データの分析と知識発見

Introduction to Data Analysis

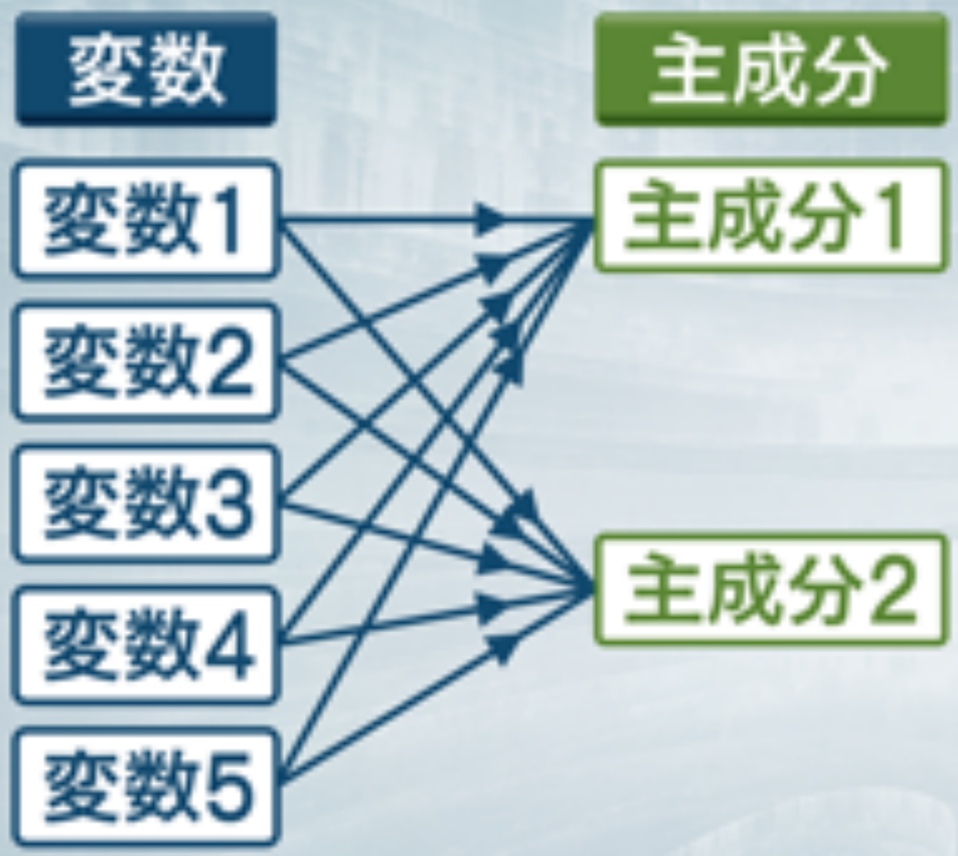
今回の構成

主成分分析について理解する

寄与率について学ぶ

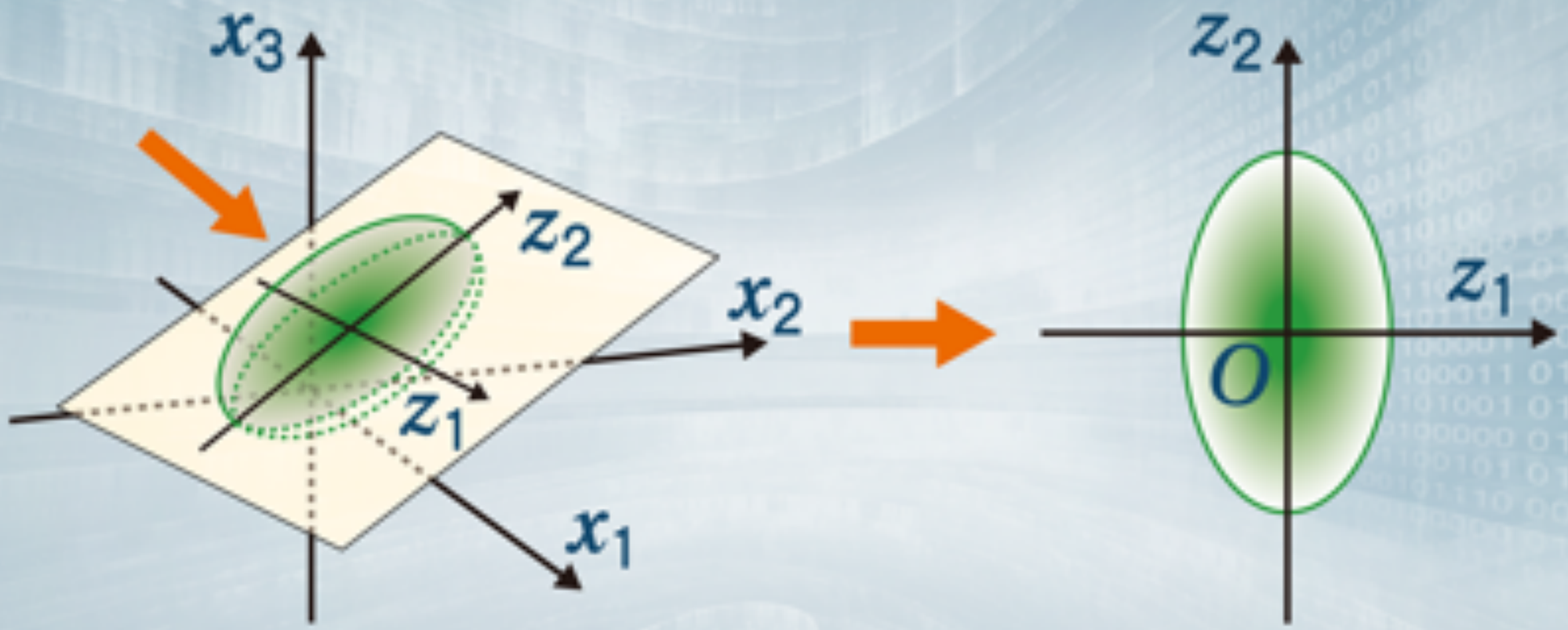
Rを用いて主成分分析を行う

主成分分析



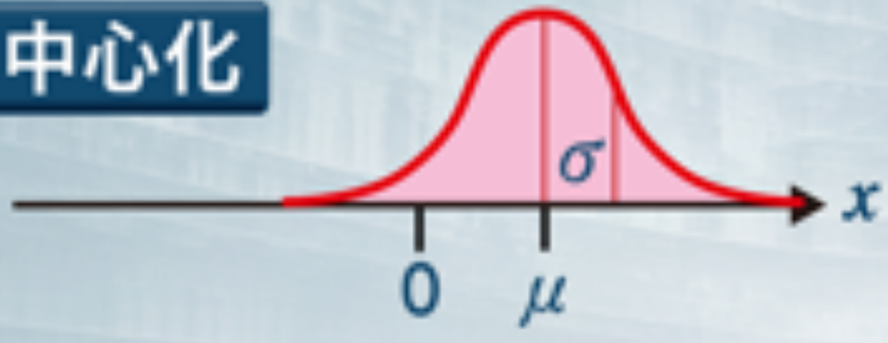
- 変数の特徴をなるべく保持したまま、より少ない変数（主成分）で表す
- 直交回転：ばらつきが大きい方向を探す
- データ間の関係
- データの圧縮

主成分分析とは



中心化と標準化

中心化



$$x_i^{(p)\prime} = (x_i^{(p)} - \mu_i)$$

標準化

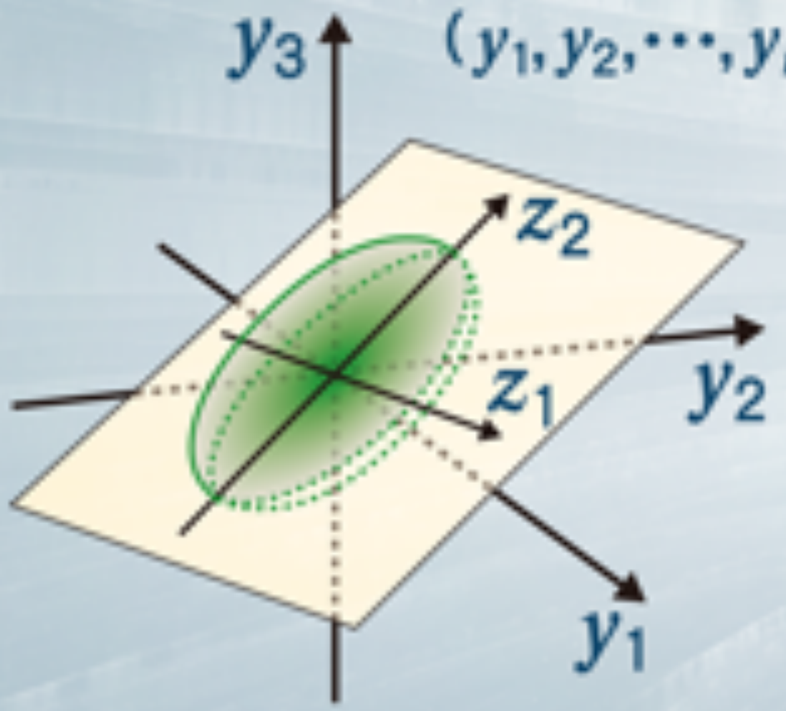
$$x_i^{(p)\prime\prime} = \frac{(x_i^{(p)} - \mu_i)}{\sigma_i}$$

主成分分析の導出

$$(x_1, x_2, \cdots, x_r)(x_1^{(1)}, x_2^{(1)}, \cdots, x_r^{(1)}), \cdots, (x_1^{(N)}, x_2^{(N)}, \cdots, x_r^{(N)})$$

↓ 中心化（または、標準化）

$$(y_1, y_2, \cdots, y_r)(y_1^{(1)}, y_2^{(1)}, \cdots, y_r^{(1)}), \cdots, (y_1^{(N)}, y_2^{(N)}, \cdots, y_r^{(N)})$$



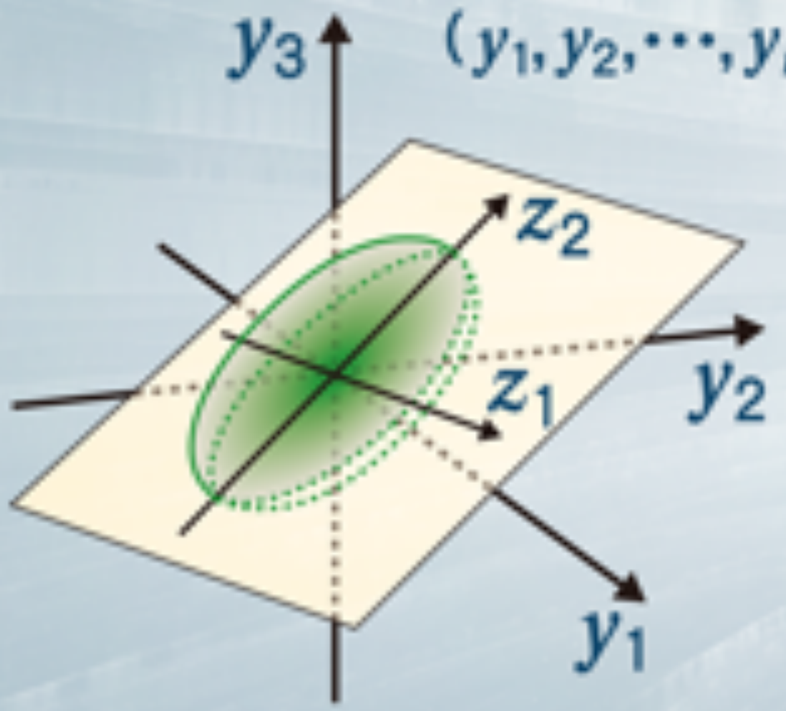
$$\begin{aligned}
z_1 &= a_{11} y_1 + a_{12} y_2 + \cdots + a_{1r} y_r \\
z_2 &= a_{21} y_1 + a_{22} y_2 + \cdots + a_{2r} y_r \\
&\vdots \qquad\qquad\qquad \vdots \\
z_m &= a_{m1} y_1 + a_{m2} y_2 + \cdots + a_{mr} y_r
\end{aligned}$$

主成分分析の導出

$$(x_1, x_2, \cdots, x_r)(x_1^{(1)}, x_2^{(1)}, \cdots, x_r^{(1)}), \cdots, (x_1^{(N)}, x_2^{(N)}, \cdots, x_r^{(N)})$$

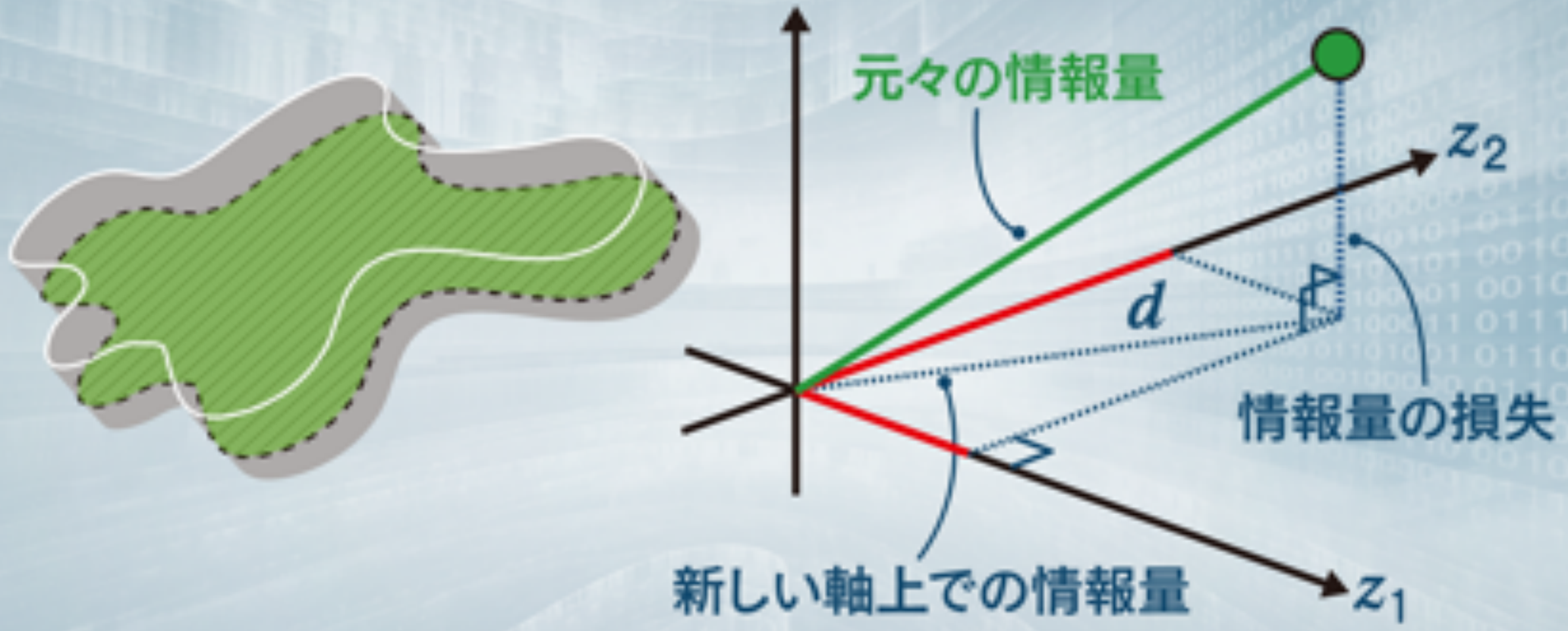
↓ 中心化（または、標準化）

$$(y_1, y_2, \cdots, y_r)(y_1^{(1)}, y_2^{(1)}, \cdots, y_r^{(1)}), \cdots, (y_1^{(N)}, y_2^{(N)}, \cdots, y_r^{(N)})$$



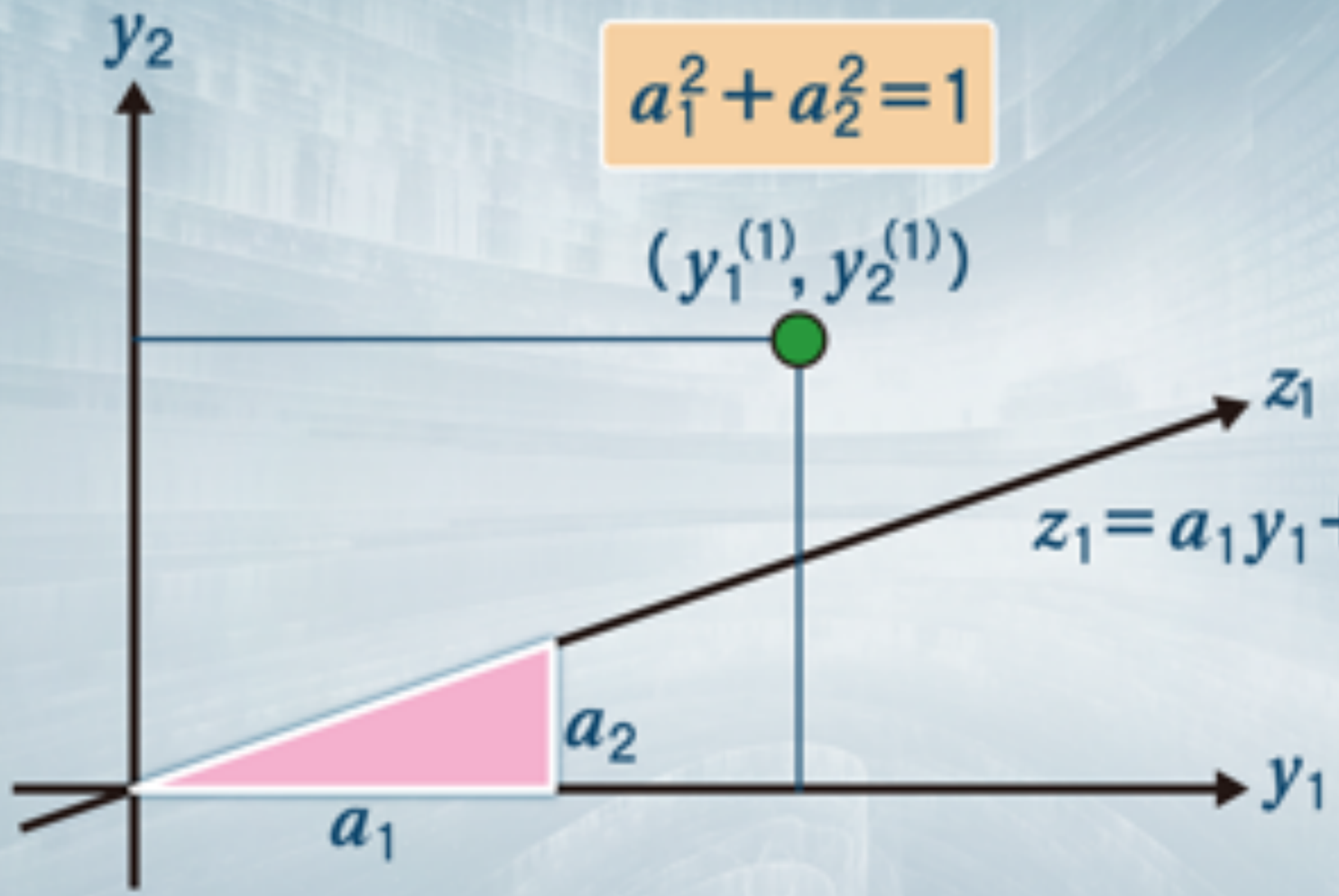
$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & \cdots & a_{1r} \\ & \ddots & \ddots & \\ a_{m1} & \cdots & \cdots & a_{mr} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{pmatrix}$$

主成分分析の導出



主成分分析の導出(1)

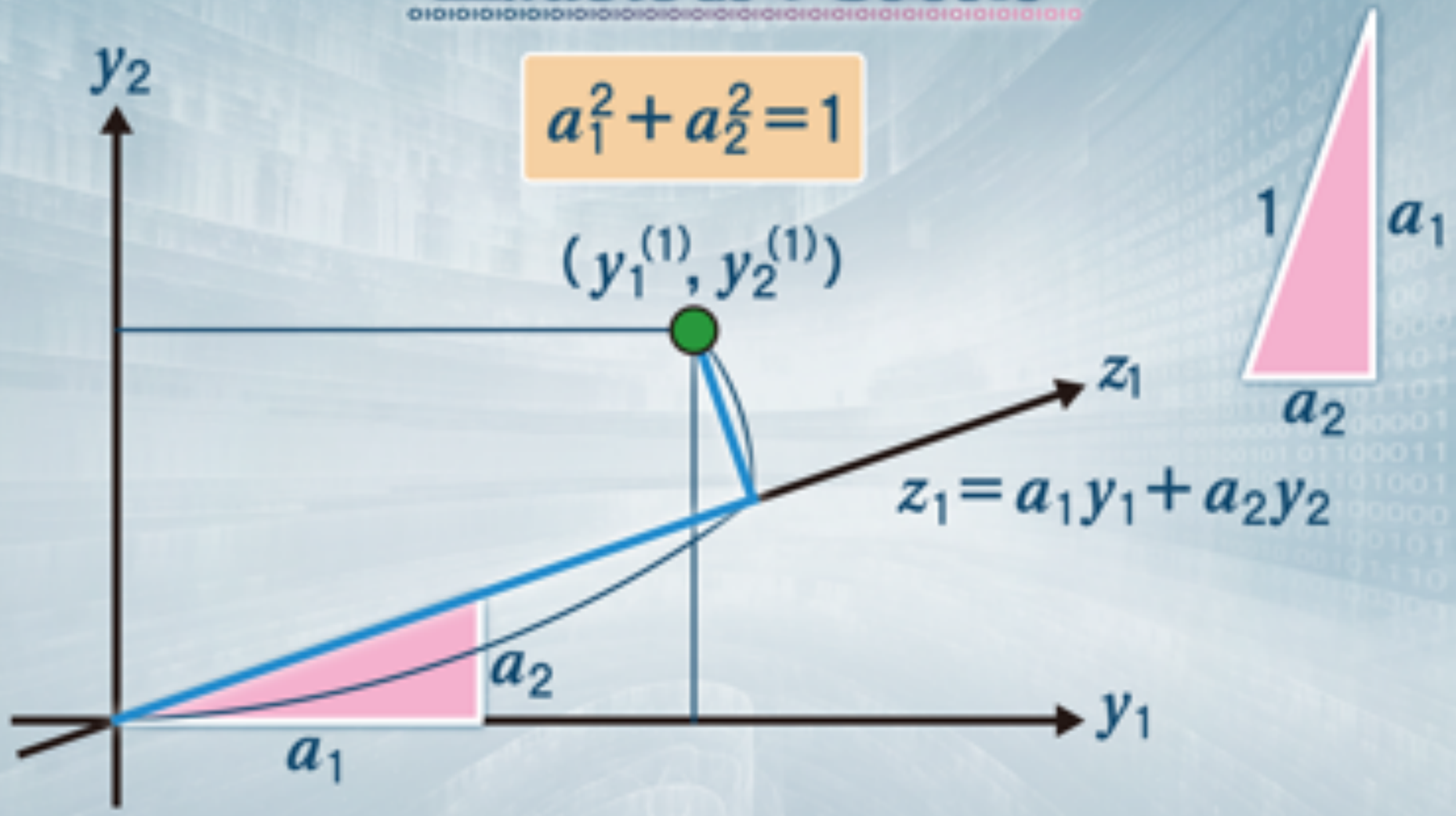
$$a_1^2 + a_2^2 = 1$$



$$z_1 = a_1 y_1 + a_2 y_2$$

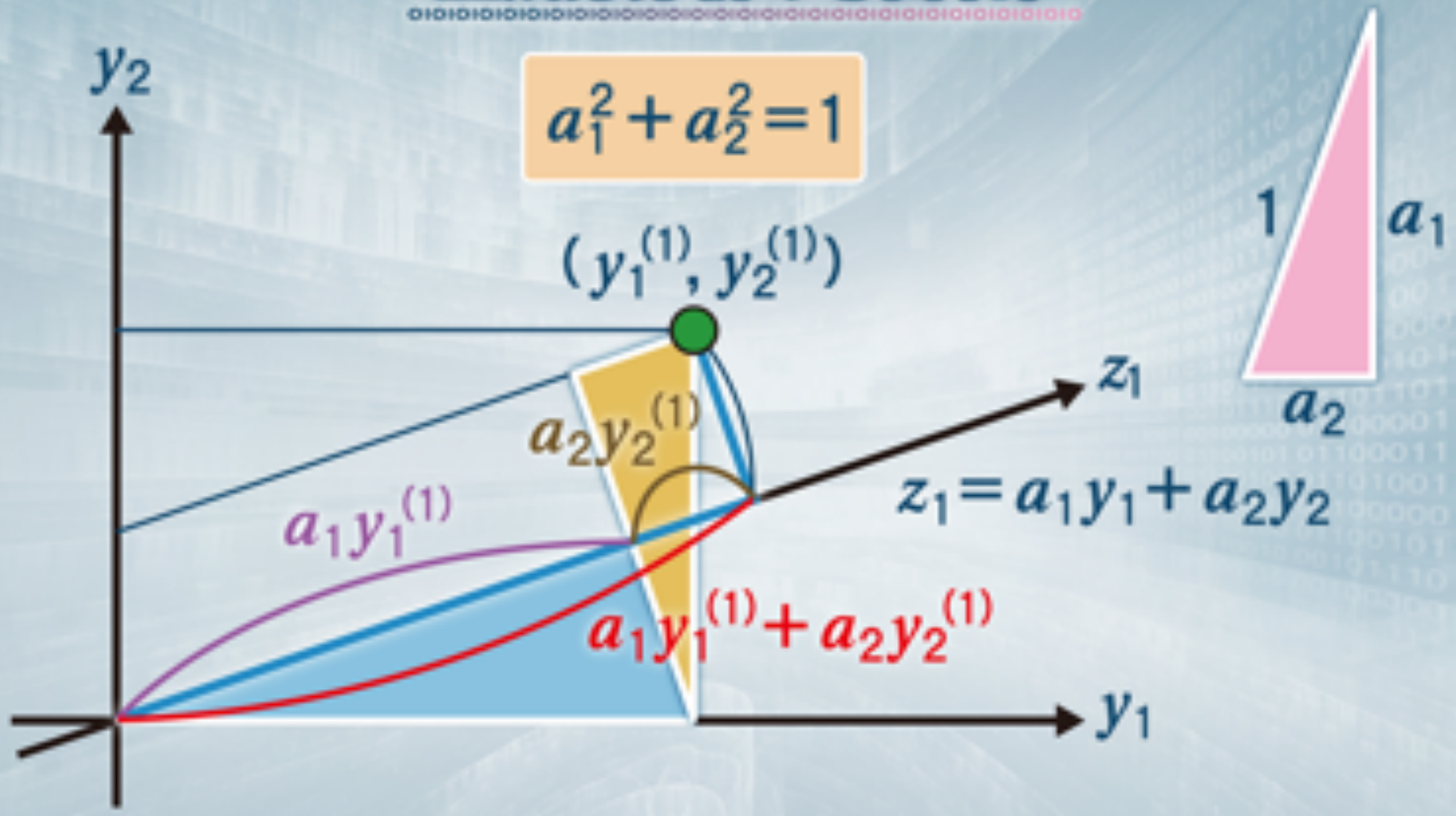
主成分分析の導出(1)

$$a_1^2 + a_2^2 = 1$$

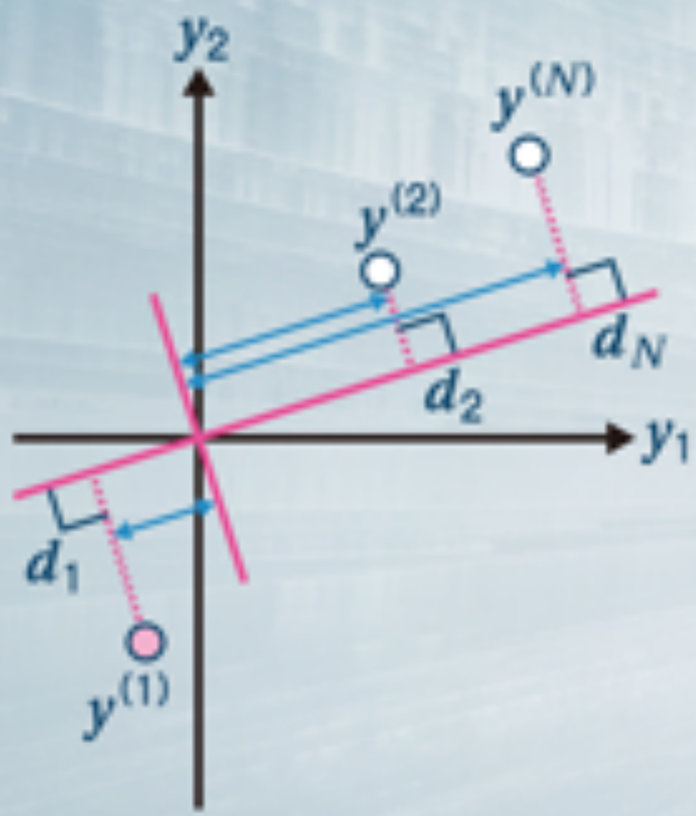


主成分分析の導出(1)

$$a_1^2 + a_2^2 = 1$$



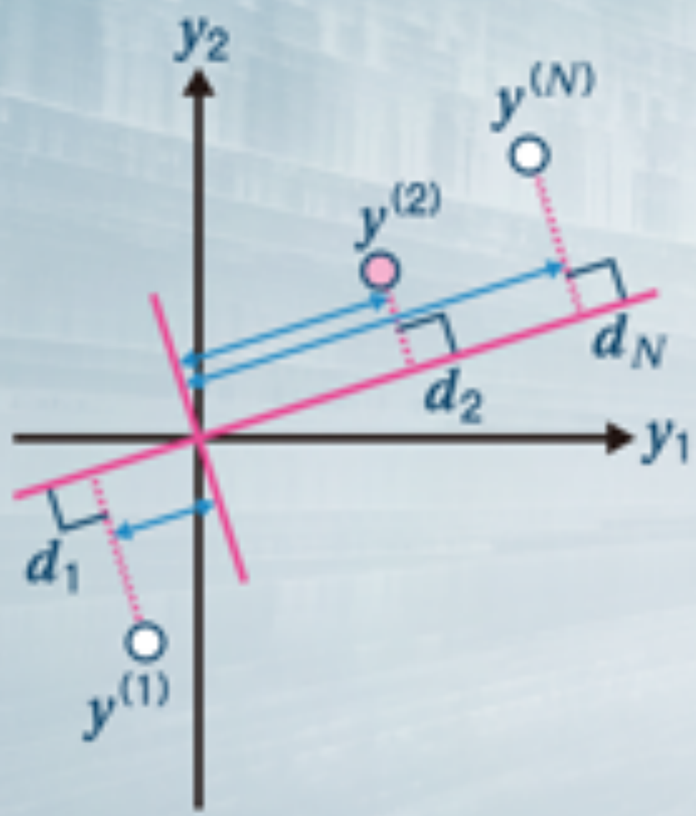
主成分分析の導出(2)



$$(d^{(p)})^2 = (a_1 y_1^{(p)} + a_2 y_2^{(p)})^2$$

$$(d^{(1)})^2 = a_1^2 (y_1^{(1)})^2 + a_2^2 (y_2^{(1)})^2 + 2a_1 a_2 (y_1^{(1)} y_2^{(1)})$$

主成分分析の導出(2)

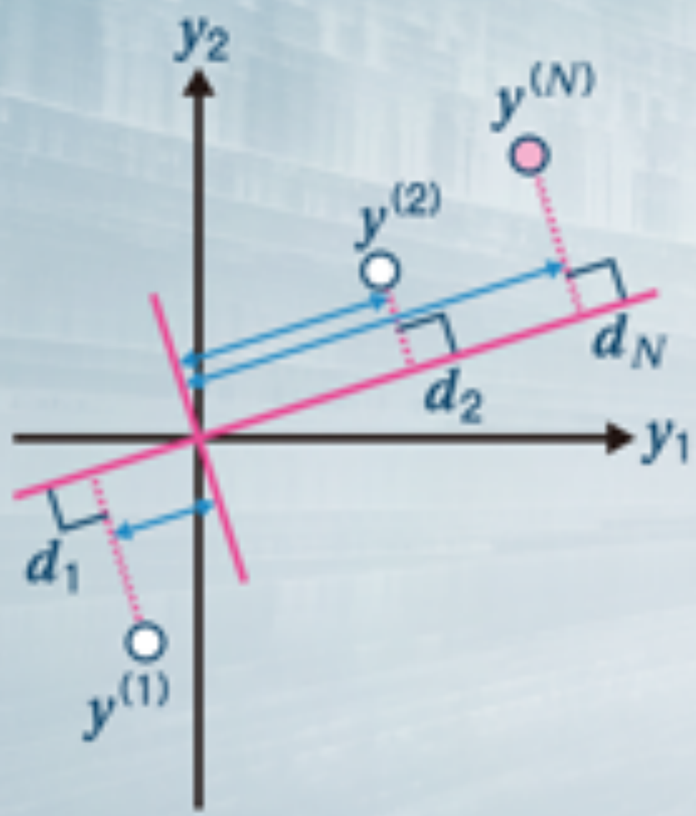


$$(d^{(p)})^2 = (a_1 y_1^{(p)} + a_2 y_2^{(p)})^2$$

$$(d^{(1)})^2 = a_1^2 (y_1^{(1)})^2 + a_2^2 (y_2^{(1)})^2 + 2a_1 a_2 (y_1^{(1)} y_2^{(1)})$$

$$(d^{(2)})^2 = a_1^2 (y_1^{(2)})^2 + a_2^2 (y_2^{(2)})^2 + 2a_1 a_2 (y_1^{(2)} y_2^{(2)})$$

主成分分析の導出(2)



$$(d^{(p)})^2 = (a_1 y_1^{(p)} + a_2 y_2^{(p)})^2$$

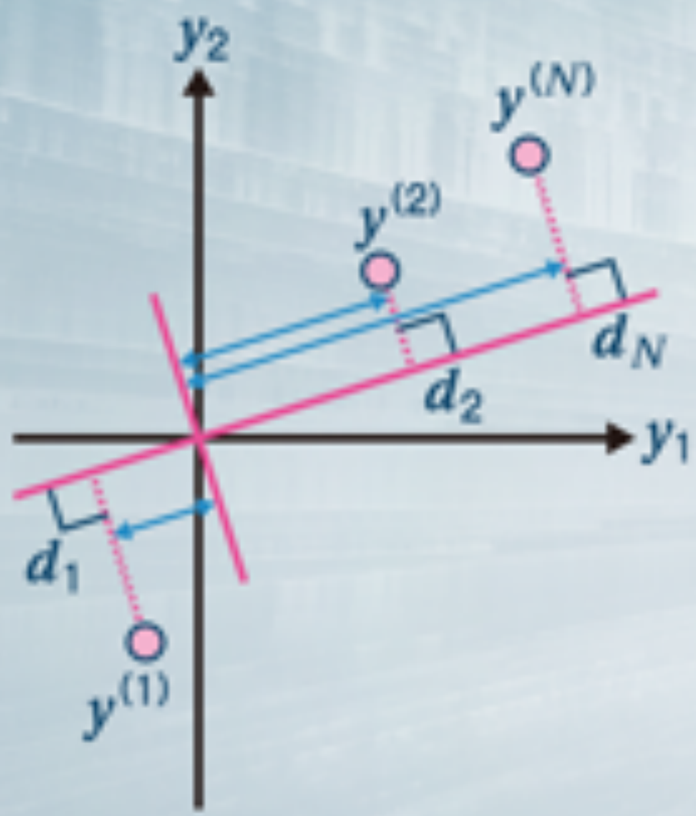
$$(d^{(1)})^2 = a_1^2 (y_1^{(1)})^2 + a_2^2 (y_2^{(1)})^2 + 2a_1 a_2 (y_1^{(1)} y_2^{(1)})$$

$$(d^{(2)})^2 = a_1^2 (y_1^{(2)})^2 + a_2^2 (y_2^{(2)})^2 + 2a_1 a_2 (y_1^{(2)} y_2^{(2)})$$

$$\vdots \qquad \vdots$$

$$(d^{(N)})^2 = a_1^2 (y_1^{(N)})^2 + a_2^2 (y_2^{(N)})^2 + 2a_1 a_2 (y_1^{(N)} y_2^{(N)})$$

主成分分析の導出(2)



$$(d^{(p)})^2 = (a_1 y_1^{(p)} + a_2 y_2^{(p)})^2$$

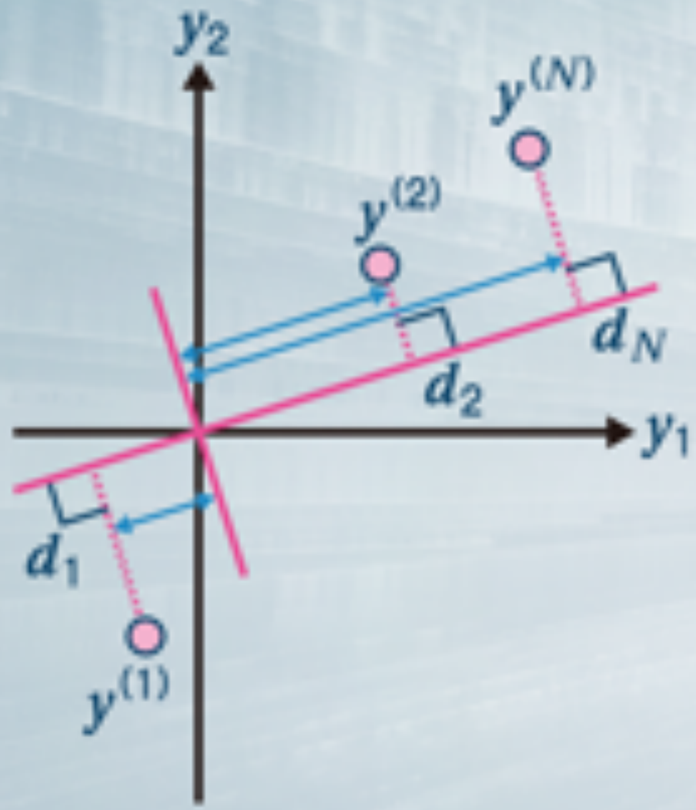
$$\begin{aligned}
(d^{(1)})^2 &= a_1^2 (y_1^{(1)})^2 + a_2^2 (y_2^{(1)})^2 + 2a_1 a_2 (y_1^{(1)} y_2^{(1)}) \\
(d^{(2)})^2 &= a_1^2 (y_1^{(2)})^2 + a_2^2 (y_2^{(2)})^2 + 2a_1 a_2 (y_1^{(2)} y_2^{(2)}) \\
&\vdots \qquad\qquad\qquad \vdots \\
+(d^{(N)})^2 &= a_1^2 (y_1^{(N)})^2 + a_2^2 (y_2^{(N)})^2 + 2a_1 a_2 (y_1^{(N)} y_2^{(N)})
\end{aligned}$$

N(または $N-1$)で割る

$$U(a_1, a_2) = s_{11} a_1^2 + 2 s_{12} a_1 a_2 + s_{22} a_2^2$$

中心化した場合

主成分分析の導出(2)



$$(d^{(p)})^2 = (a_1 y_1^{(p)} + a_2 y_2^{(p)})^2$$

$$\begin{aligned}
(d^{(1)})^2 &= a_1^2 (y_1^{(1)})^2 + a_2^2 (y_2^{(1)})^2 + 2a_1 a_2 (y_1^{(1)} y_2^{(1)}) \\
(d^{(2)})^2 &= a_1^2 (y_1^{(2)})^2 + a_2^2 (y_2^{(2)})^2 + 2a_1 a_2 (y_1^{(2)} y_2^{(2)}) \\
&\vdots \qquad\qquad\qquad \vdots \\
+(d^{(N)})^2 &= a_1^2 (y_1^{(N)})^2 + a_2^2 (y_2^{(N)})^2 + 2a_1 a_2 (y_1^{(N)} y_2^{(N)})
\end{aligned}$$

N(または $N-1$)で割る

$$U(a_1, a_2) = a_1^2 + 2r_{12} a_1 a_2 + a_2^2$$

標準化した場合

ラグランジュの未定乗数法

中心化されたデータの場合

$$U(a_1, a_2) = s_{11}a_1^2 + 2s_{12}a_1a_2 + s_{22}a_2^2$$

$a_1^2 + a_2^2 = 1$ という条件のもとで

$U(a_1, a_2)$ が最大となるような

a_1, a_2 を求める

ラグランジュの未定乗数法

中心化されたデータの場合

$$U(a_1, a_2) = s_{11}a_1^2 + 2s_{12}a_1a_2 + s_{22}a_2^2$$

$a_1^2 + a_2^2 = 1$ という条件のもとで

$U(a_1, a_2)$ が最大となるような

a_1, a_2 を求める

ラグランジュの未定乗数法

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

ラグランジュの未定乗数法

中心化されたデータの場合

$$U(a_1, a_2) = s_{11}a_1^2 + 2s_{12}a_1a_2 + s_{22}a_2^2$$

$a_1^2 + a_2^2 = 1$ という条件のもとで
$U(a_1, a_2)$ が最大となるような
a_1, a_2 を求める

a_1 で偏微分	$s_{11}a_1 + s_{12}a_2 + \lambda a_1 = 0$
a_2 で偏微分	$s_{12}a_1 + s_{22}a_2 + \lambda a_2 = 0$
λ で偏微分	$a_1^2 + a_2^2 = 1$

ラグランジュの未定乗数法

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

ラグランジュの未定乗数法

中心化されたデータの場合

$$U(a_1, a_2) = s_{11}a_1^2 + 2s_{12}a_1a_2 + s_{22}a_2^2$$

$a_1^2 + a_2^2 = 1$ という条件のもとで
$U(a_1, a_2)$ が最大となるような
a_1, a_2 を求める

ラグランジュの未定乗数法

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

a_1 で偏微分	$s_{11}a_1 + s_{12}a_2 + \lambda a_1 = 0$
a_2 で偏微分	$s_{12}a_1 + s_{22}a_2 + \lambda a_2 = 0$
λ で偏微分	$a_1^2 + a_2^2 = 1$

これをまとめると

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$a_1^2 + a_2^2 = 1$$

ラグランジュの未定乗数法

中心化されたデータの場合

$$U(a_1, a_2) = s_{11}a_1^2 + 2s_{12}a_1a_2 + s_{22}a_2^2$$

$a_1^2 + a_2^2 = 1$ という条件のもとで
$U(a_1, a_2)$ が最大となるような
a_1, a_2 を求める

ラグランジュの未定乗数法

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

a_1 で偏微分 $s_{11}a_1 + s_{12}a_2 + \lambda a_1 = 0$

a_2 で偏微分 $s_{12}a_1 + s_{22}a_2 + \lambda a_2 = 0$

λ で偏微分 $a_1^2 + a_2^2 = 1$

これをまとめると

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$a_1^2 + a_2^2 = 1$$

分散共分散行列の
固有値, 固有ベクトルを求める

固有値と寄与率

中心化後のデータの場合

$$
\begin{aligned}
U(a_1^*, a_2^*) &= s_{11}a_1^{*2} + s_{22}a_2^{*2} + 2s_{12}a_1^*a_2^* \\
&= a_1^*(s_{11}a_1^* + s_{12}a_2^*) + a_2^*(s_{12}a_1^* + s_{22}a_2^*) \\
&= \lambda(a_1^{*2} + a_2^{*2}) = \lambda
\end{aligned}
$$

標準化後のデータの場合

$$
\begin{aligned}
U(a_1^*, a_2^*) &= a_1^{*2} + a_2^{*2} + 2r_{12}a_1^*a_2^* \\
&= a_1^*(a_1^* + r_{12}a_2^*) + a_2^*(r_{12}a_1^* + a_2^*) \\
&= \lambda(a_1^{*2} + a_2^{*2}) = \lambda
\end{aligned}
$$

固有値と寄与率

中心化後のデータの場合

$$
\begin{aligned}
U(a_1^*, a_2^*) &= s_{11}a_1^{*2} + s_{22}a_2^{*2} + 2s_{12}a_1^*a_2^* \\
&= a_1^*(s_{11}a_1^* + s_{12}a_2^*) + a_2^*(s_{12}a_1^* + s_{22}a_2^*) \\
&= \lambda(a_1^{*2} + a_2^{*2}) = \lambda
\end{aligned}
$$

標準化後のデータの場合

$$
\begin{aligned}
U(a_1^*, a_2^*) &= a_1^{*2} + a_2^{*2} + 2r_{12}a_1^*a_2^* \\
&= a_1^*(a_1^* + r_{12}a_2^*) + a_2^*(r_{12}a_1^* + a_2^*) \\
&= \lambda(a_1^{*2} + a_2^{*2}) = \lambda
\end{aligned}
$$

- 値の大きい固有値から順に考える
- 値の大きい固有値に対応した軸（主成分）を順に
  第1主成分
  第2主成分
  という
- 固有ベクトルが互いに直交する

主成分分析の導出（寄与率，累積寄与率）

一般（r次元）

中心化後のデータの場合

$$A=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{r1} & s_{r2} & \cdots & s_{rr} \end{pmatrix}$$

$$A\boldsymbol{a}_i=\lambda_i\boldsymbol{a}_i,\ \boldsymbol{a}_i=\begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ir} \end{pmatrix}$$

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r\geq 0$$

寄与率

$$\frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_r}=\frac{\lambda_i}{\sum_{k=1}^{r}\lambda_k}$$

累積寄与率

$$\frac{\lambda_1+\cdots+\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_r}=\frac{\sum_{j=1}^{i}\lambda_j}{\sum_{k=1}^{r}\lambda_k}$$

分散共分散行列の固有値・固有ベクトル

主成分分析の導出（寄与率，累積寄与率）

一般（r次元）

標準化後のデータの場合

$$A=\begin{pmatrix} 1 & r_{12} & \cdots & r_{1r} \\ r_{21} & 1 & \cdots & r_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ r_{r1} & r_{r2} & \cdots & 1_{rr} \end{pmatrix}$$

$$A\boldsymbol{a}_i=\lambda_i\boldsymbol{a}_i,\ \boldsymbol{a}_i=\begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ir} \end{pmatrix}$$

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r\geq 0$$

寄与率

$$\frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_r}=\frac{\lambda_i}{\sum_{k=1}^{r}\lambda_k}$$

累積寄与率

$$\frac{\lambda_1+\cdots+\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_r}=\frac{\sum_{j=1}^{i}\lambda_j}{\sum_{k=1}^{r}\lambda_k}$$

相関行列の固有値・固有ベクトル

主成分分析のまとめ

一般（r次元）

中心化後のデータの場合

$$A=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{r1} & s_{r2} & \cdots & s_{rr} \end{pmatrix}$$

$$A\boldsymbol{a}_i=\lambda_i\boldsymbol{a}_i,\ \boldsymbol{a}_i=\begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ir} \end{pmatrix}$$

累積寄与率

$$\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\lambda_2+\cdots+\lambda_r}=\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{k=1}^{r}\lambda_k}$$

m個の固有値・固有ベクトルを計算し

$$\begin{aligned} z_1&=a_{11}y_1+a_{12}y_2+\cdots+a_{1r}y_r \\ z_1&=a_{21}y_1+a_{22}y_2+\cdots+a_{2r}y_r \\ &\vdots \\ z_m&=a_{m1}y_1+a_{m2}y_2+\cdots+a_{mr}y_r \end{aligned}$$

を求める

主成分分析のまとめ

一般(r次元)

標準化後のデータの場合

$$A=\begin{pmatrix} 1 & r_{12} & \cdots & r_{1r} \\ r_{21} & 1 & \cdots & r_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ r_{r1} & r_{r2} & \cdots & 1_{rr} \end{pmatrix}$$

$$A\boldsymbol{a}_i=\lambda_i\boldsymbol{a}_i,\ \boldsymbol{a}_i=\begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ir} \end{pmatrix}$$

累積寄与率

$$\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\lambda_2+\cdots+\lambda_r}=\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{k=1}^{r}\lambda_k}$$

m個の固有値・固有ベクトルを計算し

$$\begin{aligned} z_1&=a_{11}y_1+a_{12}y_2+\cdots+a_{1r}y_r \\ z_1&=a_{21}y_1+a_{22}y_2+\cdots+a_{2r}y_r \\ &\vdots \\ z_m&=a_{m1}y_1+a_{m2}y_2+\cdots+a_{mr}y_r \end{aligned}$$

を求める

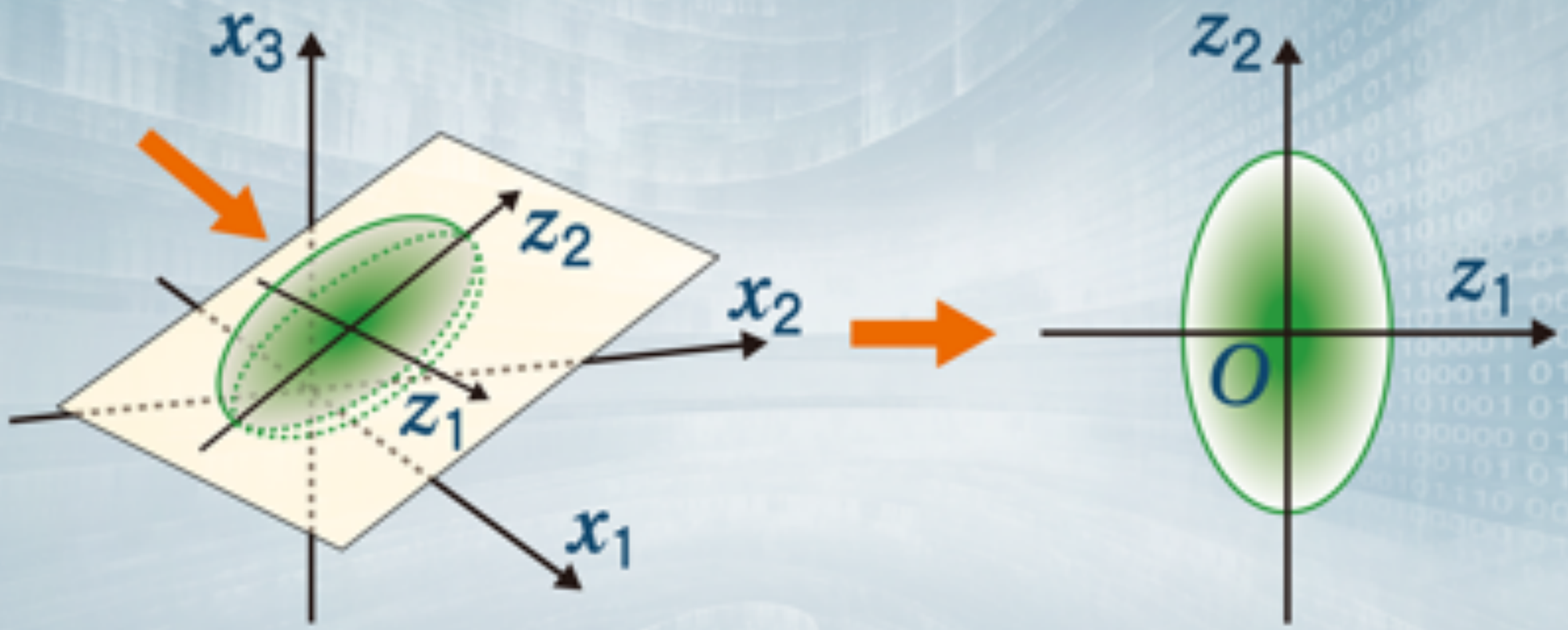
Rによる主成分分析の計算

```
> w1 <- read.table("ch081.dat", header=TRUE, row.names=1)
```

```
> w2 <- prcomp(w1, center=TRUE, scale=TRUE)
> summary(w2)
```

```
> plot(w2$x, type="n")
> text(w2$x, rownames(w2$x))
```

主成分分析とは



Rによる主成分分析の計算

```
> w1 <- read.table("ch081.dat", header=TRUE, row.names=1)
```

```
> w2 <- prcomp(w1, center=TRUE, scale=TRUE)
> summary(w2)
```

```
> plot(w2$x, type="n")
> text(w2$x, rownames(w2$x))
```

データについて

体格データ		
name	height	weight
A1	147.3	52.3
A2	149.9	53.2
A3	152.4	54.5
A4	154.9	55.9
A5	157.5	57.3
A6	160.0	58.6
A7	162.6	60.0
A8	165.1	61.4
A9	167.6	63.2
A10	170.2	64.5
A11	172.7	66.4
A12	175.3	68.2
A13	177.8	70.0
A14	180.3	72.3
A15	182.9	74.5

標準化後		
name	height	weight
A01	−1.5652476	−1.4022687
A02	−1.3416408	−1.2732255
A03	−1.1180340	−1.0796608
A04	−0.8944272	−0.8860962
A05	−0.6708204	−0.6925315
A06	−0.4472136	−0.4989668
A07	−0.2236068	−0.3054021
A08	0.0000000	−0.1118374
A09	0.2236068	0.1462489
A10	0.4472136	0.3398136
A11	0.6708204	0.5978998
A12	0.8944272	0.8559861
A13	1.1180340	1.1140723
A14	1.3416408	1.4366802
A15	1.5652476	1.7592880