Rで主成分分析/因子分析

Rによるデータサイエンス

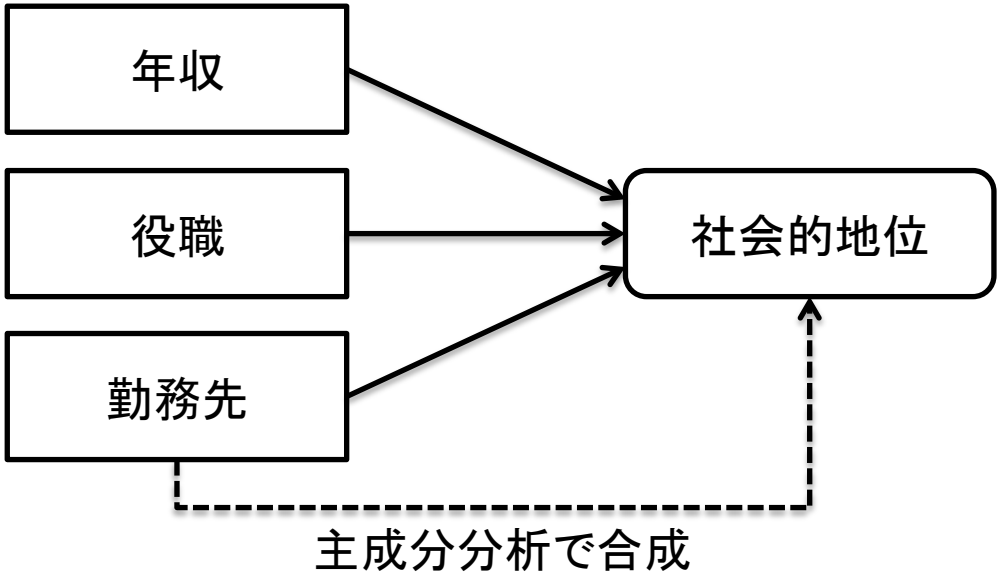
主成分分析と因子分析

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 → 身体の大きさ

例2）年収＋役職＋勤務先 → 社会的地位

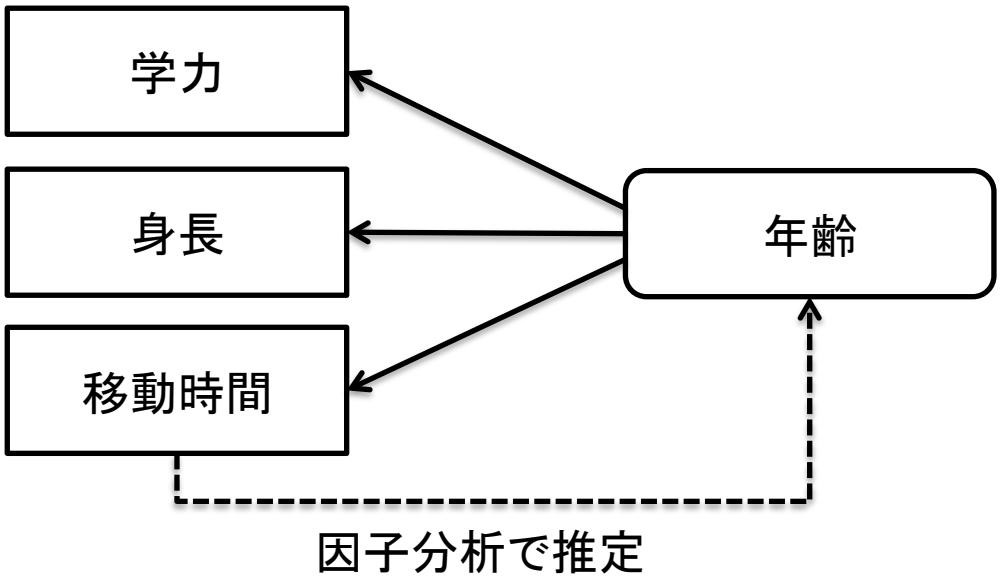


因子分析（FA）

- 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）

Rで主成分分析

［演習］準備：データを用意する

- Excelで以下のデータを入力（p.70表1.2）
  → 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

	A	B	C	D	E
1	A	B	C	D	E
2	50	57	74	94	112
3	57	50	57	74	94
4	74	57	50	57	74
5	94	74	57	50	57
6	112	94	74	57	50
7	128	112	94	74	57
8	140	128	112	94	74
9	147	140	128	112	94
10	150	147	140	128	112
11	147	150	147	140	128
12	140	147	150	147	140
13	128	140	147	150	147
14	112	128	140	147	150
15	94	112	128	140	147
16	74	94	112	128	140
17	57	74	94	112	128

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

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
絶対パス指定＝フォルダツリーのトップから指定

- 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）

※ 最近のRはMacと同じように区切りが / でも行ける･･･らしい

作業フォルダの表示と変更

- 現在の作業フォルダの表示

```
> getwd()
[1] "C:/Users/murao/Documents"
```

- 作業フォルダの変更

```
> setwd("~")
> getwd()
[1] "C:/Users/murao/Documents"
> setwd("~/../Desktop")
> getwd()
[1] "C:/Users/murao/Desktop"
```

［演習］CSVデータの読み込みと確認

- Excelで作成したCSVファイルを読み込んでみる

```
> DAT <- read.csv("~/../Desktop/circle.csv")
> DAT
     A   B   C   D   E
1   50  57  74  94 112
2   57  50  57  74  94
3   74  57  50  57  74
4   94  74  57  50  57
5  112  94  74  57  50
6  128 112  94  74  57
7  140 128 112  94  74
8  147 140 128 112  94
9  150 147 140 128 112
10 147 150 147 140 128
11 140 147 150 147 140
12 128 140 147 150 147
13 112 128 140 147 150
14  94 112 128 140 147
15  74  94 112 128 140
16  57  74  94 112 128
> 
```

オブジェクトの種類の確認

- class命令を利用

```
> v <- c(1,2,3,4,5)
> class(v)
[1] "numeric"
> m <- matrix(0,2,5)
> class(m)
[1] "matrix"
> c <- read.csv('…')
> class(c)
[1] "data.frame"
>
```

行名・列名の確認

- 列の名前

```
> colnames(DAT)
[1] "A" "B" "C" "D" "E"
>
```

- 行の名前

```
> rownames(DAT)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9" ...
>
```

部分行または部分列の指定

- 部分行の指定

```
> DAT[1,]
```
← 1行目を表示

```
> DAT[2:6,]
```
← 2～6行目を表示

- 部分列の指定

```
> DAT[,1]
```
← 1列目を表示

```
> DAT[,2:4]
```
← 2～4列目を表示

[演習] 主成分分析する

```
> PCA <- princomp(DAT)
```

【実行例】

```
> PCA <- princomp(DAT)
> PCA
Call:
princomp(x = DAT)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
65.8425316 39.5193986  3.7146160  1.2617966  0.2960714

 5  variables and  16 observations.
```

［演習］寄与率のチェック

```
> summary(PCA)
```

【実行例】

```
> summary(PCA)
Importance of components:
                            Comp.1     Comp.2      Comp.3       Comp.4       Comp.5
Standard deviation     65.8425316 39.5193986 3.714615982 1.2617966279 2.960714e-01
Proportion of Variance  0.7332328  0.2641493 0.002333763 0.0002692822 1.482593e-05
Cumulative Proportion   0.7332328  0.9973821 0.999715892 0.9999851741 1.000000e+00
>
```

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] 因子負荷量のチェック

```
> PCA$loadings
```

【実行例】

```
> PCA$loadings

Loadings:
  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
A -0.365  0.620  0.584 -0.340 -0.162
B -0.481  0.340 -0.164  0.620  0.492
C -0.521        -0.514         -0.681
D -0.481 -0.340 -0.164 -0.620  0.492
E -0.365 -0.620  0.584  0.340 -0.162
```

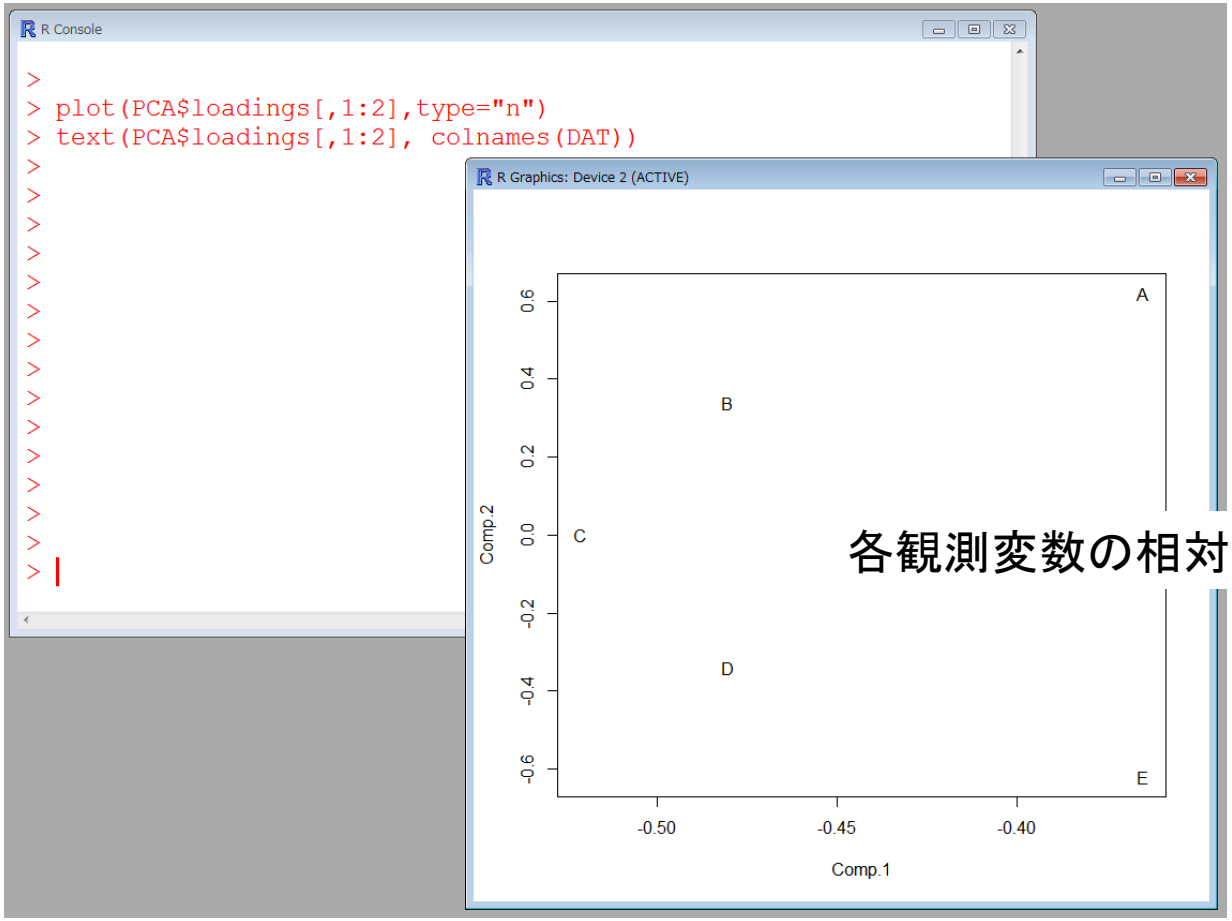
因子負荷量＝各観測変数の各主成分への影響力

→ 各観測変数の相対的な関係の指標

[演習] 因子負荷量のプロット

```
> plot(PCA$loadings[,1:2], type="n")
> text(PCA$loadings[,1:2], colnames(DAT))
```

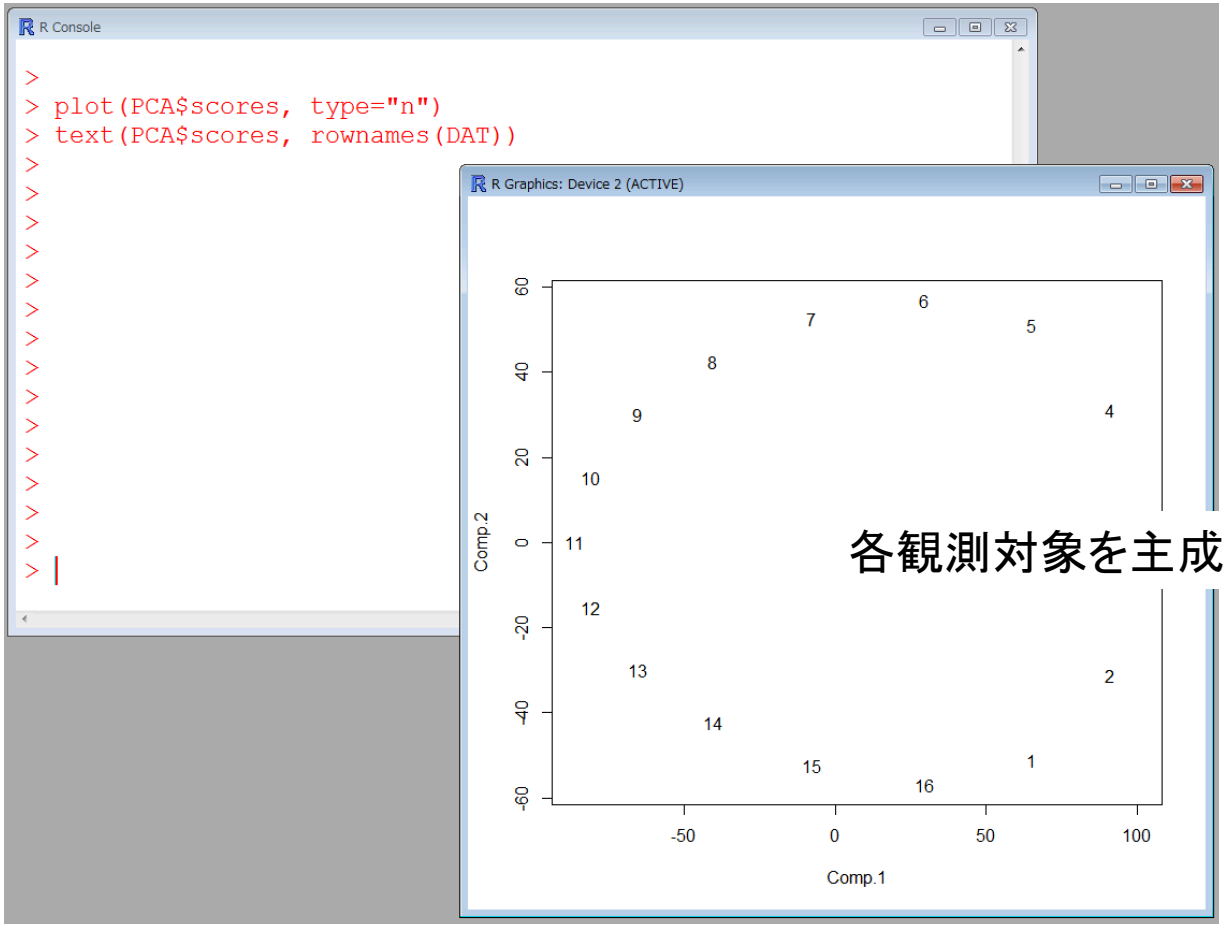
点を描画しない



各観測変数の相対的な関係が見える

[演習] 主成分得点のプロット

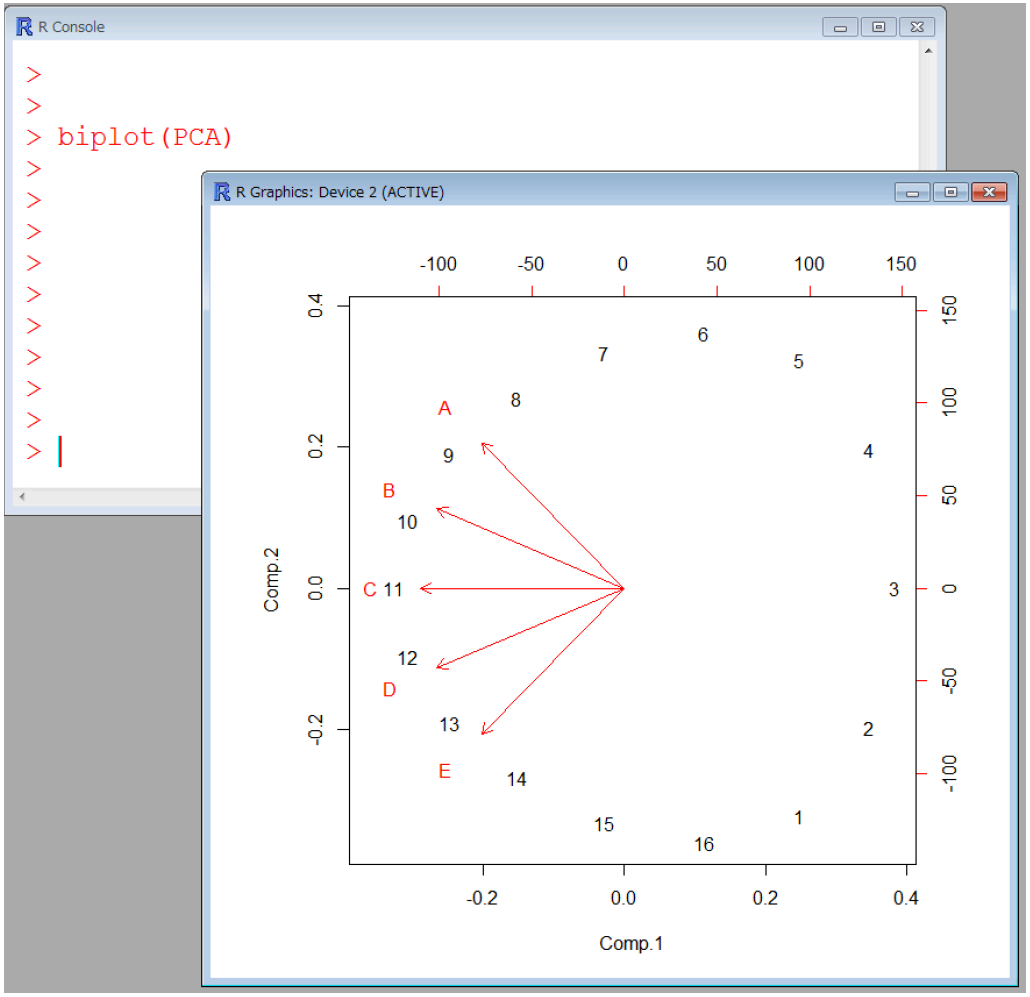
```
> plot(PCA$scores[,1:2], type="n")
> text(PCA$scores[,1:2], rownames(DAT))
```



各観測対象を主成分空間にプロット

［演習］因子負荷量＋主成分得点をプロット

```
> biplot(PCA)
```



［演習］データを用意する

- Excelで以下のデータを入力（p.78表2.1）

	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データの読み込みと確認

- Excelで作成したCSVファイルを読み込んでみる

```
> TMP <- read.csv('scores.csv')
> TMP
      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
>
```

[演習] CSVから読み込んだデータを変換

- 1列目がヘッダとして処理されないため

ncol(TMP) ＝ TMPの列数

```
> DAT <- TMP[,2:ncol(TMP)]
> colnames(DAT)
[1] "Math" "Sci"  "Lang" "Eng"  "Soc"
> rownames(DAT)
[1] "1" "2" "3" "4" "5" "6" "7"
> rownames(DAT) <- TMP[,1]
> rownames(DAT)
[1] "Tanaka"   "Sato"     "Suzuki"   "Honda"    "Kawabata" "Yoshino"
[7] "Saito"
> DAT
         Math Sci Lang Eng Soc
Tanaka     89  90   67  46  50
Sato       57  70   80  85  90
Suzuki     80  90   35  40  50
Honda      40  60   50  45  55
Kawabata   78  85   45  55  60
Yoshino    55  65   80  75  85
Saito      90  85   88  92  95
>
```

TMPの2列目以降をDATにコピー

行名は正しく処理されている

列名は読み込まれていない

列名を設定

［練習］主成分分析で分析してみよう

- 読み込んだテストの点数データを分析してみる

Rで因子分析

［演習］因子分析

```
> FA <- factanal(DAT, factors=2)
```

因子数の指定

【実行例】

R Console

```
>
> FA <- factanal(DAT, factors=2)
> FA

Call:
factanal(x = DAT, factors = 2)

Uniquenesses:
 Math   Sci  Lang   Eng   Soc
0.005 0.029 0.241 0.005 0.006
```

因子分析の結果

```
> FA

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

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

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
```

独自性＝共通因子の乏しさ

因子負荷量＝因子の変数への重み

寄与率

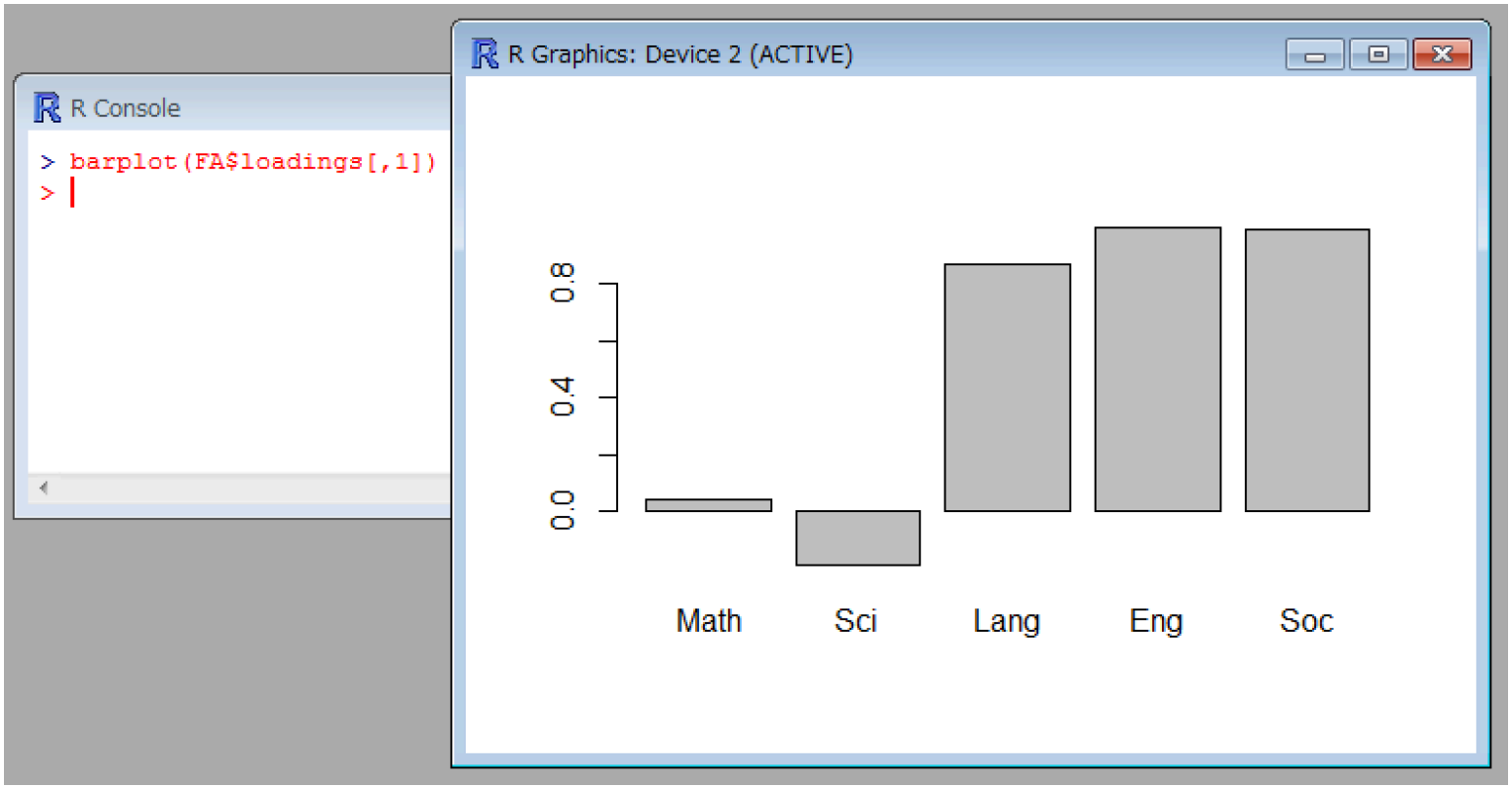
適合度＝p値が大きいと良い

［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】

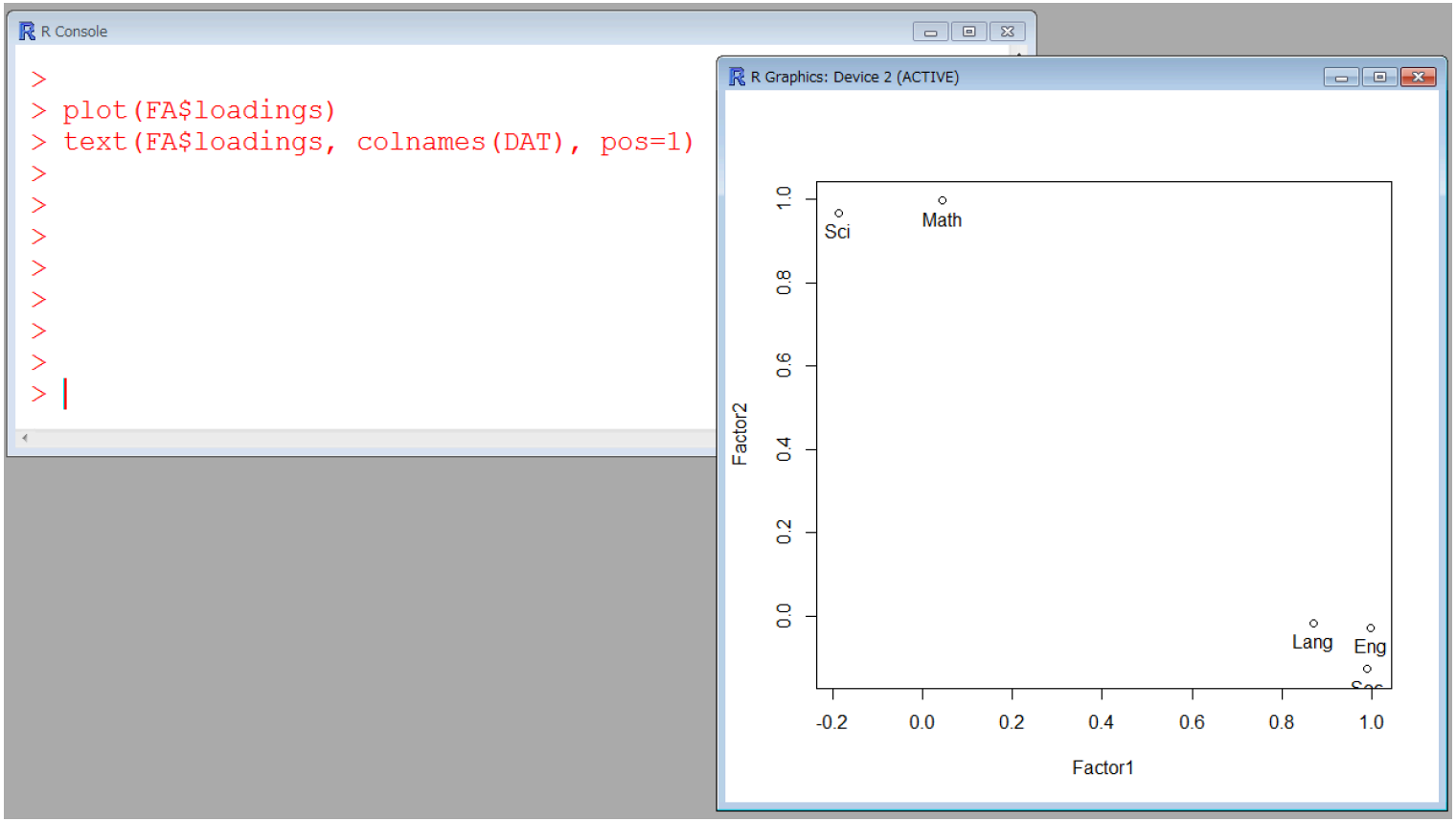


［演習］因子負荷量のプロット

```
> plot(FA$loadings)
> text(FA$loadings, colnames(DAT), pos=1)
```

ラベルの位置

【実行例】

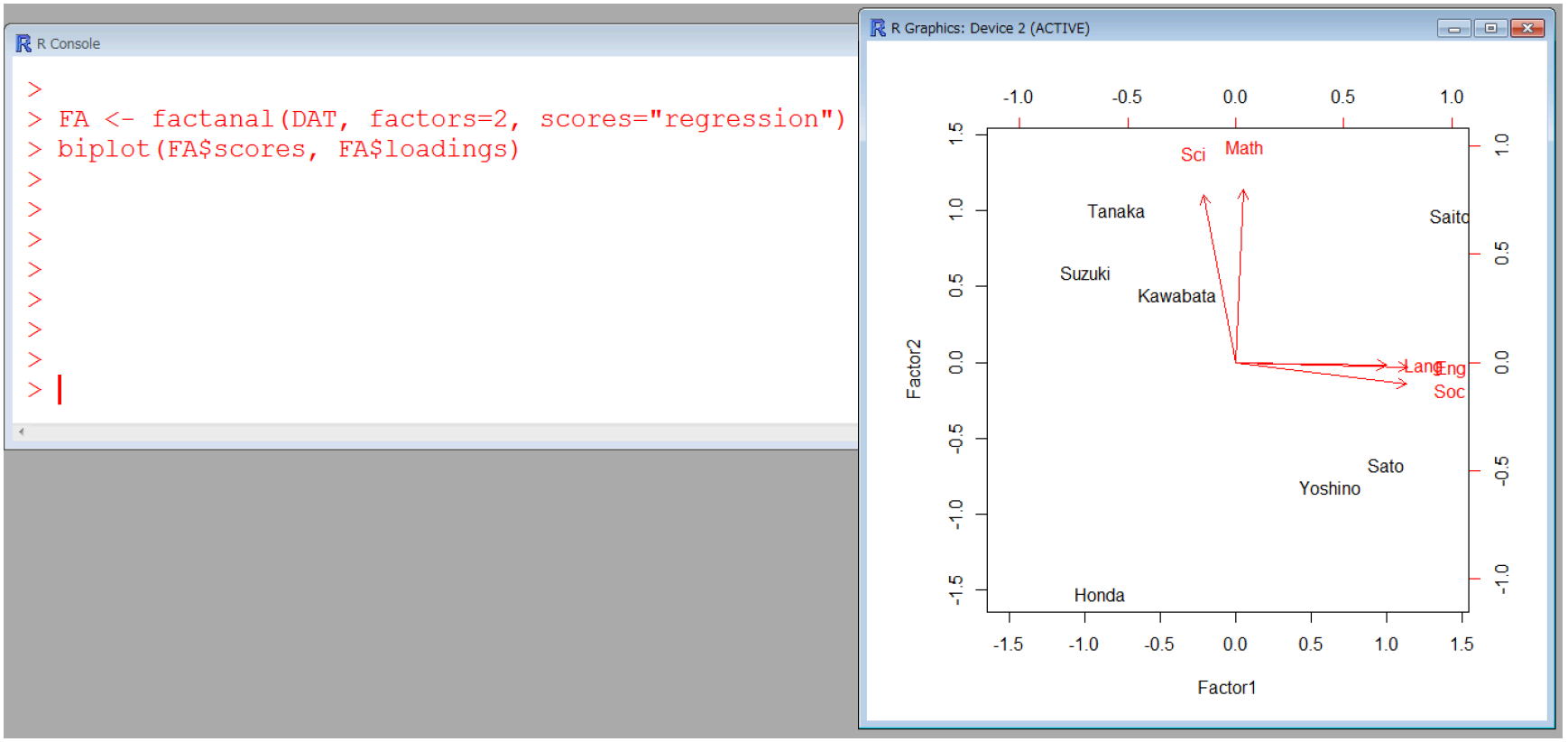


［演習］因子負荷量＋因子得点のプロット

```
> FA <- factanal(DAT, factors=2, scores="regression")
> biplot(FA$scores, FA$loadings)
```

因子得点を計算する

【実行例】



[練習] 以下のデータを分析してみよう

- 市販のお茶についての印象アンケートの結果

Name	Formal	Revolutionary	Modern	Unique	Attractive	Stylish
Suntory Oolong Tea	1.91	4.4	4.36	4.33	4.46	4.24
Ito-enn O-i Ocha	3.3	3.71	3.59	3.76	3.79	3.91
Kirin Kiki-cha	3.39	1.84	2.23	1.69	1.91	1.8
Coca-cola Soukenbicha	3.66	3.21	2.81	3.14	3.31	3.16
Kirin Nama-cha	3.46	2.83	2.46	2.91	3.09	2.87