

# Rで主成分分析/因子分析

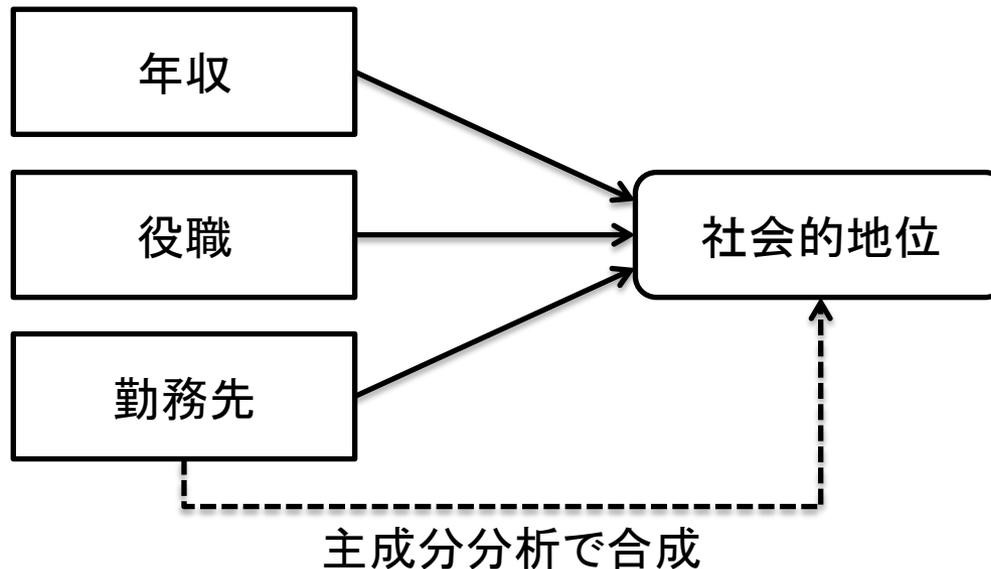
# 主成分分析と因子分析

# 主成分分析 (PCA)

- 多数の変数で説明されるデータ
  - 変数を合成
  - より少ない変数 (=主成分) でデータを説明

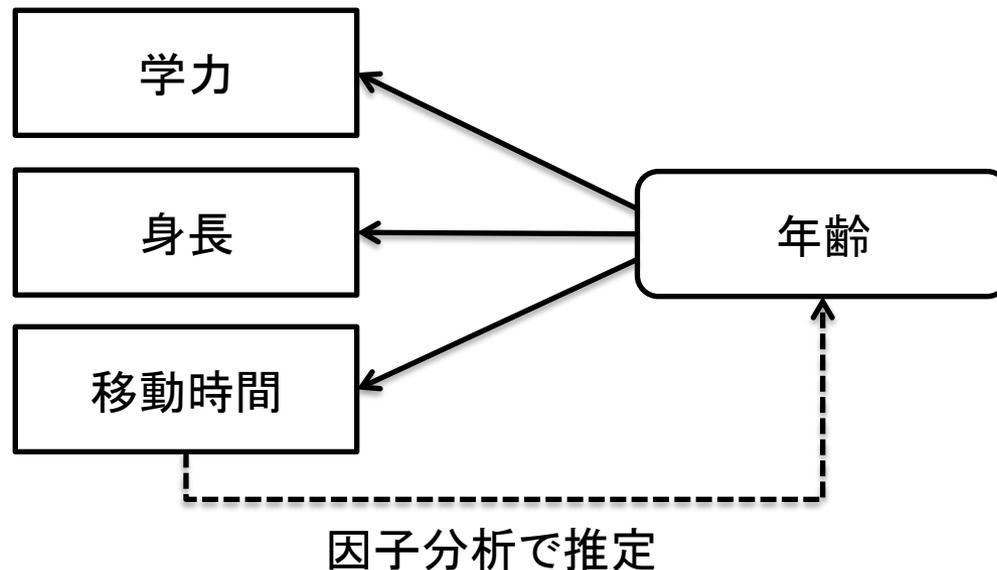
例1) 身長 + 体重 → 身体の高さ

例2) 年収 + 役職 + 勤務先 → 社会的地位



# 因子分析 (FA)

- 多数の変数で説明されるデータ
    - 共通因子を抽出
    - より少ない因子でデータを説明
- 例1) 学力・身長・移動時間 ← 年齢
- 例2) 商品売上・大気汚染度 ← 人口



# 主成分分析と因子分析の違い

- 因果関係 → 向きが違う
- 因子数：自動 (PCA) ⇔ 予め指定 (FA)
- 誤差：考慮しない (PCA) ⇔ 独自因子 (FA)

準備

## [演習] データを用意する

- Excelで以下のようなデータを入力

	A	B	C	D	E	F
1	Name	Math	Sci	Lang	Eng	Soc
2	Tanaka	89	90	67	46	50
3	Sato	57	70	80	85	90
4	Suzuki	80	90	35	40	50
5	Honda	40	60	50	45	55
6	Kawabata	78	85	45	55	60
7	Yoshino	55	65	80	75	85
8	Saito	90	85	88	92	95

→ CSV形式で保存（保存時に形式を指定）

※ファイル名や保存場所に日本語が含まれない方が良い

# CSV形式

- CSV=Comma Separated Values  
→ コンマ区切りの値
- コンマと改行で区切られたデータ
- テキストファイル

# Rで主成分分析

# Rの起動と終了

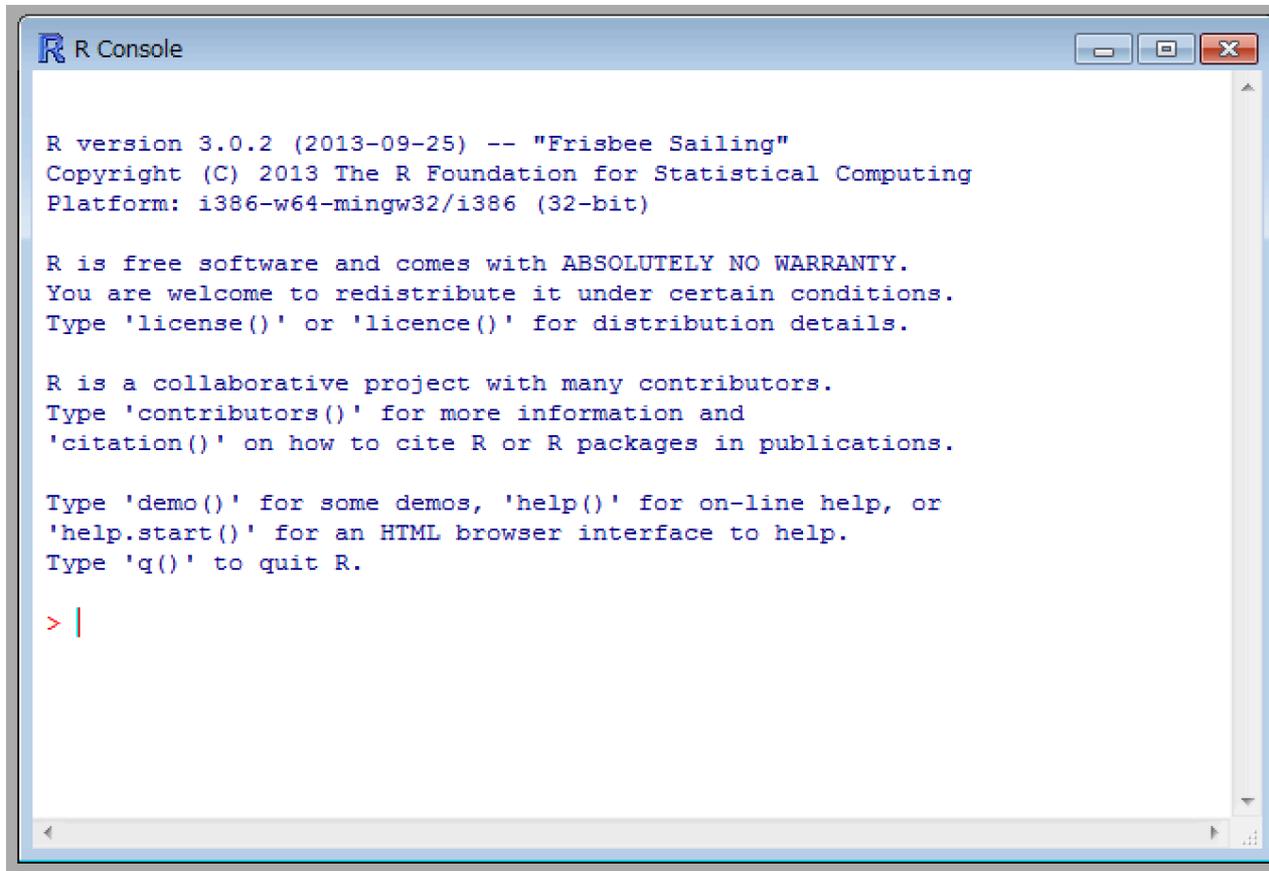
- Rの起動（GUI版）  
→ Rのアイコンをダブルクリック



- Rの終了  
→ 普通に終了  
→ 「Save workspace image?」と聞かれる  
※ 可能なら「はい」を選択  
→ 実行履歴が保存される

# Rの基本

- Rはコマンドラインのツール  
→ コンソールで命令をキーボード入力する



```
R Console

R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

# [演習] データの読み込みと確認

R Console

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

```
> DAT <- read.table('C:\\Users\\mura0\\Desktop\\test.csv', header=TRUE, sep=",")  
> DAT
```

	Name	Math	Sci	Lang	Eng	Soc
1	Tanaka	89	90	67	46	50
2	Sato	57	70	80	85	90
3	Suzuki	80	90	35	40	50
4	Honda	40	60	50	45	55
5	Kawabata	78	85	45	55	60
6	Yoshino	55	65	80	75	85
7	Saito	90	85	88	92	95

```
> |
```

# CSVファイルの読み込み

- 下記を入力

```
> DAT <- read.table('...', header=TRUE, sep=",")
```

(入力 は 青字)

CSVファイルのパスを指定

1行目をヘッダ行とするかどうか

区切りはコンマ

- 代わりに下記の命令でもOK

```
> DAT <- read.csv('...')
```

# ファイルのパス

- Windowsの場合

```
C:\\Users\\114567c\\Documents\\hoge.csv
```

区切りは \ (バックスラッシュ) または ¥ (円記号) を**2個**

- Macの場合

```
/Users/114567c/Documents/hoge.csv
```

区切りは / (スラッシュ)

# 結果の代入と表示

- 代入

```
> DAT <- ...
```

変数DATに結果を入力

- 変数の表示

```
> DAT
```

# 部分行列の指定

- 部分行の指定

> DAT[1, ] ← 1行目を表示

> DAT[2:4, ] ← 2~4行目を表示

- 部分列の指定

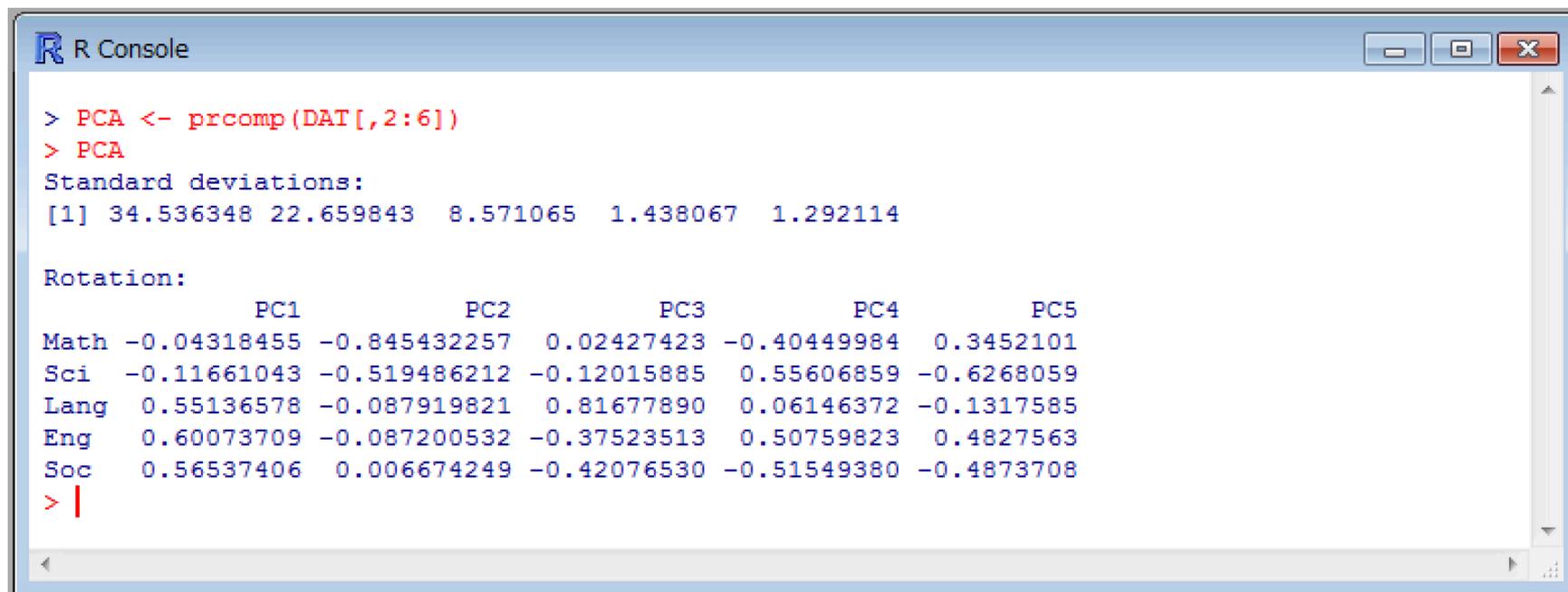
> DAT[, 1] ← 1列目を表示

> DAT[, 2:6] ← 2~4列目を表示

# [演習] 数値部分を主成分分析する

```
> PCA <- prcomp(DAT[,2:6])
```

## 【実行例】

A screenshot of an R Console window. The window title is "R Console". The console shows the execution of the command `> PCA <- prcomp(DAT[,2:6])` followed by `> PCA`. The output displays the standard deviations of the principal components and a rotation matrix. The standard deviations are: [1] 34.536348 22.659843 8.571065 1.438067 1.292114. The rotation matrix has columns labeled PC1 through PC5 and rows labeled Math, Sci, Lang, Eng, and Soc. The console ends with a red prompt character `> |`.

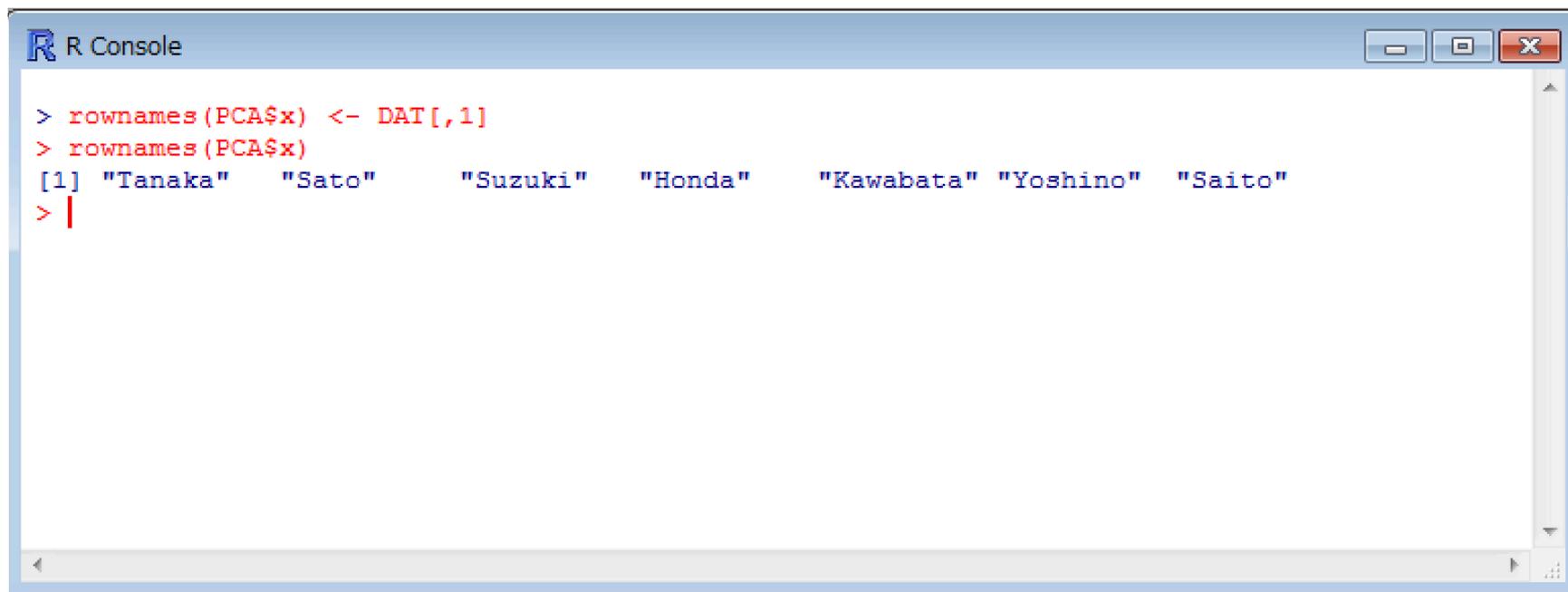
```
> PCA <- prcomp(DAT[,2:6])
> PCA
Standard deviations:
[1] 34.536348 22.659843 8.571065 1.438067 1.292114

Rotation:
      PC1      PC2      PC3      PC4      PC5
Math -0.04318455 -0.845432257 0.02427423 -0.40449984 0.3452101
Sci  -0.11661043 -0.519486212 -0.12015885 0.55606859 -0.6268059
Lang 0.55136578 -0.087919821 0.81677890 0.06146372 -0.1317585
Eng  0.60073709 -0.087200532 -0.37523513 0.50759823 0.4827563
Soc  0.56537406 0.006674249 -0.42076530 -0.51549380 -0.4873708
> |
```

## [演習] 結果にラベルをつける

```
> rownames(PCA$x) <- DAT[,1]
```

### 【実行例】



```
R Console  
> rownames(PCA$x) <- DAT[,1]  
> rownames(PCA$x)  
[1] "Tanaka" "Sato" "Suzuki" "Honda" "Kawabata" "Yoshino" "Saito"  
> |
```

# [演習] 寄与率のチェック

```
> summary(PCA)
```

## 【実行例】

R Console

```
> summary(PCA)
```

```
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5
Standard deviation	34.5363	22.6598	8.57106	1.43807	1.29211
Proportion of Variance	0.6688	0.2879	0.04119	0.00116	0.00094
Cumulative Proportion	0.6688	0.9567	0.99790	0.99906	1.00000

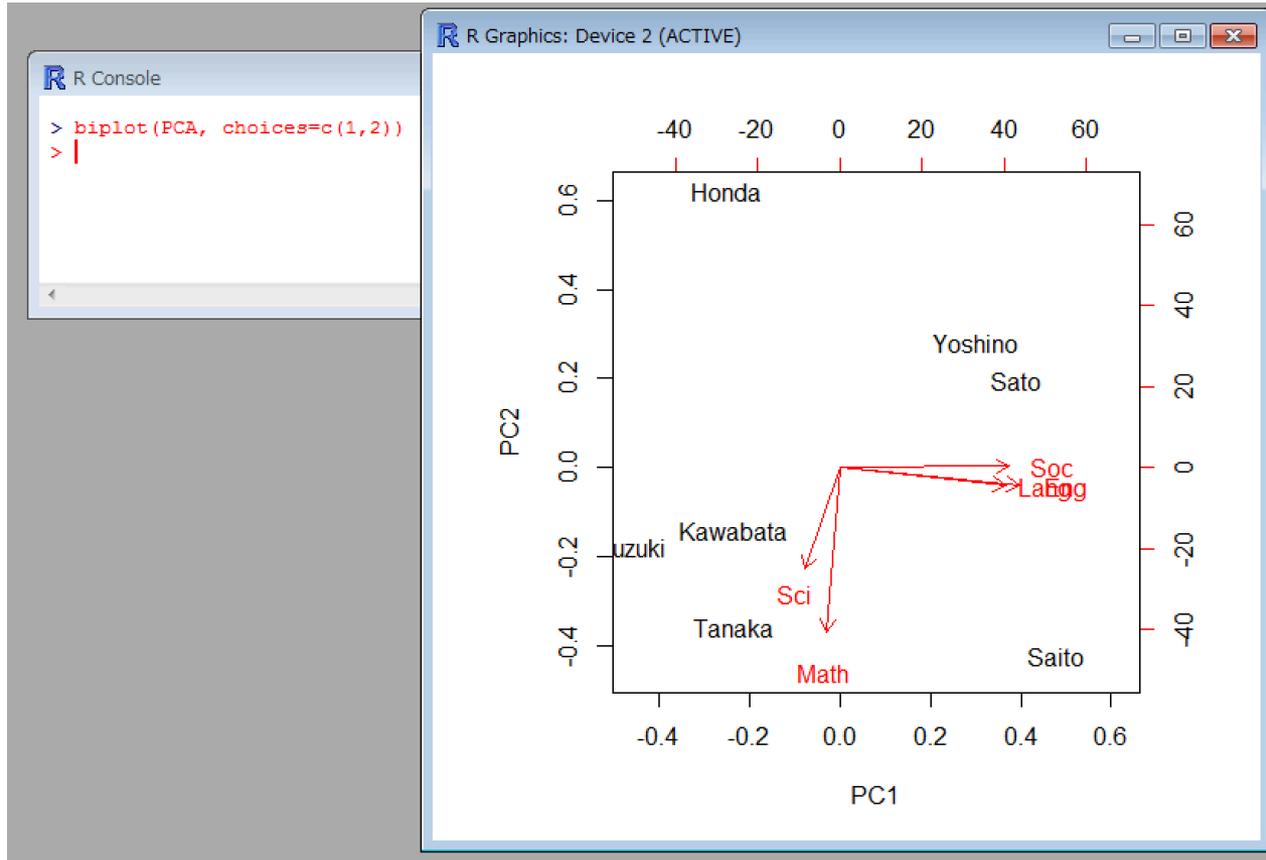
```
> |
```

累積寄与率=いくつの主成分でデータが正しく表現できるか

# [演習] プロット

```
> biplot(PCA, choices=c(1,2)) ← 主成分1と2をプロット
```

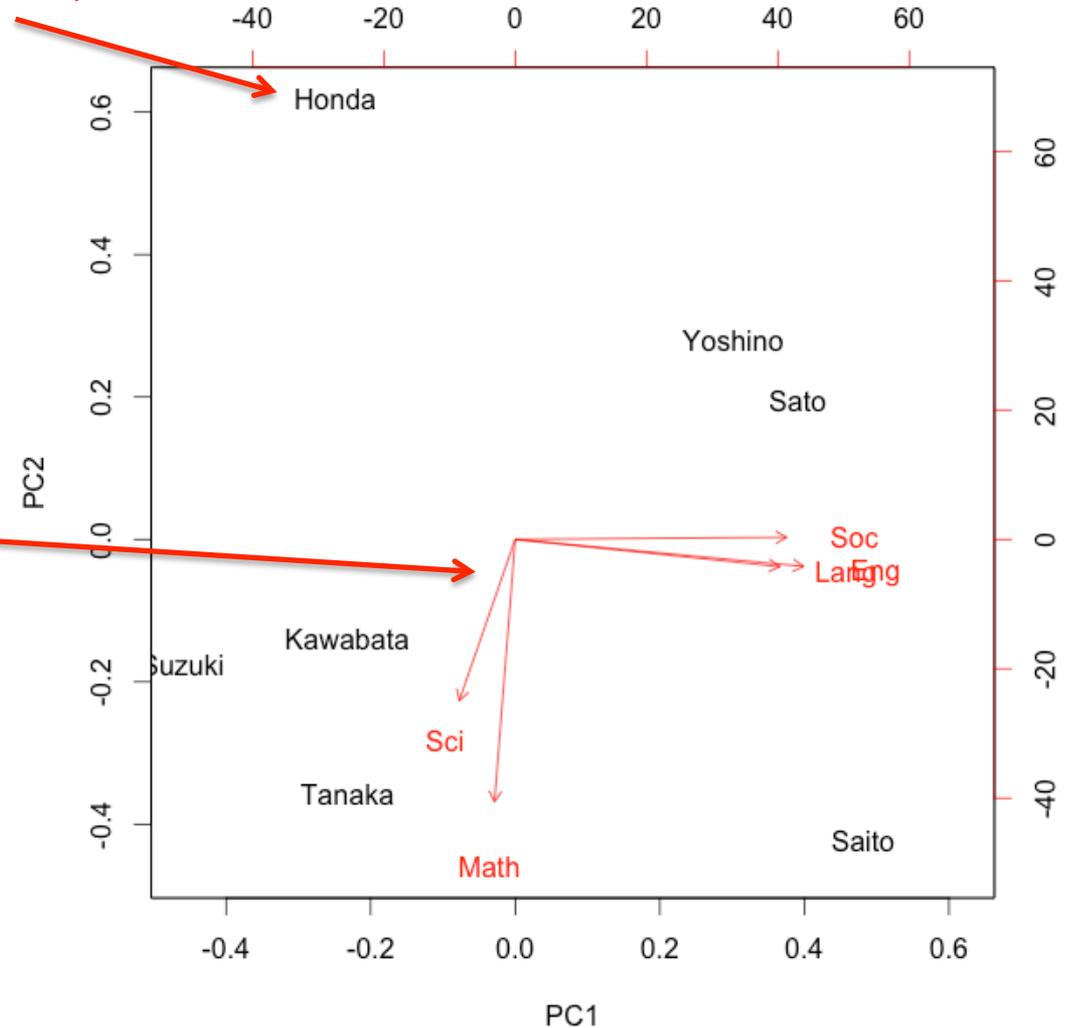
## 【実行例】



# biplotの結果

主成分空間にプロットされたデータ

各変数と主成分の関係  
(= 因子負荷量)

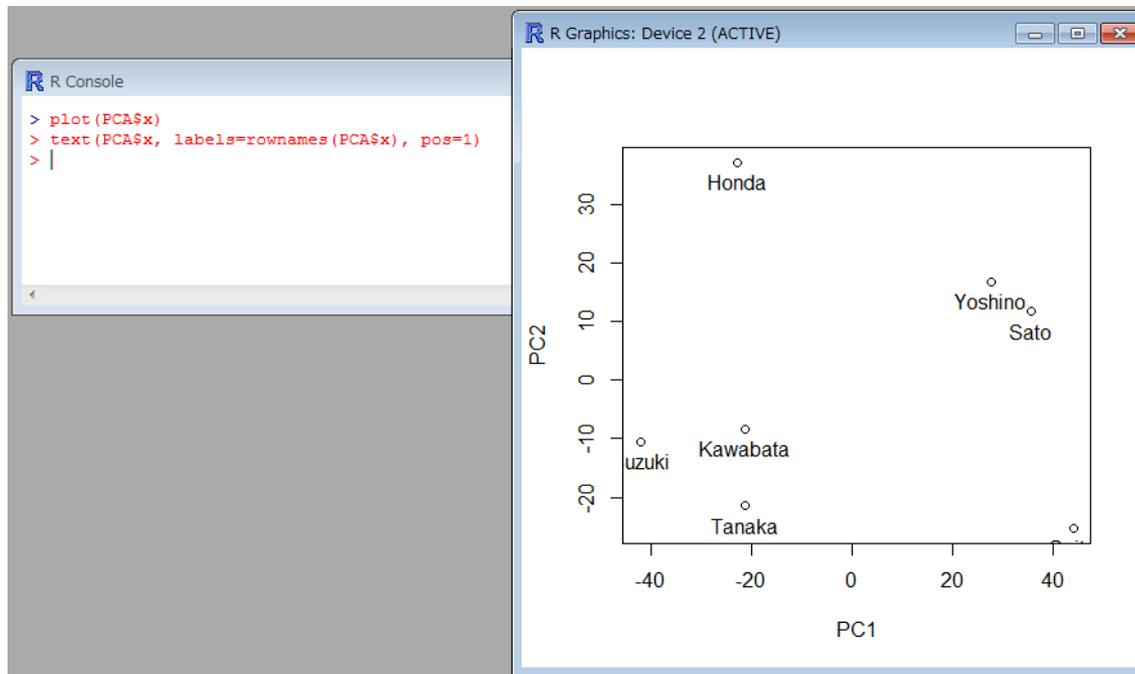


# 【演習】 データのみのプロット

```
> plot(PCA$x)  
> text(PCA$x, labels=rownames(PCA$x), pos=1)
```

ラベルの位置 = 点の下

## 【実行例】

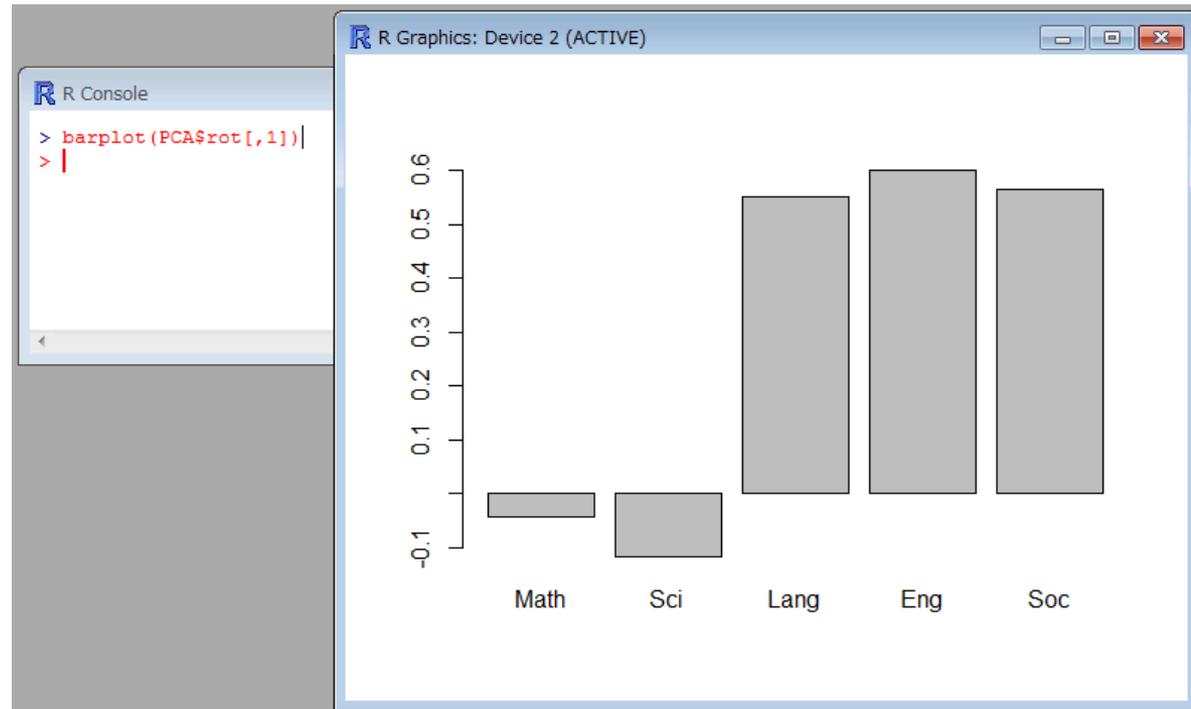


# [演習] 因子負荷量のプロット

```
> barplot(PCA$rot[,1])
```

第1因子の負荷量

【実行例】



# [演習] データの数を制限する場合

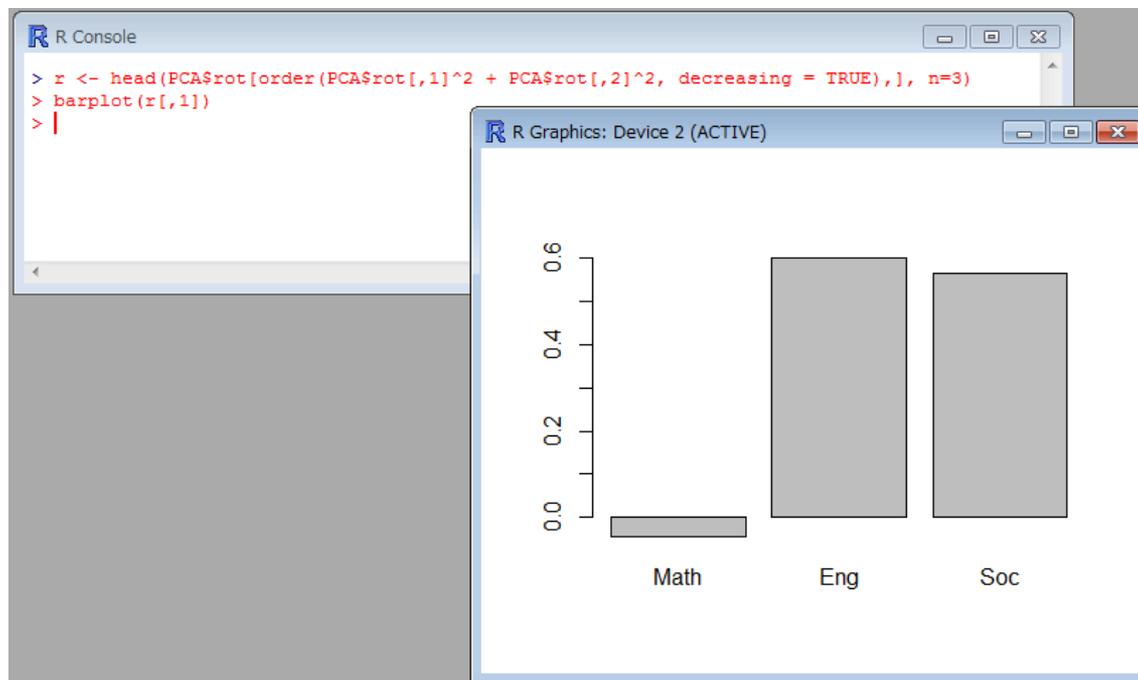
```
> r <- head(PCA$rot[order(PCA$rot[,1]^2 + PCA$rot[,2]^2, decreasing = TRUE),], n=3)  
> barplot(r[,1])
```

順番をつける大きさ

降順

個数

## 【実行例】



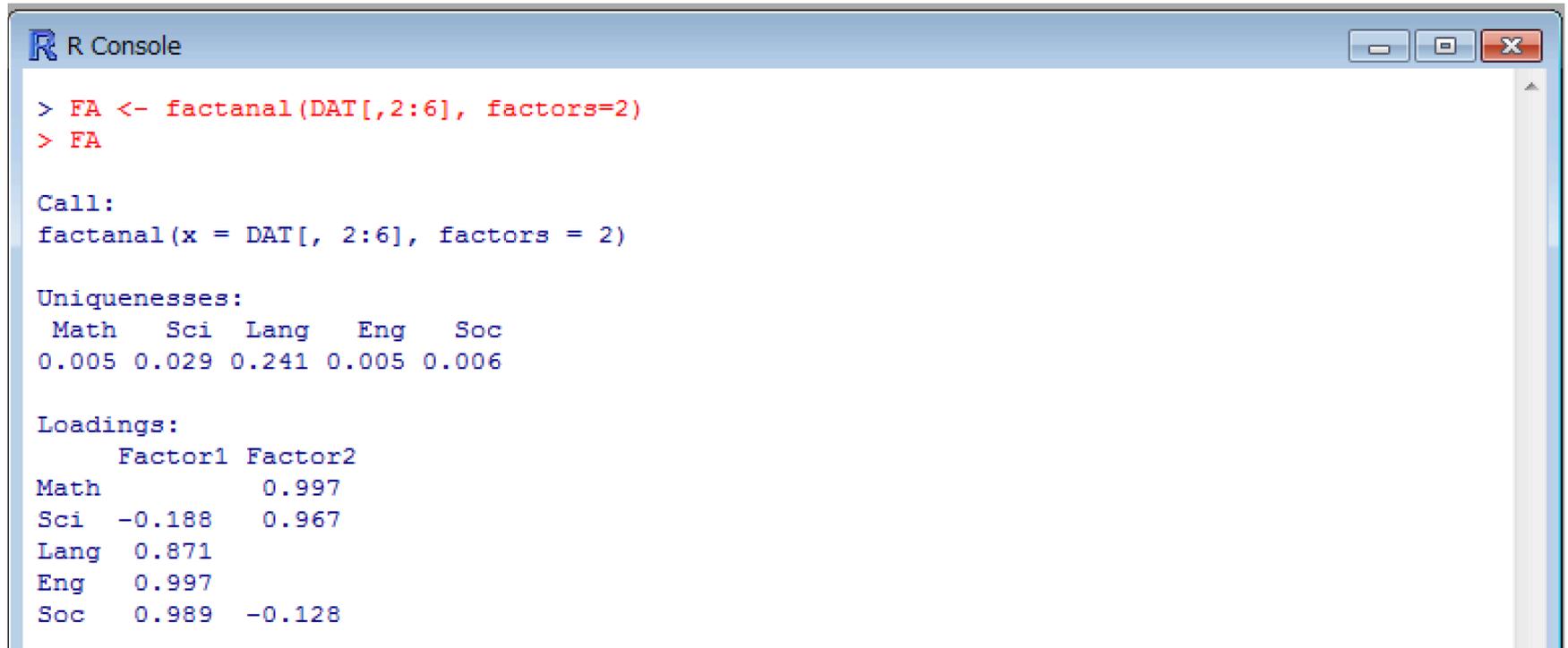
# Rで因子分析

# [演習] 数値部分を因子分析

```
> FA <- factanal(DAT[,2:6], factors=2)
```

因子数の指定

## 【実行例】



```
R Console  
> FA <- factanal(DAT[,2:6], factors=2)  
> FA  
  
Call:  
factanal(x = DAT[, 2:6], factors = 2)  
  
Uniquenesses:  
  Math   Sci  Lang   Eng   Soc  
0.005 0.029 0.241 0.005 0.006  
  
Loadings:  
      Factor1 Factor2  
Math          0.997  
Sci   -0.188   0.967  
Lang   0.871  
Eng    0.997  
Soc    0.989  -0.128
```

# 因子分析の結果

> FA

Uniquenesses:

Math	Sci	Lang	Eng	Soc
0.005	0.029	0.241	0.005	0.006

← 独自性 = 共通因子の乏しさ

Loadings:

	Factor1	Factor2
Math		0.997
Sci	-0.188	0.967
Lang	0.871	
Eng	0.997	
Soc	0.989	-0.128

← 因子負荷量 = 因子の変数への重み

← 寄与率

	Factor1	Factor2
SS loadings	2.768	1.946
Proportion Var	0.554	0.389
Cumulative Var	0.554	0.943

← 適合度 = p値が大きいと良い

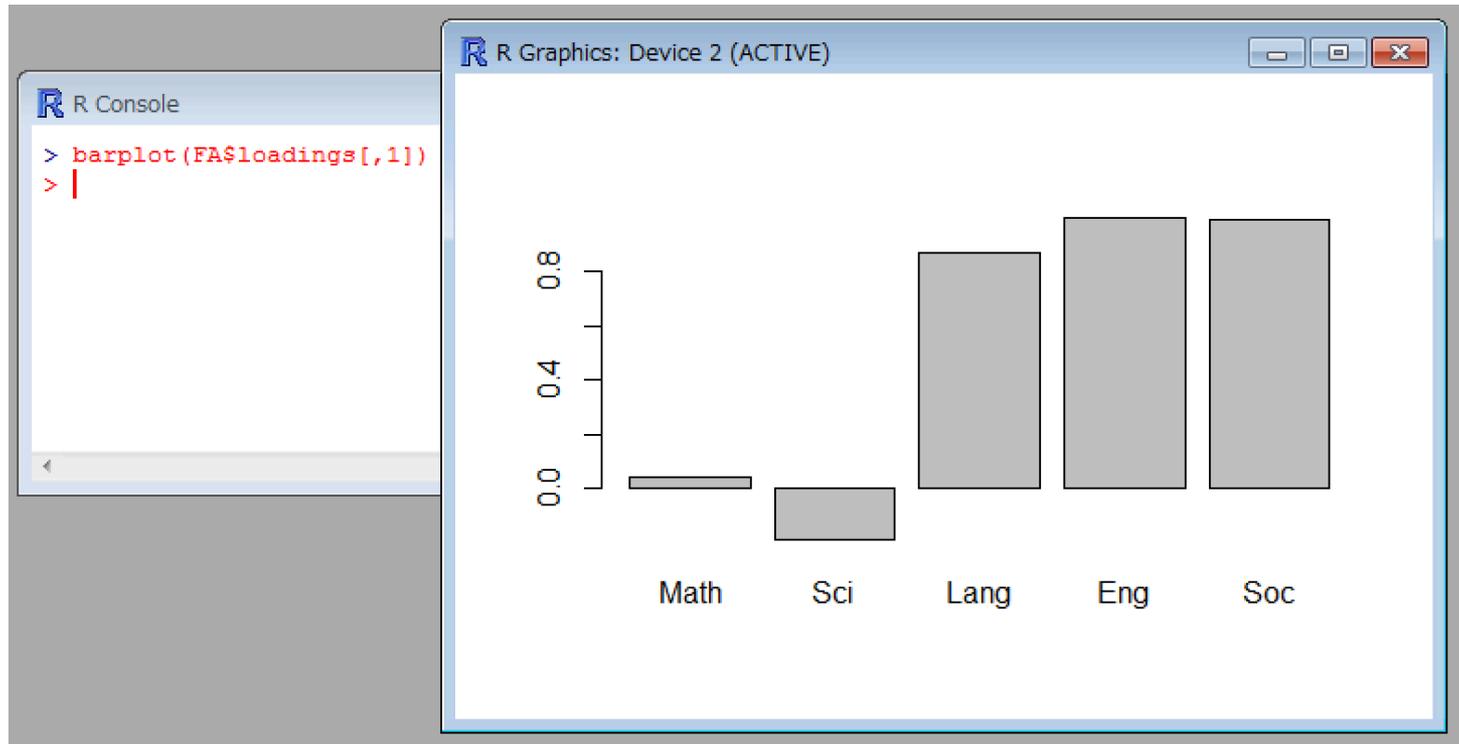
Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 1.73 on 1 degree of freedom.  
The p-value is 0.188

# [演習] 因子負荷量のプロット

```
> barplot(FA$loadings[,1])
```

第1因子の負荷量

【実行例】



# [演習] データのプロット

点を描画しない

```
> plot(FA$loadings, type="n")  
> text(FA$loadings, labels=DAT[,1])
```

## 【実行例】

