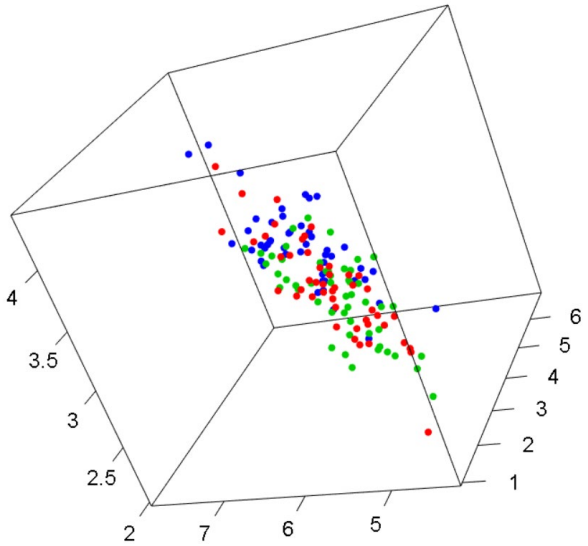
例：3次元データを2次元平面上に描くとき、どの角度から見た図がベスト？

Edgar Andersonによる3種類のアヤメの花のデータ（非常に有名）

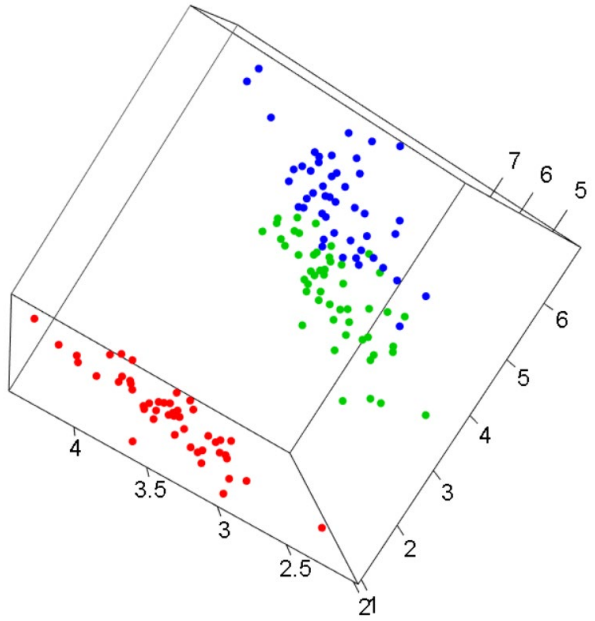
出典：R. A. Fisher, "The use of multiple measurements in taxonomic problems", Annals of Eugenics, Vol. 7, No. 2, 179–188, 1936.

※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）

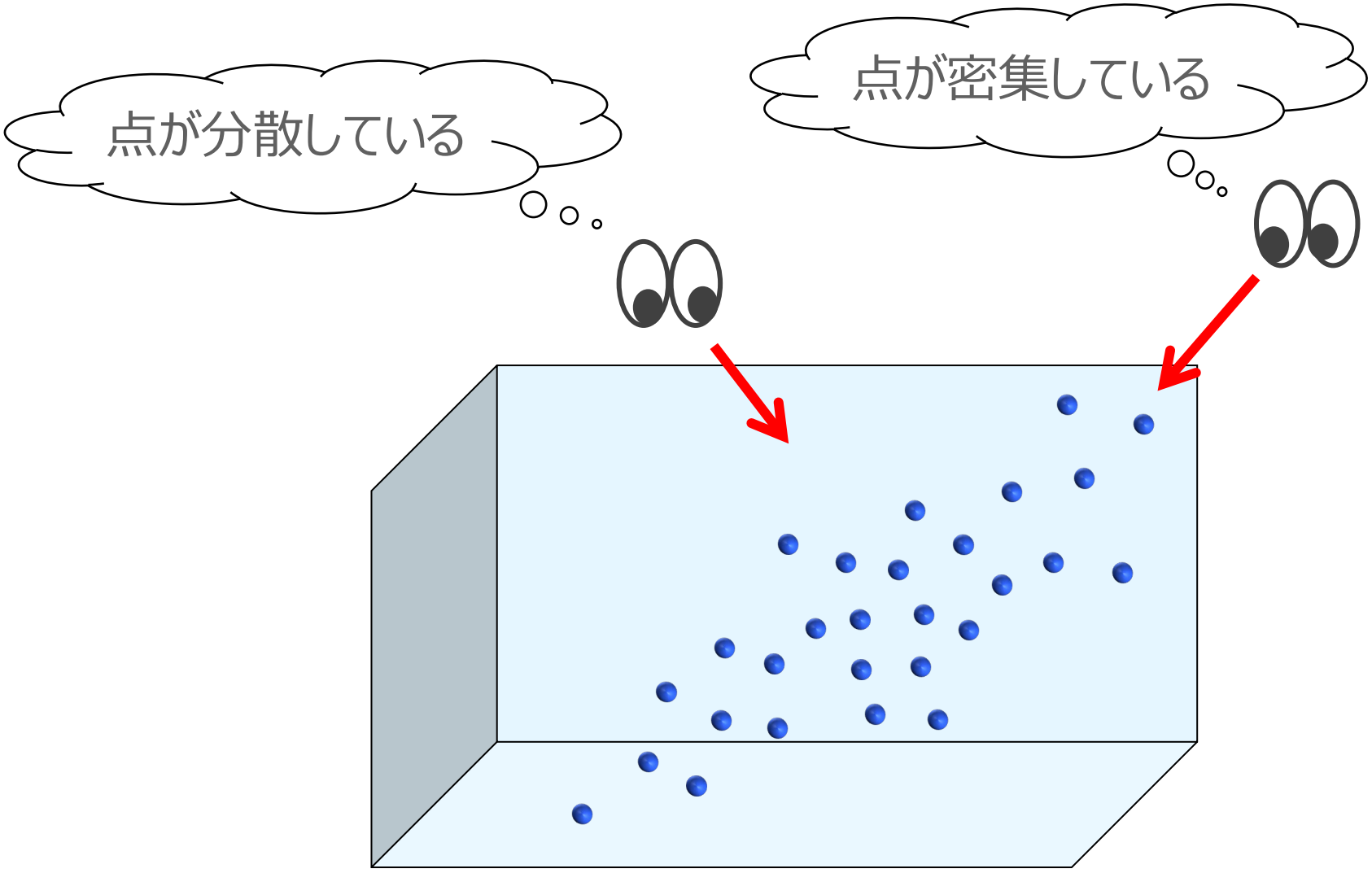
【例】アヤメの花のデータを要約したい（4次元 → 2次元）

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

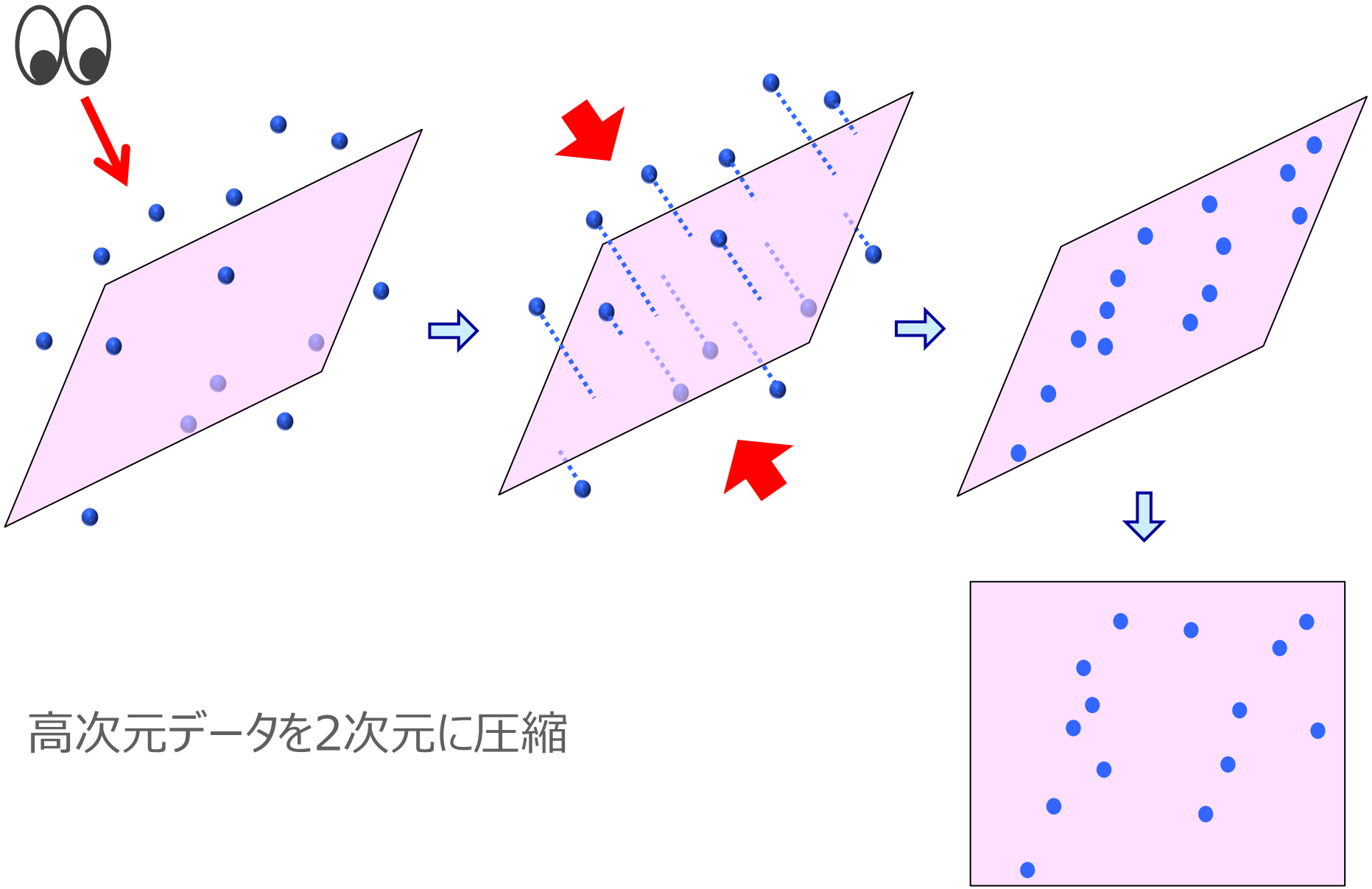
No	萼片長	萼片幅	花弁長	花弁幅	品種
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	virginica

主成分分析（目的）

【例】アヤメの花のデータを要約したい



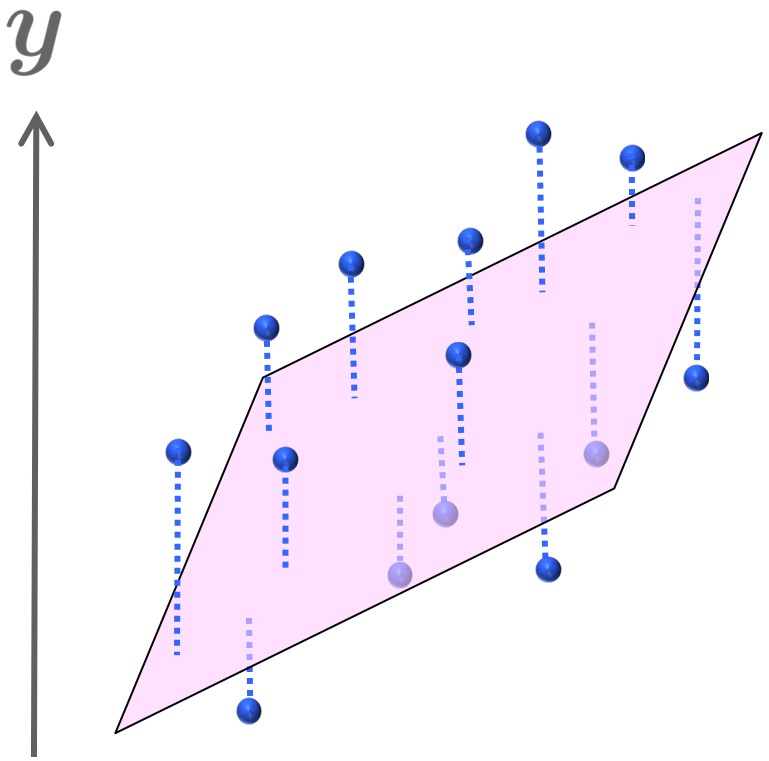
主成分分析（次元圧縮）



高次元データを2次元に圧縮

主成分分析（回帰分析との比較）

重回帰分析

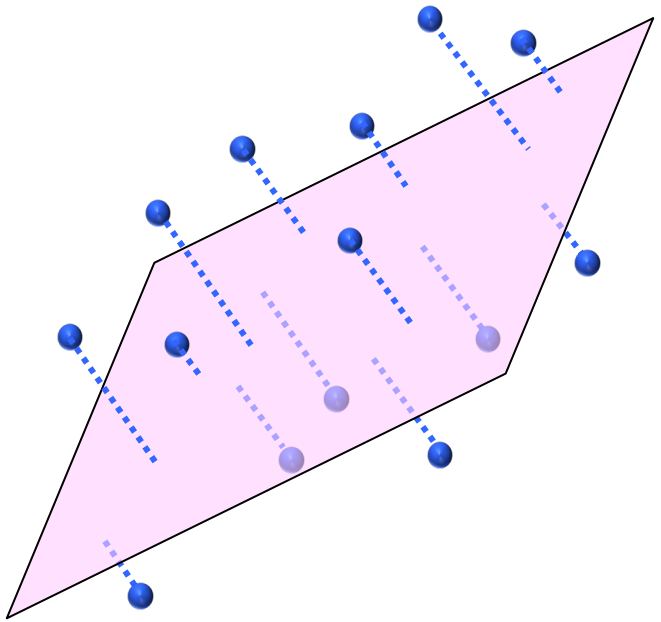


y 座標（目的変数）の値を近似

⇩

y 座標方向のズレを最小化

主成分分析



空間上の点の散らばりを近似

⇩

点と平面との距離を最小化

主成分分析（データの規格化）

No i	萼片長 x_{i1}	萼片幅 x_{i2}	…	花弁幅 x_{im}	品種
1	5.1	3.5	…	0.2	setosa
2	4.9	3.0	…	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
n	5.9	3.0	…	1.8	virginica

$k = 1, 2, \ldots, m$

標本平均 $$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$$

標本標準偏差 $$s_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}$$

規格化 $$z_{ik} = \frac{x_{ik} - \overline{x}_k}{s_k}$$ （平均 0、分散 1 ）

主成分分析（相関係数行列）

相関係数 $r_{jk} = \dfrac{1}{n-1}\sum_{i=1}^{n} z_{ij}z_{ik}$ 特に, $r_{kk} = \dfrac{1}{n-1}\sum_{i=1}^{n} z_{ik}^2 = 1$

相関係数行列（対称行列）

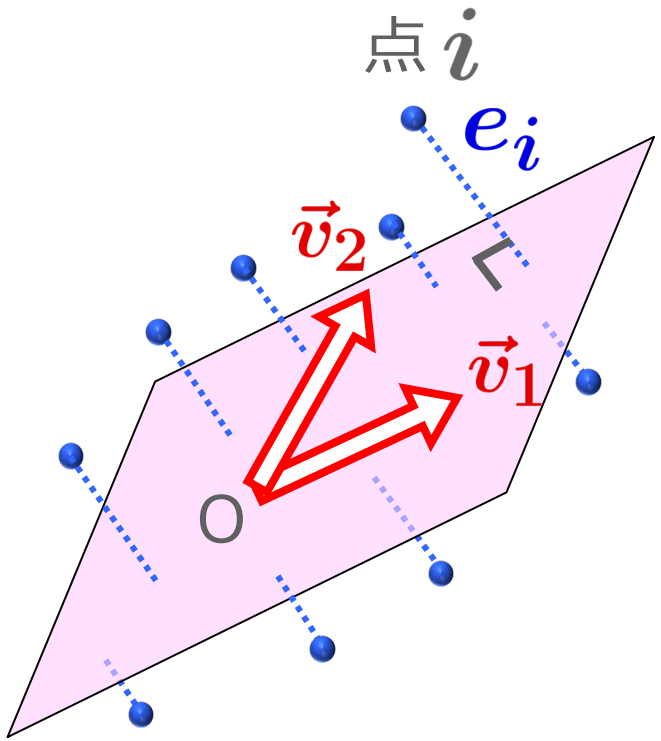
$$R = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & 1 \end{bmatrix}$$

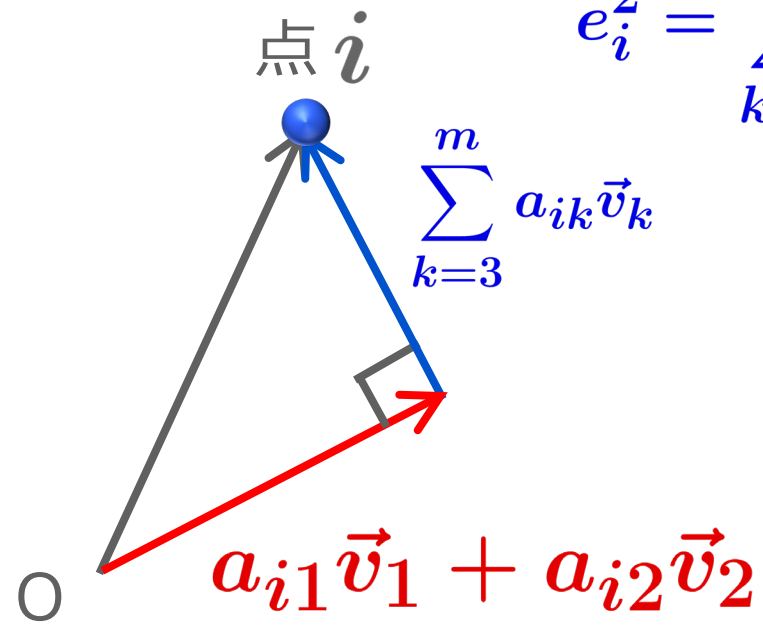
$$\vec{z}_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix} \text{とおくと} \quad R = \frac{1}{n-1}\begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix}\begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix}$$

※ ${}^t\vec{z}_i$ は $\vec{z}_i$ の転置

主成分分析（座標系の変換）

点と平面の距離の平方和 $\sum_{i=1}^{n} e_i^2$ を最小にする正規直交系 $\{\vec{v}_i\}$ は？





$$e_i^2 = \sum_{k=3}^{m} a_{ik}^2$$

$x_1 x_2 \cdots x_m$ 座標系を $\vec{v}_1 \vec{v}_2 \cdots \vec{v}_m$ 座標系に変換

主成分分析（直交分解）

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

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

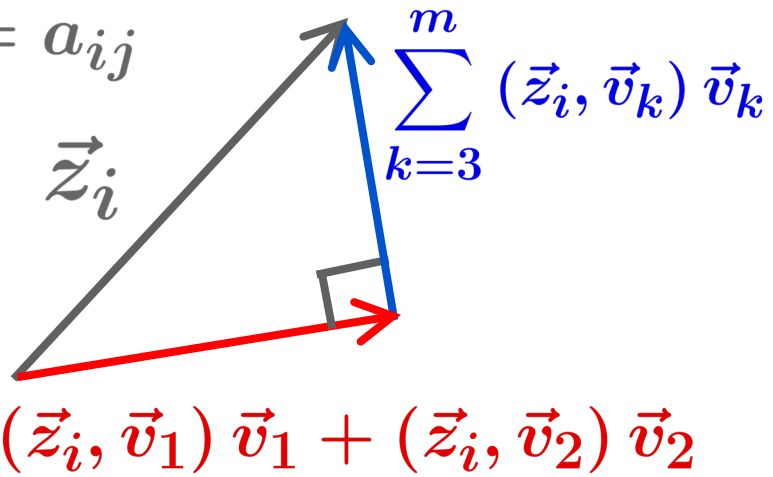
$$\vec{z}_i = \sum_{k=1}^{m} a_{ik}\vec{v}_k \quad (a_{ik} \in \mathbb{R}) \text{ と書ける}$$

$$\Downarrow$$

$$(\vec{z}_i, \vec{v}_j) = \sum_{k=1}^{m} a_{ik}(\vec{v}_k, \vec{v}_j) = a_{ij}$$

$$\Downarrow$$

$$\vec{z}_i = \sum_{k=1}^{m} (\vec{z}_i, \vec{v}_k)\vec{v}_k$$



$$\Downarrow$$

$$|\vec{z}_i|^2 = (\vec{z}_i, \vec{v}_1)^2 + (\vec{z}_i, \vec{v}_2)^2 + \sum_{k=3}^{m} (\vec{z}_i, \vec{v}_k)^2$$

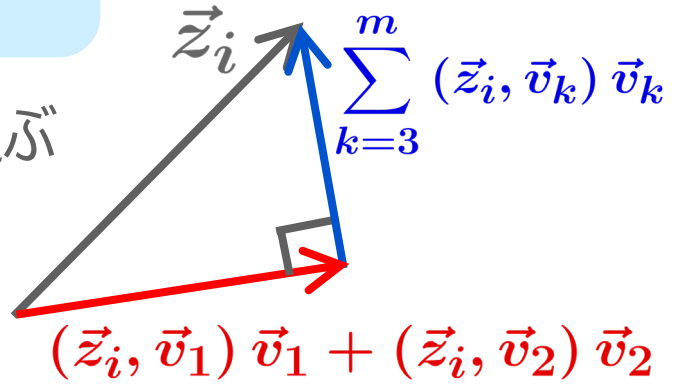
主成分分析（誤差最小化）

$$|\vec{z}_i|^2 = (\vec{z}_i, \vec{v}_1)^2 + (\vec{z}_i, \vec{v}_2)^2 + \sum_{k=3}^{m} (\vec{z}_i, \vec{v}_k)^2 \quad \leftarrow e_i^2$$

$\Downarrow$

$\sum_{i=1}^{n} \sum_{k=3}^{m} (\vec{z}_i, \vec{v}_k)^2$ が最小になるよう $\vec{v}_1, \vec{v}_2$ を選ぶ

$\Downarrow$



$\sum_{i=1}^{n} \sum_{k=1}^{2} (\vec{z}_i, \vec{v}_k)^2$ が最大になるよう $\vec{v}_1, \vec{v}_2$ を選ぶ

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

$$= \sum_{k=1}^{2} {}^t\vec{v}_k \begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix} \begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix} \vec{v}_k = (n-1)\left({}^t\vec{v}_1 R \vec{v}_1 + {}^t\vec{v}_2 R \vec{v}_2 \right)$$

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。

$|\vec{x}_1| = |\vec{x}_2| = 1, \quad (\vec{x}_1, \vec{x}_2) = 0$ という条件のもと、

${}^t\vec{x}_1 A\vec{x}_1 + {}^t\vec{x}_2 A\vec{x}_2$ を最大にする $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^n$ は、

A の最大および第2固有値に対応する固有ベクトルとなるように選べる。

つまり、$A\vec{x}_1 = \lambda_1\vec{x}_1, \quad A\vec{x}_2 = \lambda_2\vec{x}_2$

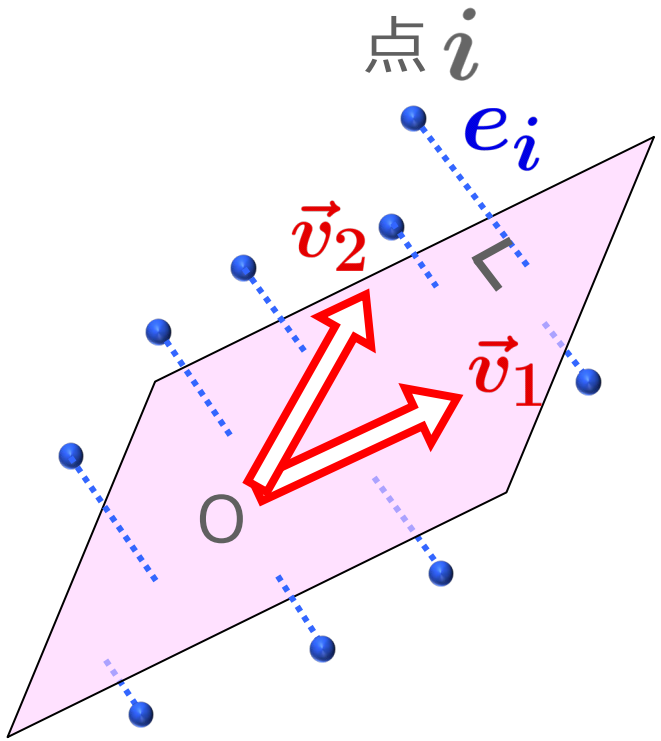
（λ_1, λ_2 は A の1番目と2番に大きい固有値）

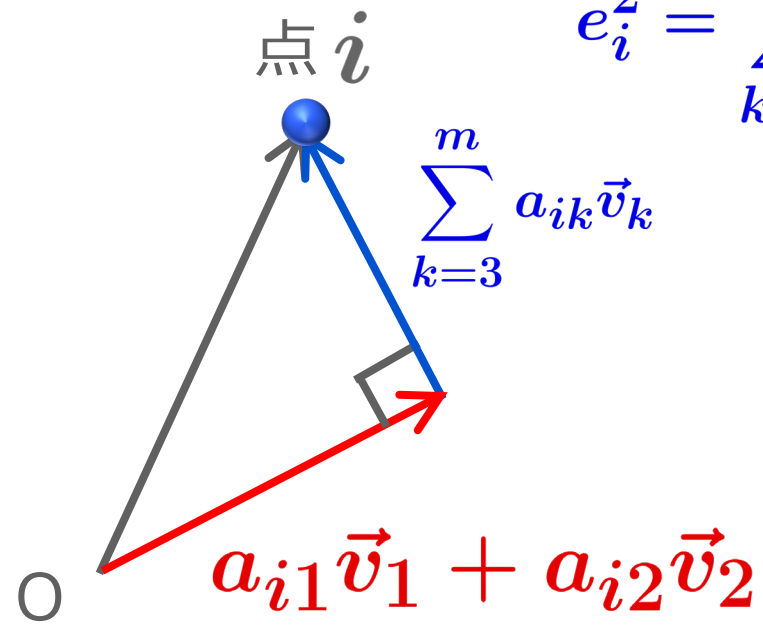
$$\sum_{i=1}^{n}\sum_{k=1}^{2}(\vec{z}_i, \vec{v}_k)^2 = (n-1)\left({}^t\vec{v}_1 R\vec{v}_1 + {}^t\vec{v}_2 R\vec{v}_2\right)$$

を最大とするには $\vec{v}_1, \vec{v}_2$ を相関係数行列 R の最大および第2固有値に対応する固有ベクトルとすればよい。

主成分分析（座標系の変換）

点と平面の距離の平方和 $\sum_{i=1}^{n} e_i^2$ を最小にする正規直交系 $\{\vec{v}_i\}$ は？





$$e_i^2 = \sum_{k=3}^{m} a_{ik}^2$$

$x_1 x_2 \cdots x_m$ 座標系を $\vec{v}_1 \vec{v}_2 \cdots \vec{v}_m$ 座標系に変換

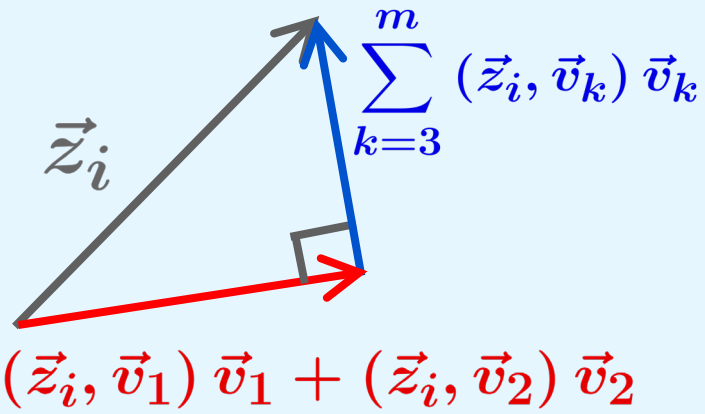
主成分分析（結論）

m次元データ ⇨ 2次元データ

$(x_{i1}, \ldots, x_{im})$ $(a_{i1}, a_{i2}) = ((\vec{z}_i, \vec{v}_1), (\vec{z}_i, \vec{v}_2))$

標本平均 $\overline{x}_k = \dfrac{1}{n}\sum_{i=1}^{n} x_{ik}$

標本標準偏差 $s_k = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}$



規格化

$$z_{ik} = \frac{x_{ik} - \overline{x}_k}{s_k},\ \vec{z}_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix}$$

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

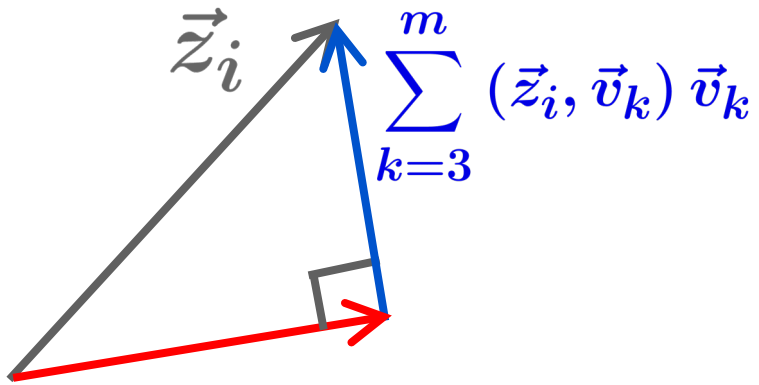
$\vec{v}_1, \vec{v}_2$: R の最大および第2固有値に対応する固有ベクトル

主成分分析（寄与率）

$$|\vec{z}_i|^2 = (\vec{z}_i, \vec{v}_1)^2 + (\vec{z}_i, \vec{v}_2)^2 + \sum_{k=3}^{m} (\vec{z}_i, \vec{v}_k)^2$$

第1主成分 $\vec{v}_1$ の寄与率

$$\frac{\sum_{i=1}^{n} (\vec{z}_i, \vec{v}_1)^2}{\sum_{i=1}^{n} |\vec{z}_i|^2}$$



第2主成分 $\vec{v}_2$ の寄与率

$$\frac{\sum_{i=1}^{n} (\vec{z}_i, \vec{v}_2)^2}{\sum_{i=1}^{n} |\vec{z}_i|^2}$$

第2主成分までの累積寄与率

$$\frac{\sum_{i=1}^{n} \{(\vec{z}_i, \vec{v}_1)^2 + (\vec{z}_i, \vec{v}_2)^2\}}{\sum_{i=1}^{n} |\vec{z}_i|^2}$$

※いずれも、0以上1以下

主成分分析（寄与率）

第1主成分の寄与率

※ tr はトレース（対角成分の和）

$$\sum_{i=1}^{n} |\vec{z}_i|^2 = \mathrm{tr}\left\{ \begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix} \begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix} \right\} = \mathrm{tr}\left\{ \begin{bmatrix} \vec{z}_1 & \cdots & \vec{z}_n \end{bmatrix} \begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix} \right\}$$

$$= \mathrm{tr}\left\{(n-1)R\right\} = (n-1)m$$

※ 一般に、
$\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$

$$\sum_{i=1}^{n} (\vec{z}_i, \vec{v}_1)^2 = (n-1)\,{}^t\vec{v}_1 R \vec{v}_1$$

$$= (n-1)\,{}^t\vec{v}_1(\lambda_1\vec{v}_1) = (n-1)\lambda_1$$

よって、 $\dfrac{\sum_{i=1}^{n}(\vec{z}_i, \vec{v}_1)^2}{\sum_{i=1}^{n}|\vec{z}_i|^2} = \dfrac{(n-1)\lambda_1}{(n-1)m} = \dfrac{\lambda_1}{m}$

第2主成分の寄与率 $\dfrac{\lambda_2}{m}$　　第2主成分までの累積寄与率 $\dfrac{\lambda_1+\lambda_2}{m}$

主成分分析（目的）

【例】アヤメの花のデータを要約したい（4次元 → 2次元）

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

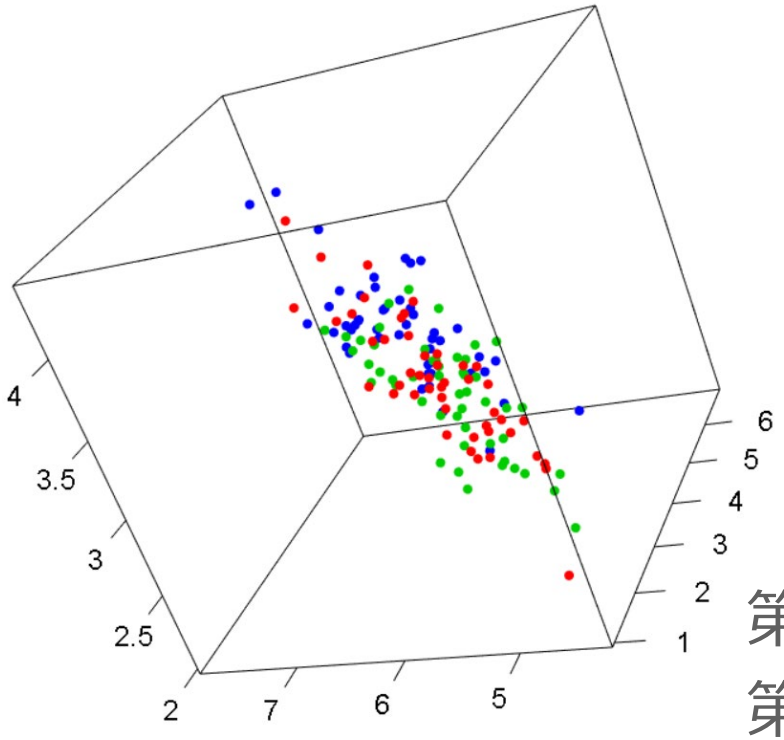
No	萼片長	萼片幅	花弁長	花弁幅	品種
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	virginica

主成分分析（解析結果）

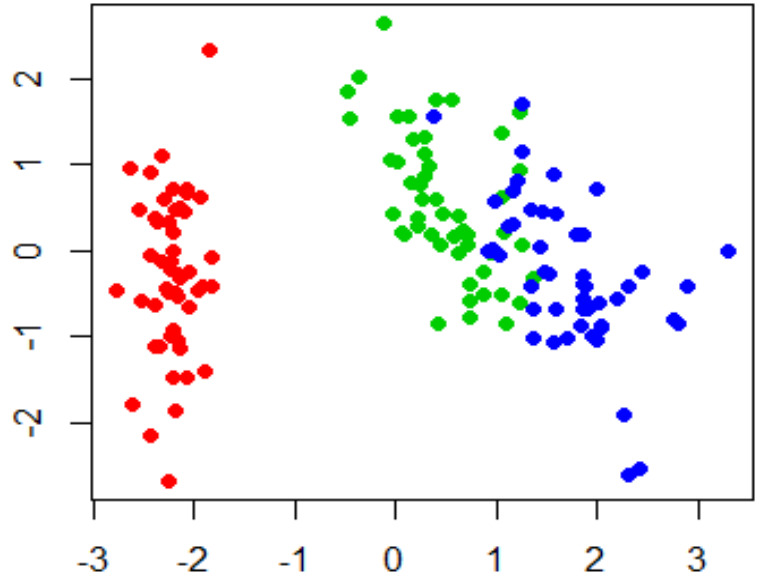
【例】アヤメの花のデータを要約したい（4次元 → 2次元）

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

元データの第1～第3成分



主成分分析の結果



第1主成分の寄与率 0.7296

第2主成分の寄与率 0.2285

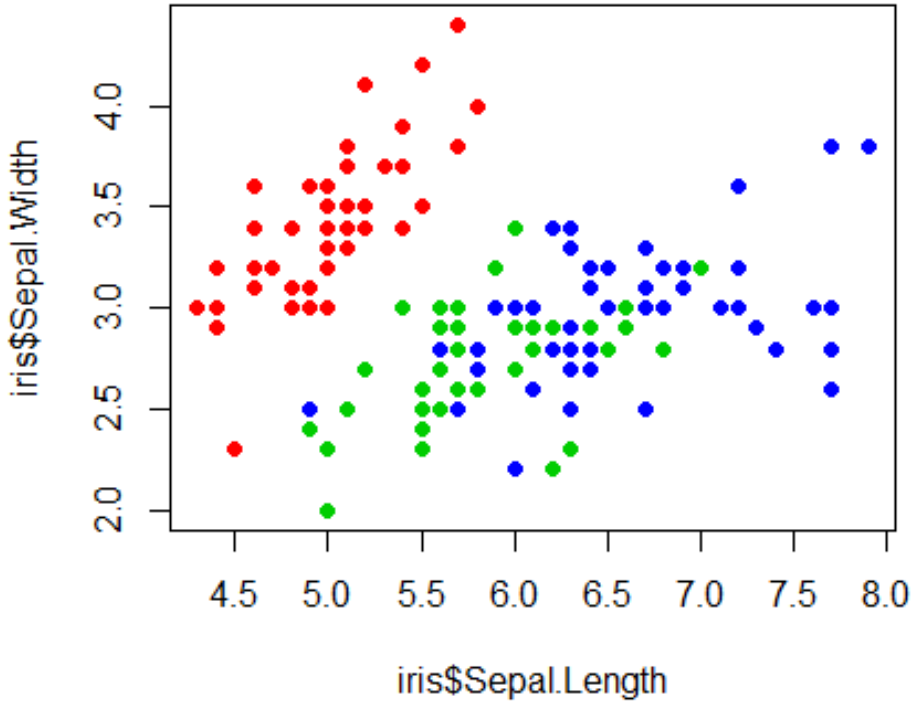
第2主成分までの累積寄与率 0.9581

主成分分析（解析結果）

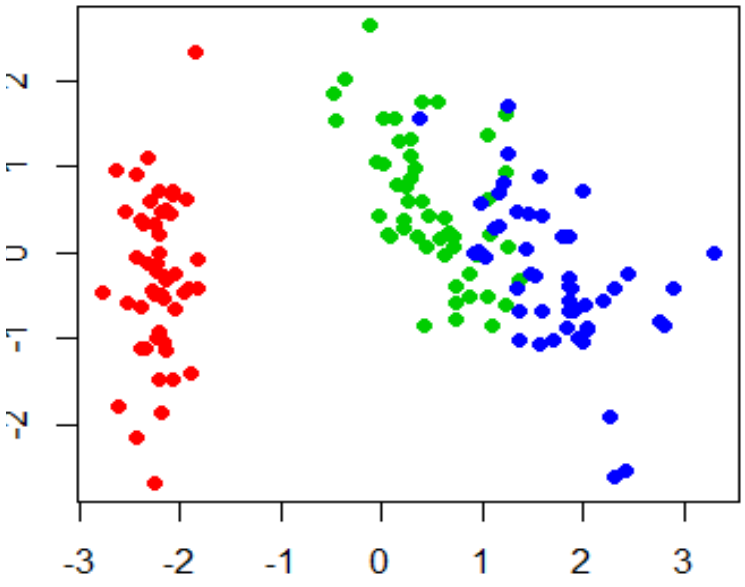
【例】アヤメの花のデータを要約したい（4次元 → 2次元）

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



主成分分析の結果



花弁のデータを無視して、萼片だけ見て2次元に落とすより、

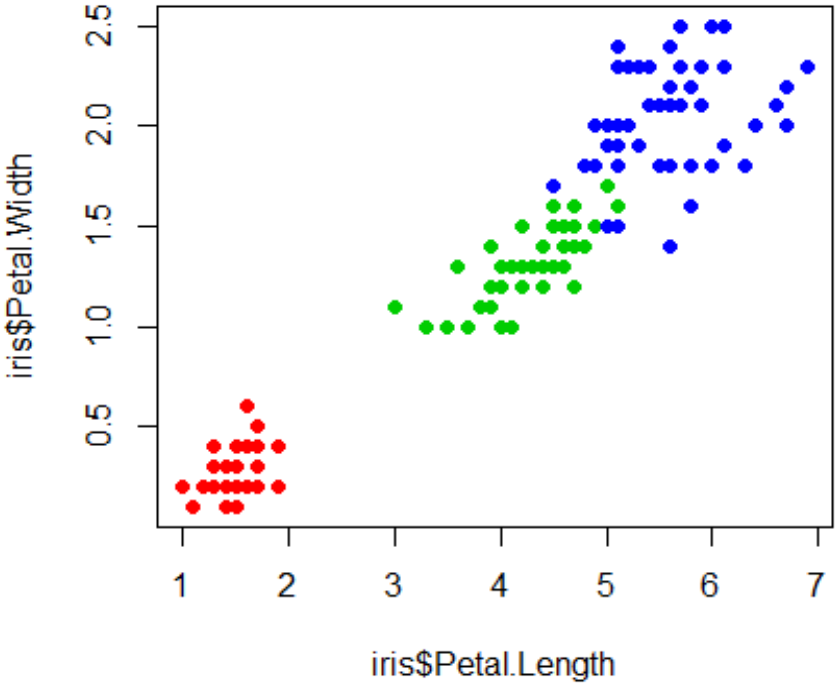
主成分分析で2次元に落とす方が良さそう。

主成分分析（解析結果）

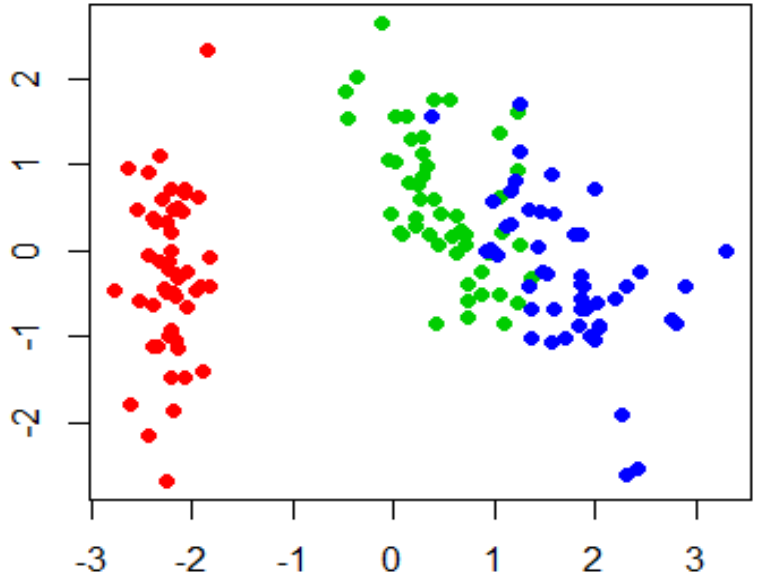
【例】アヤメの花のデータを要約したい（4次元 → 2次元）

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

花弁の長さ・幅



主成分分析の結果



でも実は、品種の判別が目的なら、花弁の長さだけで十分？

高度な解析の前に、まずは生データ（元のデータ）をしっかり観察しよう!!

主成分分析（一般論）

CDSE
Center for Data Science, Ehime University

主成分分析（一般論）

m 次元データ ⇨ ℓ 次元データ $(\ell < m)$

$(x_{i1}, \ldots, x_{im})$ $(a_{i1}, \ldots, a_{i\ell}) = ((\vec{z}_i, \vec{v}_1), \ldots, (\vec{z}_i, \vec{v}_\ell))$

標本平均

$$\overline{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}$$

標本標準偏差

$$s_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \overline{x}_k)^2}$$

規格化

$$z_{ik} = \frac{x_{ik} - \overline{x}_k}{s_k}, \quad \vec{z}_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix}$$

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

$\vec{v}_1, \ldots, \vec{v}_\ell$: R の第1～第 ℓ 固有値に対応する固有ベクトル

第k主成分の寄与率 $\dfrac{\lambda_k}{m}$　第ℓ主成分までの累積寄与率 $\dfrac{\lambda_1 + \cdots + \lambda_\ell}{m}$

（$\lambda_1, \ldots, \lambda_\ell$：$R$ の第1～第 ℓ 固有値）

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.