主成分分析

内容

- 主成分分析
  - 主成分分析について
  - 成績データの解析
  - 「R」で主成分分析
  - 相関行列による主成分分析
  - 寄与率・累積寄与率
  - 因子負荷量
  - 主成分得点

主成分分析

次元の縮小と主成分分析

主成分分析
- 次元の縮小に関する手法

- 次元の縮小
  - 国語、数学、理科、社会、英語の総合点
    ⇒5次元データから1次元データへの縮約
  - 体形評価：BMI (Body Mass Index) 判定
    肥満度の判定方法の1つで、次の式で得られる。

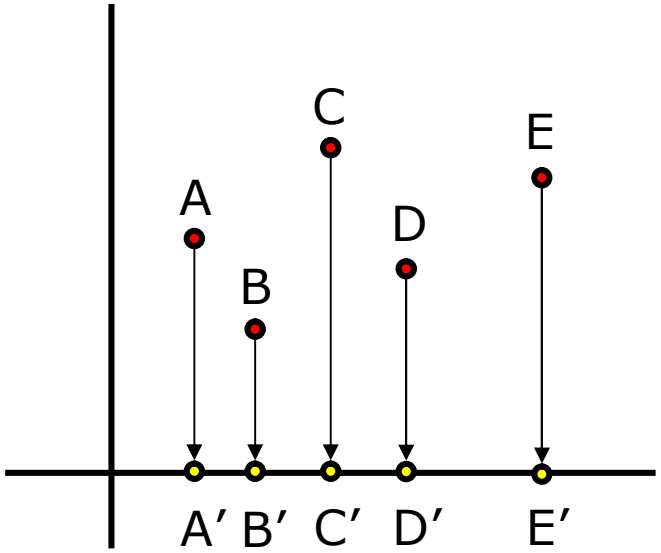
$$\mathrm{BMI} = \frac{\text{体重}\,(\mathrm{kg})}{\text{身長}\,(\mathrm{m})^2}$$ ⇒ 2次元データを1次元データに縮約

主成分分析とは

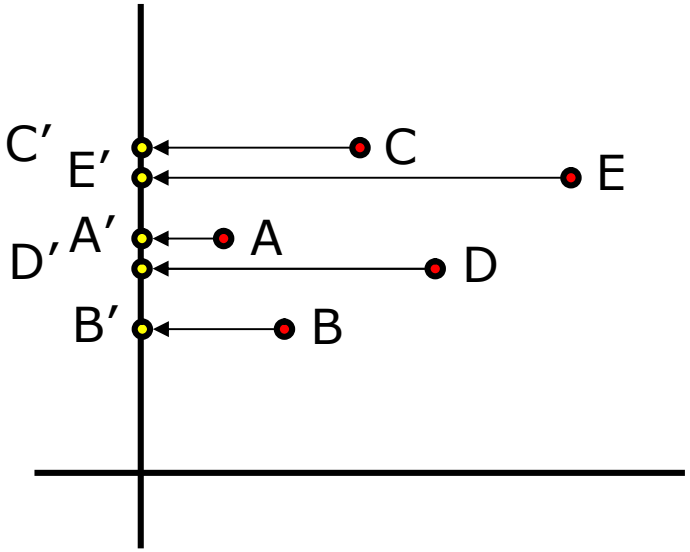
- 主成分分析
  - 多次元データのもつ情報をできるだけ損わずに
    低次元空間に情報を縮約する方法
  - 多次元データを2次元・3次元データに縮約できれば、
    データ全体の雰囲気を視覚化することができる。
    視覚化により、データが持つ情報を解釈しやすくなる。

次元の縮約と情報の損失

- ２次元のデータを１次元に縮約することを考える。



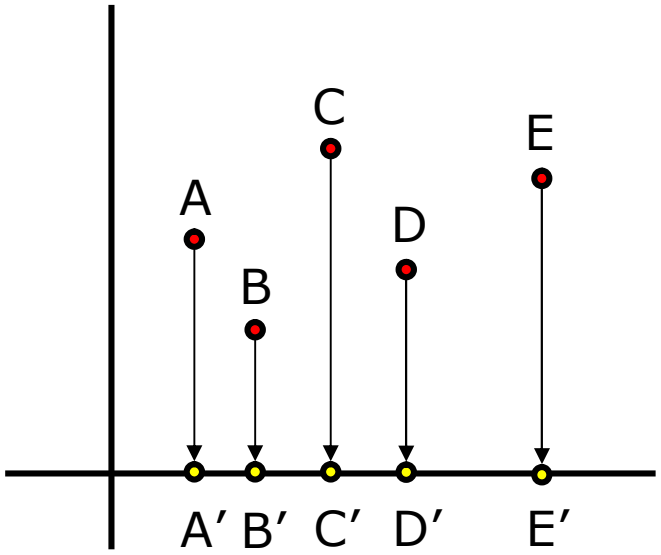
縮約の方法　①
縦軸の情報の損失



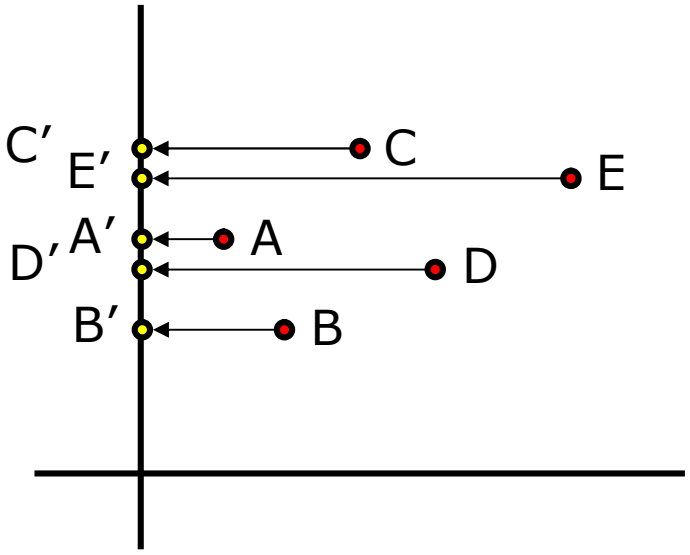
縮約の方法　②
横軸の情報の損失

情報量と分散

- 射影したデータのバラツキが大きいほど、
  もとのデータの情報を多く含んでいると考えられる。



個体差が現れやすい

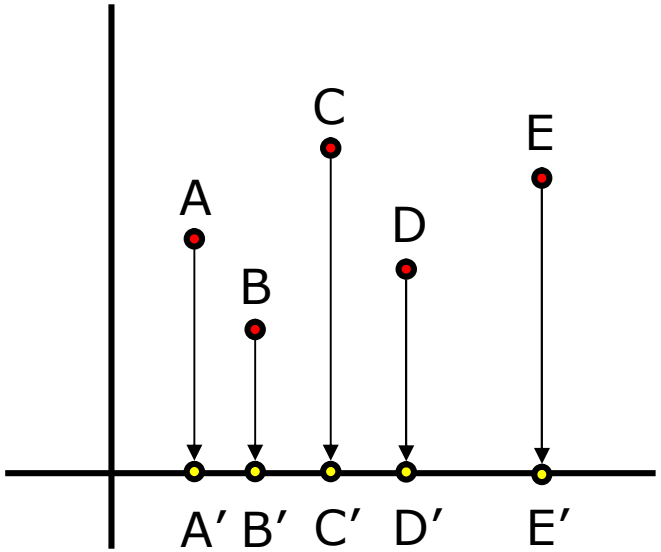


個体差が現れにくい

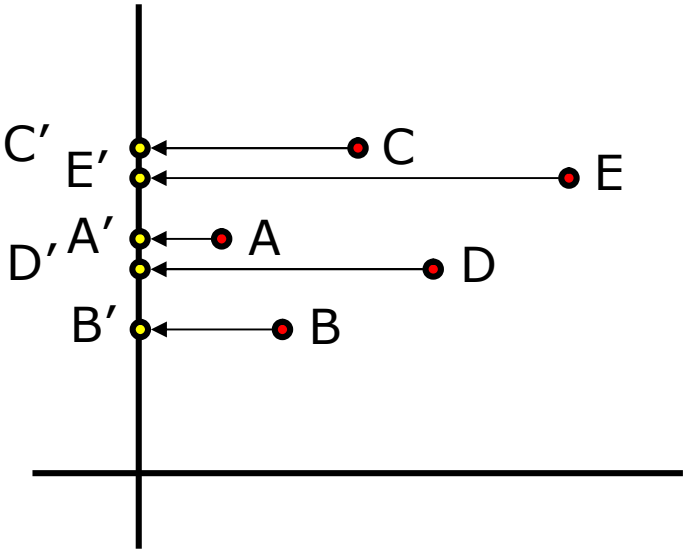
情報量 ⟷ 分散

主成分分析の目的

- もとのデータの情報の損失ができるだけ小さくなるような軸を探したい。



情報の損失が少ない

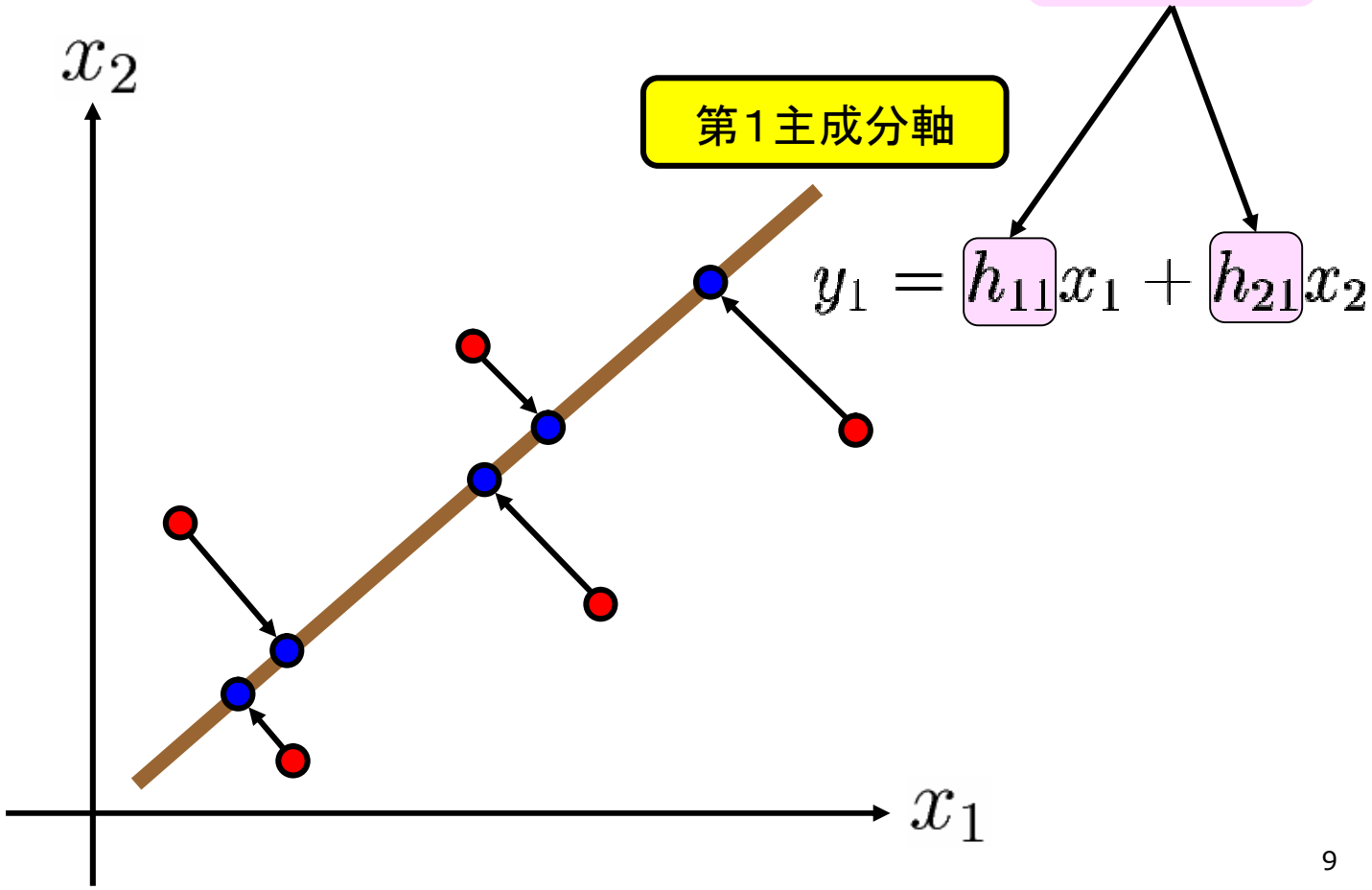


情報の損失が多い

射影したデータの分散が最大となる軸を探す

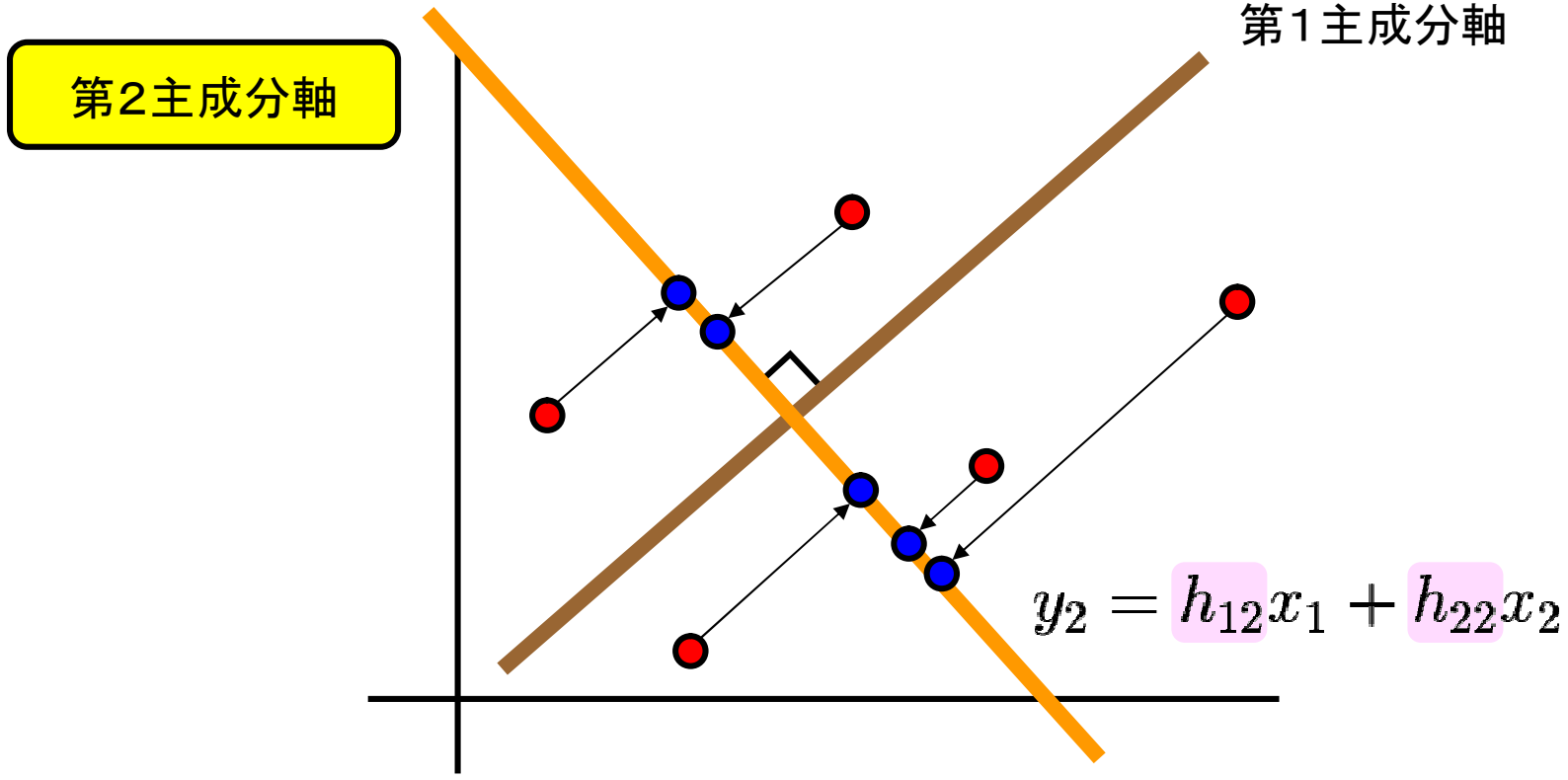
第1主成分

- 射影したデータの分散が最大となるような軸を探す



第2主成分

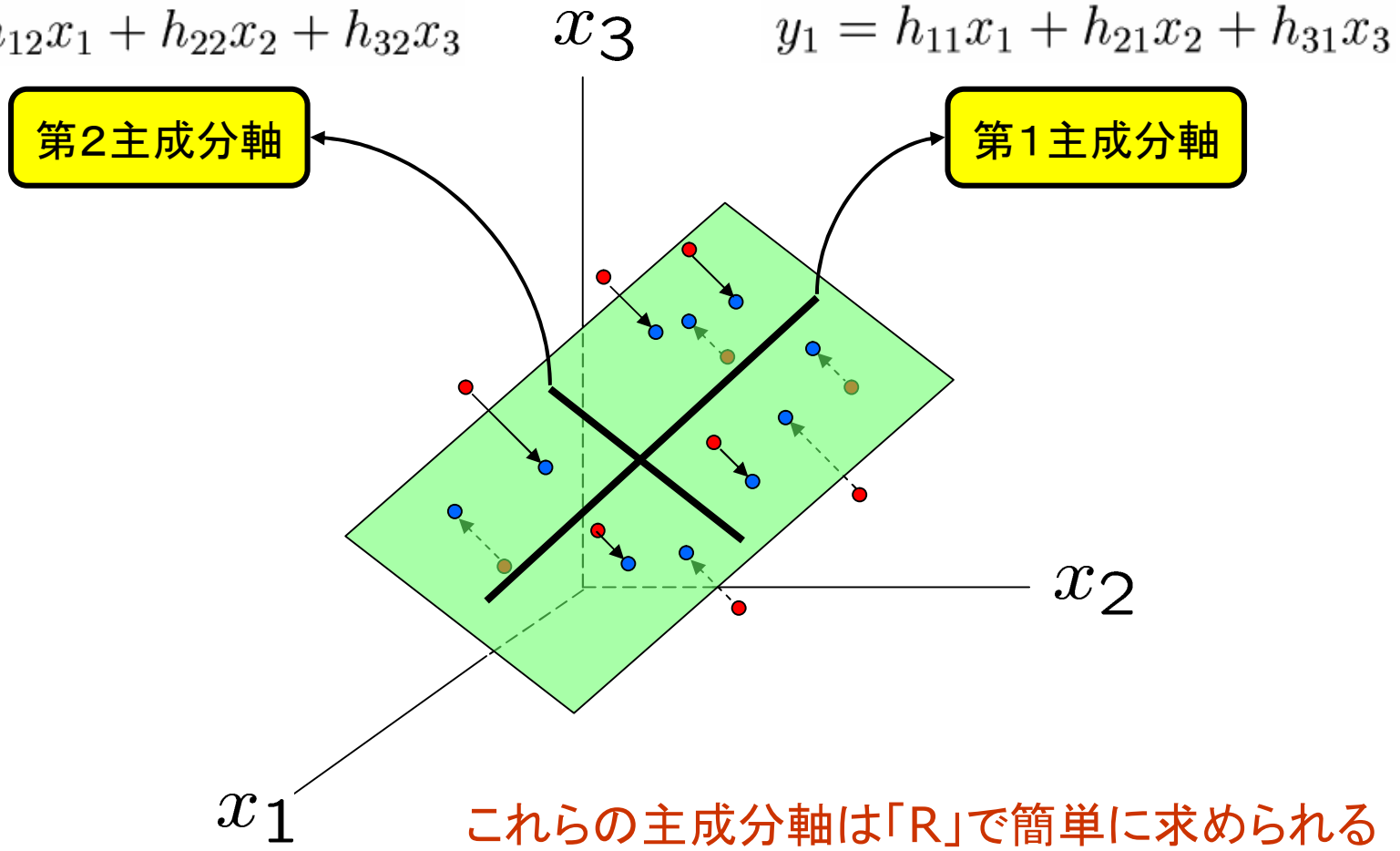
- 第1主成分と直交する軸の中で、軸上に射影したデータの分散が最大となる軸を探す



3次元空間から2次元空間への縮約

$y_2 = h_{12}x_1 + h_{22}x_2 + h_{32}x_3$

$y_1 = h_{11}x_1 + h_{21}x_2 + h_{31}x_3$



これらの主成分軸は「R」で簡単に求められる

主成分分析の要点

- 主成分の分散：主成分がもつ情報量
  ＝分散共分散行列（相関行列）の固有値
- 主成分を構成する係数は、分散共分散行列（相関行列）の固有ベクトルを求める手続きにより得られる
- 主成分分析では、分散共分散行列から分析を行う場合と、相関行列から行う場合で結果が異なる。
- データが異なる尺度（単位）で測定されている場合には、変数を基準化して分析を行う必要がある。

具体例：成績データ　杉山 髙一 著「多変量データ解析入門」

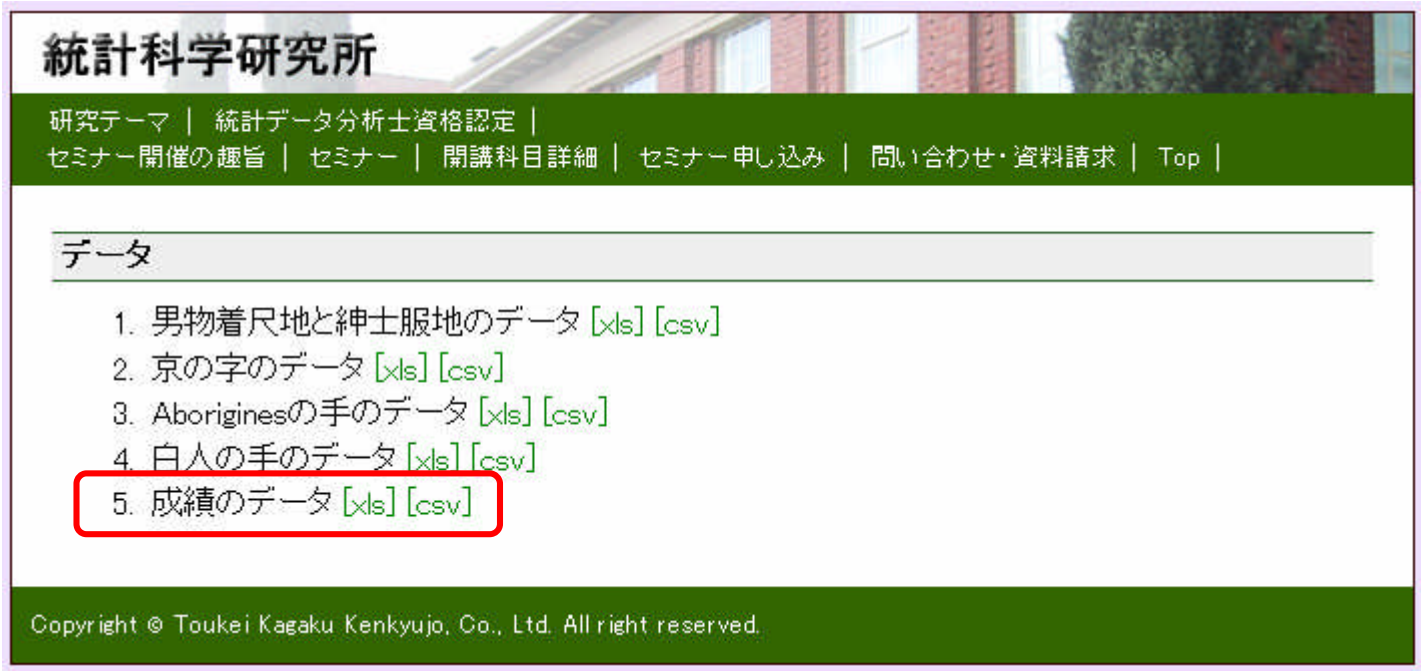
- 中学2年生の成績データ
  - 標本数 ： 166
  - 変数の数 ： 科目数＝9
    - 国語、社会、数学、理科、音楽、英語、体育、技家、英語
  - ダウンロードしたファイルに記載されている最後の3列のデータ（変数名：「4year」「5year」「6year」）を削除して分析を行う
  - 9科目の得点を適当に組み合わせた変数を作り、できるだけ少ない変数で生徒の特徴を捉えたい

データのダウンロード

- 統計科学研究所のウェブサイト
  - http://www.statistics.co.jp/index.htm

データのダウンロード

- 成績のデータの[csv]を右クリック
  ⇒名前を付けて保存



主成分分析を行うプログラム

```
seiseki <-read.csv("seiseki.csv", header=T)
result <- prcomp(seiseki, scale=T)
summary(result)
biplot(result)
```

- プログラムの概要
  - 1行目：データの読み込み
  - 2行目：主成分分析を行う関数 "prcomp" を適用
  - 3行目：主成分分析の結果の要約の出力
  - 4行目：主成分得点をプロットする関数 "biplot" を適用

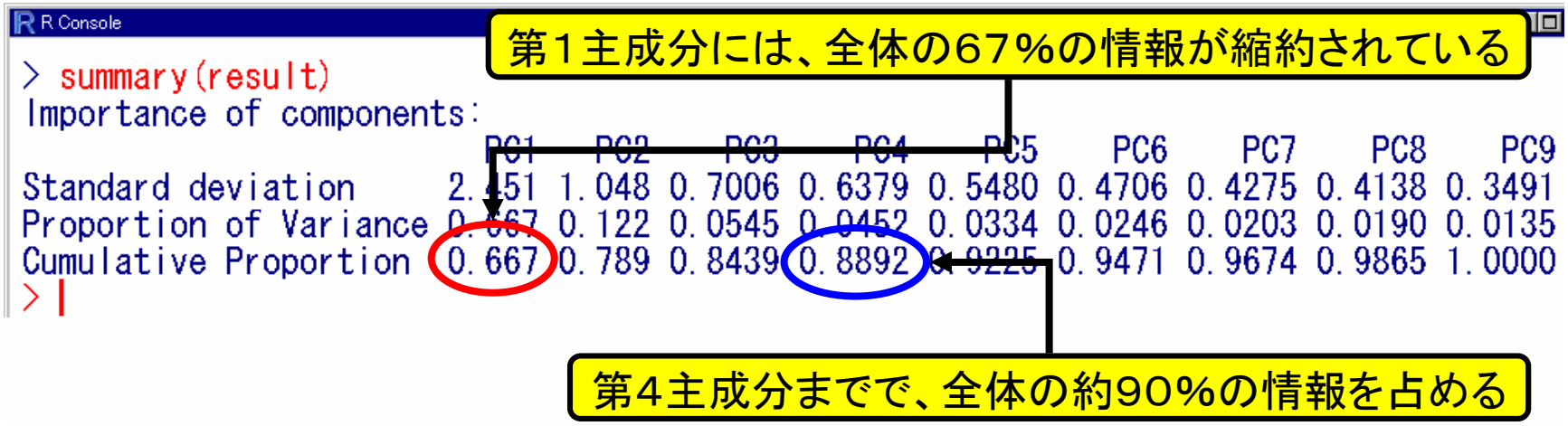
相関行列から主成分分析を行う

```
result <- prcomp(seiseki, scale=T)
```

- 引数 “scale” について
  - 関数 “prcomp” に、引数 “scale=T” を指定
    ⇒ 相関行列から主成分分析を行う
  - 関数 “prcomp” に、引数 “scale=F” を指定
    ⇒ 分散共分散行列から主成分分析を行う

分析結果の要約

- 分析結果に関数 “summary” を適用
  - Standard deviation (標準偏差)
  - Proportion of Variance (寄与率)
  - Cumulative Proportion (累積寄与率)



寄与率と累積寄与率

- **標準偏差**　該当する主成分がもつ情報量
  - 第 i 主成分の標準偏差。これらは第 i 固有値の平方根となっている。
- **第 i 主成分の寄与率**
  - $\dfrac{l_i}{\text{総分散}}$, 総分散 $= \sum_{i=1}^{p} l_i$, l_i：第 i 固有値

    全情報量のうち、該当する主成分が占める情報量の割合
- **第 m 主成分までの累積寄与率**
  - $\dfrac{l_1 + l_2 + \cdots l_m}{\text{総分散}}$　選択した主成分が占める情報量の割合

    次元の縮約により失う情報量を測ることができる

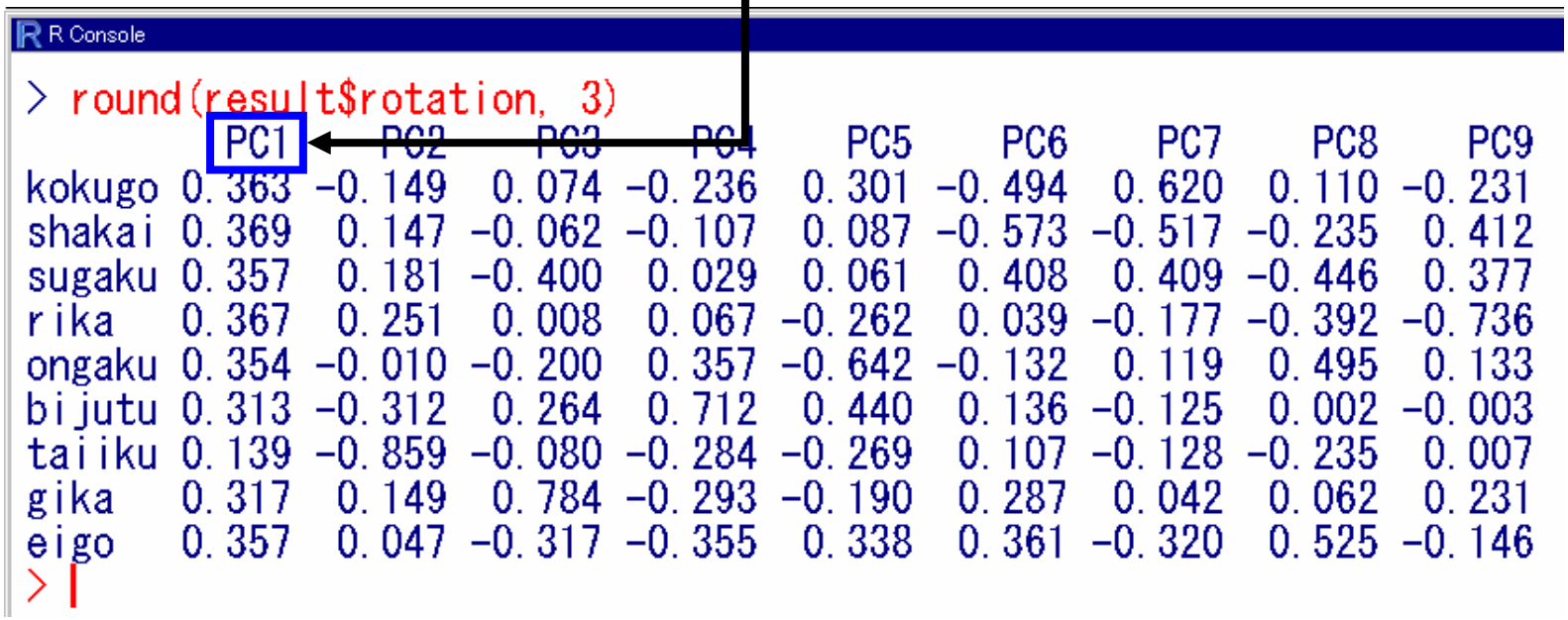
分析結果の出力

- 次のようにして、関数 ”prcomp” で得られたオブジェクトから、分析結果を得ることができる
- 今回のプログラムの場合
  - result$rotation ： 固有ベクトル（主成分軸の係数）
  - result$x ： 主成分得点
- 関数 “round” を使って出力結果を適当な桁数で丸めると見やすくなる
  - round(result$x, disits=3) ： 主成分得点を小数点3桁で表示

固有ベクトルの出力

- round(result$rotation, 3) の出力

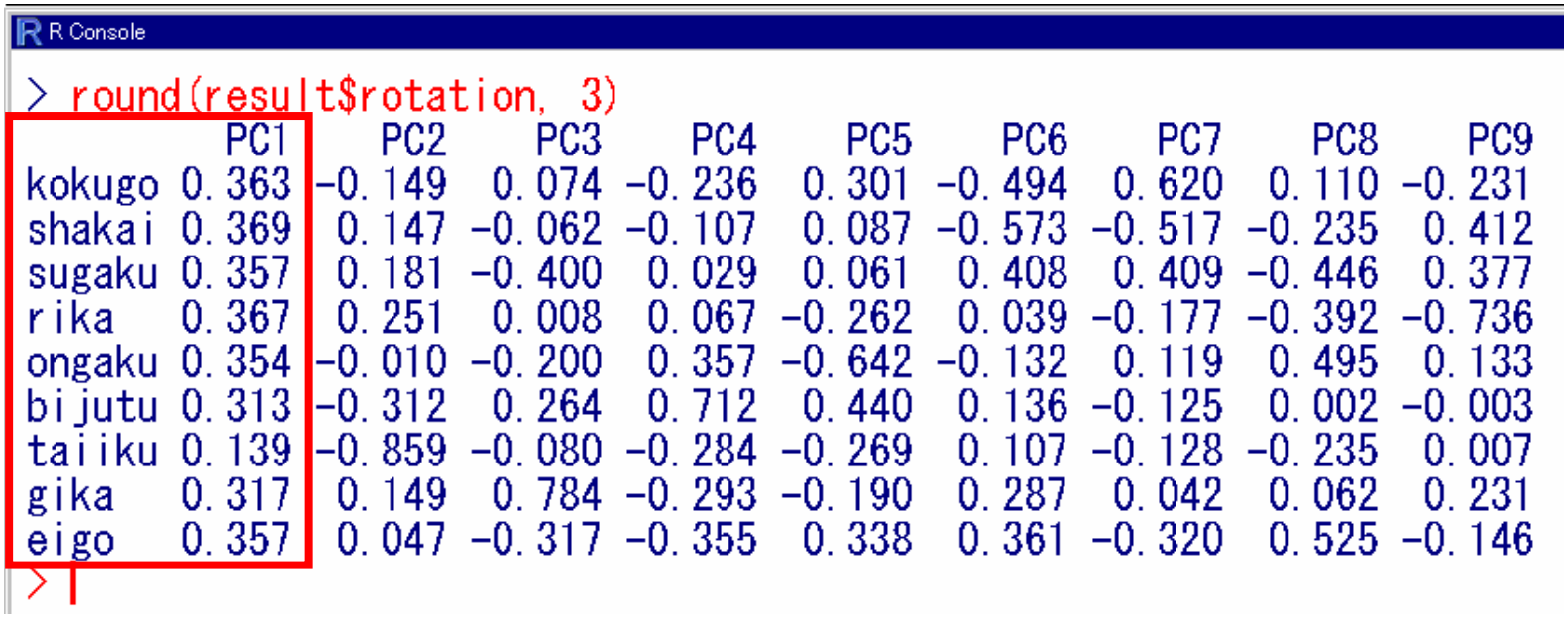
PC：Principal Component ⟷ 主成分



第1主成分の構成

- 第1主成分 ＝

  0.363×国語 ＋ 0.369×社会 ＋ 0.357×数学＋
  0.367×理科 ＋ 0.354×音楽 ＋ 0.313×美術＋
  0.139×体育 ＋ 0.317×技家 ＋ 0.357×英語



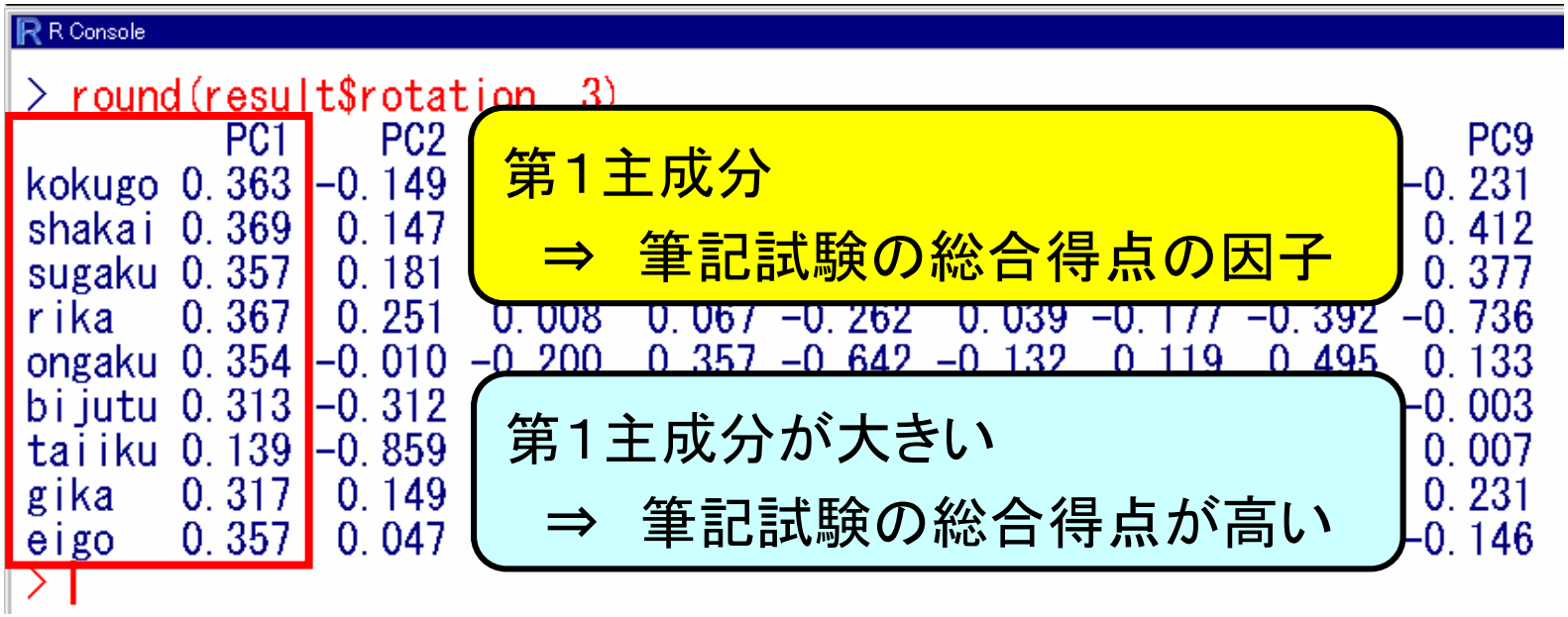
第1主成分の構成

- 第1主成分 ＝

  0.363×国語 ＋ 0.369×社会 ＋ 0.357×数学＋

  0.367×理科 ＋ 0.354×音楽 ＋ 0.313×美術＋

  小 0.139×体育 ＋ 0.317×技家 ＋ 0.357×英語



第2主成分の解釈

- 第2主成分 ＝

  −0.149×国語 ＋ 0.147×社会 ＋ 0.181×数学

  +0.251×理科 − 0.010×音楽 − 0.312×美術

  −0.859×体育 ＋ 0.149×技家 ＋ 0.047×英語

R Console

```
> round(result$rotation, 3)
          PC1    PC2    PC3                    PC7    PC8    PC9
kokugo  0.363 -0.149  0.074                  0.620  0.110 -0.231
shakai  0.369  0.147 -0.062                  0.517 -0.235  0.412
sugaku  0.357  0.181 -0.400                  0.409 -0.446  0.377
rika    0.367  0.251  0.008  0.067 -0.262  0.039 -0.177 -0.392 -0.736
ongaku  0.354 -0.010 -0.200  0.357 -0.642 -0.132  0.119  0.495  0.133
bijutu  0.313 -0.312  0.264
taiiku  0.139 -0.859 -0.080
gika    0.317  0.149  0.784
eigo    0.357  0.047 -0.317
>
```

第2主成分
⇒ 体育の因子

第2主成分が小さい（符号に注意）
⇒ 体育の得点が優れている

因子負荷量

- 各主成分の意味づけ
  - 主成分に強く寄与している変数を見つけることが重要
- 因子負荷量
  - 主成分と各変数との相関係数
    相関行列から分析を始めた場合の因子負荷量

$$\mathrm{Cor}(x_i, y_j) = \sqrt{l_j} h_{ij}, \quad x_i : i \text{ 番目の変数}, \quad y_j : \text{第 } j \text{ 主成分}$$

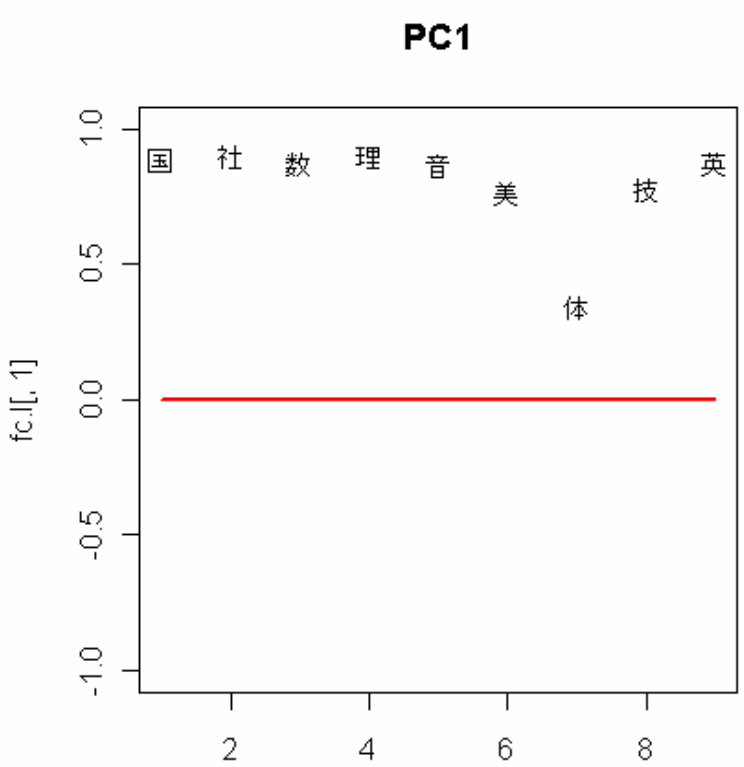
    参考 ： 奥野 忠一著「多変量解析法 改訂版」日科技連
  - 因子負荷量が1か-1に近い因子ほど、主成分に強く寄与している
  - 因子負荷量をプロットすることにより、
    主成分に寄与している因子を視覚的に捉えることができる

因子負荷量に関するプログラム

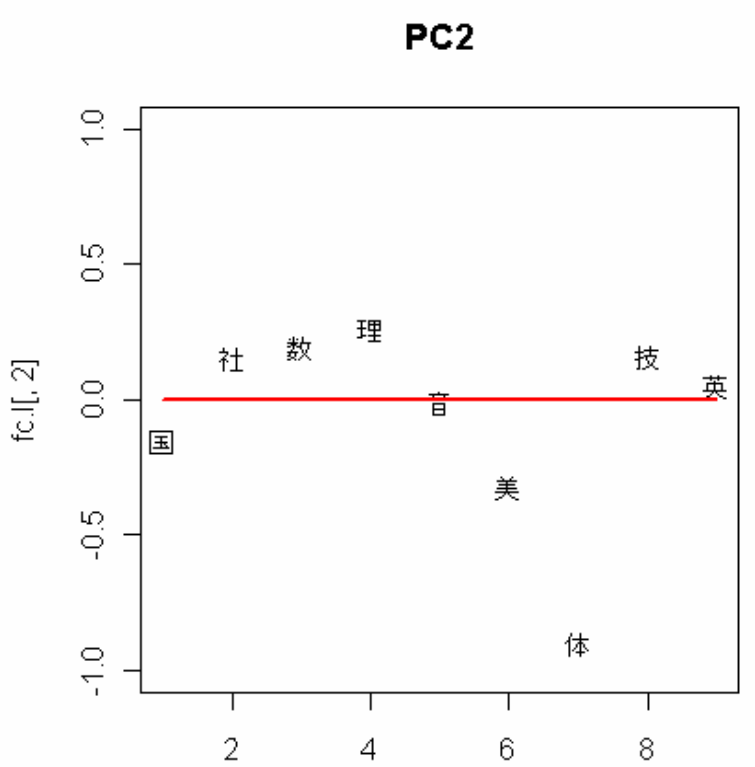
```
fc.l <- sweep(result$rotation, MARGIN=2, result$sdev, FUN="*")
subject <- c("国", "社", "数", "理", "音", "美", "体", "技", "英")
plot(fc.l[,1], pch=subject, ylim=c(-1,1), main="PC1")
plot(fc.l[,2], pch=subject, ylim=c(-1,1), main="PC2")
```

- プログラムの概略
  - 1行目 ： 因子負荷量の計算
    - 固有ベクトル（result$rotation）と、
      対応した固有値の平方根（result$sdev）との積をとる
  - sweep 関数の使い方は、apply 関数とよく似ている
    - 参考URL：R-Tips 24節「applyファミリー」
      http://cse.naro.affrc.go.jp/takezawa/r-tips/r/24.html

因子負荷量のプロット（1次元）

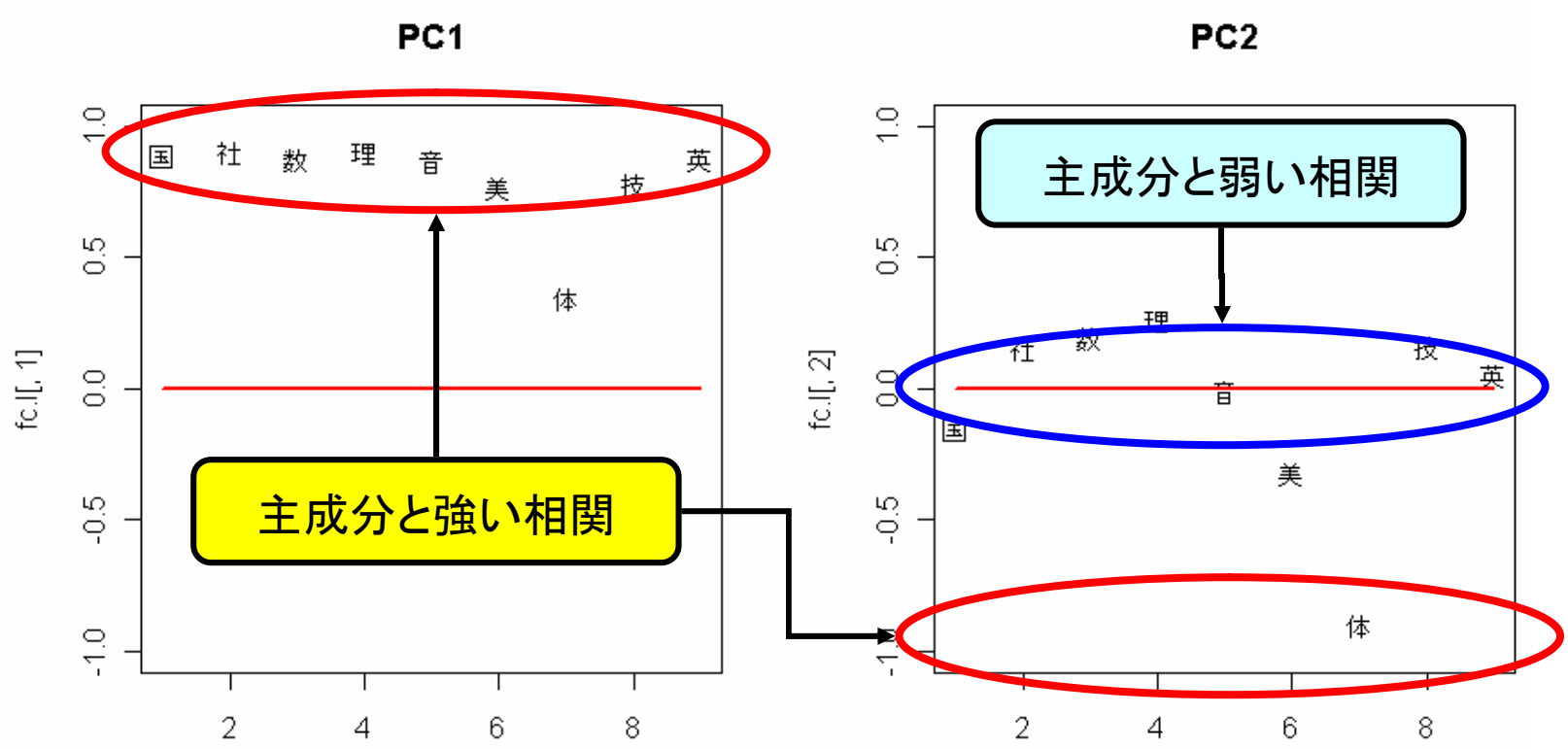


3行目のプログラムの出力



4行目のプログラムの出力

因子負荷量の解釈（1次元）

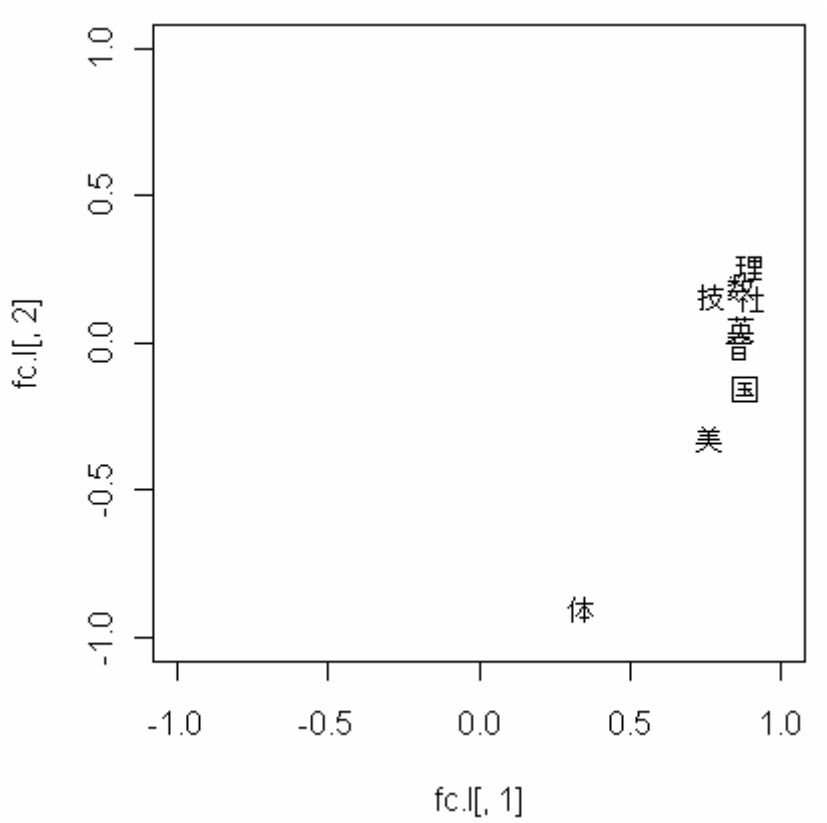


筆記試験の因子

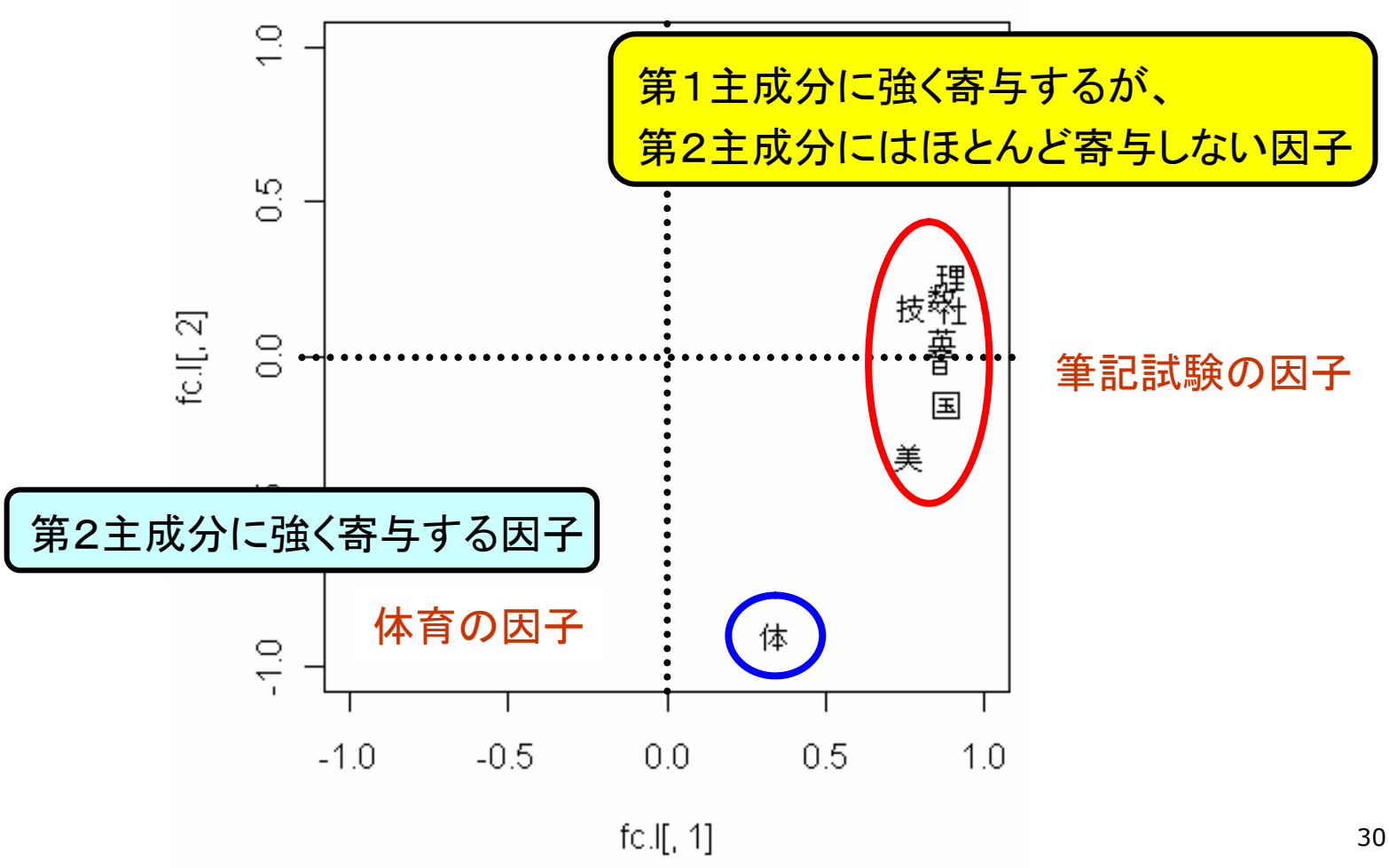
体育の因子

因子負荷量のプロット（2次元）

```
plot(fc.l[,1],  fc.l[,2], pch=subject,
        xlim=c(-1,1), ylim=c(-1,1), main=title)
```



因子負荷量の解釈（2次元）



主成分得点の定義

- 主成分得点の定義
  - 主成分 $y_i = h_{i1}x_1 + h_{i2}x_2 + \cdots + h_{ip}x_p$ に
    個々のデータを代入したもの

result$x

- 成績データの例
  - 第1主成分 ＝

$h_{13}x_3$

0.363×国語 ＋ 0.369×社会 ＋ 0.357×数学＋
0.367×理科 ＋ 0.354×音楽 ＋ 0.313×美術＋
0.139×体育 ＋ 0.317×技家 ＋ 0.357×英語

国語	社会	数学	理科	音楽	美術	体育	技家	英語
95	87	77	100	77	82	78	96	87

相関行列から分析を行う場合は、全ての変数を基準化たものを代入する

主成分得点

- 4人目の成績

国語	社会	数学	理科	音楽	美術	体育	技家	英語
95	87	77	100	77	82	78	96	87

第1主成分得点：5.107

第2主成分得点 :0.228

- 130人目の成績

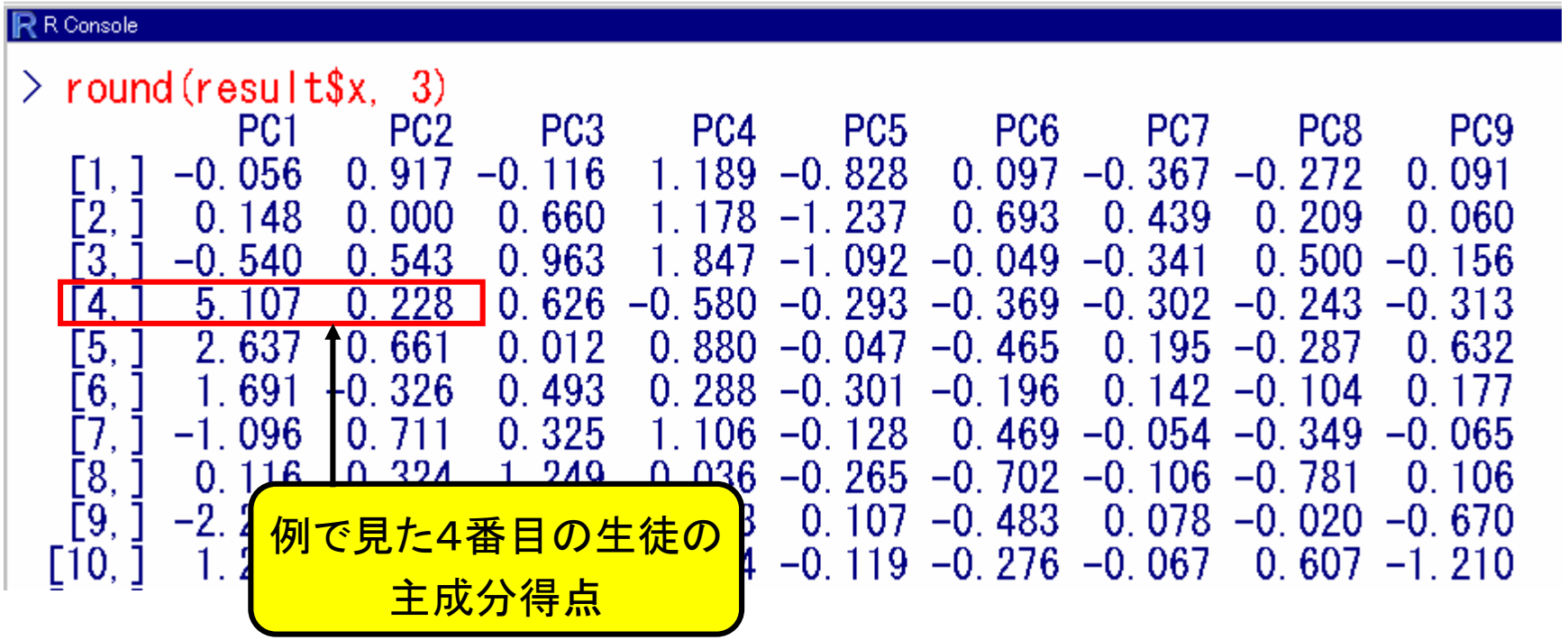
国語	社会	数学	理科	音楽	美術	体育	技家	英語
64	36	20	31	53	68	99	7	26

第1主成分得点：-0.812

第2主成分得点：-2.244

主成分得点の出力

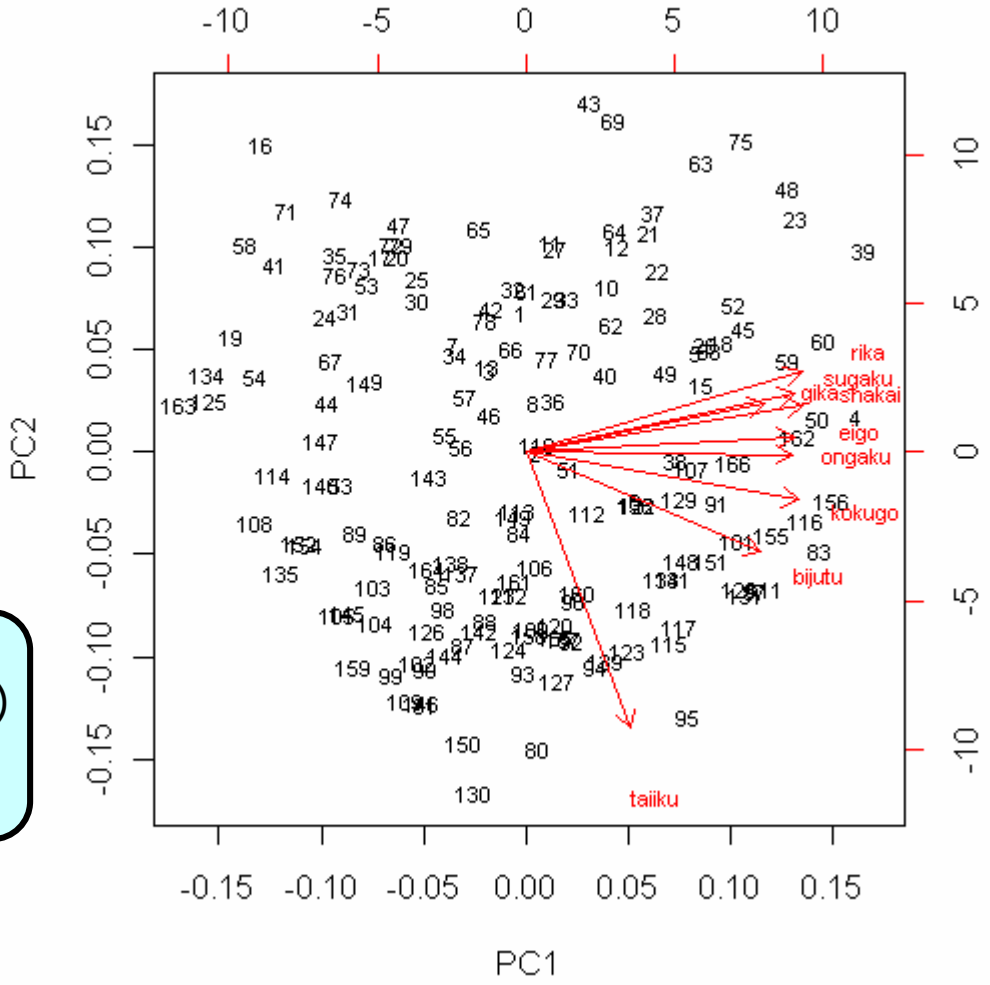
- result$x：主成分得点を出力する



主成分得点のプロット（biplot）

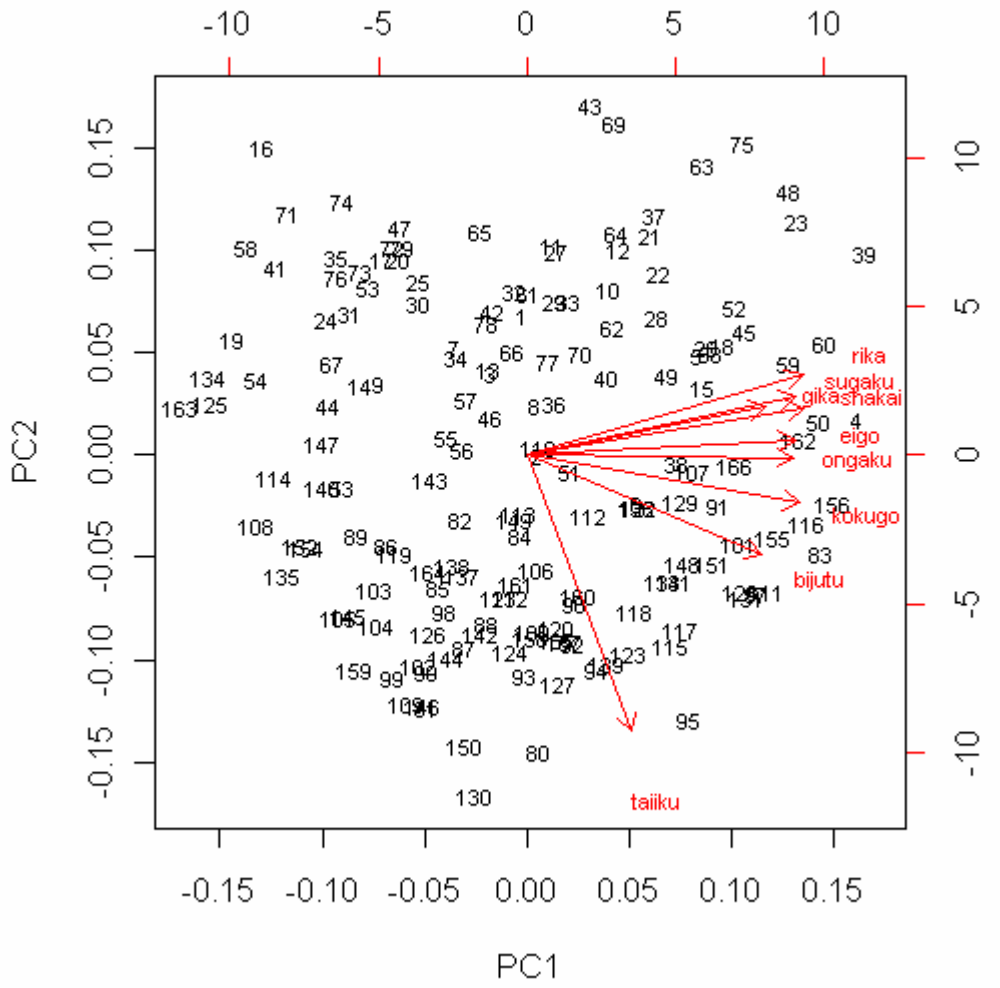
- 主成分得点を低次元空間にプロットすると、個体の特徴や位置を把握しやすくなる
- 「R」では、biplot 関数を適用することで、解釈しやすい形で主成分得点のプロットを得ることができる

```
result <- prcomp(seiseki, scale=T)
biplot(result)
```



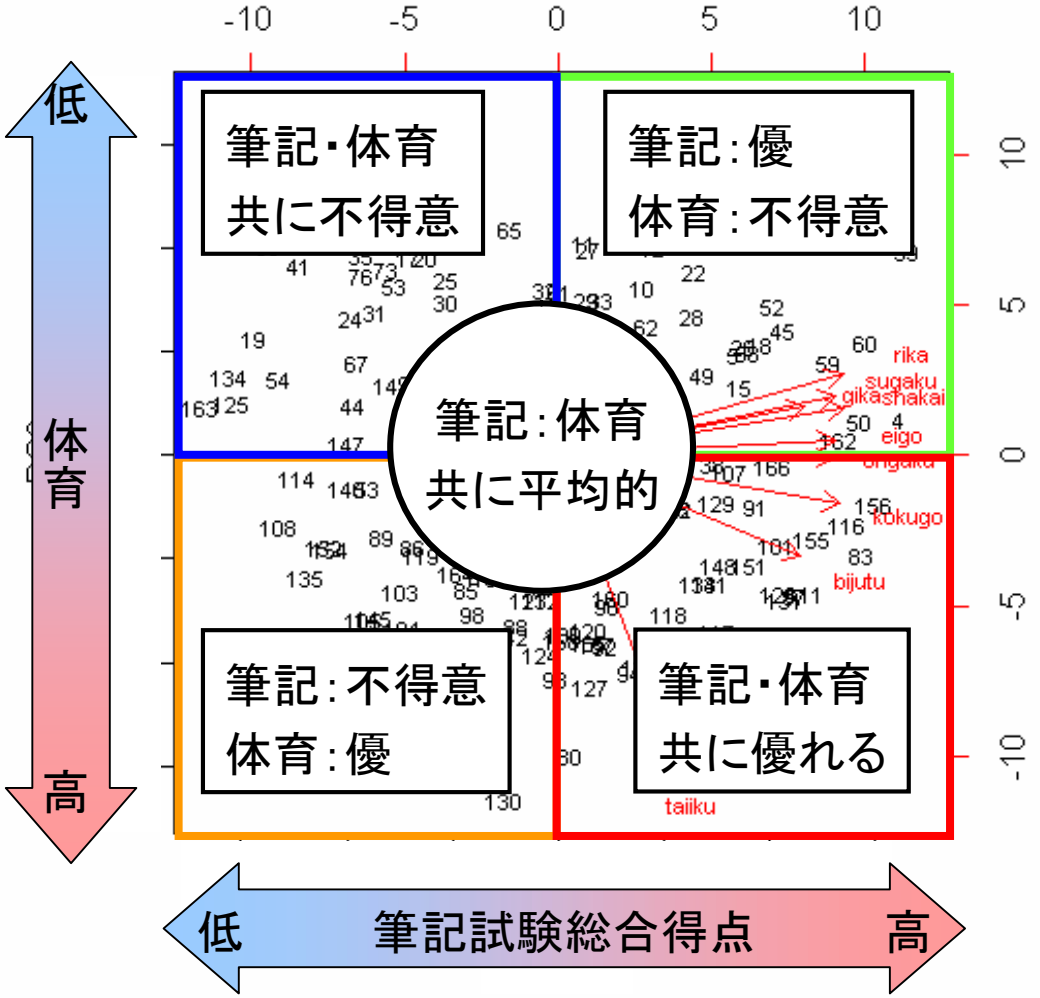
biplotの解釈

- 第一主成分の解釈
  - 筆記試験の総合得点
    ⇒右にあるデータほど筆
    試験の総合点が高い
- 第二主成分の解釈
  - 体育の因子
    ⇒下にあるデータほど
    体育の成績が良い
  - 主成分の符号や
    ベクトルの向きに注意す



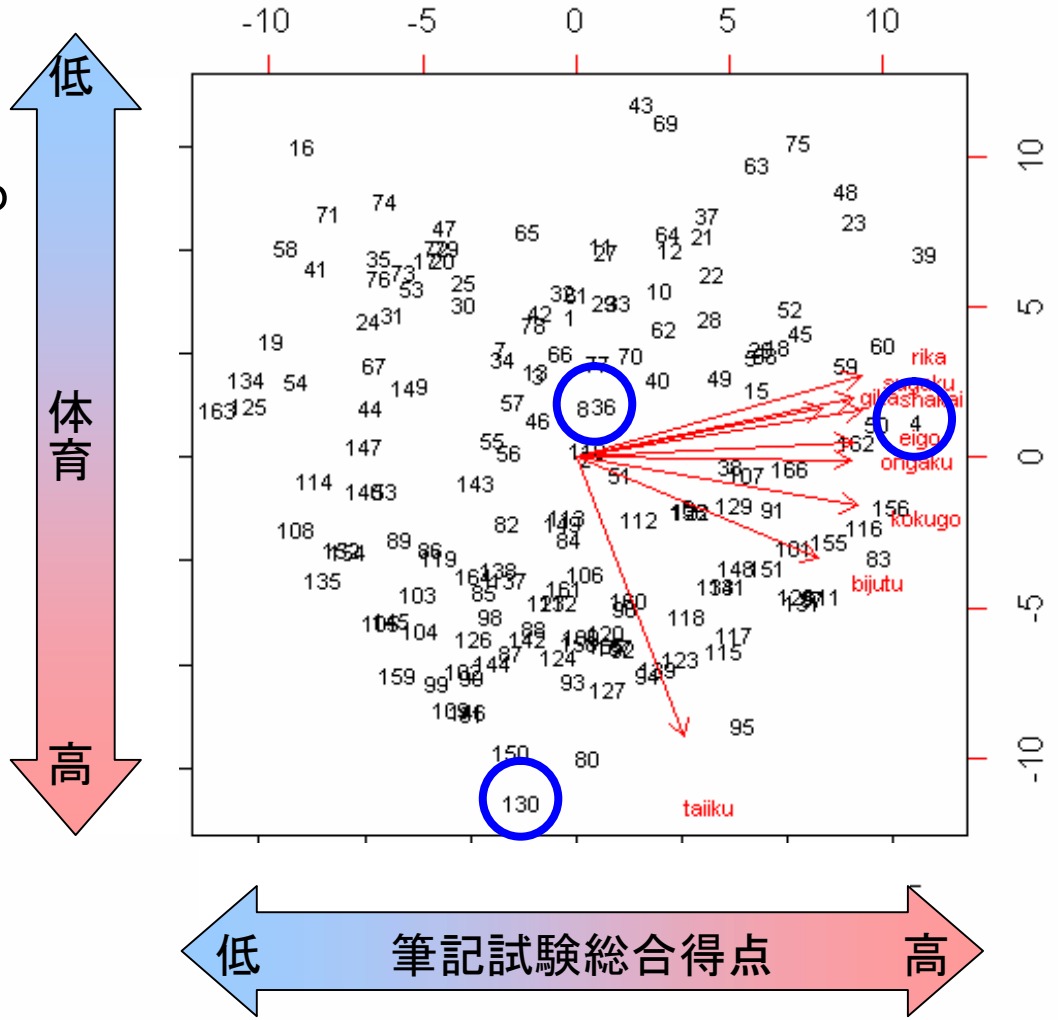
主成分によるデータの位置づけ

- 主成分の解釈から、各個体を右図のように分類して考えることができる
- 主成分得点の順にデータを並べ替えることである特性値について順位付けを行うこともできる



主成分によるデータの位置づけ

- biplotの見方
  - 4番
    筆記試験が優れている
    体育は平均程度
  - 130番
    筆記試験は平均程度
    体育得意
  - 8番
    筆記試験も体育も
    平均程度



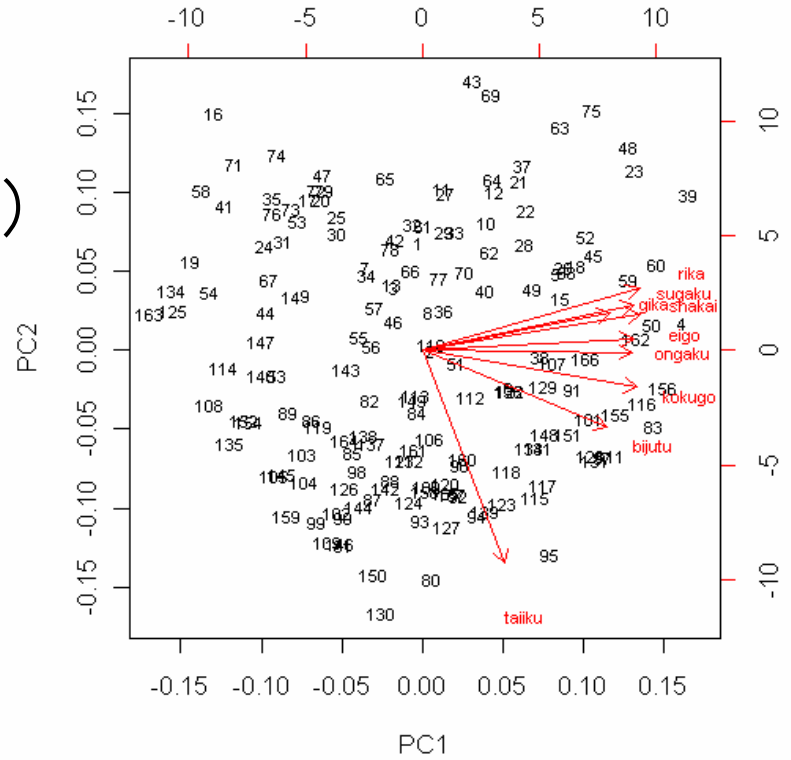
主成分得点とデータ

- 例で挙げた生徒の成績と主成分得点
  - PC1：第1主成分の主成分得点
  - PC2：第2主成分の主成分得点

	国語	社会	数学	理科	音楽	美術	体育	技家	英語	PC1	PC2
4	95	87	77	100	77	82	78	96	87	5.1	0.2
8	56	54	37	59	35	64	53	67	7	0.1	0.3
130	64	36	20	31	53	68	99	7	26	-0.8	-2.2
平均	57.5	39.6	45.6	49.9	42.6	62.5	57.7	47.3	39.1	0	0

まとめ

- 主成分分析を行う関数 "prcomp" の使い方
- 主成分得点の出力の仕方
  - obj$x
- 固有ベクトル（主成分軸の係数）の出力の仕方
  - obj$rotation
- 因子負荷量の求め方と解釈
- 関数 "biplot"の使い方と解釈



参考URL

- 統計科学研究所のウェブサイト
  http://www.statistics.co.jp/index.htm
- R-Tips
  http://cse.naro.affrc.go.jp/takezawa/r-tips/r2.html
- JIN'S PAGE
  http://www1.doshisha.ac.jp/~mjin/R/