因子分析

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

回帰分析と主成分分析（再考）

CDSE
Center for Data Science, Ehime University

重回帰分析（再考）

体脂肪率のデータ 出典：The Data And Story LibraryのBodyfat
https://dasl.datadescription.com/datafile/bodyfat/

人物 i	体重 x_{i1}	身長 x_{i2}	体脂肪率 y_i
1	154.25	67.75	12.3
2	173.25	72.25	6.1
⋮	⋮	⋮	⋮
n	207.5	70	31.9

重回帰モデル

人物に依存しない未知の定数（これを推定）

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \quad (1 \leq i \leq n)$$

y_i：人物 i の体脂肪率
x_{i1}：人物 i の体重
x_{i2}：人物 i の身長
e_i：残差

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2$ が最小になるように $\beta_0, \beta_1, \beta_2$ を定める

重回帰分析（再考）

回帰モデル $$y_i = \beta_0 + \sum_{j=1}^{r} \beta_j x_{ij} + e_i \qquad (1 \le i \le n)$$

規格化 $$z_i = \sum_{j=1}^{r} b_j f_{ij} + e_i$$

平均0、分散1

$$\begin{cases} y_i & \to & z_i \\ x_{ij} & \to & f_{ij} \end{cases}$$

目的変数を m 種類に $$z_{ik} = \sum_{j=1}^{r} b_{jk} f_{ij} + e_{ik} \qquad (1 \le k \le m)$$

$$\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} = \begin{bmatrix} f_{11} & \cdots & f_{1r} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nr} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{r1} & \cdots & b_{rm} \end{bmatrix} + \begin{bmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{bmatrix}$$

重回帰分析（再考）

$$\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} = \begin{bmatrix} f_{11} & \cdots & f_{1r} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nr} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{r1} & \cdots & b_{rm} \end{bmatrix} + \begin{bmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{bmatrix}$$

目的変数

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

m種類の目的変数

残差

$$E = \begin{bmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{bmatrix}$$

m種類の目的変数

説明変数

$$F = \begin{bmatrix} f_{11} & \cdots & f_{1r} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nr} \end{bmatrix}$$

n個（人）の個体

r種類の説明変数

係数

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{r1} & \cdots & b_{rm} \end{bmatrix}$$

（個体に依存しない）

回帰式の行列表現

$$Z = FB + E$$

重回帰分析（再考）

$$\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} = \begin{bmatrix} f_{11} & \cdots & f_{1r} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nr} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{r1} & \cdots & b_{rm} \end{bmatrix} + \begin{bmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{bmatrix}$$

$n \times m$ Z = $n \times r$ F $r \times m$ B + $n \times m$ E

目的変数（既知）　説明変数（既知）　係数（推測）　残差（最小化）

↓ 近似

$n \times m$ Z ≈ $n \times r$ F $r \times m$ B

※ $\mathrm{rank}\ FB \leq r$ に注意

主成分分析（再考）

アヤメのデータ

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

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（再考）

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

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

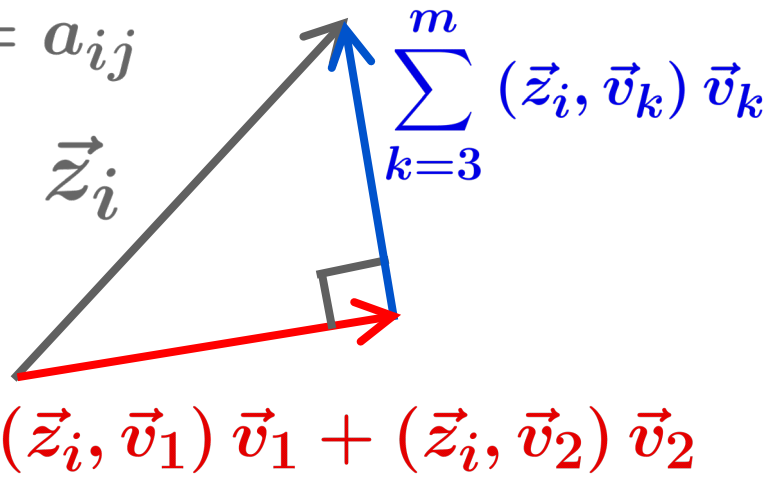
$$\vec{z}_i = \sum_{k=1}^{m} a_{ik}\vec{v}_k \quad (a_{ik} \in \mathbb{R}) \text{ と書ける}$$

$$\Downarrow$$

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

$$\Downarrow$$

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



$$\Downarrow$$

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

主成分分析（再考）

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

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

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

$$= \sum_{k=1}^{r} a_{ik}\vec{v}_k + \sum_{k=r+1}^{m} a_{ik}\vec{v}_k$$

r 次元　　m−r 次元

$$\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

m 次元 → r 次元

$$\begin{bmatrix} {}^t\vec{z}_1 \\ \vdots \\ {}^t\vec{z}_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nr} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_1 \\ \vdots \\ {}^t\vec{v}_r \end{bmatrix} + \begin{bmatrix} a_{1\,r+1} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n\,r+1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_{r+1} \\ \vdots \\ {}^t\vec{v}_m \end{bmatrix}$$

主成分分析（再考）

$$\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nr} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_1 \\ \vdots \\ {}^t\vec{v}_r \end{bmatrix} + \begin{bmatrix} a_{1\,r+1} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n\,r+1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_{r+1} \\ \vdots \\ {}^t\vec{v}_m \end{bmatrix}$$

◆ 用語・記号を回帰分析に合わせる

「目的変数」

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

「残差」

$$E = \begin{bmatrix} a_{1\,r+1} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n\,r+1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_{r+1} \\ \vdots \\ {}^t\vec{v}_m \end{bmatrix}$$

「説明変数」

$$F = \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nr} \end{bmatrix}$$

「係数」（個体に依存しない）

$$B = \begin{bmatrix} {}^t\vec{v}_1 \\ \vdots \\ {}^t\vec{v}_r \end{bmatrix}$$

主成分分析の行列表現

$$Z = FB + E$$

主成分分析（再考）

$$\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nr} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_1 \\ \vdots \\ {}^t\vec{v}_r \end{bmatrix} + \begin{bmatrix} a_{1\,r+1} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n\,r+1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} {}^t\vec{v}_{r+1} \\ \vdots \\ {}^t\vec{v}_m \end{bmatrix}$$

$n \times m$ $\quad$ $n \times r$ $\quad$ $r \times m$ $\quad$ $n \times m$

$Z = F\,B + E$

目的変数（既知） $\quad$ 説明変数（推測） $\quad$ 係数（推測） $\quad$ 残差（最小化）

観測変数 $\quad$ 主成分得点 $\quad$ 主成分

⇩ 近似

$n \times m$ $\quad$ $n \times r$ $\quad$ $r \times m$

$Z \approx F\,B$

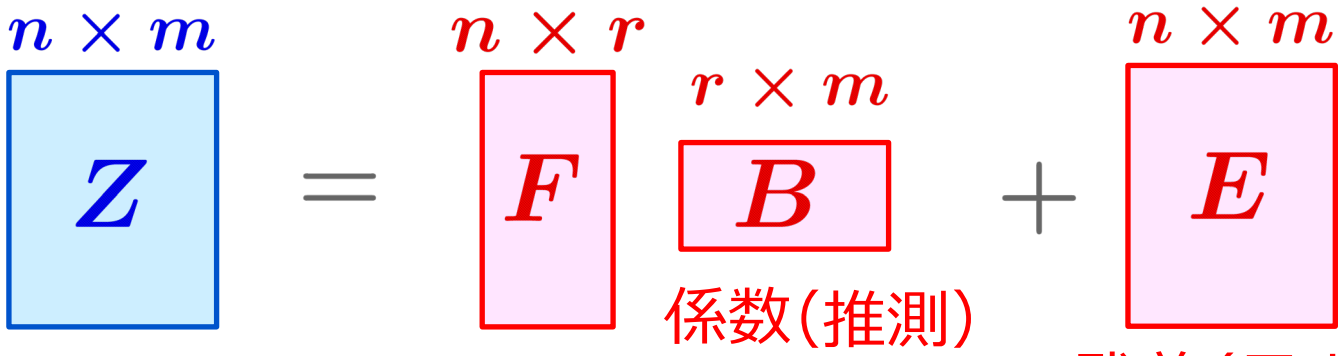
※ $\mathrm{rank}\ BF \leq r$ に注意

因子分析

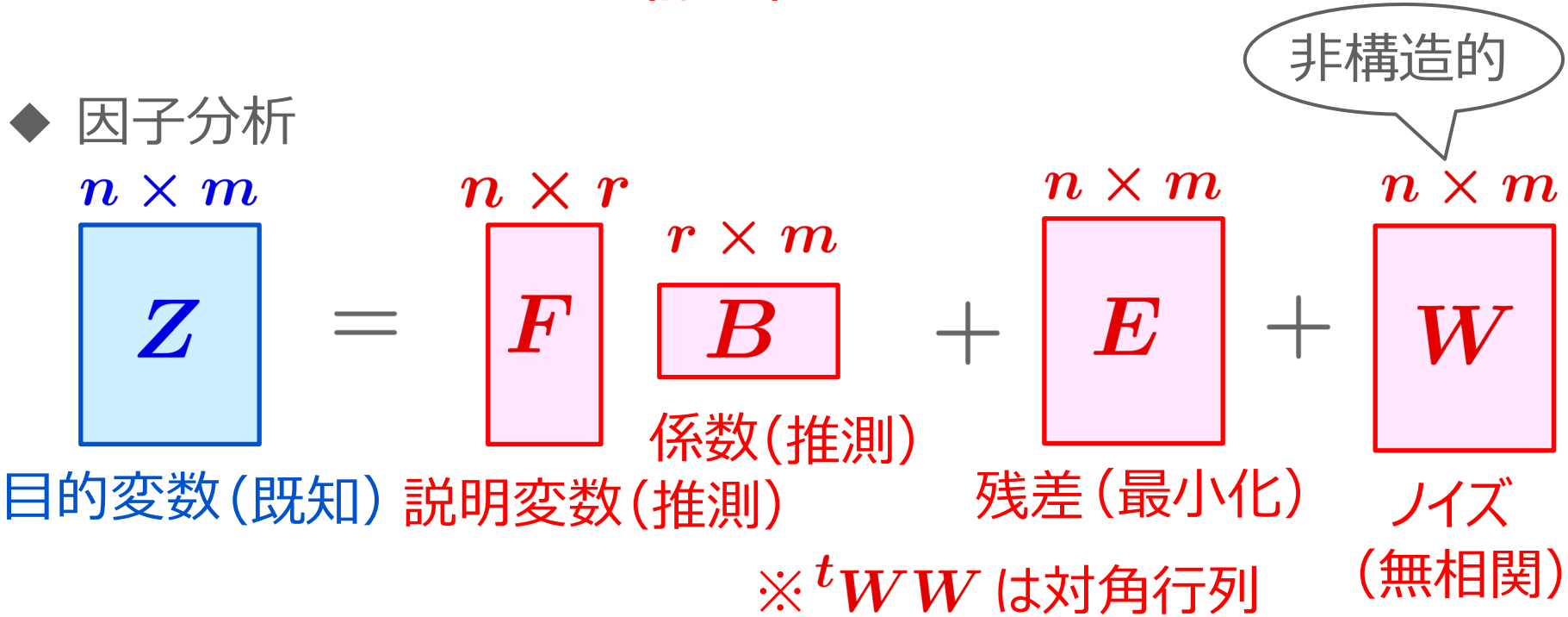
CDSE
Center for Data Science, Ehime University

因子分析（主成分分析との比較）

◆ 主成分分析



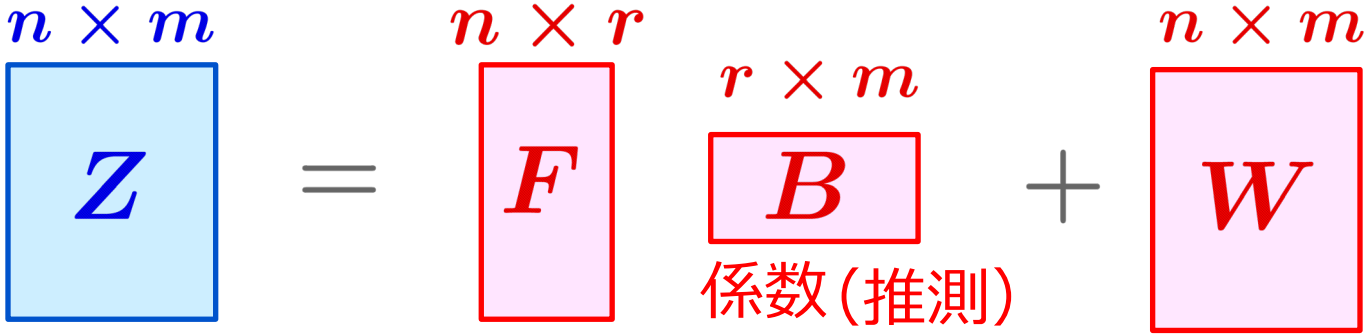
◆ 因子分析



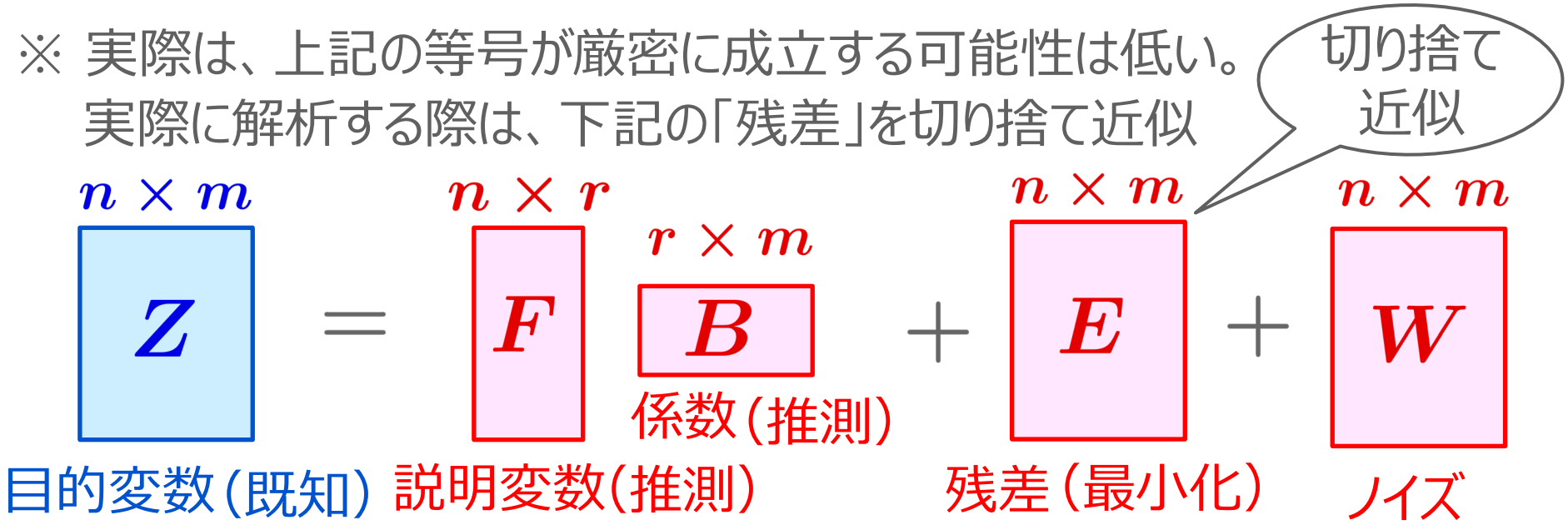
因子分析（モデル）

※ よくある説明：下記の「モデル」から出発

※ ${}^tWW$ は対角行列



※ 実際は、上記の等号が厳密に成立する可能性は低い。
実際に解析する際は、下記の「残差」を切り捨て近似



因子分析（計算手順）

$$Z = F B + E + W$$

観測変数 Z

因子得点 F

因子負荷量 B

残差（最小化） E

独自因子 W

※ ${}^tWW$ は対角行列

◆ 計算手順（イメージ）

(1) 独自因子 W を推測して、観測変数 Z から除去

(2) $Z - W$ に対し、主成分分析と同様の分解を実行

(3) B, F を具体的に求める

※ 独自因子の推測が最大の難関（完璧な方法はない）

因子分析（独自因子の推定）

$$Z = F B + E + W$$

観測変数　因子得点　因子負荷量　残差（最小化）　独自因子

※ ${}^tWW$ は対角行列

◆ 独自因子の推定法の例（イメージ）

(1) 観測変数 Z_i を、残りの観測変数 Z_k $(k \neq i)$ に回帰

(2) 回帰の残差を W_i とし、これを並べた行列を W とする

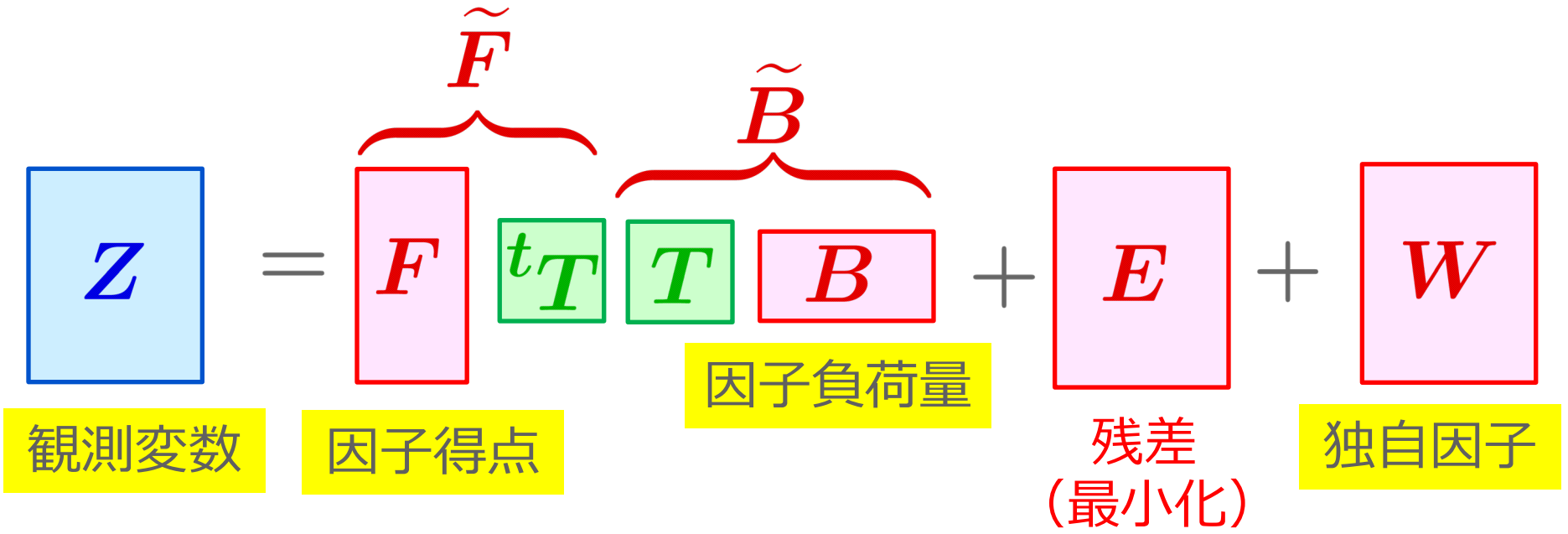
(3) 一般に、${}^tWW$ は対角行列にならないが、非対角成分は無視

$$Z = \begin{matrix} Z_1 & \cdots & Z_m \end{matrix} \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}$$

$$W = [W_1, \ldots, W_m]$$

因子分析（回転不定性）

Tを直交行列とすると、$F^t TTB = FB$ なので、



$$\begin{cases} F & \Rightarrow & \widetilde{F} = F^t T \\ B & \Rightarrow & \widetilde{B} = TB \end{cases}$$

因子を変換（例えば回転）可能

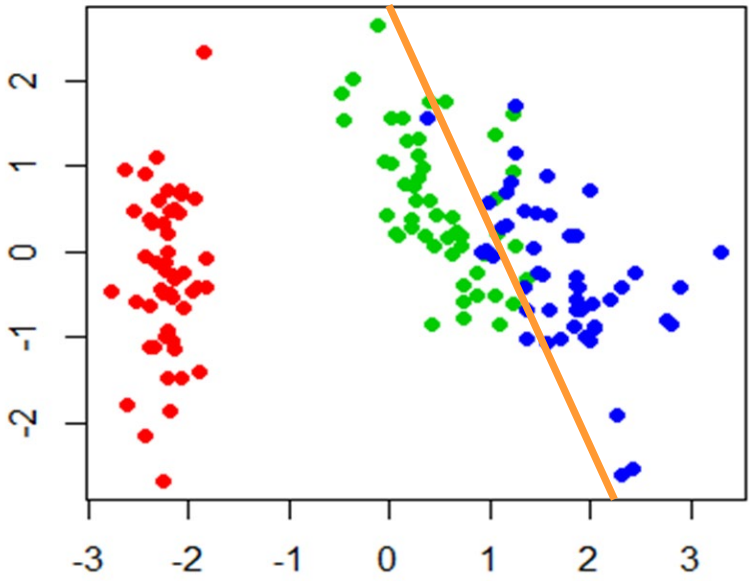
※ T を上手く選ぶことで、因子の解釈をしやすくする

主成分分析の回転不定性

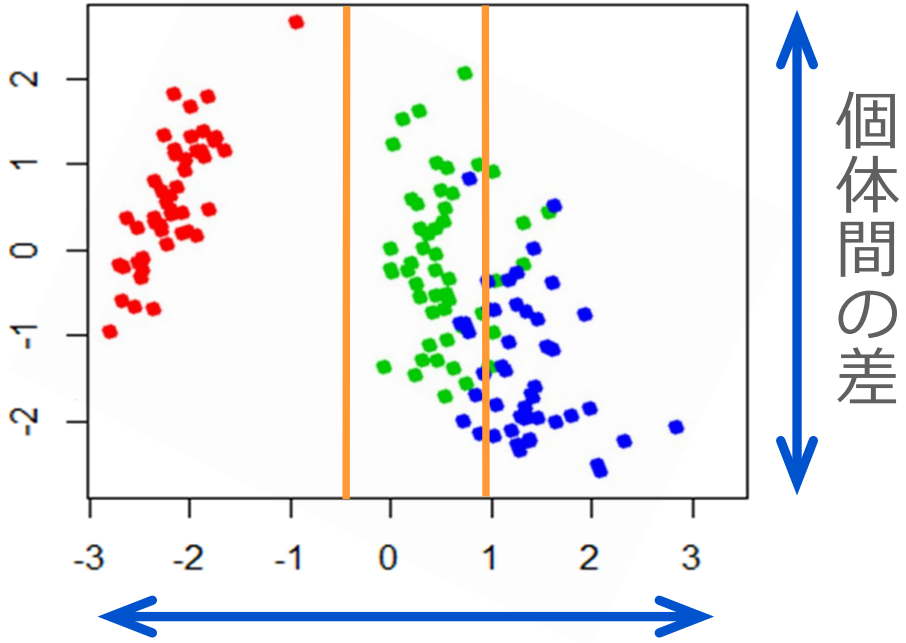
【例】アヤメの花のデータを要約したい（4次元 → 2次元）

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

主成分分析の結果



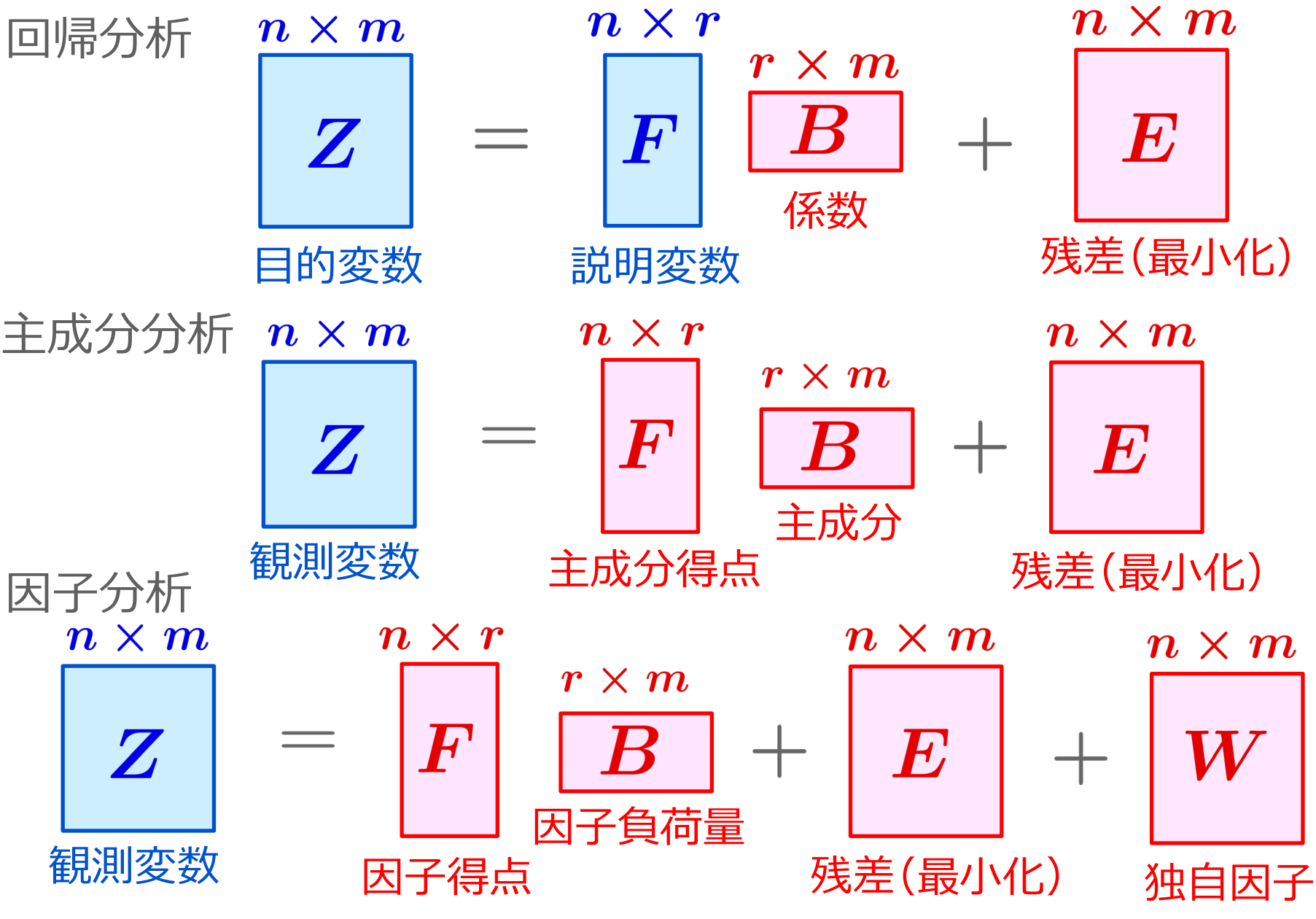
主成分分析の結果（回転後）



主成分分析でも要約を回転可能

「因子分析は主成分分析と違い、回転不定性がある」は誤解の元

回帰・主成分・因子分析（m 次元 → r 次元）



因子分析（解析例）

【例】アヤメの花のデータは1因子で説明可能？（4次元 → 1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.

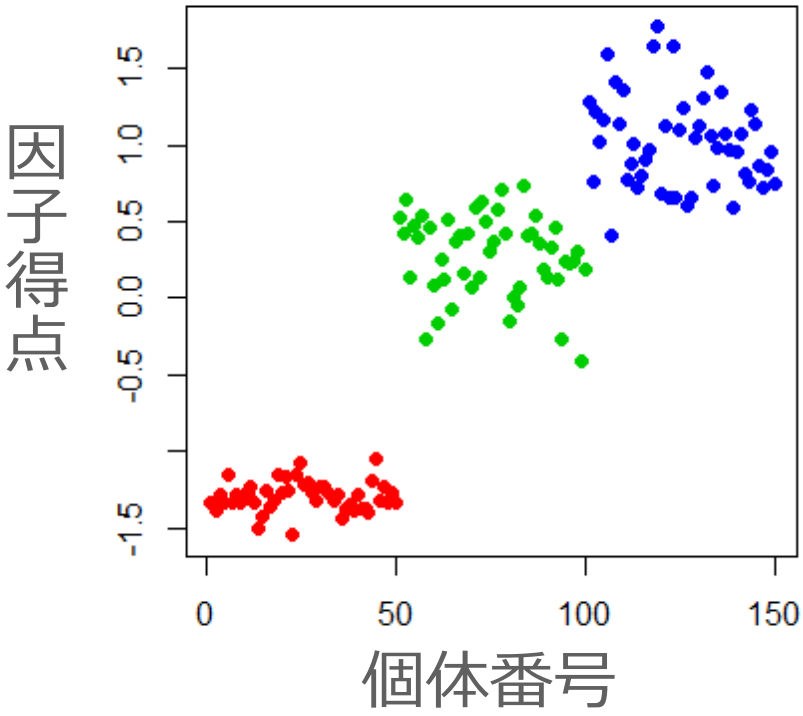
No	萼片長	萼片幅	花弁長	花弁幅	品種
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	virginica

因子分析（解析例）

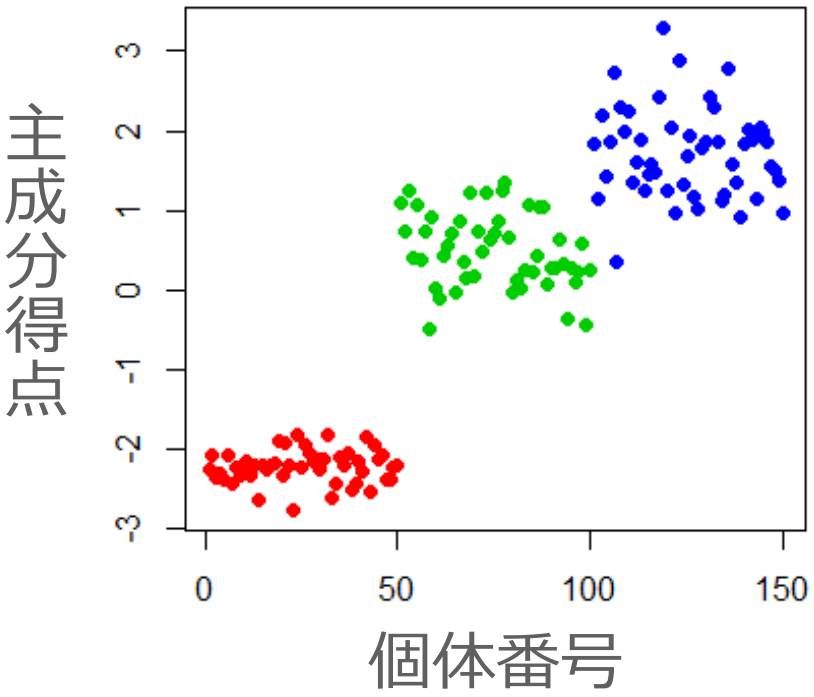
【例】アヤメの花のデータは1因子で説明可能？（4次元 → 1次元）

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

因子分析の結果



主成分分析の結果



因子分析（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.

・体脂肪率のデータ

The Data And Story Library, Bodyfat, https://dasl.datadescription.com/datafile/bodyfat/
（最終閲覧日 2022年9月4日）