多変量解析

第11回　主成分分析

萩原•篠田
情報理工学部

主成分分析

学力の特徴(分布)を少ない変数(主成分)で表現できないか？

第1主成分 $z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$ 総合的学力

第2主成分 $z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$ 文系・理系志向

u_1, u_2, u_3, u_4 は x_1, x_2, x_3, x_4 を標準化した変数

生徒No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

寄与率　第1主成分: 0.680

第2主成分: 0.306

累積: 0.986

↓

第2主成分までで4次元データの98.6%までが表現できる

keywords

説明変数、総合的指標、主成分、主成分得点、寄与率、情報損失量、固有値、固有ベクトル

主成分分析

学力の特徴(分布)を少ない変数(主成分)で表現できないか？

第1主成分 $z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$ 総合的学力

第2主成分 $z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$ 文系・理系志向

u_1, u_2, u_3, u_4 は x_1, x_2, x_3, x_4 を標準化した変数

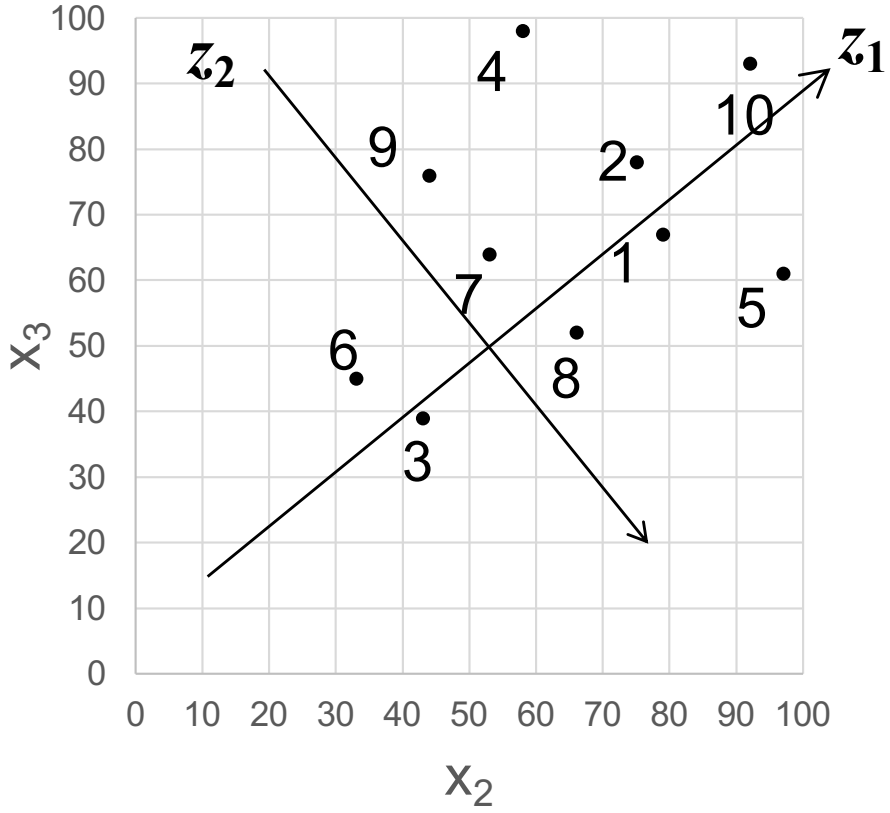
寄与率　第1主成分: 0.680

第2主成分: 0.306

累積: 0.986

↓

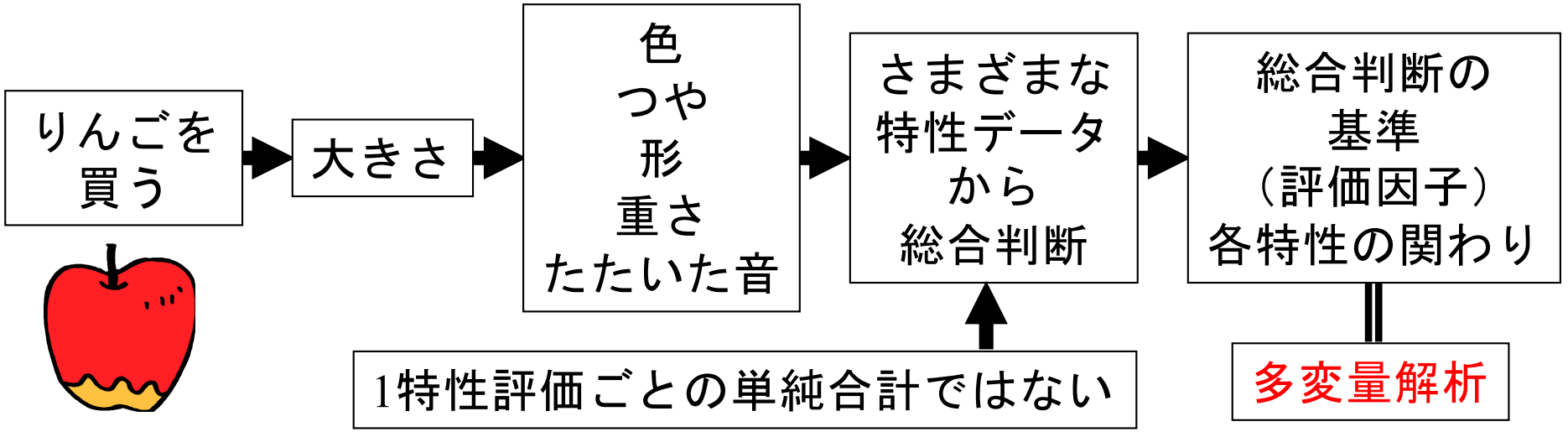
第2主成分までで4次元データの98.6%までが表現できる



keywords

説明変数、総合的指標、主成分、主成分得点、寄与率、情報損失量、固有値、固有ベクトル

多変量情報の解析法



主成分分析

りんご

総合特性

↓

総合的に
取り扱う

大きさ、色、形、つや、・・・　（説明変数）

要因($x_1, x_2, \cdots, x_p$)

$$z = a_1x_1 + a_2x_2 + \cdots + a_px_p$$

・・・主成分

主成分分析

多くの変数($x_1, x_2, \cdots, x_p$)の値をできるだけ情報の損失なしに
1個または互いに独立な少数の総合的指標($z_1, z_2, \cdots, z_m$)で表す

独立とは？

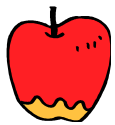
$m \leq p$

$$z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \qquad \text{第1主成分}$$

$$z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \qquad \text{第2主成分}$$

$$\vdots \qquad\qquad \vdots$$

$$z_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mp}x_p \qquad \text{第}m\text{主成分}$$



説明変数 (大きさ x_1, 色 x_2)

総合指標（主成分） (z_1, z_2)

データ

北海道のリンゴ $(x_{1北}, x_{2北})$

青森のリンゴ $(x_{1青}, x_{2青})$

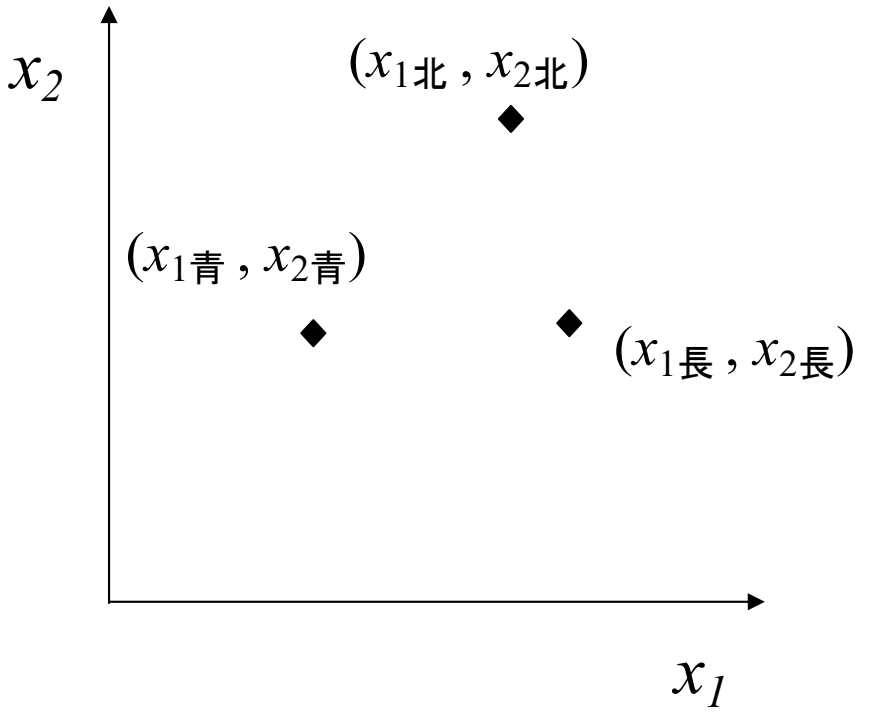
長野のリンゴ $(x_{1長}, x_{2長})$

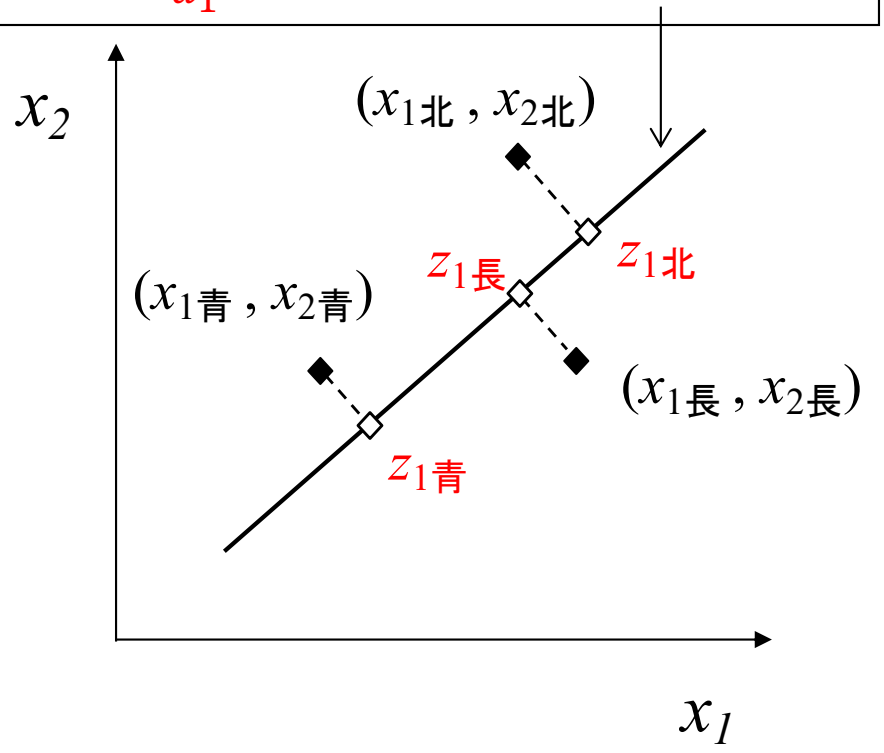
食べたくなるリンゴ z_1

描きたくなるリンゴ z_2

主成分得点 $z_1 = a_1x_1 + a_2x_2$

↳ 傾き$\frac{a_2}{a_1}$の直線上に射影された位置

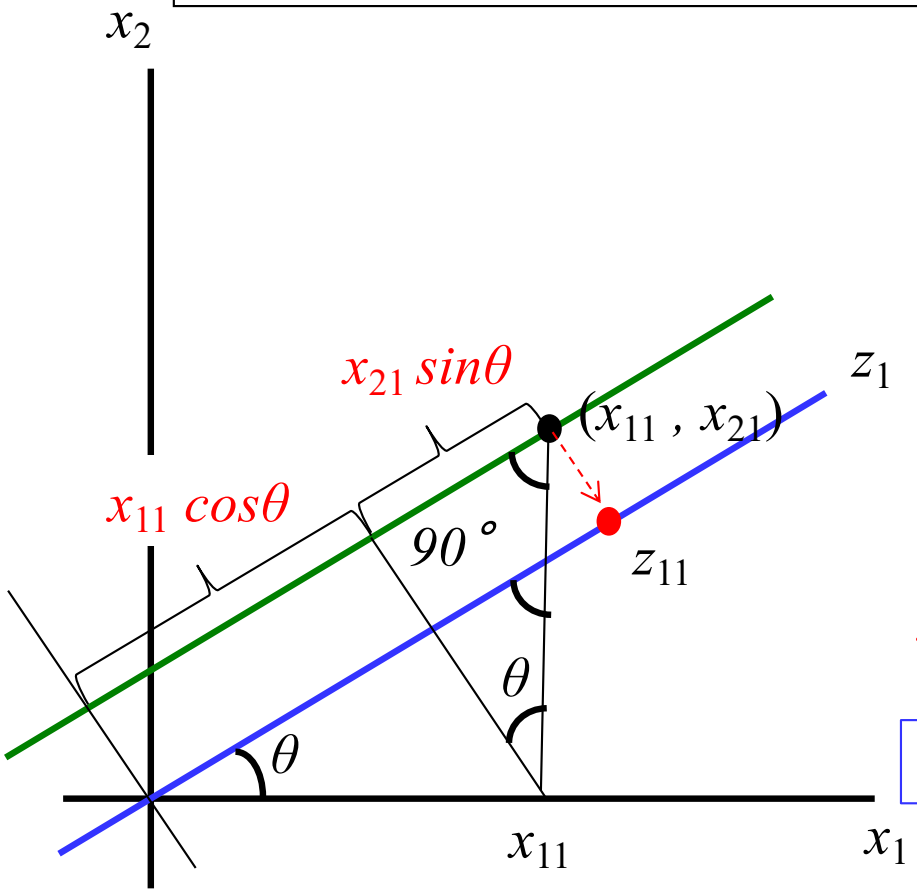




訂正

主成分得点　$z_1 = a_1x_1 + a_2x_2$

↳ 傾き$\frac{a_2}{a_1}$の直線上に射影された位置



直線の傾き（方向比）

$$tan\theta = \frac{sin\theta}{cos\theta}\left(= \frac{a_2}{a_1}\right)$$

点$(x_{11}\,,\,x_{21})$は，z_1軸上では

$$x_{11}cos\theta + x_{21}sin\theta = a_1x_{11} + a_2x_{21} = z_{11}$$

z_1軸を平行移動してもz_1の値は変わらない

主成分の求め方（その1）情報の損失量を最も少なくする

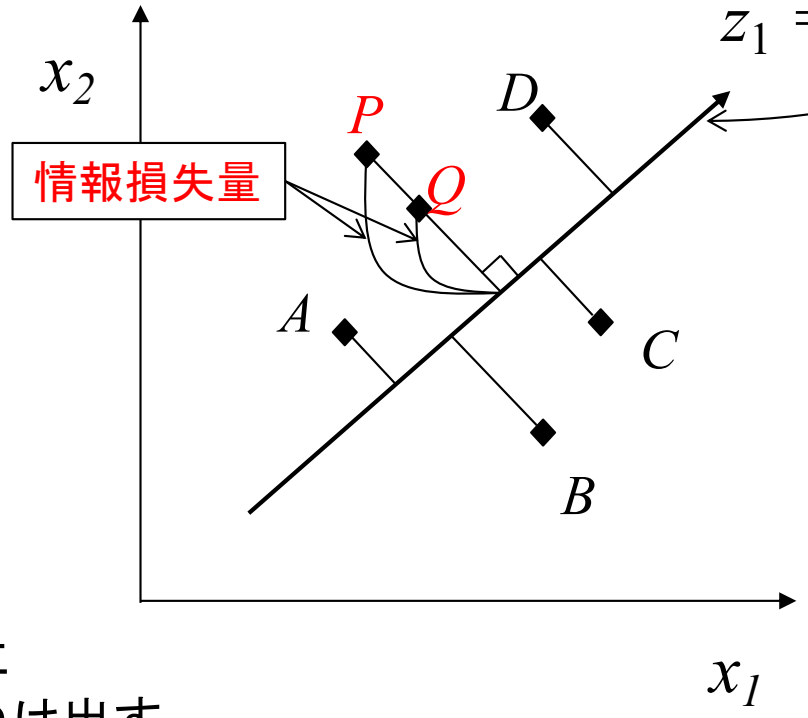
情報の損失とは

データP, Qは主成分z_1上で考えると同じ点に移動

従ってデータから主成分z_1に下ろした垂線の長さは
主成分z_1上では考慮されない

総合的指標

$$z_1 = a_1x_1 + a_2x_2$$



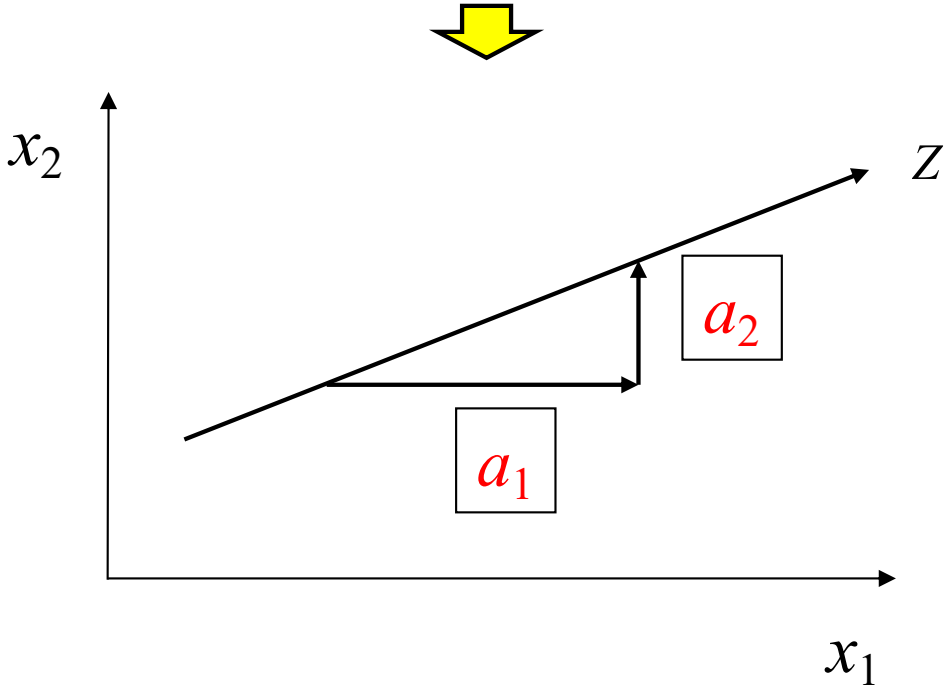
z_1 軸上では
垂線の長さ＝情報の損失

できるだけ情報の損失なしに
最も良い方向比 $a_1 : a_2$ を見つけ出す
＝情報損失量の和が最小になる方向比 $a_1 : a_2$ を見つける
＝垂線の長さの和が最小になる方向比 $a_1 : a_2$ を見つける

主成分の求め方（その1）情報の損失量を最も少なくする

情報の損失量を少なくする係数a_1 , a_2 を決める

直線 z は、方向比が a_1 ： a_2 なので



直線 z の式

$$x_2 = \frac{a_2}{a_1} x_1 + \text{切片}$$

$$a_1 x_2 = a_2 x_1 + a_1 \cdot \text{切片}$$

$a_1 \cdot$ 切片 $\rightarrow a_0$

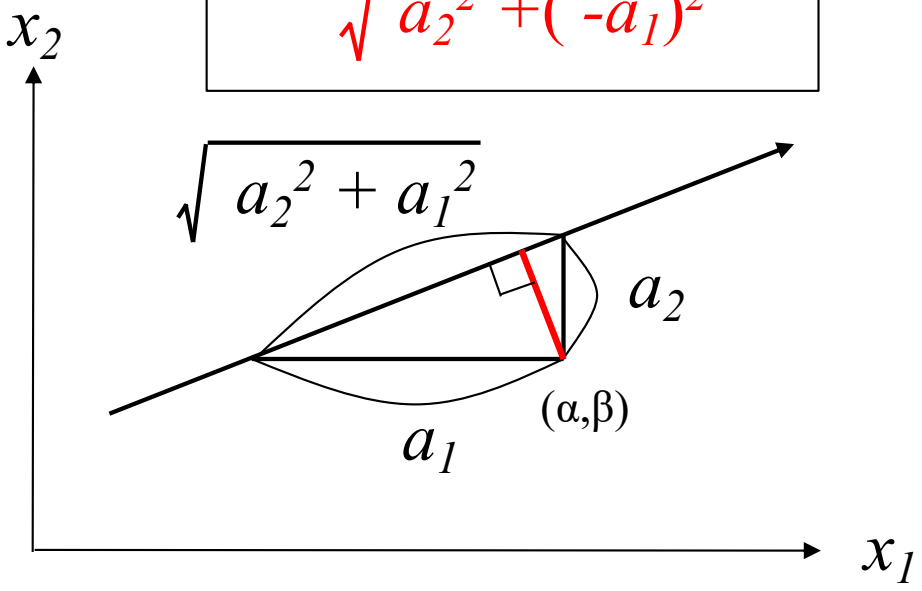
$$a_2 x_1 - a_1 x_2 + a_0 = 0$$

主成分の求め方（その1）情報の損失量を最も少なくする

直線 z は、$a_2x_1 - a_1x_2 + a_0 = 0$

平面上の点(α , β)からこの直線におろした垂線の長さは

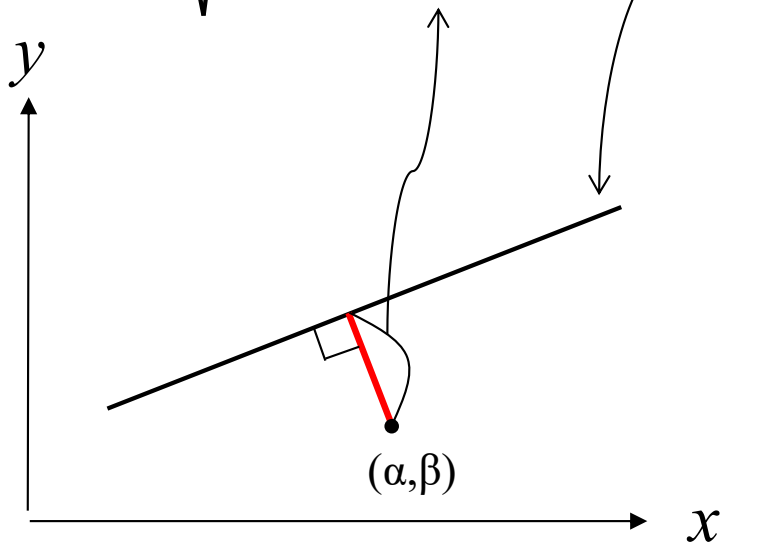
$$\frac{|\, a_2\alpha - a_1\beta + a_0 \,|}{\sqrt{a_2^2 + (-a_1)^2}}$$



＜ヘッセの標準形＞

点(α , β)から直線 $ax + by + c = 0$ におろした垂線の長さは

$$\frac{|\, a\alpha + b\beta + c \,|}{\sqrt{a^2 + b^2}}$$



できるだけ情報の損失なしに
最も良い方向比 $a_1 : a_2$ を見つけ出す
＝垂線の長さの和が最小になる方向比 $a_1 : a_2$ を見つける

（例題） 説明変数x_1：スポーツ施設数
説明変数x_2：教育施設数

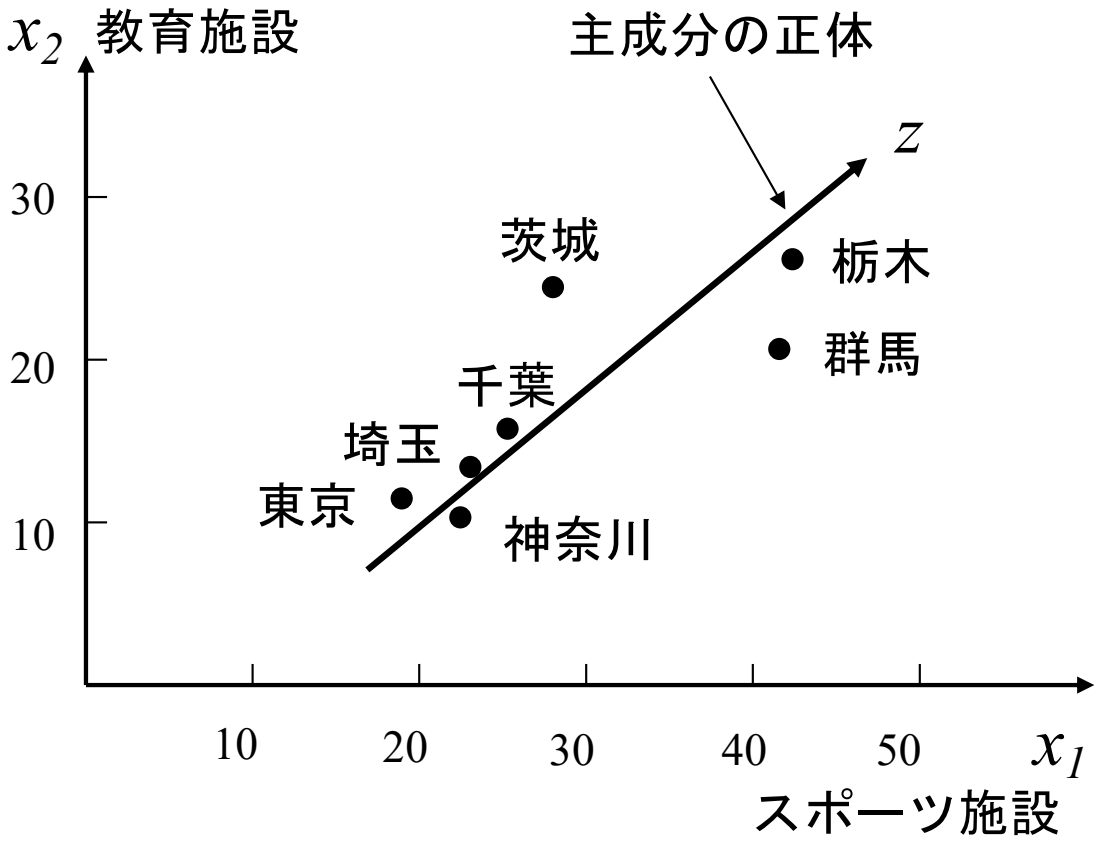
表1

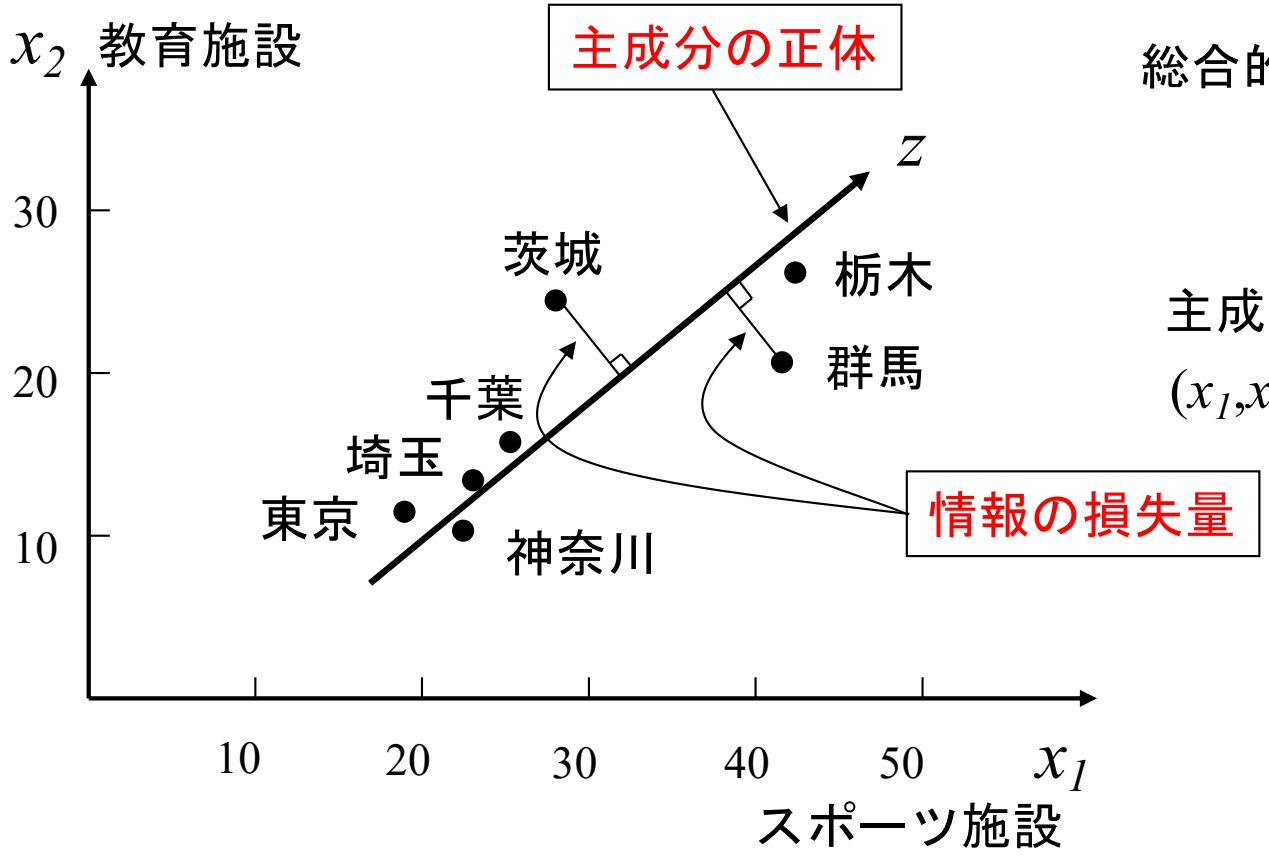
変量 / サンプル	人口10万人当りスポーツ施設数 x_1	人口10万人当り教育施設数 x_2
埼玉	22.9	13.7
千葉	24.9	16.2
東京	19.3	11.3
神奈川	22.0	10.4
茨城	28.6	24.9
栃木	42.6	26.5
群馬	41.3	20.3
平均	28.8	17.614
偏差平方和	533.2	248.5
偏差積和	292.4	
分散	88.87	41.41
共分散	48.73	

主成分を求めるとは：

2つの説明変数の総合特性を求めること

$$z_1 = a_1 x_1 + a_2 x_2$$





総合的特性を表す1次式

$$z = a_1 x_1 + a_2 x_2$$

主成分 = 総合特性 z

(x_1, x_2)の総合特性値が求まる

情報の損失量：各点から z におろした垂線の長さ

情報の損失量を少なくする係数a_1 , a_2 を決める

平面上の点(α , β)から直線 z におろした垂線の長さ：$\dfrac{|a_2\alpha - a_1\beta + a_0|}{\sqrt{a_2^2 + (-a_1)^2}}$

知りたいのは直線 z の方向、つまり a_1 と a_2 の比

従って　$a_2^2 + a_1^2 = 1$　という条件を付けると

	スポーツ施設 x_1	教育施設 x_2	情報損失量
埼玉	22.9	13.7	$\lvert 22.9a_2 - 13.7a_1 + a_0 \rvert$
千葉	24.9	16.2	$\lvert 24.9a_2 - 16.2a_1 + a_0 \rvert$
東京	19.3	11.3	$\lvert 19.3a_2 - 11.3a_1 + a_0 \rvert$
神奈川	22.0	10.4	$\lvert 22.0a_2 - 10.4a_1 + a_0 \rvert$
茨城	28.6	24.9	$\lvert 28.6a_2 - 24.9a_1 + a_0 \rvert$
栃木	42.6	26.5	$\lvert 42.6a_2 - 26.5a_1 + a_0 \rvert$
群馬	41.3	20.3	$\lvert 41.3a_2 - 20.3a_1 + a_0 \rvert$

7県の情報損失量の平方和

$$U(a_2, a_1, a_0) = \sum_{i=1}^{7} (a_2 x_{1i} - a_1 x_{2i} + a_0)^2$$

$$= (22.9a_2 - 13.7a_1 + a_0)^2 + (24.9a_2 - 16.2a_1 + a_0)^2 \vdots + (41.3a_2 - 20.3a_1 + a_0)^2$$

$$= 6339a_2^2 + 2420a_1^2 - 7687a_1a_2 + 403.2a_2a_0 - 246.6a_1a_0 + 7a_0^2$$

$a_2^2 + a_1^2 = 1$　という条件のもと　$U(a_2 , a_1 , a_0)$の最小値を与える a_1 , a_2を求める

ラグランジュの乗数法

関数$U(a_2, a_1, a_0)$が条件$a_2^2 + a_1^2 = 1$のもとに、点(α,β,γ)で極値を取るならば
関数$F(a_2, a_1, a_0, \lambda)$を

$$F(a_2, a_1, a_0, \lambda) = U(a_2, a_1, a_0) - \lambda(a_2^2 + a_1^2 - 1)$$

とおいたとき、極値をとる点(α,β,γ)は連立方程式

$$\frac{\partial F}{\partial a_2} = 0, \quad \frac{\partial F}{\partial a_1} = 0, \quad \frac{\partial F}{\partial a_0} = 0, \quad a_2^2 + a_1^2 - 1 = 0$$

の解となる

$$F(a_2, a_1, a_0, \lambda) = U(a_2, a_1, a_0) - \lambda(a_2^2 + a_1^2 - 1)$$

$$= 6339a_2^2+2420a_1^2-7687a_1a_2+403.2a_2a_0-246.6a_1a_0+7a_0^2-\lambda(a_2^2 + a_1^2 - 1)$$

$$\frac{\partial F}{\partial a_2} =12678a_2-7687a_1+403.2a_0-2\lambda a_2= 0 \quad \text{-------} \quad ①$$

$$\frac{\partial F}{\partial a_1} =4840a_1-7687a_1-246.6a_0-2\lambda a_1= 0 \quad \text{-------} \quad ②$$

$$\frac{\partial F}{\partial a_0} =403.2a_2-246.6a_1+14a_0= 0 \quad \text{-------} \quad ③$$

③より

$$a_0 = -28.8a_2 + 17.614a_1 \quad \text{-------} \quad ④$$

④を①、②に代入すると

$$\begin{cases} (533.2-\lambda)a_2 - 292.4a_1 = 0 & \text{-------} \quad ⑤ \\ -292.4a_2 + (248.5-\lambda)a_1 = 0 & \text{-------} \quad ⑥ \end{cases}$$

$$\begin{bmatrix} 533.2-\lambda & -292.4 \\ -292.4 & 248.5-\lambda \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad ⑦$$

固有値固有ベクトル問題を解けば良い

(a_2 , a_1)が$(0 , 0)$以外の解をもつためには

$$\begin{vmatrix} 533.2-\lambda & -292.4 \\ -292.4 & 248.5-\lambda \end{vmatrix} = 0$$

従って、λの2次方程式

$$(533.2 - \lambda)(248.5 - \lambda) - 292.4^2 = 0$$

を解いて

$$\lambda_1 = 65.65 \qquad \lambda_2 = 716.1 \quad \text{-------} \quad ⑧$$

$\lambda_1 = 65.65$を⑤に代入

最小の固有値
→第1主成分

$$(533.2 - 65.65)a_2 - 292.4a_1 = 0$$

$$467.59a_2 - 292.4a_1 = 0$$

（$\lambda_1 = 65.65$を⑥に代入しても良い

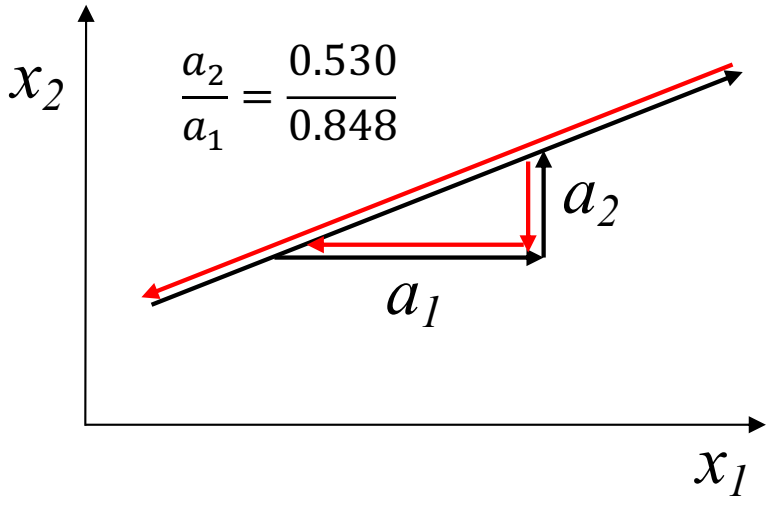
$$-292.4a_2 + (248.5-65.65)a_1 = 0$$

$$-292.4a_2 + 182.84a_1 = 0$$ ）

$a_2^2 + a_1^2 = 1$　という条件で解くと

$$\begin{cases} (a_1 , a_2) = (0.848 , 0.530) & \text{-------} \quad ⑨ \\ (a_1 , a_2) = (-0.848 , -0.530) & \text{-------} \quad ⑩ \end{cases}$$

の2つの解を得る。直線としては同一。



$\lambda_2 = 716.1$を⑤に代入した場合は

$$(533.2 - 716.1)a_2 - 292.4a_1 = 0$$
$$-182.8a_2 - 292.4a_1 = 0$$

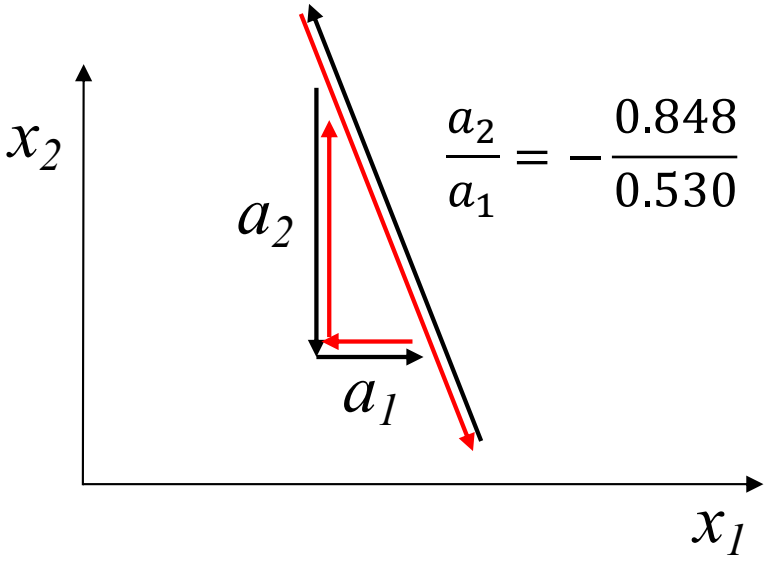
（$\lambda_2 = 716.1$を⑥に代入しても良い

$$-292.4a_2 + (248.5 - 716.1)a_1 = 0$$
$$-292.4a_2 - 467.6a_1 = 0$$ ）

$a_2^2 + a_1^2 = 1$　という条件で解くと

$$\begin{cases} (a_1, a_2) = (0.530, -0.848) & \text{------- ⑪} \\ (a_1, a_2) = (-0.530, 0.848) & \text{------- ⑫} \end{cases}$$

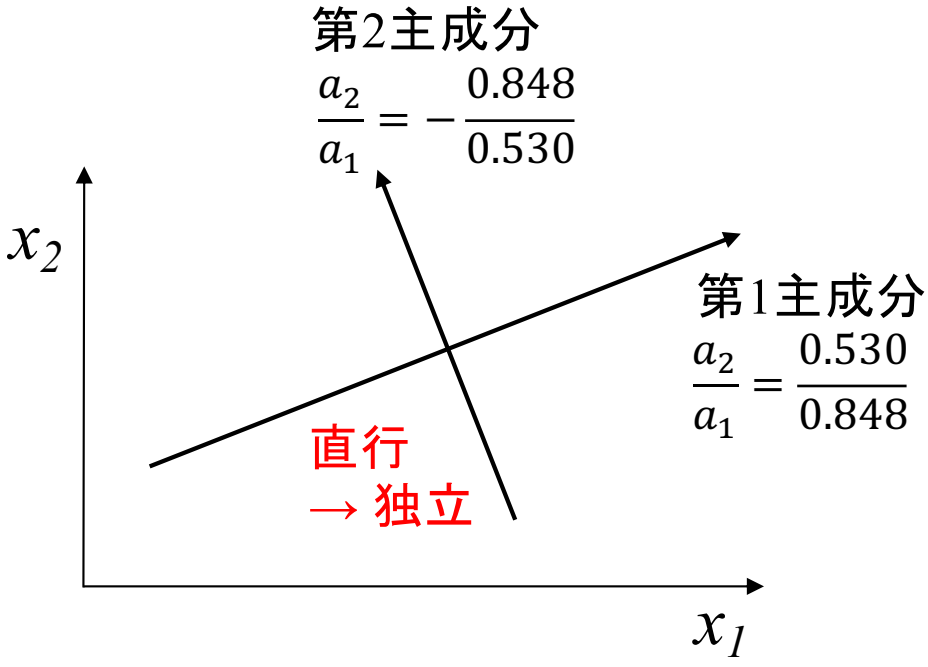
の2つの解を得る。直線としては同一。



$$\begin{bmatrix} 533.2-\lambda & -292.4 \\ -292.4 & 248.5-\lambda \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad ⑦$$

固有値固有ベクトル問題を解いて得た2つの固有値$\lambda_1 = 65.65$と$\lambda_2 = 716.1$に対する固有ベクトルは，第1主成分と第2主成分に相当する．

最小の固有値
→第1主成分



主成分得点（1）

主成分 z を表す直線 z は

$a_2x_1 - a_1x_2 + a_0 = 0$

④より

$a_0 = -28.8a_2 + 17.614a_1$

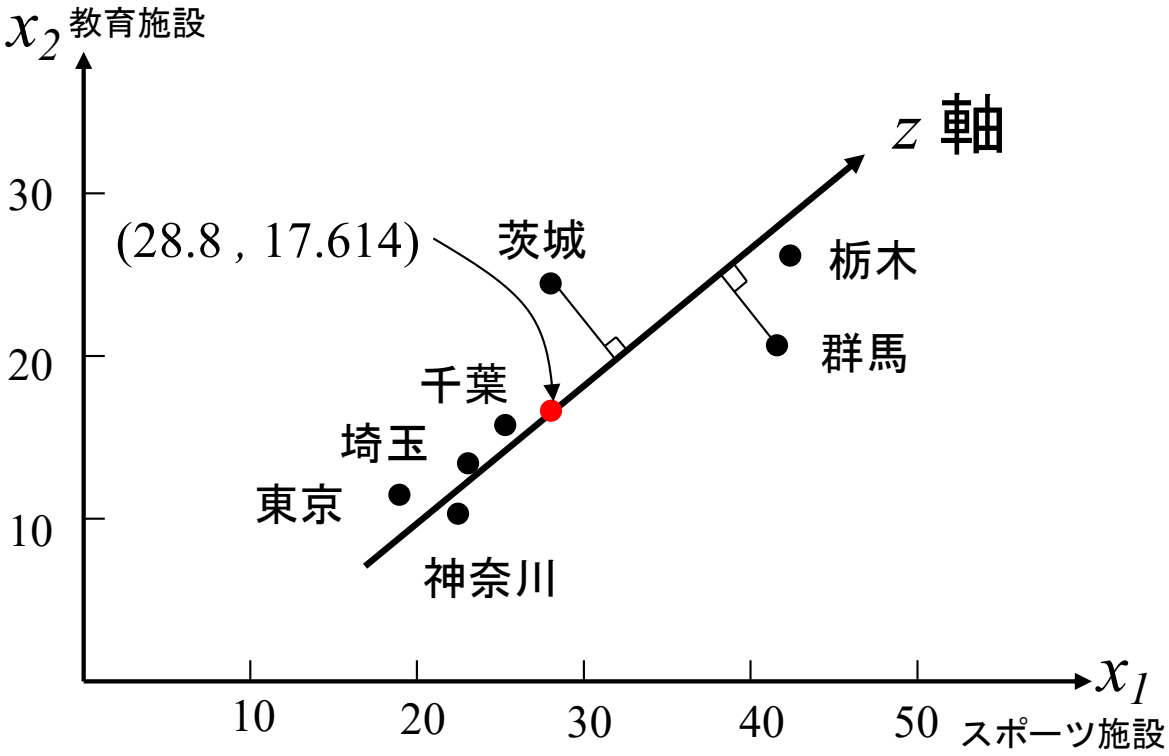
従って

$a_2x_1 - a_1x_2 - 28.8a_2 + 17.614a_1 = 0$

$a_2(x_1 - 28.8) - a_1(x_2 - 17.614) = 0$

説明変数x_1, x_2の平均値は直線z上にある

$(\overline{x_1}, \overline{x_2}) = (28.8, 17.614)$



	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2	$a_1x_1 + a_2x_2$	平均値で z = 0 となる主成分得点
埼玉	22.9	13.7	26.68	-7.08
千葉	24.9	16.2	29.70	-4.06
東京	19.3	11.3	22.36	-11.40
神奈川	22.0	10.4	24.17	-9.59
茨城	28.6	24.9	37.45	3.69
栃木	42.6	26.5	50.17	16.41
群馬	41.3	20.3	45.78	12.02
平均	28.8	17.614	**33.76**	**0**

主成分得点（2）

主成分の解釈

主成分の意味するもの ←－－－ x_1 , x_2の係数から判断

例題の場合

・x_1 , x_2の係数は0.848 , 0.530と共に正の値で1に近い
・スポーツ施設、教育施設が多いほど z の値大（栃木、群馬）
・スポーツ施設、教育施設が少ないほど z の値小（東京、神奈川）

主成分 z は、各県における施設の充実度を意味している

命名（任意）

主成分得点（3）

主成分得点を各点から z 軸におろした垂線との交点の z 軸上での値と定義する

$$z(x_1, x_2) = 0.848x_1 + 0.530x_2$$

さらに平均値をz軸の原点とすると

$$z(x_1, x_2) = 0.848x_1 + 0.530x_2 - 33.76$$

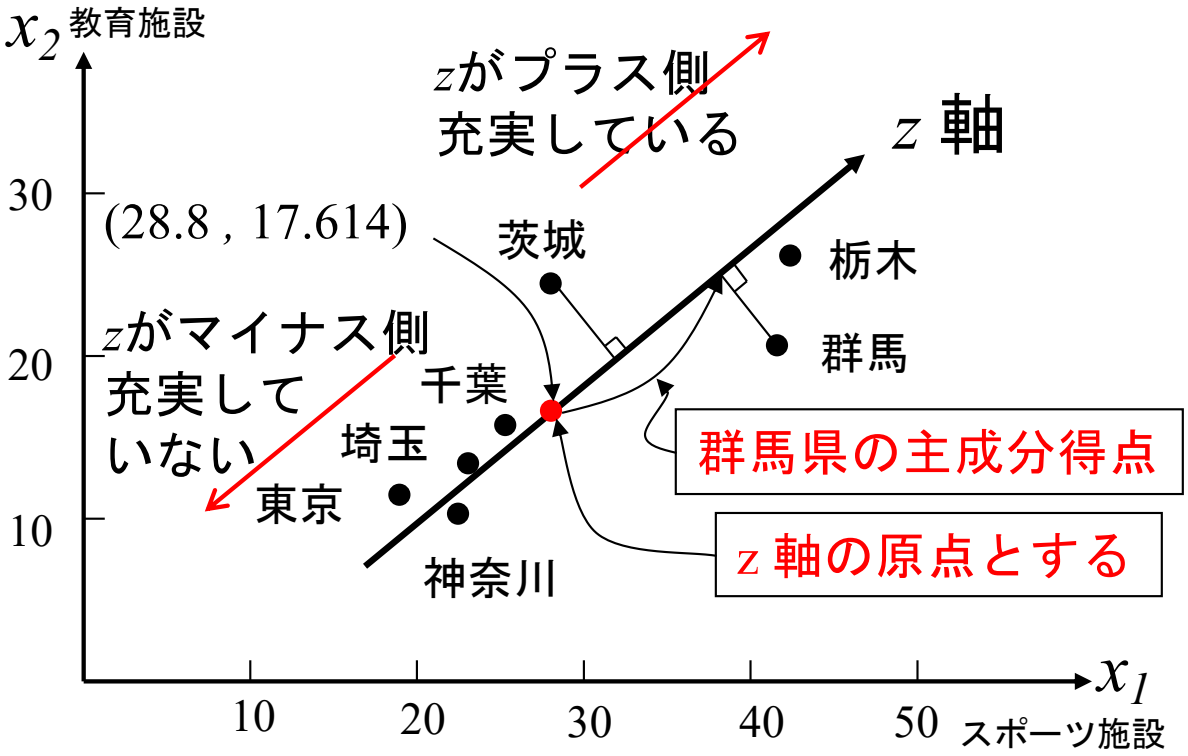
この式で主成分得点が得られる

例

山梨の
スポーツ施設　x_1　30.0
教育施設　x_2　20.0

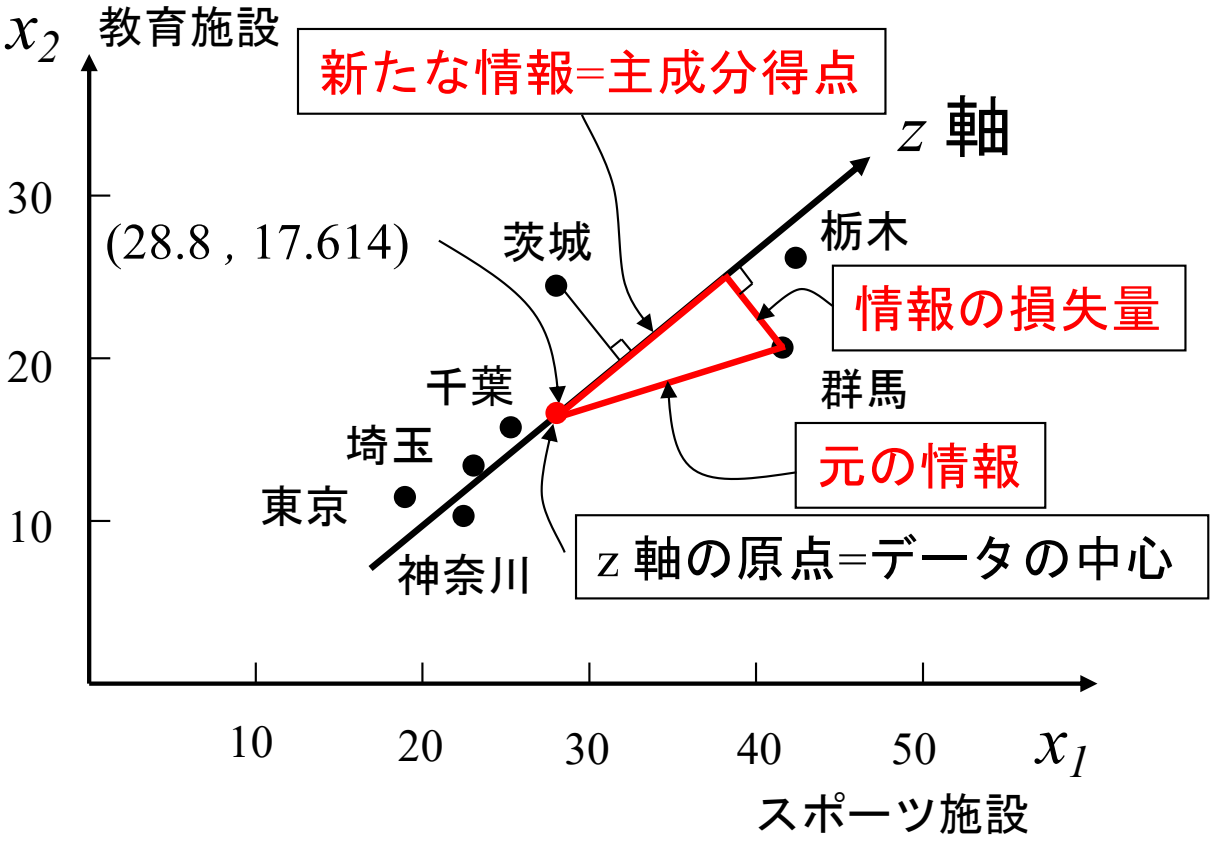
$z(30.0, 20.0) = 2.28$

山梨の施設充実度 = 2.28



	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2	$a_1x_1 + a_2x_2$	平均値で z = 0 となる主成分得点
埼玉	22.9	13.7	26.68	-7.08
千葉	24.9	16.2	29.70	-4.06
東京	19.3	11.3	22.36	-11.40
神奈川	22.0	10.4	24.17	-9.59
茨城	28.6	24.9	37.45	3.69
栃木	42.6	26.5	50.17	16.41
群馬	41.3	20.3	45.78	12.02
平均	28.8	17.614	**33.76**	**0**

寄与率



$(\text{元の情報})^2 = (\text{新たな情報})^2 + (\text{情報の損失量})^2$ 変動、偏差平方和

$$\text{主成分の寄与率} = \frac{\sum(\text{新たな情報})^2}{\sum(\text{元の情報})^2} = \frac{\text{主成分の変動}}{\text{元情報の変動}} = \frac{\text{主成分の分散}}{\text{元情報の分散}}$$

$$\text{例題の主成分の寄与率} = \frac{Var(z_1)}{Var(x_1) + Var(x_2)} = 0.916$$

第1主成分と第2主成分の関係

$\lambda = 65.65$についてa_1, a_2は
$(a_1, a_2) = (0.848, 0.530)$
$z_1 = 0.848x_1 + 0.530x_2$　・・・第1主成分

⑧のもう一つの$\lambda = 716.1$についてa_1, a_2を求めると　$(a_1, a_2) = (-0.530, 0.848)$
$z_2 = -0.530x_1 + 0.848x_2$　・・・第2主成分

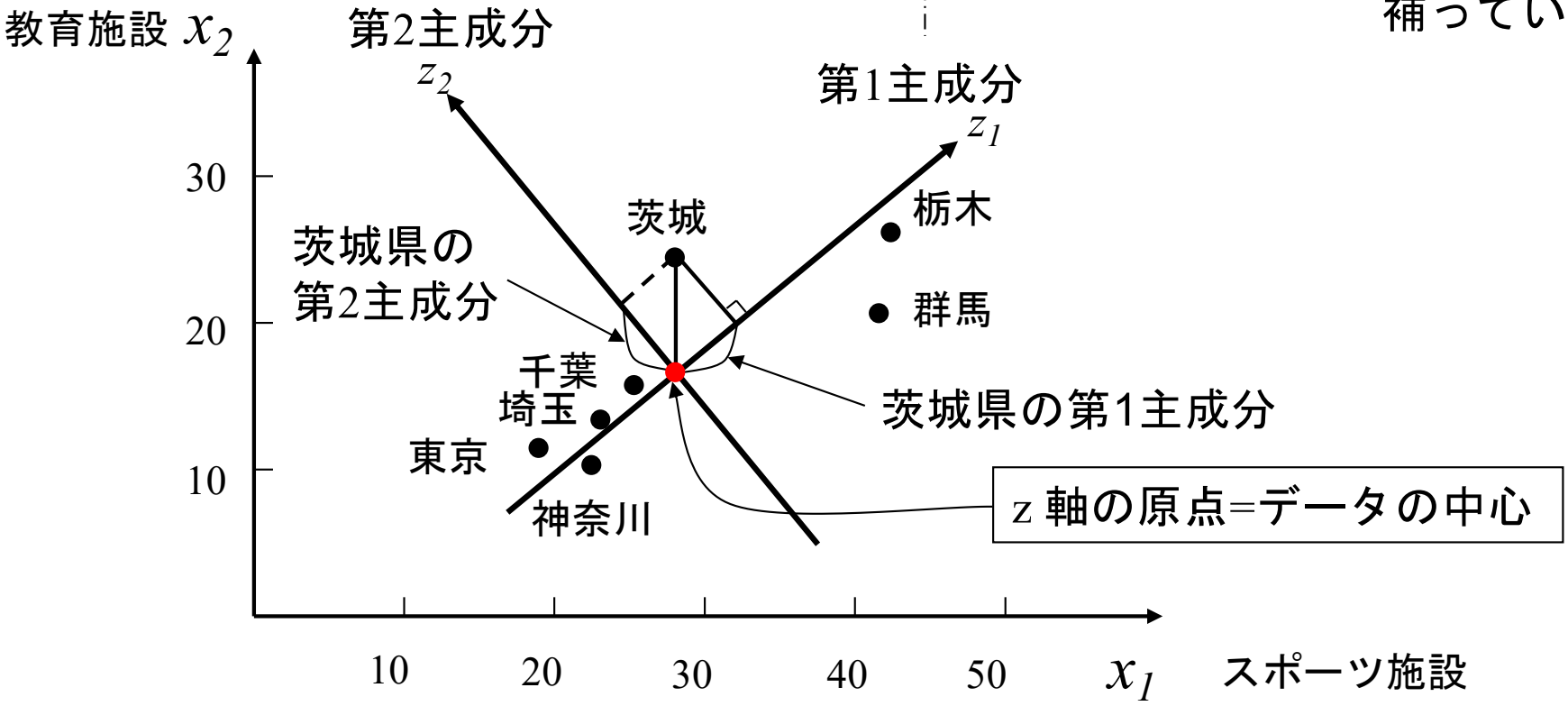
z_1の寄与率 = 0.916

z_2の寄与率 = 0.084

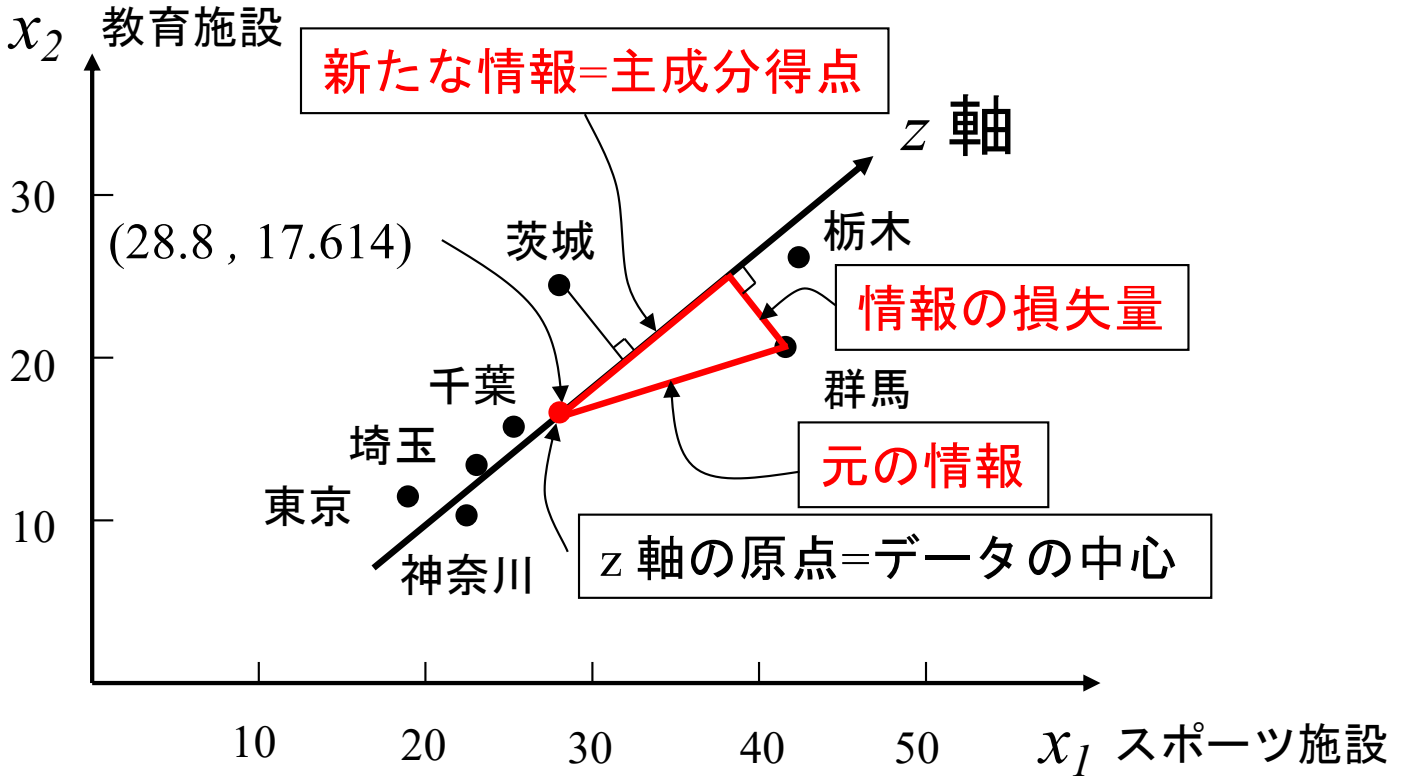
累積寄与率 = 0.916 + 0.084 = 1

第1主成分と第2主成分は直交（互いに独立）

第1主成分の損失を第2主成分が補っている



主成分の求め方（その2）　分散の最大化



$$(\text{元の情報})^2 = (\text{新たな情報})^2 + (\text{情報の損失量})^2$$

変動、偏差平方和

$$\frac{\Sigma(\text{元の情報})^2}{n-1} = \frac{\Sigma(\text{新たな情報})^2}{n-1} + \frac{\Sigma(\text{情報の損失量})^2}{n-1}$$

分散

$$\frac{\Sigma(\text{元の情報})^2}{n-1} = \text{主成分得点 } z \text{ の分散} + \frac{\Sigma(\text{情報の損失量})^2}{n-1}$$

情報の損失量を最小にする　➡　主成分 z の分散を最大にする

分散共分散行列の固有値、固有ベクトルの問題を解くことになる

主成分分析のプログラムはこの方法で主成分を求めている

主成分 $z = a_1x_1 + a_2x_2$ の求め方（分散$Var(z)$の最大化）

$$Var(z) = Var(a_1x_1 + a_2x_2) = a_1{}^2Var(x_1) + a_2{}^2Var(x_2) + 2a_1a_2Cov(x_1, x_2)$$

ラグランジュの乗数法

$$G(a_1, a_2, \lambda') = a_1{}^2Var(x_1) + a_2{}^2Var(x_2) + 2a_1a_2Cov(x_1, x_2) - \lambda'(a_1^2 + a_2^2 - 1)$$

サンプル＼変量	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2
埼玉	22.9	13.7
千葉	24.9	16.2
東京	19.3	11.3
神奈川	22.0	10.4
茨城	28.6	24.9
栃木	42.6	26.5
群馬	41.3	20.3
平均	28.8	17.614
分散	88.87	41.41
共分散	48.73	

$$\begin{cases} \dfrac{\partial G}{\partial a_1} = 2(a_1Var(x_1) + a_2Cov(x_1, x_2) - \lambda' a_1) = 0 \\ \dfrac{\partial G}{\partial a_2} = 2(a_2Var(x_2) + a_1Cov(x_1, x_2) - \lambda' a_2) = 0 \end{cases}$$

分散共分散行列の固有値固有ベクトル問題

$$\begin{bmatrix} Var(x_1) & Cov(x_1, x_2) \\ Cov(x_1, x_2) & Var(x_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda' \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\begin{bmatrix} 88.87 & 48.73 \\ 48.73 & 41.41 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda' \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

固有値 $\lambda' = 119.3, 10.94$

固有ベクトル $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.848 \\ 0.530 \end{bmatrix}, \begin{bmatrix} -0.530 \\ 0.848 \end{bmatrix}$

分散 $Var(z_1) = \lambda'_1, Var(z_2) = \lambda'_2$

寄与率 $$\frac{Var(z_1)}{Var(x_1) + Var(x_2)} = \frac{\lambda'_1}{Var(x_1) + Var(x_2)} = \frac{119.3}{88.87 + 41.41} = 0.916$$

主成分の求め方

表1

サンプル＼変量	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2
埼玉	22.9	13.7
千葉	24.9	16.2
東京	19.3	11.3
神奈川	22.0	10.4
茨城	28.6	24.9
栃木	42.6	26.5
群馬	41.3	20.3
平均	28.8	17.614
偏差平方和	533.2	248.5
偏差積和	292.4	
分散	88.87	41.41
共分散	48.73	

情報損失量Uの最小化

偏差平方和偏差積和行列の固有値固有ベクトル問題

$$\begin{bmatrix} (n-1)Var(x_1) & -(n-1)Cov(x_1,x_2) \\ -(n-1)Cov(x_1,x_2) & (n-1)Var(x_2) \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \end{bmatrix} = \lambda \begin{bmatrix} a_2 \\ a_1 \end{bmatrix}$$

$$\begin{bmatrix} 533.2 & -292.4 \\ -292.4 & 248.5 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \end{bmatrix} = \lambda \begin{bmatrix} a_2 \\ a_1 \end{bmatrix}$$

固有値 $\lambda = 65.65, 716.1$

固有ベクトル $\begin{bmatrix} a_2 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0.530 \\ 0.848 \end{bmatrix}, \begin{bmatrix} 0.848 \\ -0.530 \end{bmatrix}$

情報損失量 $U_1 = \lambda_1, U_2 = \lambda_2$

寄与率

$$\frac{Var(z_1)}{Var(x_1)+Var(x_2)} = 1 - \frac{U_1}{(n-1)\left(Var(x_1)+Var(x_2)\right)} = 1 - \frac{65.65}{533.2+248.5} = 0.916$$

分散$Var(z)$の最大化

分散共分散行列の固有値固有ベクトル問題

$$\begin{bmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_1,x_2) & Var(x_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda' \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\begin{bmatrix} 88.87 & 48.73 \\ 48.73 & 41.41 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda' \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

固有値 $\lambda' = 119.3, 10.94$

固有ベクトル $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.848 \\ 0.530 \end{bmatrix}, \begin{bmatrix} -0.530 \\ 0.848 \end{bmatrix}$

得られる主成分は同じ

分散 $Var(z_1) = \lambda'_1, Var(z_2) = \lambda'_2$

寄与率

$$\frac{Var(z_1)}{Var(x_1)+Var(x_2)} = \frac{\lambda'_1}{Var(x_1)+Var(x_2)} = \frac{119.3}{88.87+41.41} = 0.916$$

主成分分析は単位の影響を受ける

	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2
埼玉	22.9	13.7
千葉	24.9	16.2
東京	19.3	11.3
神奈川	22.0	10.4
茨城	28.6	24.9
栃木	42.6	26.5
群馬	41.3	20.3
平均	28.8	17.614
分散	88.87	41.41
共分散	48.73	

固有値 $\lambda_1 = 65.65$

固有ベクトル $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.848 \\ 0.530 \end{bmatrix}$

	人口1万人当りスポーツ施設x_1	人口10万人当り教育施設x_2
埼玉	2.29	13.7
千葉	2.49	16.2
東京	1.93	11.3
神奈川	2.20	10.4
茨城	2.86	24.9
栃木	4.26	26.5
群馬	4.13	20.3
平均	2.88	17.614
分散	0.8887	41.41
共分散	4.873	

固有値 $\lambda_1 = 41.99$

固有ベクトル $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.118 \\ 0.993 \end{bmatrix}$

主成分はその変数の単位の影響を受ける

変数の単位を変えると方向比 $a_1 : a_2$ も変わるため、
主成分の解釈も変わってしまう。
説明変量が身長と体重のように一方がcmで他方がkgであるような場合もある。

単位の影響を受けない主成分分析はあるか。

⬇

変数の単位の影響を取り除く統計手法：データの標準化

データの標準化　分散 → 1
共分散 → 相関係数

相関係数　$r = \dfrac{x_1 と x_2 の共分散}{\sqrt{x_1の分散} \cdot \sqrt{x_2の分散}}$

単位を変えない場合		
	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2
分散	88.87	41.41
共分散	48.73	

単位を変えた場合		
	人口1万人当りスポーツ施設x_1	人口10万人当り教育施設x_2
分散	0.8887	41.41
共分散	4.873	

$$r = \frac{48.73}{\sqrt{88.87} \cdot \sqrt{41.42}} = 0.8033$$

$$r = \frac{4.873}{\sqrt{0.8887} \cdot \sqrt{41.42}} = 0.8033$$

同じ値になる

データの標準化　　分散　→　1
共分散　→　相関係数

従って

分散共分散行列　　標準化　　相関行列

$$\begin{bmatrix} \text{分散} & \text{共分散} \\ \text{共分散} & \text{分散} \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & \text{相関係数} \\ \text{相関係数} & 1 \end{bmatrix}$$

となる。

4ページ前のスライドでは

主成分の求め方 ➡ 主成分 z の分散を最大にする

分散共分散行列の固有値、固有ベクトルの問題を解くことになる

2ページ前のスライドでは

主成分分析は単位の影響を受ける

単位の影響を受けない主成分分析はあるか

⬇

変数の単位の影響を取り除く統計手法：データの標準化

標準化されたデータの分散共分散行列による主成分分析は
相関行列による主成分分析である

主成分$z = a_1u_1 + a_2u_2$の求め方（標準化→分散$Var(z)$の最大化）

$$Var(z) = Var(a_1u_1 + a_2u_2) = {a_1}^2Var(u_1) + {a_2}^2Var(u_2) + 2a_1a_2Cov(u_1, u_2)$$
$$= {a_1}^2 + {a_2}^2 + 2a_1a_2r$$

ラグランジュの乗数法　$G(a_1, a_2, \lambda') = a_1^2 + a_2^2 + 2a_1a_2r - \lambda'(a_1^2 + a_2^2 - 1)$

$$\begin{cases} \dfrac{\partial G}{\partial a_1} = 2(a_1 + a_2r - \lambda' a_1) = 0 \\ \dfrac{\partial G}{\partial a_2} = 2(a_2 + a_1r - \lambda' a_2) = 0 \end{cases}$$

標準化　$\bar{u}_1 = \bar{u}_2 = 0$, $Var(u_1) = Var(u_2) = 1$, $Cov(u_1, u_2) = r$

サンプル＼変量	人口10万人当りスポーツ施設x_1	人口10万人当り教育施設x_2	$u_1 = \dfrac{x_1 - \bar{x}_1}{\sqrt{Var(x_1)}}$	$u_2 = \dfrac{x_2 - \bar{x}_2}{\sqrt{Var(x_2)}}$
埼玉	22.9	13.7	-0.626	-0.608
千葉	24.9	16.2	-0.414	-0.220
東京	19.3	11.3	-1.008	-0.981
神奈川	22.0	10.4	-0.721	-1.121
茨城	28.6	24.9	-0.021	1.132
栃木	42.6	26.5	1.464	1.381
群馬	41.3	20.3	1.326	0.417
平均	28.8	17.614	0.000	0.000
分散	88.87	41.41	1.000	1.000
共分散	48.73		r =0.803	

相関行列の固有値固有ベクトル問題

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda'\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0.803 \\ 0.803 & 1 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda'\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

固有値　$\lambda' = 1.803, 0.197$

固有ベクトル　$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}, \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix}$

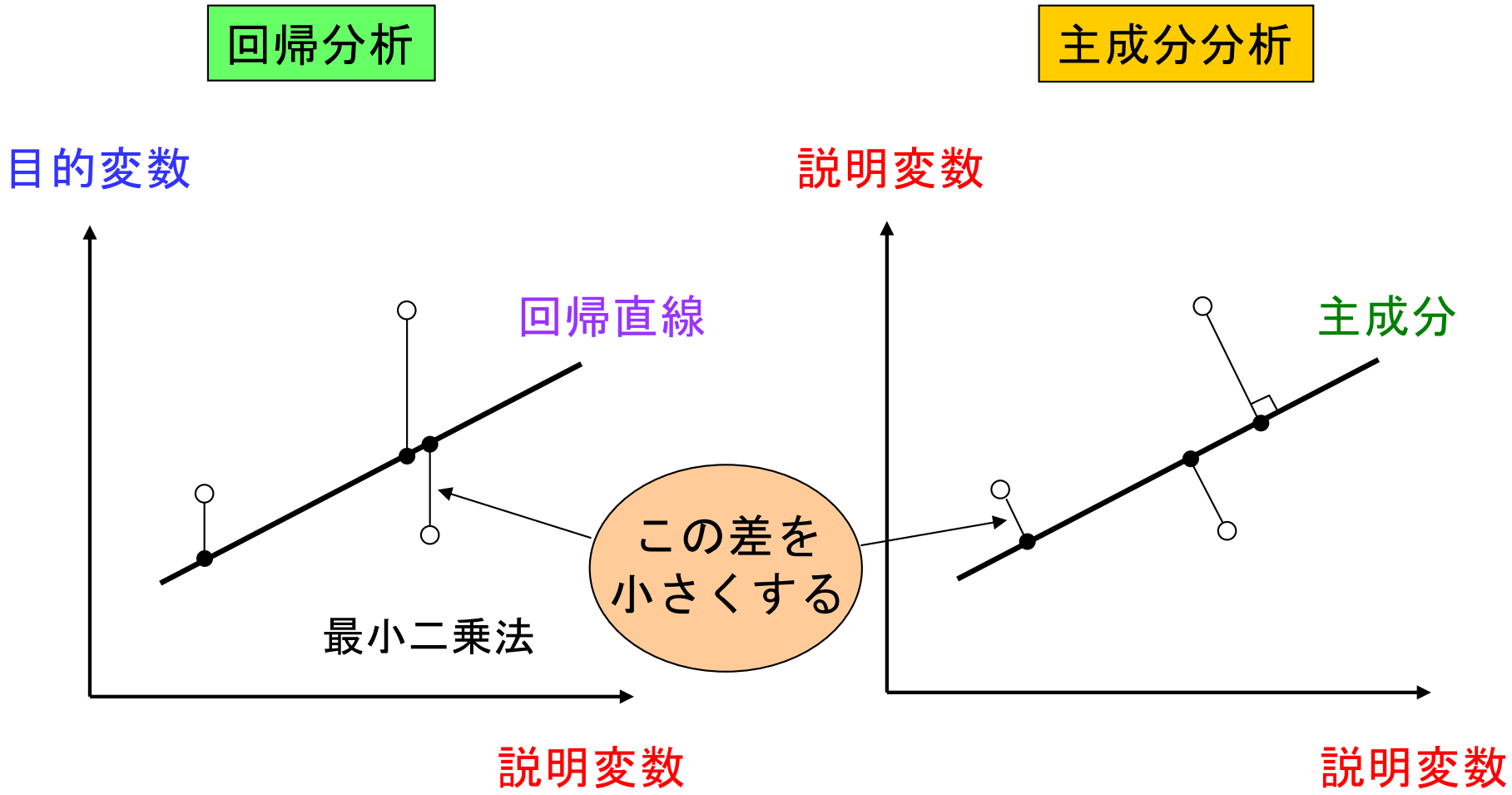
寄与率　$\dfrac{Var(z_1)}{Var(u_1) + Var(u_2)} = \dfrac{1.803}{1+1} = 0.902$

$\dfrac{Var(z_2)}{Var(u_1) + Var(u_2)} = \dfrac{0.197}{1+1} = 0.098$

累積寄与率 = 1

主成分分散=固有値: $Var(z_1) = {\lambda'}_1, Var(z_2) = {\lambda'}_2$

回帰分析と主成分分析



主成分分析

①主成分分析とは何か。
　式も用いて説明せよ。
②主成分の寄与率とは何か。
　式も用いて説明せよ。
③主成分分析は単位の影響を受ける。
　単位の影響を受けないようにする
　ためにはどのようにしたらよいか。
④回帰分析と主成分分析の求め
　方の違いを図を用いて説明せよ。