多変量解析の基礎（主成分分析）
―理論とRによる演習―

[本稿のWebページ]

古橋武

多変量解析

- 主成分分析

について基礎理論を解説し,
Rによる演習を行います.

3. 主成分分析
3.1 基礎理論

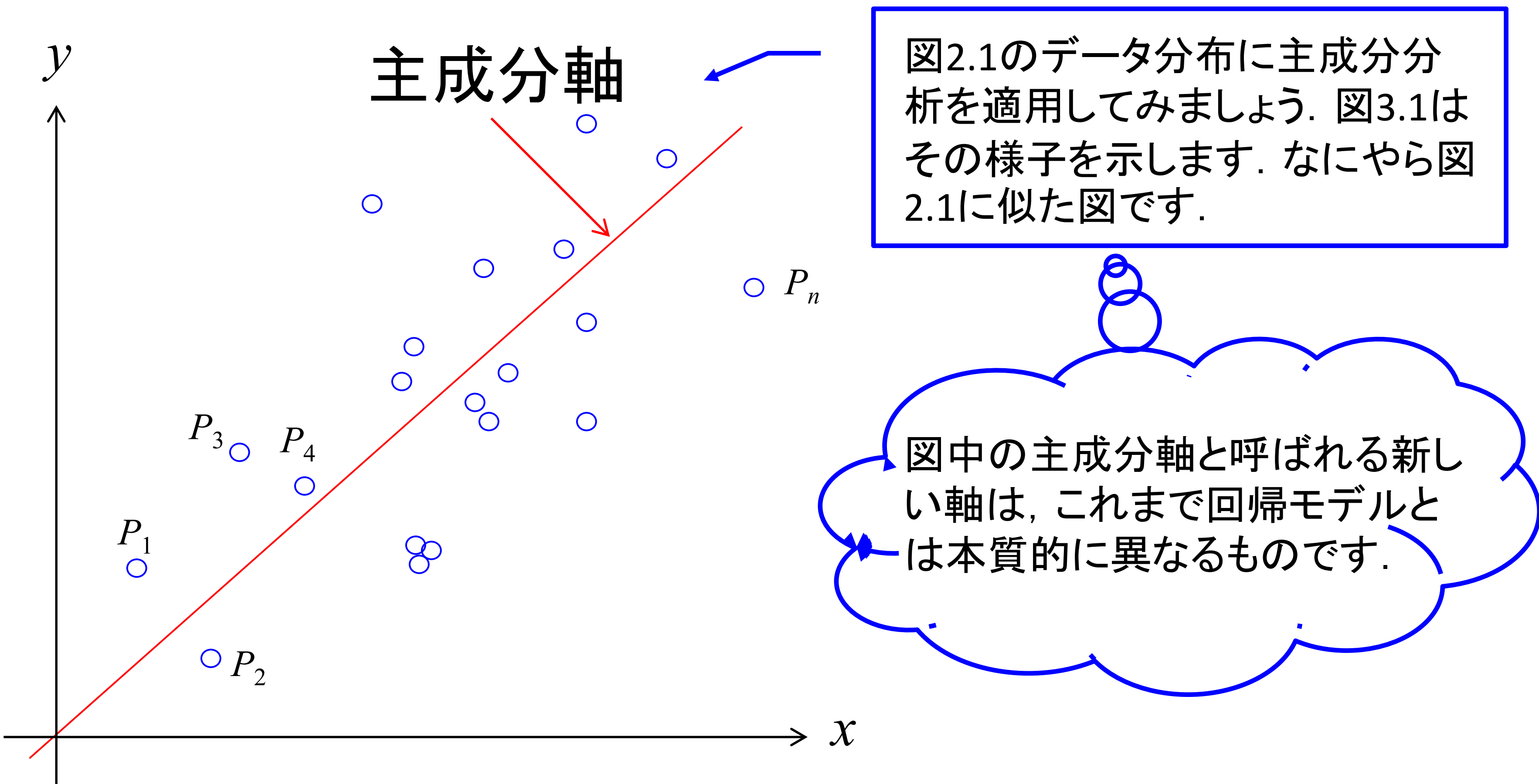


図3.1 主成分分析の様子

主成分軸は次式の値を最小とする軸です.

$$E = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 \qquad (3.1)$$

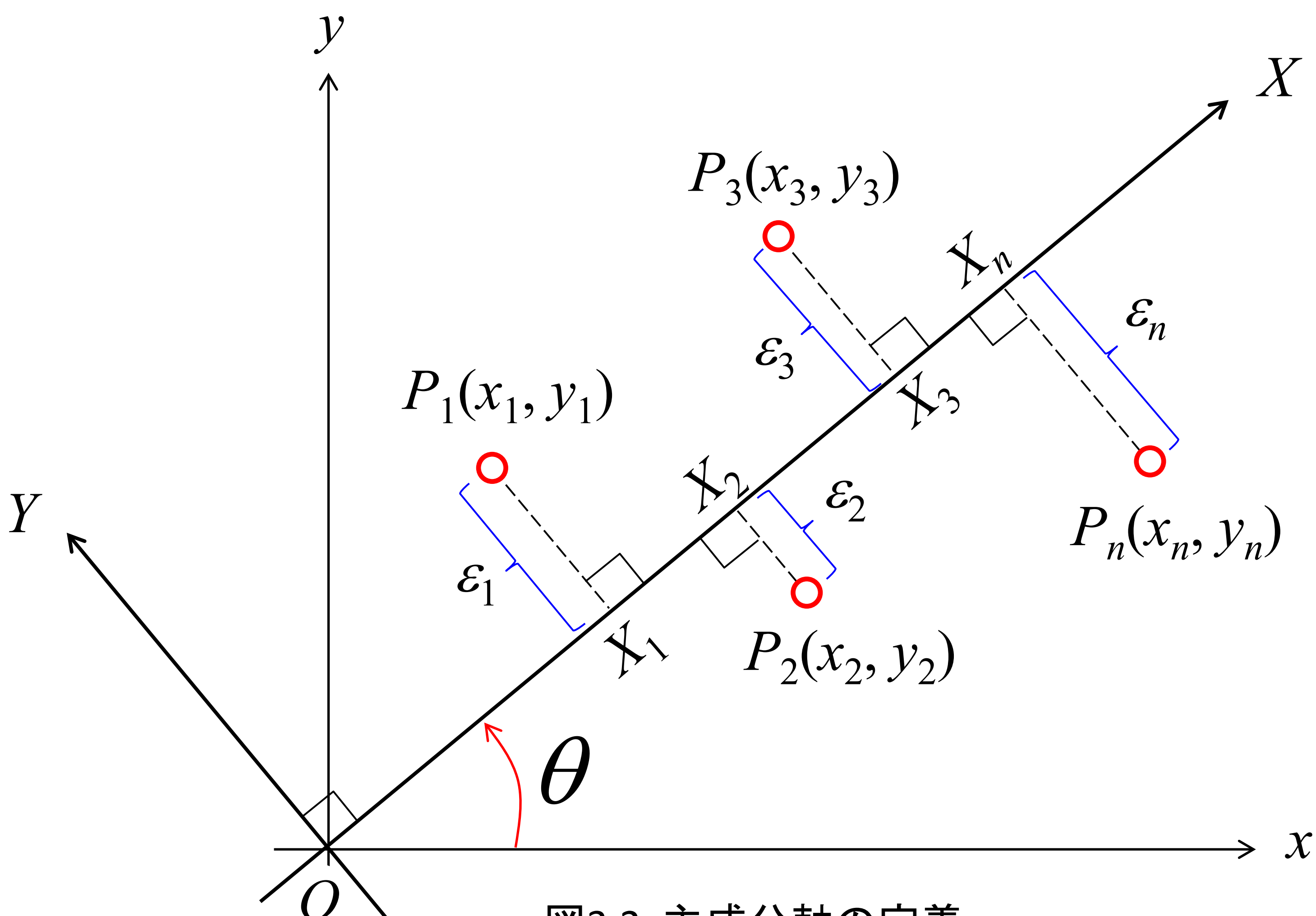


図3.2 主成分軸の定義

ピタゴラスの定理により，図3.3の直角三角形において

$$\overline{OP_1}^2 = \varepsilon_1^2 + \overline{OX_1}^2 \tag{3.2}$$

の関係が成立するので，(3.1)式の右辺は

$$E = \overline{OP_1}^2 - \overline{OX_1}^2 + \cdots + \overline{OP_n}^2 - \overline{OX_n}^2 \tag{3.3}$$

と変形できます．

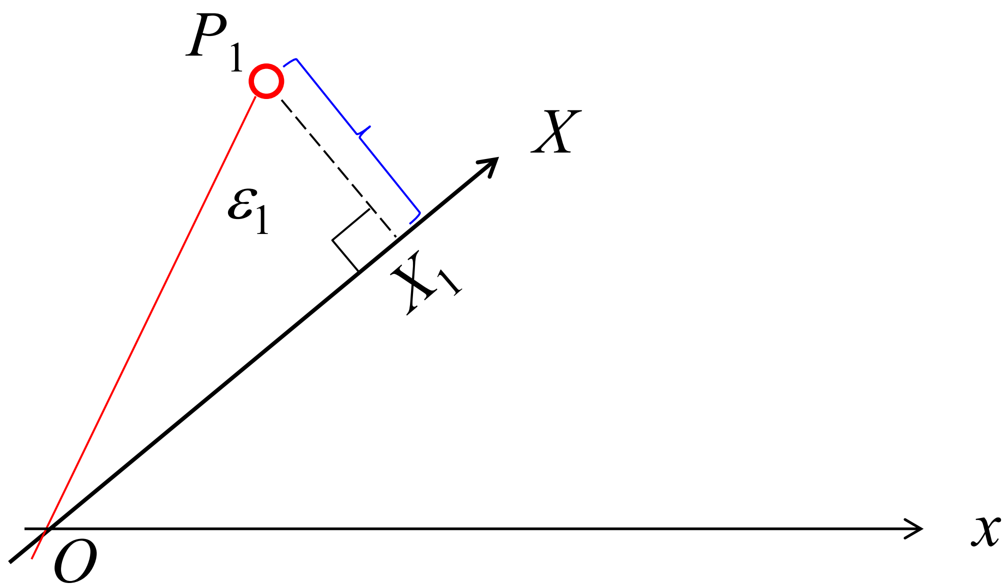


図3.3 ピタゴラスの定理

$\overline{OP_1}, \overline{OP_2}, \cdots, \overline{OP_n}$ は角度q に無関係なので，Eを最小化することは

$$\overline{OX_1}^2 + \cdots + \overline{OX_n}^2 \qquad (3.4)$$

を最大化することと等価です．主成分軸は上式の値を最大とする軸です．言い換えると**データの座標が最も「ばらつく」軸が主成分軸**です．

X_i と x_i, y_iの関係を以下のように表します．

$$\begin{aligned} X_1 &= x_1 l_1 + y_1 l_2 \\ X_2 &= x_2 l_1 + y_2 l_2 \\ &\vdots \\ X_n &= x_n l_1 + y_n l_2 \end{aligned} \qquad (3.6)$$

行列・ベクトルで表すと

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \qquad (3.7)$$

となります．ここで

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \qquad Z = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} \qquad (3.8)$$

$$\boldsymbol{l} = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix}$$

とおくと，(3.7)式は

$$X = \mathrm{Z}\boldsymbol{l} \qquad (3.9)$$

と簡潔に表せます．$\boldsymbol{l}$は**主成分軸を表すベクトル**です．

$$\overline{OX_1}^2 + \cdots + \overline{OX_n}^2 \qquad (3.4)$$

(3.4)式を最大化することは

$$X^t X \rightarrow 最大 \qquad (3.10)$$

と表現できます． $X^t X$ は**目的関数**と呼ばれます．ただし，

$$l_1^2 + l_2^2 = 1 \qquad (3.11)$$

の**制約条件**の下で最大化する必要があります．

制約条件下での最大化問題の解法にラグランジュの未定定数法があります.

$$
\begin{aligned}
F &= \frac{1}{n-1} X^t X - \lambda(\boldsymbol{l}^t \boldsymbol{l} - 1) \\
&= \frac{1}{n-1} \boldsymbol{l}^t Z^t Z \boldsymbol{l} - \lambda(\boldsymbol{l}^t \boldsymbol{l} - 1) \\
&= \boldsymbol{l}^t S \boldsymbol{l} - \lambda(\boldsymbol{l}^t \boldsymbol{l} - 1)
\end{aligned}
\qquad (3.12)
$$

ただし, $S = \frac{1}{n-1} Z^t Z$ です. λ は未定定数と呼ばれます.

このFを最大化する $\boldsymbol{l}$ は

$$
S\boldsymbol{l} = \lambda \boldsymbol{l} \qquad (3.14)
$$

を満たします. 主成分軸 $\boldsymbol{l}$ を求める問題は行列 $\boldsymbol{S}$ の固有値問題に帰着されます.

3.2 Rによる計算

Rを立ち上げて「学問塾」フォルダにある「**主成分分析_身体測定_身長_座高.R**」ファイルを開いてください．2.2項と同じ「身体測定_身長_座高.csv」のデータを用いて主成分分析を行います．

Rエディタの表示

```
x_身体測定 <- read.csv(“C:/Users/Furuhashi/Documents/主成分分析/身体測定_身長_座高.csv")

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)

XX <- as.matrix(x_身体測定)                    #データフレームを行列へ変換

n <- nrow(XX)                                  # 行数のカウント

S <- (1/(n-1))*t(XX) %*% XX                    #S = X^t Xの計算

Eigen_value <- eigen(S)                        # Sの固有値計算

abline(0, Eigen_value$vectors[2,1]/Eigen_value$vectors[1,1])#第1主成分軸の描画
```

plot(x_身体測定$身長 ,x_身体測定$座高,col="red“, pch=1)

により図2.5と同じ結果が表示されます.

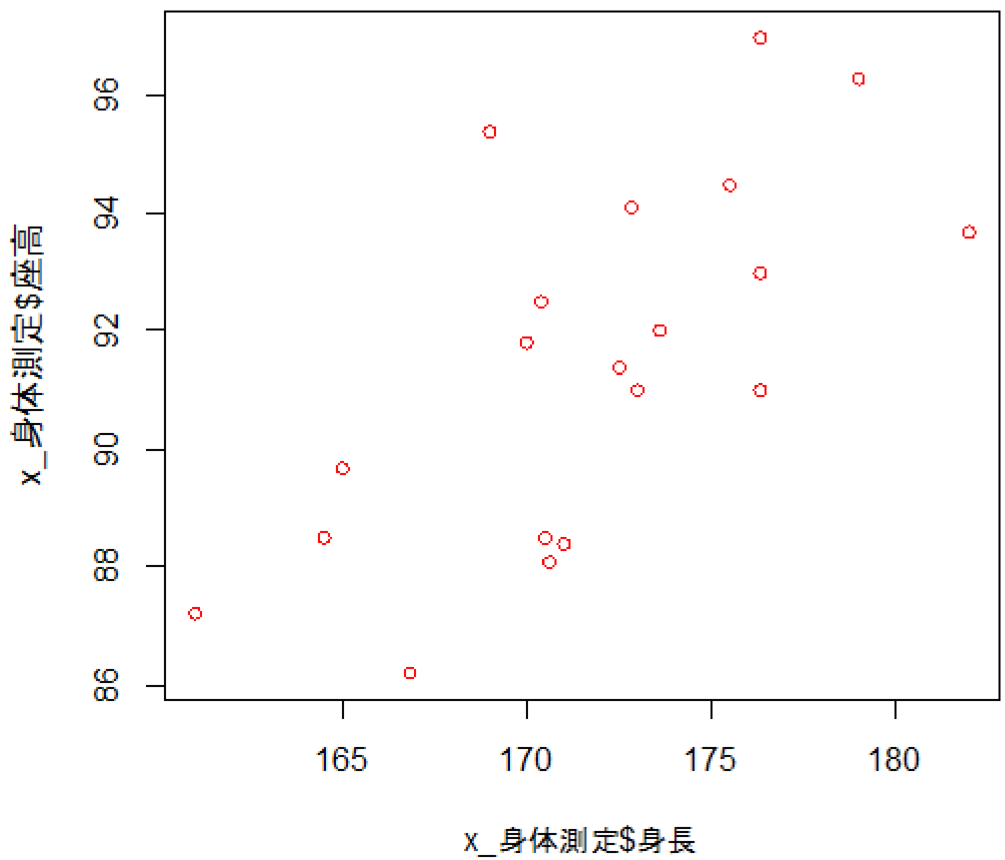


図2.5 スクリプトの3行目の実行結果

S <- (1/(n-1))*t(XX) %*% XX

により行列 S が求められます.

$S = \frac{1}{n-1} X^t X$ の計算

R Consoleの表示

- S <- (1/(n-1))*t(XX) %*% XX
- > S

	身長	座高
身長	31096.51	16561.029
座高	16561.03	8825.373

```
Eigen_value <- eigen(S)
```

*S*の固有値計算

により，行列 S の固有値が求められます．

R Consoleの表示

```
> Eigen_value <- eigen(S)          # Sの固有値計算
> Eigen_value
$values
[1] 39917.606272    4.273728

$vectors
           [,1]        [,2]
[1,] -0.8826062  0.4701131
[2,] -0.4701131 -0.8826062
```

固有値

第1主成分軸の固有値

固有ベクトル

第1主成分軸の固有ベクトル

図3.6 eigen(S)の実行結果

```
abline(0, Eigen_value$vectors[2,1]/Eigen_value$vectors[1,1])
```

により図3.7のように主成分軸が表示されます.

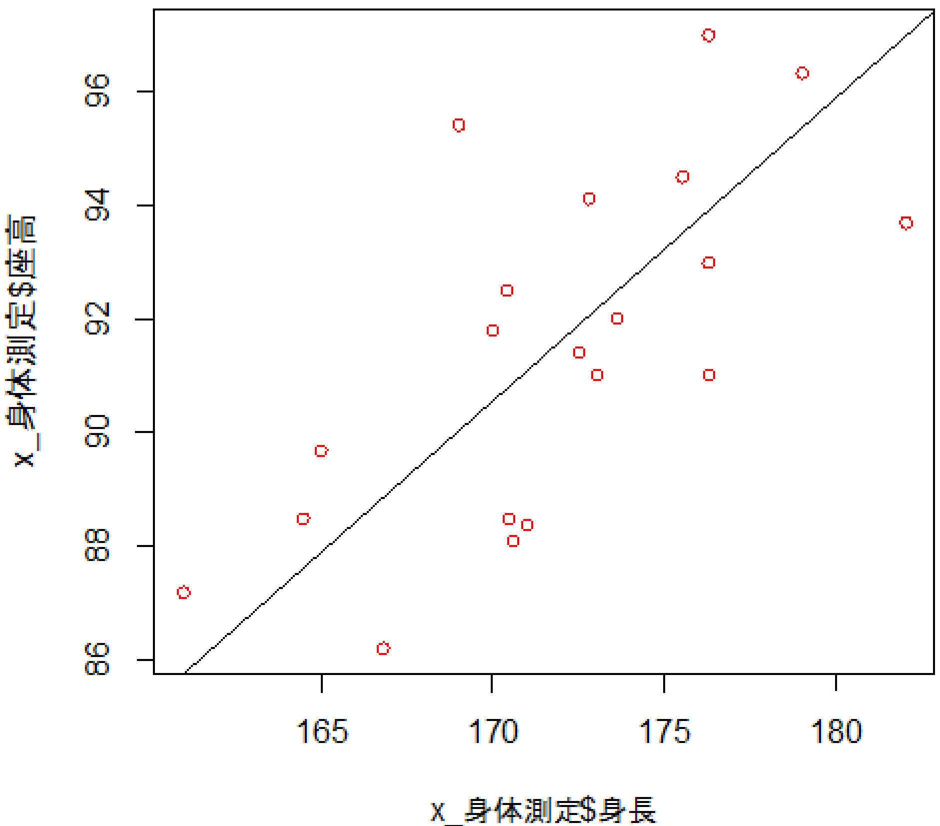


図3.7 主成分分析結果(主成分軸の表示)

3.3 データの正規化

図3.8は前項と同じ「**主成分分析_身体測定_身長_座高.R**」ファイルのスクリプトにおいて第1行目を書き換えて，「**身体測定_身長_座高（変形版）.csv**」のデータを読み込んで主成分分析を行った結果です．

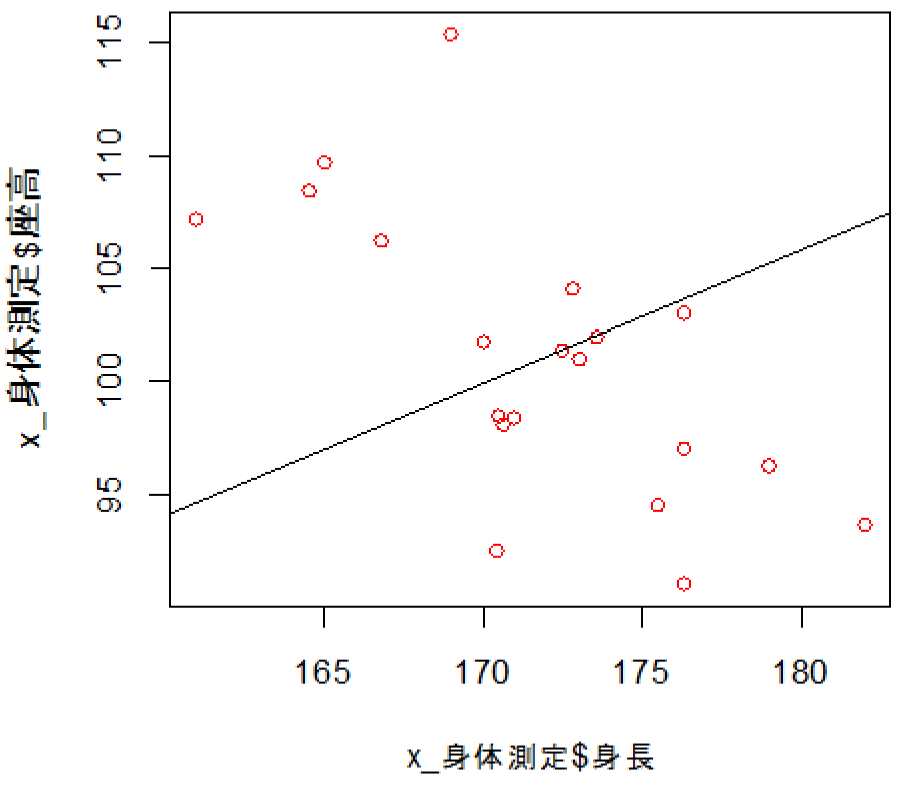
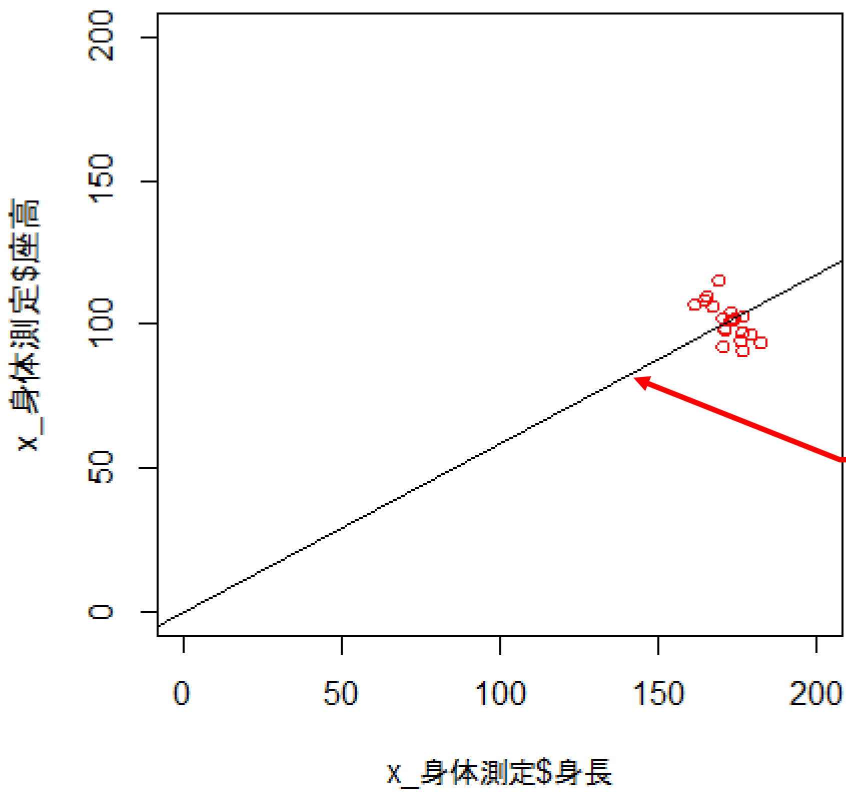


図3.8 規格化前のデータに対する主成分分析結果

図3.9は図3.8の縦軸，横軸を原点を含めて表示したグラフです．



この主成分軸はデータの「ばらつき」を捉えていると言えるでしょうか？言い難いですね．この問題に対応するには，主成分軸の定義を変えないといけないのでしょうか？

図3.9 規格化前のデータに対する主成分分析結果（原点を含む図）

この疑問に対しては，データの前処理で対応できるという答えになります．必要な前処理は**規格化**です．主成分分析の考え方，定式化を見直す必要はありません．

データの規格化は

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{v_e^2}} \qquad (3.17)$$

ただし，$\bar{x}$ は平均，v_e^2 は不偏分散です．

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$v_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (3.16)$$

規格されたデータ $\tilde{x}_i$ の平均は0，不偏分は1となります

図3.10は図3.8のデータを規格化して得られた分布です．大きく変わったのは軸の値です．縦軸，横軸のいずれも0を中心として標準偏差が1に近い分布が得られています．

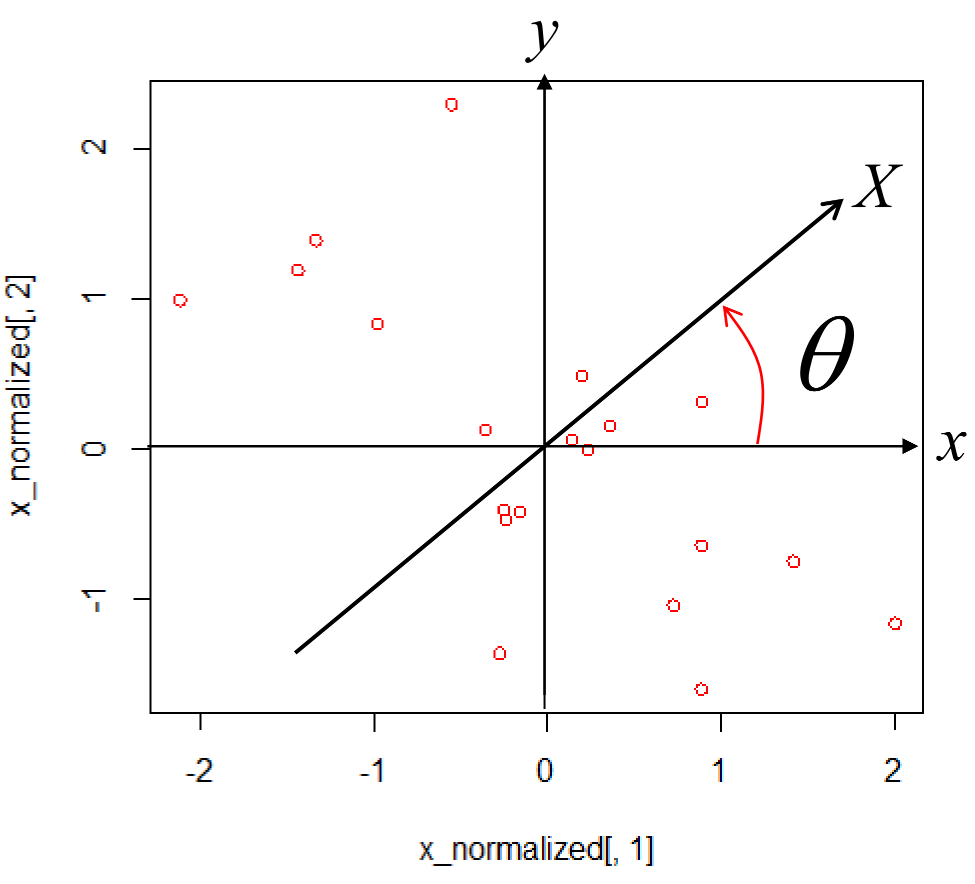


図3.10 規格化後のデータ分布

新しい原点を中心として，主成分軸を回転させて

$$\frac{1}{n-1}X^tX \tag{3.18}$$

を最大とするθを探します．
これにより，データの座標の「**不偏分散が最も大きくなる軸**」が主成分軸となります．

3.4 Rによる計算（データの正規化）

「学問塾」フォルダにある「**主成分分析_身体測定_身長_座高（規格化）.R**」ファイルを開いてください．

Rエディタの表示

```
x_身体測定 <- read.csv(“C:/Users/Furuhashi/Documents/主成分分析/身体測定_身長_座高（変形版）.csv")

plot(x_身体測定$身長,x_身体測定$座高,col="red", pch=1)

x_normalized <- scale(x_身体測定, apply(x_身体測定, 2, mean), apply(x_身体測定, 2, sd))
                                                  #列ごとに平均0，分散1に正規化

plot(x_normalized[,1],x_normalized[,2],col="red", pch=1)

XX <- as.matrix(x_normalized)                     # データフレームを行列へ変換

n <- nrow(XX)                                     # 行数のカウント

S <- (1/(n-1))*t(XX) %*% XX                       # S = X^t Xの計算

Eigen_value <- eigen(S)                           # Sの固有値計算
Eigen_value

abline(0, Eigen_value$vectors[2,1]/Eigen_value$vectors[1,1])  # 第1主成分軸の描画
```

```
x_normalized <- scale(x_身体測定, apply(x_身体測定, 2,
                        mean), apply(x_身体測定, 2, sd))
```

により，身体測定のデータが列ごと（身長，座高ごと）に規格化されます．meanが平均，sdが不偏標準偏差(standard deviation, $=\sqrt{v_e^2}$)

```
plot(x_normalized[,1],x_normalized[,2],col="red“,  pch=1)
```

を実行すると図3.10のデータ分布が得られます．

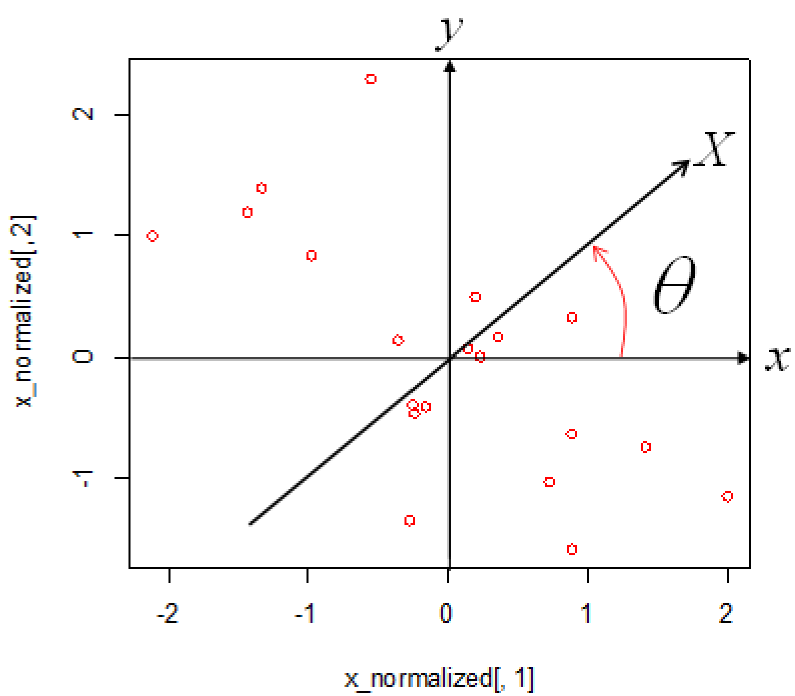


図3.10 規格化後のデータ分布

規格化以降の主成分分析の計算は3.2項と全く同じです．図3.11の主成分軸が得られます．

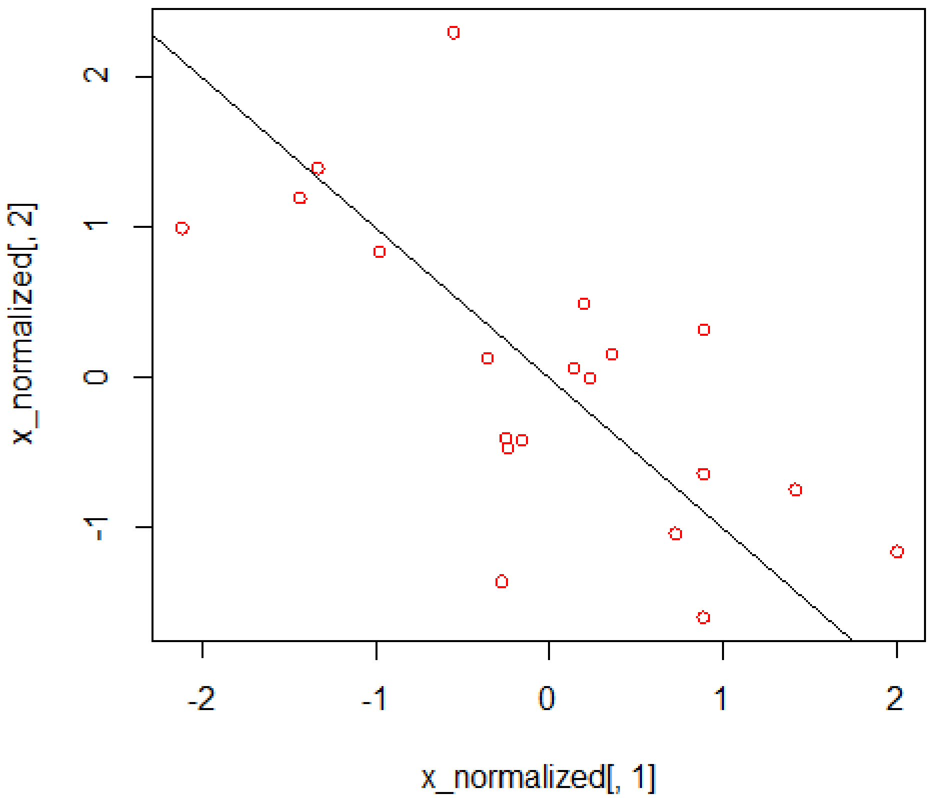


図3.11 規格化後のデータに対する主成分分析結果

3.5 多次元化

これまでのように2次元データに対して主成分分析を適用して得られた結果は，目で見て分かっていた結果とほとんど変わりません．

主成分分析は多次元データに適用されたときに威力を発揮します．主成分分析は，多次元空間の中で，データの分散の最も大きな軸を見つけ出すことができます．x_1〜x_4の4変数からなるデータP_i (x_{1i}, x_{2i}, x_{3i}, x_{4i})を想定した場合，各データの主成分軸上の座標は

$$
\begin{aligned}
X_1 &= x_{11}l_1 + x_{21}l_2 + x_{31}l_3 + x_{41}l_4 \\
X_2 &= x_{12}l_1 + x_{22}l_2 + x_{32}l_3 + x_{42}l_4 \\
&\vdots \\
X_n &= x_{1n}l_1 + x_{2n}l_2 + x_{3n}l_3 + x_{4n}l_4
\end{aligned}
\qquad (3.19)
$$

となります．

行列・ベクトルの表式に置き換えると

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & x_{3n} & x_{4n} \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \end{pmatrix} \tag{3.20}$$

となります．ここで

$$\begin{gathered} Z = \begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & x_{3n} & x_{4n} \end{pmatrix} \\ \boldsymbol{l} = \begin{pmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \end{pmatrix} \end{gathered} \tag{3.21}$$

とおき，

$S = \frac{1}{n-1} Z^t Z$ とすると，3.1項と同様にして

$$Sl = \lambda l \qquad (3.22)$$

を得ることができます．

大きな固有値から順に$\lambda_1, \lambda_2, \ldots$ とし，対応する固有ベクトルを$\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots$ とします．$\boldsymbol{l}_1$が第1主成分軸，$\boldsymbol{l}_2$が第2主成分軸と呼ばれます．

第i主成分軸上の座標は(3.9)式より

$$\boldsymbol{Z}\boldsymbol{l}_i \qquad (3.23)$$

により求めることができます．これは**主成分得点**と呼ばれます．

主成分軸への元の軸の「貢献度合い」を

$$\sqrt{\lambda}\boldsymbol{l}_i \qquad (3.24)$$

により定義します．これは**主成分負荷量**と呼ばれます．

3.6 Rによる計算(4次元データ)

「主成分分析」フォルダにある「**主成分分析_身体測定_4変数.R**」ファイルを開いてください．表2.1の身体測定の(4変数の)データを読み込んで主成分分析を行うスクリプトです．

表 2.1 身体測定結果

身長	体重	胸囲	座高
170.6	57	83	88.1
164.5	58.5	83.5	88.5
161	63	86	87.2
170.5	60.5	84	88.5
171	62	84	88.4
170	61.5	83	91.8
165	63.3	89.5	89.7
173	65	84	91
166.8	70	93.5	86.2
173.6	62.5	88.2	92
176.3	63.5	83.5	93
172.5	66	91.4	91.4
182	67.5	85.1	93.7
179	63	87	96.3
176.3	64	87	97
175.5	69.5	92.5	94.5
169	77.5	94.5	95.4
170.4	79	98	92.5
176.3	73.5	102.8	91
172.8	88	101	94.1

主成分分析_身体測定_4変数.R（その1）

```
par(mfrow=c(1,2))                                            # グラフ表示画面を2×2分割

x_身体測定 <- read.csv(“C:/Users/Furuhashi/Documents/主成分分析/身体測定_4変数.csv")

plot(x_身体測定$身長,x_身体測定$体重,col="red", pch=1)
plot(x_身体測定$胸囲,x_身体測定$座高,col="red", pch=1)

x_normalized <- scale(x_身体測定, apply(x_身体測定, 2, mean), apply(x_身体測定, 2, sd))
                                                             #列ごとに平均0，分散1に正規化

XX <- as.matrix(x_normalized)                                # データフレームを行列へ変換

n <- nrow(XX)                                                # 行数のカウント

S <- (1/(n-1))*t(XX) %*% XX                                  # S = X^t Xの計算

Eigen_value <- eigen(S)                                      # Sの固有値計算
```

主成分分析_身体測定_4変数.R（その2）

```
princip_score <- XX %*% Eigen_value$vectors                 #主成分得点の計算

princip_score_data = as.data.frame(princip_score)           #行列をデータフレームに変換

plot(princip_score_data$V1, princip_score_data$V2,col="red", pch=1)       #主成分得点の描画

Root_mean_eigen <- sqrt(Eigen_value$values)      #  固有値の平方根

princip_loading <- cbind(Root_mean_eigen[1] * Eigen_value$vectors[,1],
                         Root_mean_eigen[2] * Eigen_value$vectors[,2])
                         # 第1，2主成分負荷量の計算

princip_loading_data = as.data.frame(princip_loading)                  #行列をデータフレームに変換

loading_label <- c(1:4)                  # プロットのマーカー指定  1: ○(身長), 2:△(体重), 3:+(胸囲), 4: × (座高)

princip_loading_data = data.frame(princip_loading_data, loading_label)

plot(princip_loading_data$V1, princip_loading_data$V2,col=princip_loading_data$loading_label,
          pch=princip_loading_data$loading_label, xlim=c(-0.9, 0))        #主成分負荷量の描画
```

Zl_i の計算

$\sqrt{\lambda_i}$ の計算

$\sqrt{\lambda_i}\, l_i$ の計算

1行目の

```
par(mfrow=c(1,2))
```

はグラフ表示画面を分割する関数です．

図3.12には身長ー体重間の関係，胸囲ー座高間の関係が並べて表示されています．ただし，このニつのグラフを眺めても**データ分布の特徴はよく分かりません**．

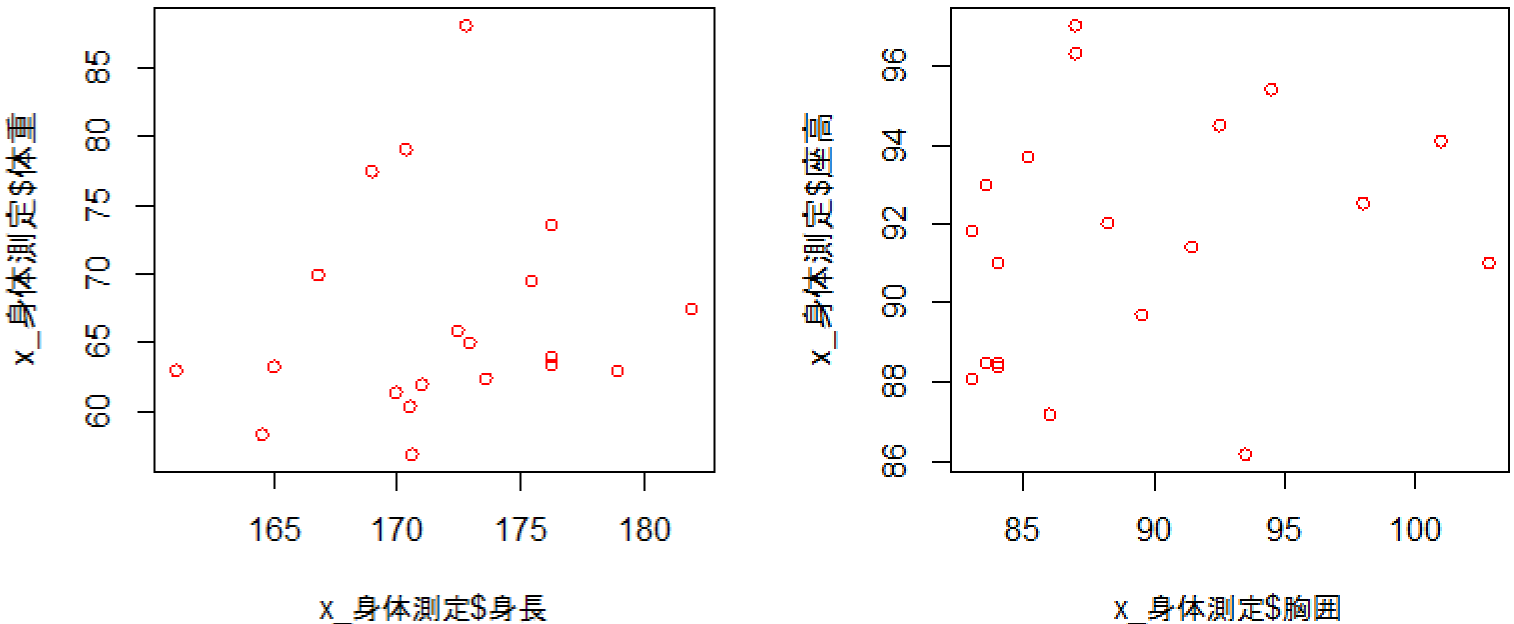


図3.12 身体測定データ（4変数）の表示

```
Root_mean_eigen <- sqrt(Eigen_value$values)
```

$\sqrt{\lambda_i}$ の計算

によりベクトル表現された固有値Eigen_value $valuesの各要素の平方根が求められ，Root_mean_　eigenに格納されます．

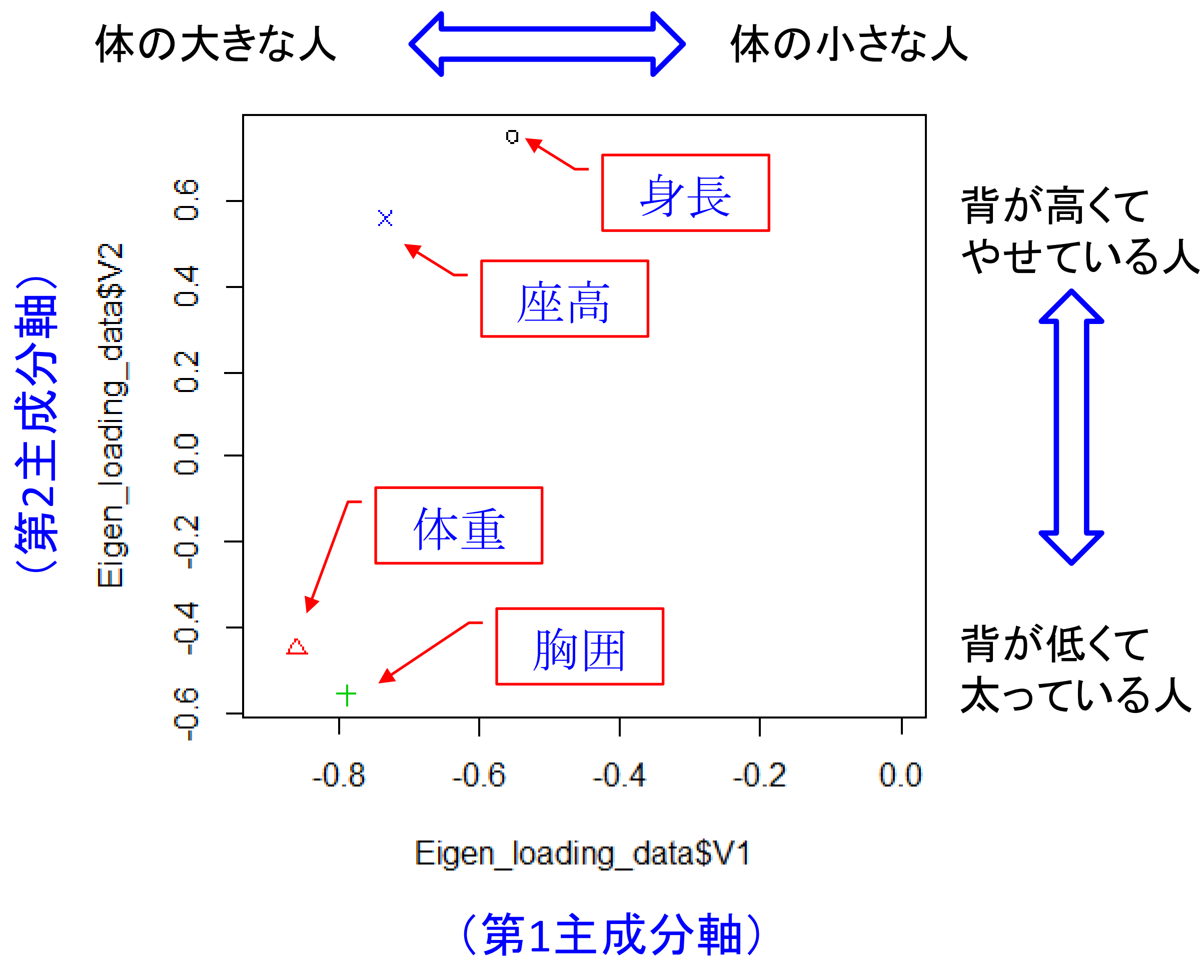
```
princip＿loading
<- cbind(Root_mean_eigen[1] *  princip_value$vectors[,1],
         Root_mean_eigen[2] * princip_value$vectors[,2])
```

$\sqrt{\lambda_i} l_i$ の計算

により，(3.24)式の第1，2**主成分負荷量**が計算されます．cbind(a, b)はベクトルa, bを結合して行列を生成する関数です．princip＿loadingは行列です．

```
plot(princip_loading_data$V1, princip_loading_data$V2,col=princip_loading_data$loading_label,
pch=princip_loading_data$loading_label, xlim=c(-0.9, 0))
```

主成分負荷量の描画



身体測定データの4次元空間内の分布においては，

体の大きさに関する分散が最も大きい（**第1主成分軸**），

（背が高くてやせている）ー（背が低くて太っている）ことに関する分散が2番目に大きい（**第2主成分軸**）

と捉えています．

図3.15 身体測定データ（4変数）の主成分負荷量の意味

```
princip_score <- XX %*% Eigen_value$vectors
```

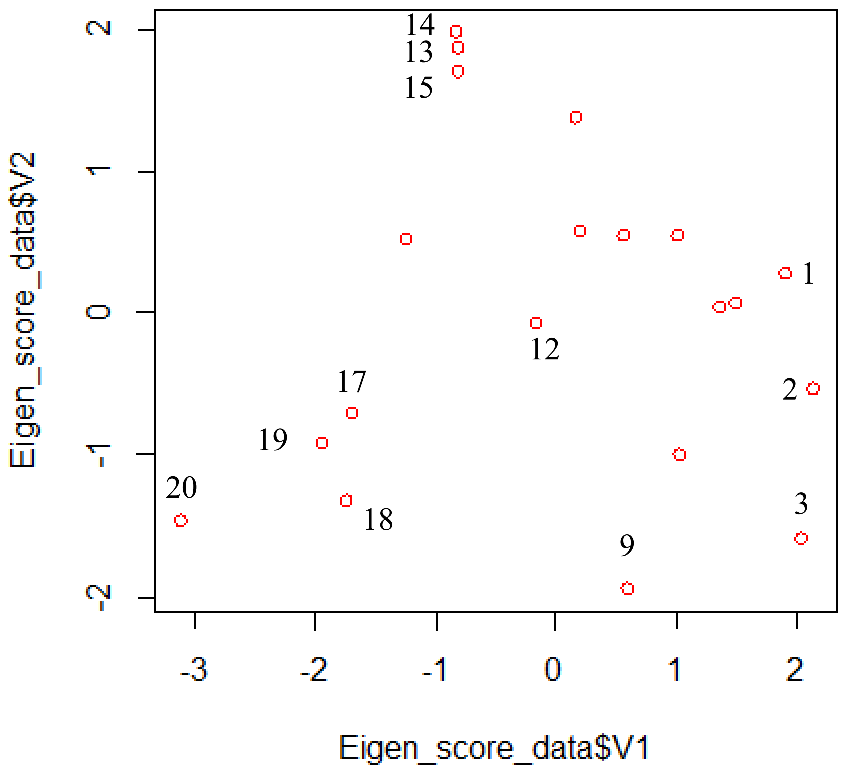
Zl_i の計算

```
plot(princip_score_data$V1, princip_score_data$V2,col="red", pch=1)
```

主成分得点の描画

体の大きな人 ⇔ 体の小さな人

（第2主成分軸）



（第1主成分軸）

背が高くてやせている人

背が低くて太っている人

i) 1～3番目の人たちは小柄である．ただし，3番目の人は若干背が低く太めである．
ii) 9番目の人は背が低く太めである．
iii) 12番目の人は平均的な体型の人である．
iv) 13～15番目の人たちは背が高くやせている．
v) 17～20番目の人たちは体が大きい．ただし若干背が低く太めである．

図3.16 身体測定データ(4変数)の主成分得点の意味

3.7 Rの組み込み関数(princomp())による計算

Rには便利な関数が用意されています．「主成分分析」フォルダにある「**主成分分析_身体測定_組込関数.R**」ファイルを開いてください．主成分分析は次の**princomp()関数**により実行できます．

```
PC <- princomp(XX, cor=TRUE)
```

XXには規格化前のデータ用いることができます．cor = TUREとすることで，(3.16)式の規格化がなされ，主成分分析が実行されます．

PCには固有ベクトル，固有値の平方根の情報が格納されています．固有ベクトルは

```
eigen_vectors <- unclass(loadings(PC))
```

によりeigen_vectorsに格納できます．また，固有値の平方根は

```
PC$sd
```

により取り出すことができます．

3.6項の主成分分析_身体測定_4変数.R

```
n <- nrow(XX)                                        # 行数のカウント
S <- (1/(n-1))*t(XX) %*% XX                          # S = X^t Xの計算
Eigen_value <- eigen(S)                              # Sの固有値計算
Root_mean_eigen <- sqrt(Eigen_value$values)          #  固有値の平方根
```

3.7項の主成分分析_身体測定_組込関数.R

```
PC <- princomp(XX, cor=TRUE)
eigen_vectors <- unclass(loadings(PC))
```

4行分を
2行にで
きました

おわりに

主成分分析について解説しました．Rのノウハウ書としないために基礎理論を述べ，その理論展開に沿ったRの計算例を紹介しました．princomp()関数を利用する方が実践的ではありますが，**理論を理解してこそ**，これらの関数を使いこなせることと思います．

なお，本スライドの内容の詳細は

「[多変量解析の基礎II（主成分分析）[kindle版]]」

にまとめて，Amasonより出版しています．