データ分析基礎
主成分分析＋演習の手順

京都大学 国際高等教育院 附属データ科学イノベーション教育研究センター

せきど ひろと
關戸 啓人

sekido.hiroto.7a@kyoto-u.ac.jp

主成分分析

主成分分析の概要

★ 主成分分析（**Principal Component Analysis, PCA**）

★ **次元の縮約の観点から，新しい座標を構成する**もの

★ 例えば，世界500都市の1時間おきの気温20年分のデータが有るとする

★ データ数は500×175320程度

★ 都市の緯度，経度，人口密度，内陸度，…，などの別の座標を導入することで，全ての気温のデータを保存しなくても良い？

★ このような新しい座標の導入をデータのみから自動的に算出

★ 筋の良い座標の取り方がわかる

★ データ容量，計算量削減，***ノイズ除去***

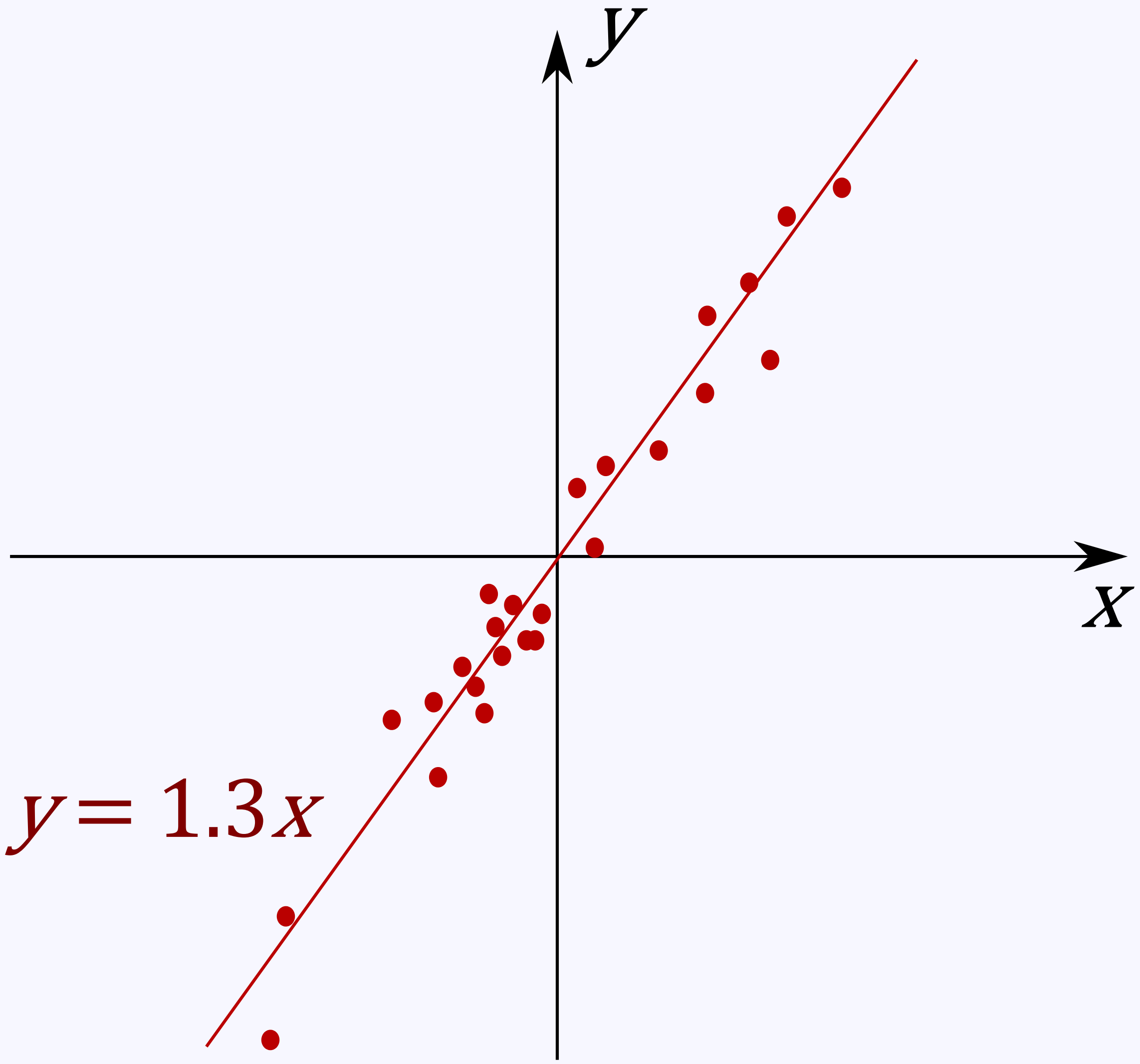
主成分分析の概要

- ★ 登場する確率変数は全て平均が0になるように，定数を足したり引いたりしているとする
  - ★ データは中心化されている

- ★ 最初は，2変数の簡単な例で主成分分析の考え方を述べる

- ★ 確率変数Xは体重，Yは身長を表すとし，データ(x_k, y_k)が与えられたとしよう
  - ★ XとYには関係があって，近似的に$Y = 1.3X$ぐらいであるとする

主成分分析の概要



主成分分析の概要

★ データ (x_k, y_k) は直線 $y = 1.3x$ の付近に散らばっている

- ★ $(\alpha + \varepsilon, 1.3\alpha + \delta)$ で α を適当に取ると，ε や δ は小さいことが多い

★ そこには，（理論的に説明できるかどうかはわからないが）何らかの力が働いていると考えることができる

- ★ その何らかの力は，以下の確率変数 Z で表されるであろう．

$$Z = \frac{X + 1.3Y}{\sqrt{1^2 + 1.3^2}}$$

- ★ 確率変数 X, Y を直交変換で Z, U に移すとしたら，U は以下のようになる

$$U = \frac{1.3X - Y}{\sqrt{1.3^2 + (-1)^2}}$$

★ 主成分分析は，このように，確率変数を**直交変換**することである

主成分分析の概要

- ★ この例では，Z は体の大きさ，U は肥満度を表しているように思える
- ★ どちらが重要な確率変数かは置いておいて，データの散らばりをより良く説明している確率変数は Z である
  - ★ $z_k = x_k + 1.3y_k$ の値を見れば，大体の x_k, y_k の値がわかるという意味である
- ★ このように，直交変換した後の確率変数で，元のデータを1番良く説明している変数を第1主成分，2番目により良く説明している変数を第2主成分，などと呼ぶ.

- ★ **主成分分析は，次元の縮小**に用いられる
  - ★ 直交変換した後の全ての確率変数を用いれば，元のデータは完全に復元できる
  - ★ しかし，それなりに小さい k について，第1主成分から第 k 主成分までのみを用いても，ほぼデータは復元できるようになる
  - ★ よって，いくつかの主成分のみを考えても支障がなくなる（支障が出ないように次元を減らす）

主成分の定義A

★ 主成分の定義は（ここで紹介するのは）2種類あるが，1つ目の定義を述べる

- ★ 1つ目の定義での考え方は，**ばらつきとは，分散である**
- ★ そして，**ばらつきをより良く説明する，とは分散が大きいこと**と考える

★ 元々の確率変数を $X_1, X_2, \ldots, X_n$ とする

- ★ 第1主成分 Z_1 は

$$
Z_1 = w_1X_1 + w_2X_2 + \cdots + w_nX_n, \\
w_1^2 + w_2^2 + \ldots + w_n^2 = 1
$$

  と書けるものの中で，分散が最も大きいものである

- ★ また，第 k 主成分は，上の形で書け，第 $k-1$ 主成分までと直交するものの中で，分散が最大となる確率変数

主成分の定義B

★ 2つ目の定義では，第1主成分 Z_1 を元々の変数との相関係数の2乗和を最大化する確率変数と取る

- ★ 第 k 主成分は，同様に，第 $k-1$ 主成分までと直交する中で，元々の変数との相関係数の2乗和を最大化するように取る

★ これは**前処理としてデータの標準化を行った後に定義Aで主成分を定義している**とも思える

- ★ データの標準化：平均が0，分散が1になるように定数を足したりかけたりする

★ 前処理の問題なので，スライドでは以下定義Aで説明する

- ★ どちらの定義も行列の固有値問題に帰着されるが，定義Aは分散共分散行列，定義Bは相関行列の固有値問題になる

分散共分散行列と相関行列

★ 分散共分散行列の (i,j) 成分は，X_i と X_j の共分散

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])]$$

である．

★ 相関行列の (i,j) 成分は，X_i と X_j の相関係数

$$\frac{X_i \text{ と } X_j \text{ の共分散}}{(X_i \text{ の標準偏差})(X_j \text{ の標準偏差})}$$

である．相関係数の絶対値は 1 以下となる．

★ 確率変数 X_1 は分散が大きいが，確率変数 X_2 は分散が小さい，となれば全体の結果は確率変数 X_1 の影響が強くなる．これを防ぐため，各変数を標準化して考えたものが，相関行列を用いたものだと考えることができる．

- ★ 実際にはほとんどの場合において標準化を行う定義 B で主成分分析を行う
- ★ ただし，データの「標準化」の仕方を考えた方が良い場合も多い

主成分

★ 以降，定義Aで述べる

★ 共分散行列の固有値を大きい順に以下とする

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$$

★ また，λ_k に対応する固有ベクトルを以下とする

$$(w_{1,k}, w_{2,k}, \ldots, w_{n,k})^{\mathrm{T}}$$

★ 第 k 主成分は

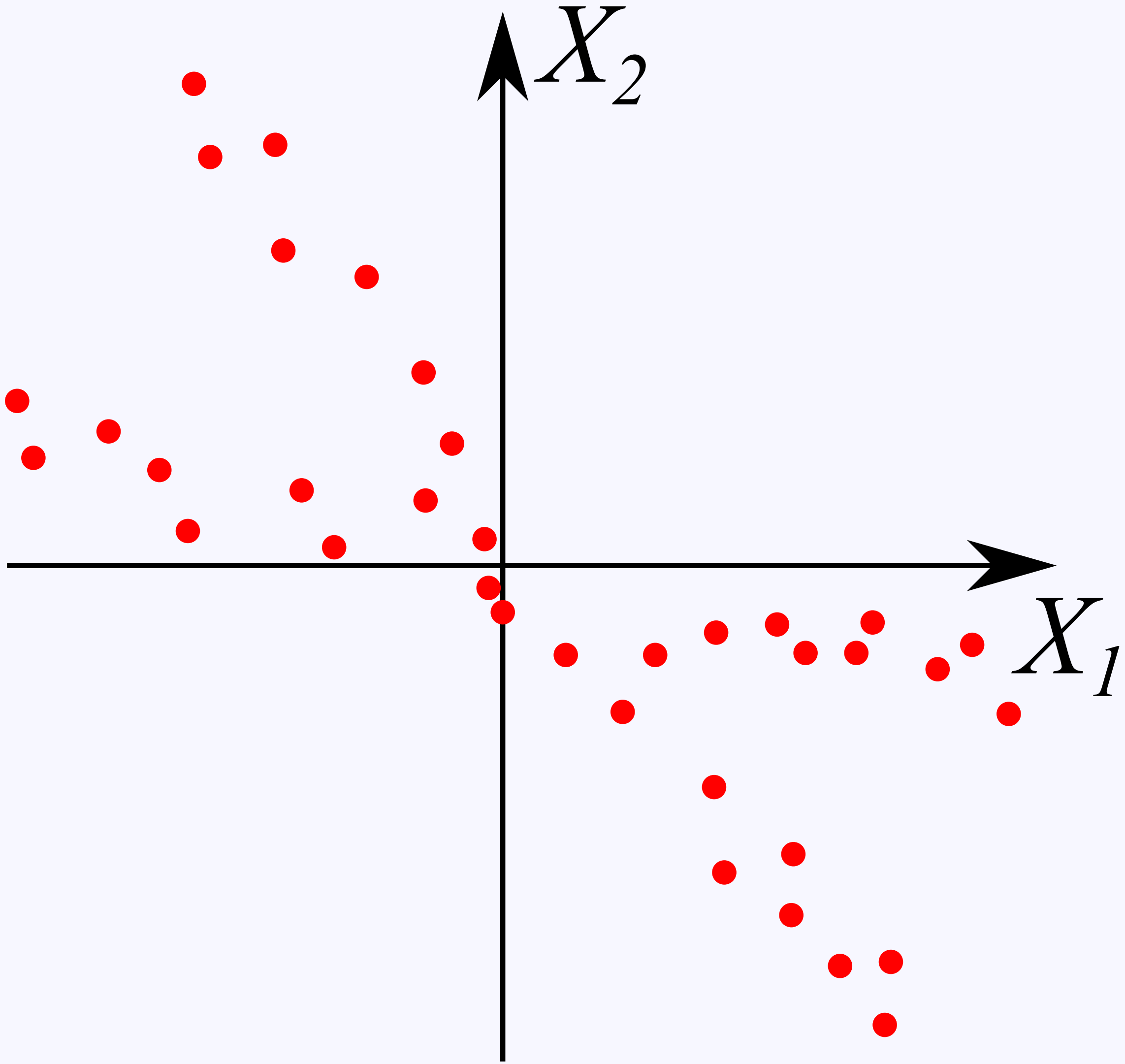
$$Z_k = w_{1,k}X_1 + w_{2,k}X_2 + \cdots + w_{n,k}X_n$$

となり，その分散は $\mathrm{V}[Z_k] = \lambda_k$ となる

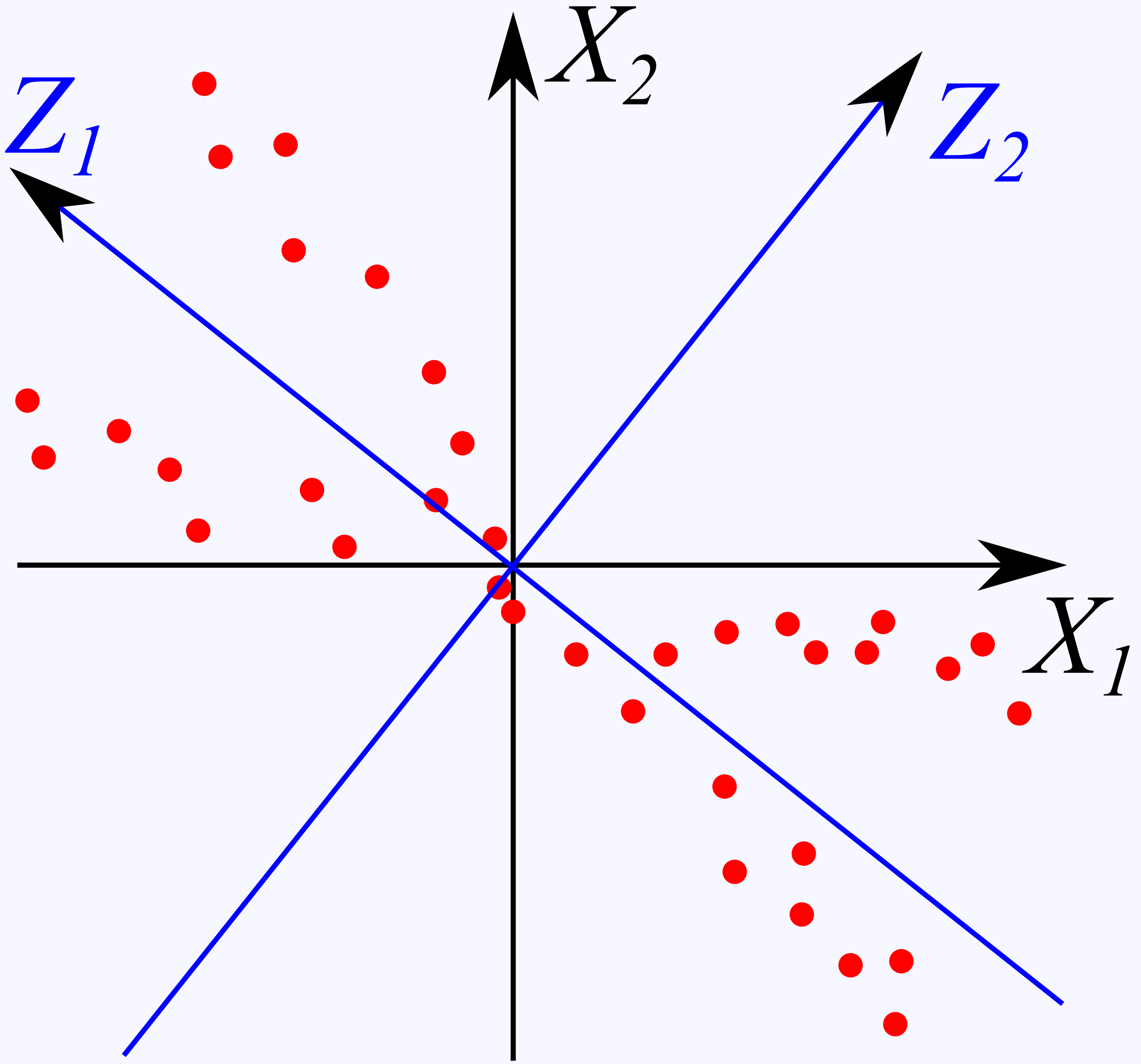
★ 証明は例えばラグランジュの未定乗数法を用いる（ここでは略）

★ データ行列が $A = UDV^{\mathrm{T}}$ と特異値分解されているとき，AV の第 k 列目が第 k 主成分

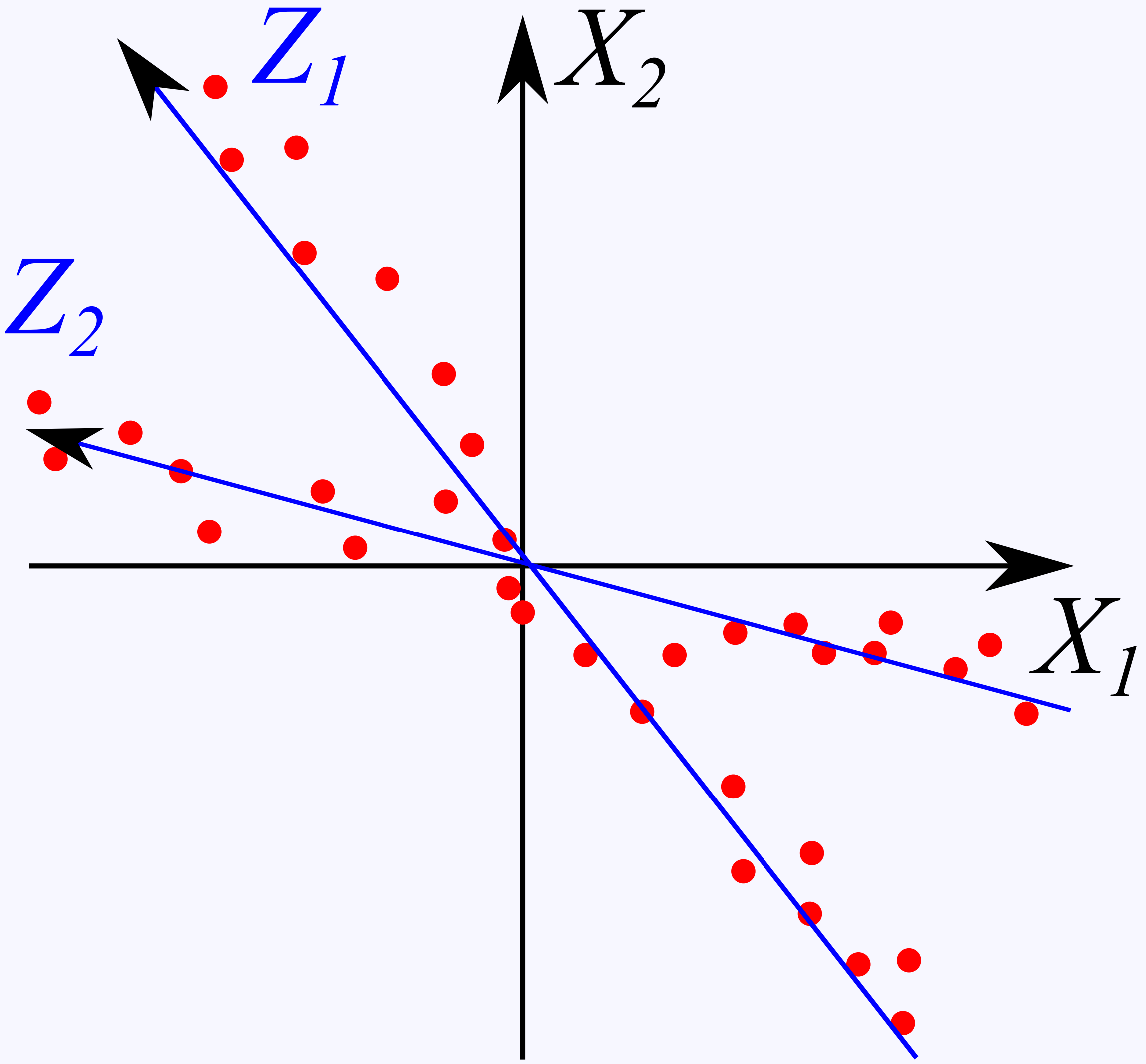
例：データ



例：主成分分析の結果



例：他の手法（因子分析，独立成分分析など？）を用いると



例に対する補足

- ★ 直交変換という制約が故に，隠れた要因を発見できないかもしれない

- ★ データについて，個々の構成要素を得ようとする方法として，別の方法で，因子分析がある
  - ★ 因子分析のやり方は，いろいろな定義があり，それぞれ結果も一致しない

- ★ 独立成分分析では，各確率変数ができるだけ独立になるように定める
  - ★ これも，いろいろな定義がある

- ★ 対して，主成分分析は，少ない主成分でデータを説明する，データの総合的なスコアを定める，ということに特化している

データで表す

★ 確率変数の数を n，標本サイズを m とする

★ データ行列を以下とする

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \in M_{m,n}(\mathbb{R})$$

★ X_i の標本平均は $\displaystyle\sum_{k=1}^{m} x_{k,i} = 0$

★ X_i と X_j の不変共分散は $\displaystyle\frac{1}{m-1}\sum_{k=1}^{m} x_{k,i}x_{k,j}$

★ 不偏共分散行列は $\displaystyle\frac{1}{m-1}A^{\mathrm{T}}A$

主成分分析を計算機で行うには

★ 主成分を求めるには

　★ **共分散行列の（大きい方から数十個の）固有値と固有ベクトルを求める**

★ または

　★ **データ行列の（大きい方から数十個の）特異値と右特異ベクトルを求める**

★ 特異値：$\sqrt{AA^{\mathrm{T}}\text{の固有値}}$

★ 右特異ベクトル：$A^{\mathrm{T}}A$ の固有ベクトル

★ 左特異ベクトル：AA^{T} の固有ベクトル

★ 主成分分析は行列の特異値分解そのもので数学的に性質が良く知られており，ある意味で自然な分析

★ 寄与率：分散の割合

★ 元々の確率変数での分散の和と，主成分での分散の和は等しい（直交変換だから）

$$\sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2 = \sum_{i=1}^{n}\sum_{k=1}^{m} z_{k,i}^2 = \mathrm{tr}(A^T A)$$

★ 主成分 Z_i の**寄与率**

$$\sum_{k=1}^{m} z_{k,i}^2 \Big/ \sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2$$

★ 主成分 $Z_1, Z_2, \ldots, Z_s$ の**累積寄与率**

$$\sum_{i=1}^{s}\sum_{k=1}^{m} z_{k,i}^2 \Big/ \sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2$$

★ 累積寄与率が，ある程度大きくなるように，使用する主成分の数を決めることが多い

（補足）無相関

★ $A = UDV^{\mathrm{T}}$ とする（特異値分解）

★ $U^{T}U = I$，$V^{\mathrm{T}}V = I$ を満たし，D は対角成分以外0

★ すると，主成分からなるデータ行列は $AV = UD$ で，以下が成り立つ

$$\mathrm{const} \times \mathrm{Cov}(UD) = (UD)^{\mathrm{T}}UD = D^{\mathrm{T}}U^{\mathrm{T}}UD = D^{\mathrm{T}}D$$

★ $D^{\mathrm{T}}D$ は対角行列：つまり主成分同士は無相関

★ ここでは，分散を最大化する方針で主成分を定義したが，「直交変換である，かつ，無相関にする」という方針でも同じ結果が得られる

（補足）残差最小化

★ 第k主成分のみを用いてデータ行列を復元したとき

- ★ 復元されたデータ行列を $\tilde{X} = (\tilde{x}_{i,j})$ とすると
- ★ $\sum(x_{i,j} - \tilde{x}_{i,j})^2$ が最小化されている

★ 主成分分析はデータ行列をできるだけ良く近似するように次元の縮小を行っている

- ★ これを定義と思っても主成分分析が得られる
- ★ **低ランク近似**

因子負荷量

★ 元々の各変数と主成分との相関係数

★ 定義Bで行った場合（データの正規化を行った場合）は

$$\mathrm{cor}(X_i, Z_j) = \sqrt{\lambda_j} w_{i,j}$$

★ この主成分は，元々のどの変数の影響を強く受けているか？ という指標

★ 主成分の意味を理解するのに使うことがある

主成分得点

★各データを主成分で表したときの値を主成分得点という

- ★i個目のデータの第j主成分得点は$(AV)_{i,j}$

主成分分析の概略：まとめ

- ★ **主成分分析**とは，数学的には**確率変数の直交変換（データの回転）**
  - ★ $X_1, X_2, \ldots, X_n$ から $Z_1, Z_2, \ldots, Z_n$ に変換
  - ★ $Z_1, Z_2, \ldots, Z_n$ は**無相関**，分散は $V[Z_1] \geq V[Z_2] \geq \cdots \geq V[Z_n]$
- ★ 主成分分析は，**次元の縮約**に用いられる
  - ★ データのばらつきを説明するには，大きい方からいくつかの主成分だけで十分かもしれない．累積寄与率を参考にする
- ★ 主成分は，データの裏に隠れた要素，要因を表しているかも
  - ★ 希望するものが得られているかはわからない，説明がつくとも限らない
  - ★ そういうものの解析をしたいのなら，因子分析，独立成分分析なども視野にいれる
- ★ 計算方法は，共分散行列の固有値分解，または，**データ行列の特異値分解**
  - ★ 実際には，固有値，あるいは，特異値の大きい方から数個だけ必要

主成分分析 - 演習

演習に使うソフトウェア

- ★ R を用いて演習を行います
- ★ R は以下のホームページよりダウンロードできる
  `https://cran.r-project.org/`
- ★ R で込み入ったプログラムなどを書く場合は RStudio も使いこなせるようになると便利です
  `https://rstudio.com/products/rstudio/download/`

- ★ Windows でユーザー名に日本語を使用している場合，うまくインストールできない可能性があります
  - ★ うまくインストールできなかった場合は，ポータブル版の R や，RStudio Cloud を利用するとうまくいくかもしれません

- ★ Excel ではかなり無理矢理行うか，有償のアドインなどを使用しないと主成分分析を行うことができない
- ★ 別ファイルの補助教材「R 言語の基礎」も参考にしてください

演習に使うソフトウェア

★以下の3連の動作を行った際のプログラムと結果に関して以下のように表記することにします

- ★Rで 5 + 4 と入力して結果の 9 が得られた
- ★Rで 9 * 3 + 9 と入力して結果の 36 が得られた
- ★Rで mean(c(1,3,6)) と入力して結果の 3.333333 が得られた

```
> 5 + 4
[1] 9
> 9 * 3 + 9
[1] 36
> mean(c(1,3,6))
[1] 3.333333
```

★実際にRで上記の3連の動作を行ったら以下のようになります

```
> 5 + 4
[1] 9
> 9 * 3 + 9
[1] 36
> mean(c(1,3,6))
[1] 3.333333
> |
```

株価の可視化 - 演習

概要

★ 株価のデータの可視化をしてみよう

★ 株価の可視化は，ローソク足チャートやツリーマップによるものなど色々ある

★ ここでは，主成分分析を用いて，また別の可視化を考える

  ★ 扱うデータの区間において，特徴的な値動きをした業種などが明らかになるかも

★ 主成分分析の利用方法としては，「次元の縮小を行う」ではなく，「面白い軸を見つける」の方に近い内容です

株価のデータ

- ★ 適当に選んだ49銘柄の株価
  - ★ 2012年6月4日〜2013年6月3日，246日分の終値のデータ
  - ★ データを作るときの都合上，新しいデータが上になっています

- ★ PandAにある `stock.csv` を使用します
  - ★ MacやLinuxの場合は `stock-utf8.csv` を使用します（日本語の文字コードの問題，クラウド環境の場合もたぶんこちら）
  - ★ 日本語がうまく処理できずエラーになる場合は `stock-nonjp.csv` を使用してください．ただし，企業のカテゴリをうまく表示することはできなくなります

ファイルに関して

- ★ stock.csv
  - ★ http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/stock.csv
- ★ stock-utf8.csv
  - ★ http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/stock-utf8.csv
- ★ stock-nonjp.csv
  - ★ http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/stock-nonjp.csv

- ★ 京都大学の授業中であれば，PandA の授業のサイトよりダウンロードすることもできます

作業フォルダの確認とデータのダウンロード

- ★ 以下のコマンドを打ち，作業フォルダを確認します
  - ★ 必須ではないですが，ダウンロードした csv データは作業フォルダに入れておくとファイル名のみで楽に読み込みができます

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

- ★ Windows 環境の場合は，上のようにマイドキュメントがデフォルトの作業フォルダに指定されています

データの読み込み

★ 以下のコマンドを打ち，データを読み込み，z に格納します

- ★ 作業フォルダにデータが有る場合は以下のようにファイル名のみを指定すればよく，そうでない場合は，作業フォルダからの相対パスかフルパスを指定します

```
> z <- read.csv('stock.csv')
```

★ うまく読み込めたときは何も表示されません．失敗した場合はエラーメッセージが表示されます

★ フルパスを指定する場合は以下のような感じになります

```
> z <- read.csv('C:/Users/admin/Documents/stock.csv')
```

★ GUI を利用してファイルを指定する場合は以下のようにします

```
> z <- read.csv(file_choose())
```

エラーメッセージ

★ 以下のエラーメッセージが表示された場合はファイルの指定がうまくできていない場合です

- ★ file_choose() などを試してください

```
> z <- read.csv('stock.csv')
Error in file(file, "rt") :  cannot open the connection
In addition:  Warning message:
In file(file, "rt") :
  cannot open file 'stock.csv':  No such file or directory
```

★ 以下のエラーメッセージが表示された場合は日本語の処理がうまくできていない場合です

- ★ 使用するデータのファイルを変えて試してみてください

```
> z <- read.csv('stock.csv')
Error in make.names(col.names, unique = TRUE) :
  invalid multibyte string 4
```

読み込んだデータの確認

★ 以下のコマンドを打ち，読み込んだファイルの最初の崇敬を確認してみます

★ 実行結果は一部省略しています

```
> head(z)
  X1332.水産.農林業.日本水産 X1334.水産.農林業.マルハニチロ
1                        192                           180
2                        197                           184
3                        193                           179
4                        206                           192
5                        207                           191
6                        206                           191
```

主成分分析の実行

★ 主成分分析は以下のようにできます．実行結果を `res` に格納すると同時に表示しています

```
> (res <- prcomp(z))
Standard deviations (1, .., p=49):
 [1] 2.649943e+04 1.331842e+03 2.416455e+02 1.984524e+02 1.491866e+02
 [6] 1.111116e+02 9.161494e+01 9.053421e+01 7.434570e+01 6.101655e+01
[11] 5.825535e+01 4.349672e+01 4.042907e+01 3.599758e+01 3.108760e+01
[16] 2.835799e+01 2.279229e+01 2.267060e+01 2.029700e+01 1.991010e+01
[21] 1.734893e+01 1.656616e+01 1.371287e+01 1.298084e+01 1.156221e+01
[26] 1.110537e+01 9.762126e+00 8.611755e+00 8.260597e+00 6.918735e+00
[31] 6.834421e+00 6.179626e+00 5.928968e+00 5.803638e+00 5.048183e+00
[36] 4.947635e+00 4.659742e+00 4.500691e+00 4.124163e+00 3.830279e+00
[41] 3.231607e+00 3.139308e+00 2.839911e+00 2.686924e+00 2.448193e+00
[46] 2.131499e+00 1.873844e+00 1.128023e+00 8.912664e-01
```

★ 実行結果の前半は各主成分の標準偏差（データ行列の特異値）が表示されます

- ★ xey は $x \times 10^y$ という意味です

主成分分析の実行

★ 実行結果の続き

```
Rotation (n x k) = (49 x 49):
                                            PC1          PC2
X1332.水産.農林業.日本水産              -7.326671e-05 -0.004734409
X1334.水産.農林業.マルハニチロ          -5.165897e-04 -0.016612736
X1605.鉱業.国際石油開発帝石             -9.995136e-01  0.030672943
X1721.建設業.コムシスホールディングス   -2.310898e-03 -0.067031298
X1801.建設業.大成建設                   -5.939505e-04 -0.029927406
X1802.建設業.大林組                     -1.476168e-03 -0.054370341
X1803.建設業.清水建設                   -5.952337e-04 -0.029941668
X1812.建設業.鹿島                       -5.270716e-04 -0.023167777
X1925.建設業.大和ハウス工業             -7.614451e-03 -0.238071913
X1928.建設業.積水ハウス                 -4.907271e-03 -0.167394106
X1963.建設業.日揮                       -4.233828e-04 -0.151379550
X2002.食料品.日清製粉                   -2.897776e-03 -0.077281556
X2269.食料品.明治                       -6.728514e-03 -0.204439539
```

★ 実行結果の後半はデータ行列の右特異ベクトルが表示されます

特異ベクトルの意味

★ とある日の第1主成分の値（主成分スコア）は

$$
\begin{aligned}
&(-7.326671\times10^{-5})\times \text{その日の日本水産の株価}\\
&+(-5.165897\times10^{-4})\times \text{その日のマルハニチロの株価}\\
&+(-9.995136\times10^{-1})\times \text{その日の国際石油開発帝石の株価}\\
&+(-2.310898\times10^{-3})\times \text{その日のコムシスホールディングスの株価}\\
&+\cdots
\end{aligned}
$$

と計算できます

寄与率の確認

★ 以下のコマンドを打ち，寄与率を確認します

```
> summary(res)
Importance of components:
                             PC1       PC2       PC3       PC4       PC5
Standard deviation     2.650e+04 1.332e+03 241.64545 198.45236 149.18656
Proportion of Variance 9.972e-01 2.520e-03   0.00008   0.00006   0.00003
Cumulative Proportion  9.972e-01 9.998e-01   0.99984   0.99989   0.99993
                             PC6      PC7      PC8      PC9     PC10
Standard deviation     111.11159 91.61494 90.53421 74.34570 61.01655
Proportion of Variance   0.00002  0.00001  0.00001  0.00001  0.00001
Cumulative Proportion    0.99994  0.99996  0.99997  0.99998  0.99998
```

★ Proportion of Variance が全体の分散に対するその主成分の分散（寄与率）で，Cumulative Proportion が累積寄与率

★ 第1主成分だけで，「データの 99.7% ほどを説明できている」ということになっている

- ★ 株価は為替などに依存して同じように動く傾向は強いが，流石におかしいのでは？

考察

- ★ 第1主成分だけでデータをほぼ説明できている
  - ★ 第1主成分を確認すると，ほぼ「国際石油開発帝石」の株価そのまま
  - ★ データを確認してみると，「国際石油開発帝石」の株価の値が大きい
    - ★ 標準化して主成分分析を行うのが適切（定義Aではなく定義B）

主成分分析の実行

★ データを標準化して主成分分析するには以下のようにします

```
> (res <- prcomp(z, scale=TRUE))
Standard deviations (1, .., p=49):
 [1] 6.31063577 1.85920876 1.24329032 1.16367231 0.87792853
 [6] 0.65963442 0.50602945 0.48657499 0.44556829 0.34156387
[11] 0.32263711 0.30850962 0.27830524 0.26106827 0.23797584
[16] 0.21547630 0.20567634 0.19234967 0.18649644 0.16050762
[21] 0.14649219 0.13864373 0.13529567 0.12859066 0.12378839
[26] 0.10786291 0.10571236 0.10367303 0.09839855 0.09578361
[31] 0.09188565 0.08898413 0.08662859 0.07997819 0.07248399
[36] 0.06900042 0.06705625 0.06223110 0.06079621 0.05971388
[41] 0.05408175 0.05149214 0.04908353 0.04796670 0.04514064
[46] 0.04350043 0.04097994 0.03740315 0.03415902
```

主成分分析の実行

★ 実行結果の続き

```
Rotation (n x k) = (49 x 49):
                                            PC1         PC2
X1332.水産.農林業.日本水産              -0.07026615 -0.430896339
X1334.水産.農林業.マルハニチロ          -0.15438743  0.076745414
X1605.鉱業.国際石油開発帝石             -0.07800879  0.156621069
X1721.建設業.コムシスホールディングス   -0.13818907  0.158446015
X1801.建設業.大成建設                   -0.14844262  0.017616711
X1802.建設業.大林組                     -0.14985324  0.055287227
X1803.建設業.清水建設                   -0.14962615 -0.036576051
X1812.建設業.鹿島                       -0.14754408  0.009313891
X1925.建設業.大和ハウス工業             -0.15507715  0.071272307
X1928.建設業.積水ハウス                 -0.15324553  0.064673359
X1963.建設業.日揮                       -0.09640020  0.139315375
X2002.食料品.日清製粉                   -0.14891363  0.134880285
X2269.食料品.明治                       -0.15010370  0.086383841
```

寄与率の確認

★ 以下のコマンドを打ち，寄与率を確認します

```
> summary(res)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     6.3106 1.85921 1.24329 1.16367 0.87793
Proportion of Variance 0.8127 0.07054 0.03155 0.02764 0.01573
Cumulative Proportion  0.8127 0.88328 0.91483 0.94246 0.95819
                           PC6     PC7     PC8     PC9    PC10
Standard deviation     0.65963 0.50603 0.48657 0.44557 0.34156
Proportion of Variance 0.00888 0.00523 0.00483 0.00405 0.00238
Cumulative Proportion  0.96707 0.97230 0.97713 0.98118 0.98356
```

★ 第1主成分のみでは，データの分散の81.2%ほどを説明している

★ 第5主成分までで，データの分散の95%以上が説明できている

主成分の意味

★ 例えば，第1主成分は，右特異ベクトルの全ての成分の負なので，全体的に株価が高くなると，第1主成分は小さくなる

　★ 第1主成分は，全体的な株価の安さ，のようなものを表していると考えられる

★ このように右特異ベクトルを見て判断することもできるが，因子負荷量を見てみよう

第1〜3主成分の因子負荷量

★ 以下のようにコマンドを打つ

★ コマンドを打つたびに図が表示される

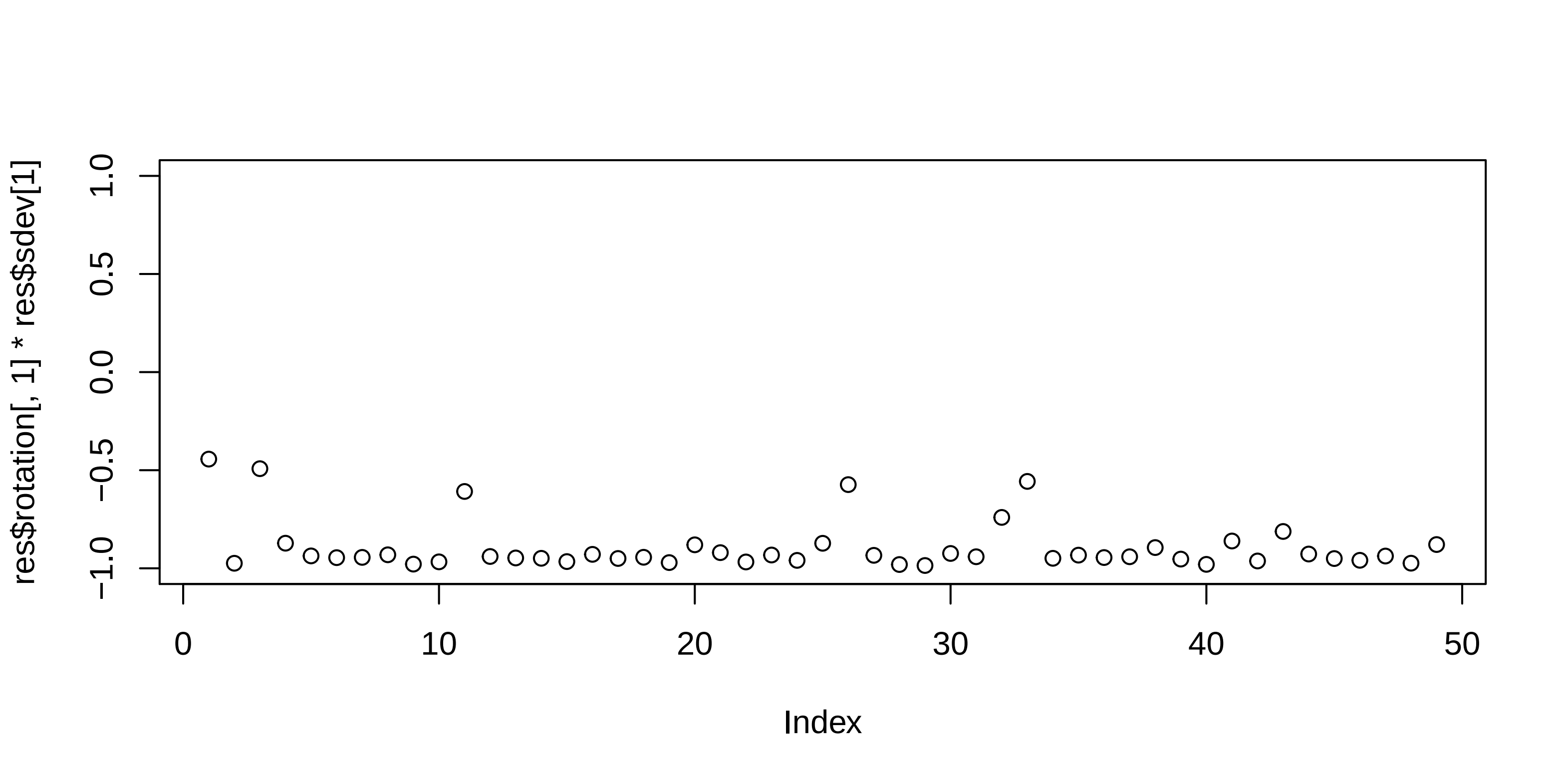
```
> plot(res$rotation[,1]*res$sdev[1], ylim=c(-1,1))
> plot(res$rotation[,2]*res$sdev[2], ylim=c(-1,1))
> plot(res$rotation[,3]*res$sdev[3], ylim=c(-1,1))
```

★ 何番がどの企業かを確認する場合は以下のコマンドを打っても良い

```
> names(z)
 [1] "X1332.水産.農林業.日本水産"
 [2] "X1334.水産.農林業.マルハニチロ"
 [3] "X1605.鉱業.国際石油開発帝石"
```

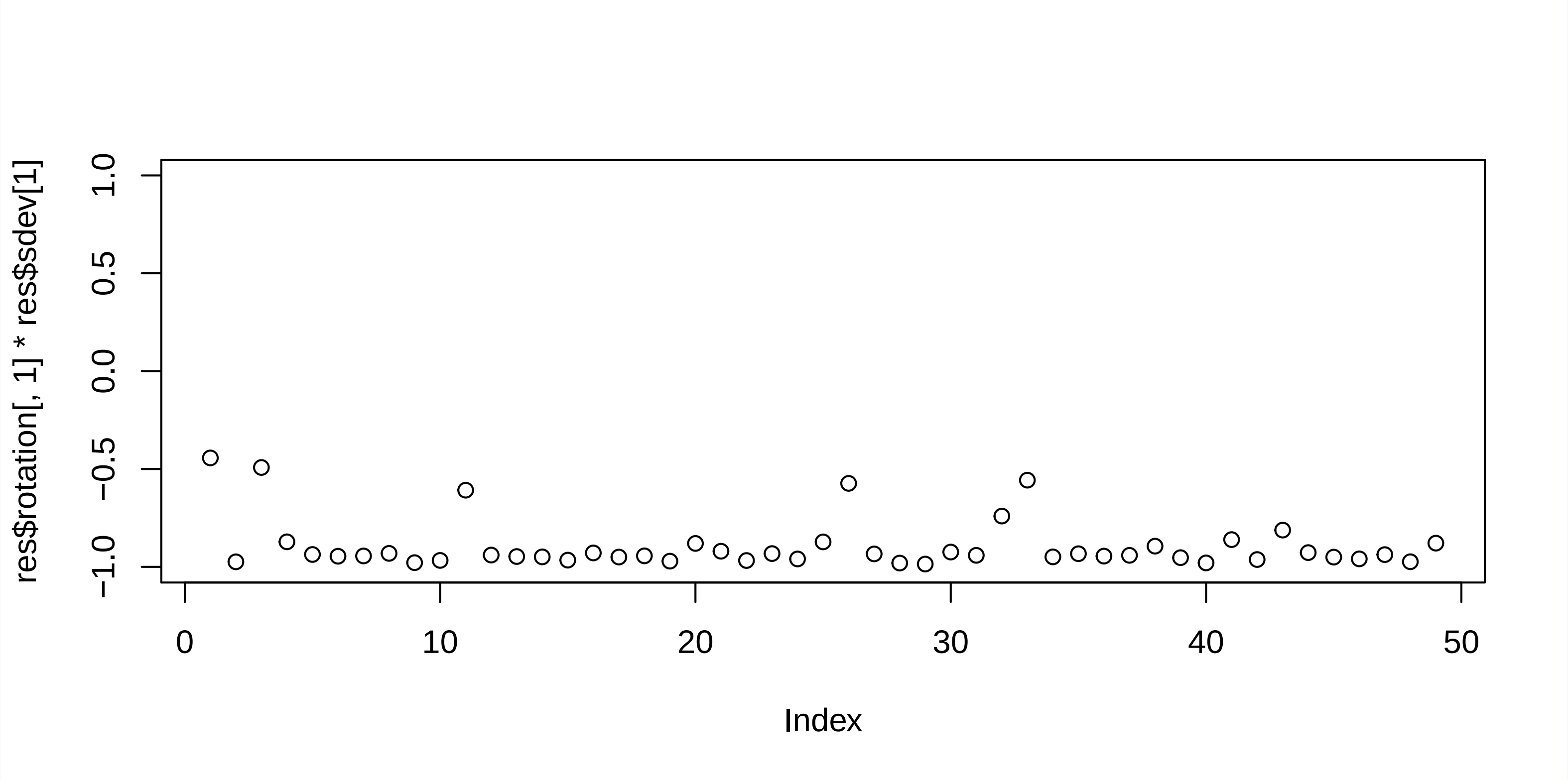
第1主成分の因子負荷量

- ★ 第1主成分の因子負荷量をプロットした結果が以下
  - ★ 第1主成分は全体的な株価の傾向を表すだろう
  - ★ 強い負の相関（ほぼ−1の企業）は全体的な傾向でほぼ価格の変動が説明できる



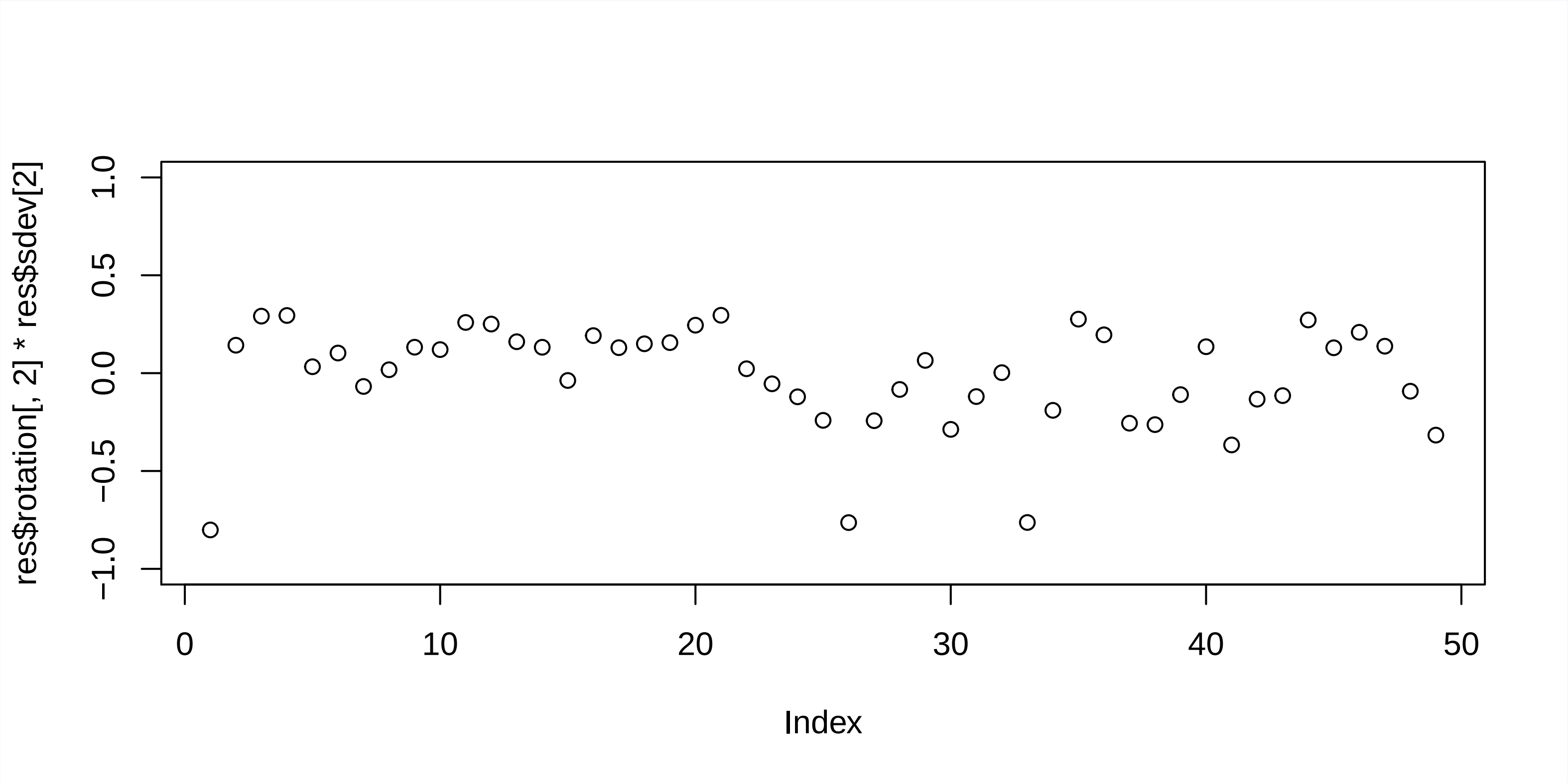
第1主成分の因子負荷量

- ★ そこまで強い相関がない5つぐらいの企業は，このデータの範囲内で，全体的な価格の変動だけではない要因があったのではないかと推測される
  - ★ 大ヒット商品を生み出した，粉飾決算が明るみに出た，等
  - ★ もともとあまり全体的な価格の変動に添わない業種



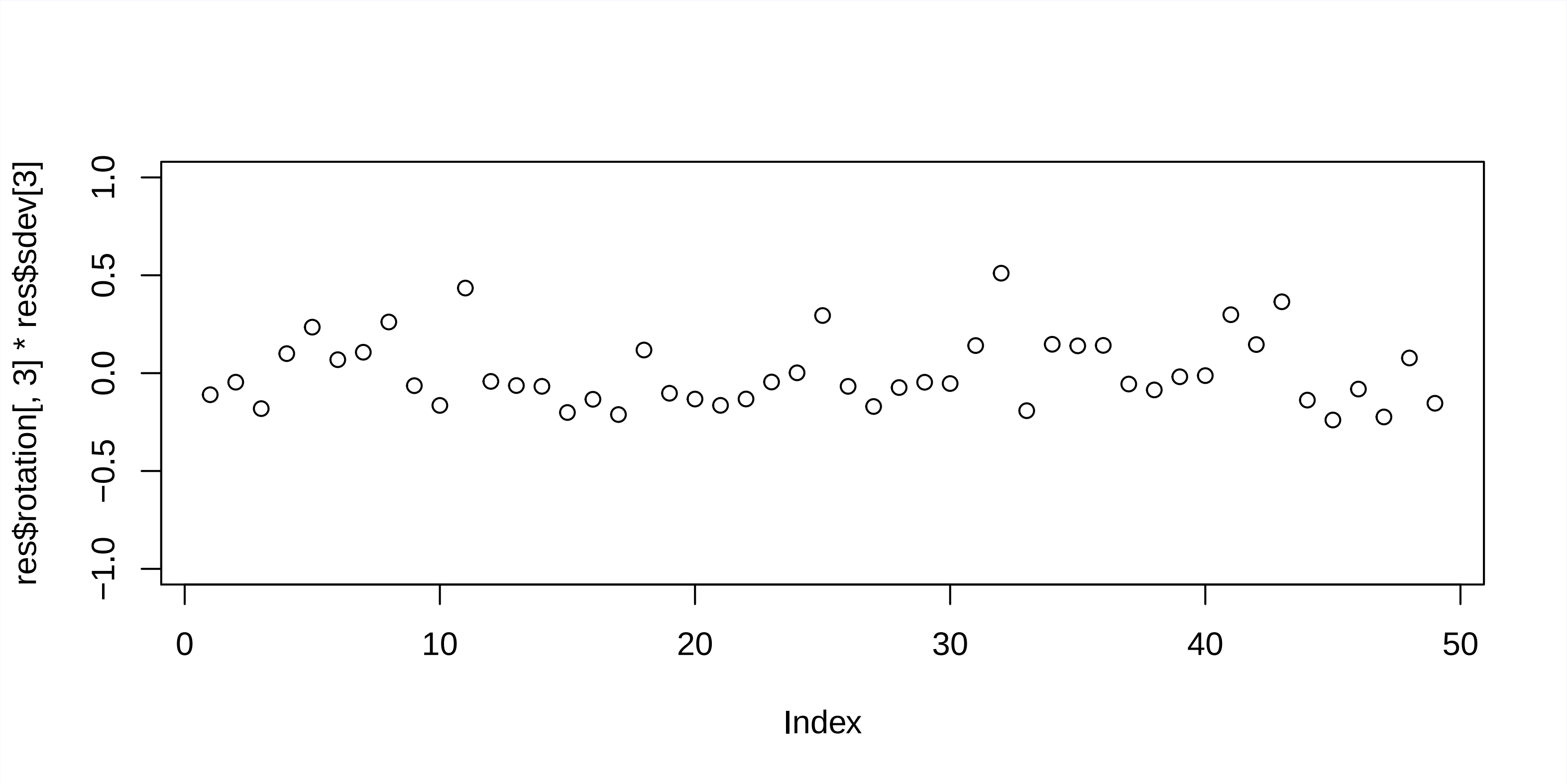
第2主成分の因子負荷量

- ★ 第2主成分以降は，特定の企業，業種などの株価の傾向を表しているのではないかと推測されるが因子負荷量を見てもよくわからない
  - ★ 第1主成分の因子負荷量が -1 に近くなかった企業のうち3つぐらいの企業は第2主成分の因子負荷量が -1 に近い



第3主成分の因子負荷量

★ 第2主成分以降は，特定の企業，業種などの株価の傾向を表しているのではないかと推測されるが因子負荷量を見てもよくわからない



二次元にプロットしてみよう

★各企業の，第1主成分，第2主成分の因子負荷量を求めて2次元にプロットしてみよう

★楽に書くために，最初に各企業，各主成分の因子負荷量を全て求めておく

```
> f <- sweep(res$rotation, MARGIN=2, res$sdev, FUN='*')
```

★プロットする際に，各企業の業種がわかるように，各企業の業種の頭文字を求めておく

```
> s <- substr(names(z),7,7)
```

★第1主成分の因子負荷量を横軸，第2主成分の因子負荷量を縦軸として各企業をプロット

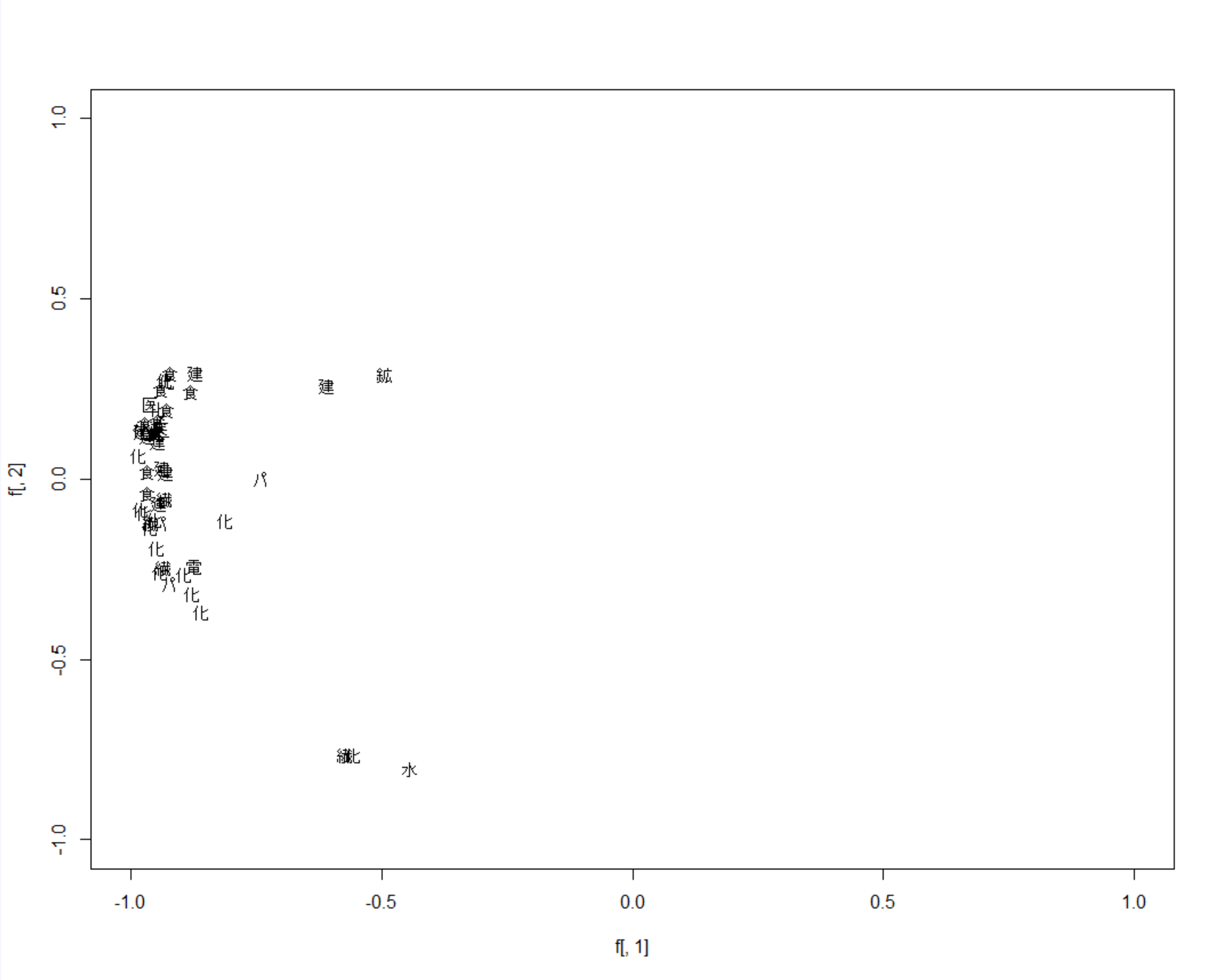
```
plot(f[,1], f[,2], xlim=c(-1,1), ylim=c(-1,1), pch=s)
```

★第2主成分の因子負荷量を横軸，第3主成分の因子負荷量を縦軸として各企業をプロット

```
plot(f[,2], f[,3], xlim=c(-1,1), ylim=c(-1,1), pch=s)
```

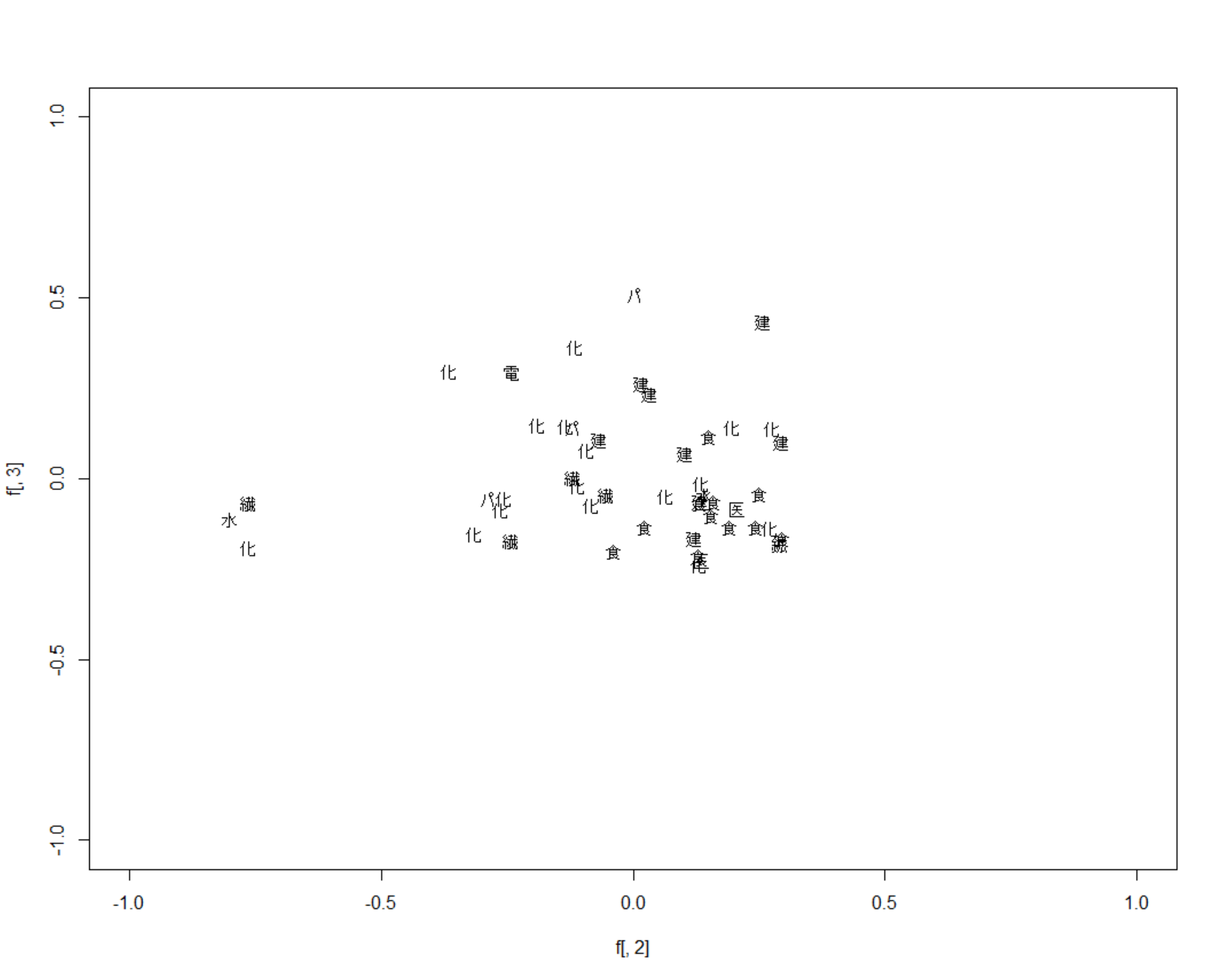
第1主成分・第2主成分の因子負荷量

★ 横軸：第1主成分の因子負荷量， 縦軸：第2主成分の因子負荷量



第２主成分・第３主成分の因子負荷量

★ 横軸：第２主成分の因子負荷量，縦軸：第３主成分の因子負荷量



文字化けする場合

- ★pch で日本語を指定した場合，環境によっては文字化けする場合があります
  - ★Windows 以外だと難しいようです

- ★文字化けする場合は，意味がわかりにくくなりますが，漢字ではなく半角英数字などに置き換えても良い

```
> s[s=='水'] <- 'I'; s[s=='鉱'] <- 'M'; s[s=='建'] <- 'C'
> s[s=='食'] <- 'F'; s[s=='パ'] <- 'P'; s[s=='化'] <- 'S'
> s[s=='繊'] <- 'B'; s[s=='医'] <- 'A'; s[s=='電'] <- 'E'
```

第2主成分・第3主成分の意味

- ★ 業種だけ見ても入り組んでいたりであまり正確なことはよくわからない
- ★ 漠然とした傾向をなんとなく読み取ると
  - ★ 第2主成分が大きい：建築業など重厚長大的な企業
  - ★ 第3主成分が小さい：食品・衣類など生活必需品に関わる企業
    - ★ 第2主成分については，あまり業種的な意味はなく，小さい値の3つの企業に共通する変動を表しているのかもしれない

時系列と主成分スコア

★ さて企業について見てきたが，時系列の方にも興味を向けてみよう

★ それぞれの日における主成分の値（主成分スコア）は以下のようにすると見ることができる

```
> res$x
             PC1         PC2          PC3         PC4
 [1,]  -8.1165794 -0.41797412  0.6371397184  3.374565497
 [2,]  -9.7428150 -1.05768507  0.8099716083  3.022418092
 [3,]  -9.3655223 -1.16547341  0.8515070369  2.793812058
 [4,] -12.1551017 -1.62468546  1.3856033873  2.867412121
 [5,] -11.8022176 -1.78037609  1.6308283401  2.916190308
 [6,] -11.7851605 -1.59718863  1.5565081370  2.863963620
 [7,] -13.9733610 -2.21627590  2.5574983967  2.908722414
 [8,] -13.3169193 -2.29250116  2.4543404653  2.135245455
 [9,] -17.6395327 -2.56295901  3.0519981597  2.300551852
[10,] -16.9945371 -2.81249762  2.2464982719  1.688080852
```

★ PC1（第１主成分）の値が小さい：その日は全体的に株価が高い，などとなっている

biplot

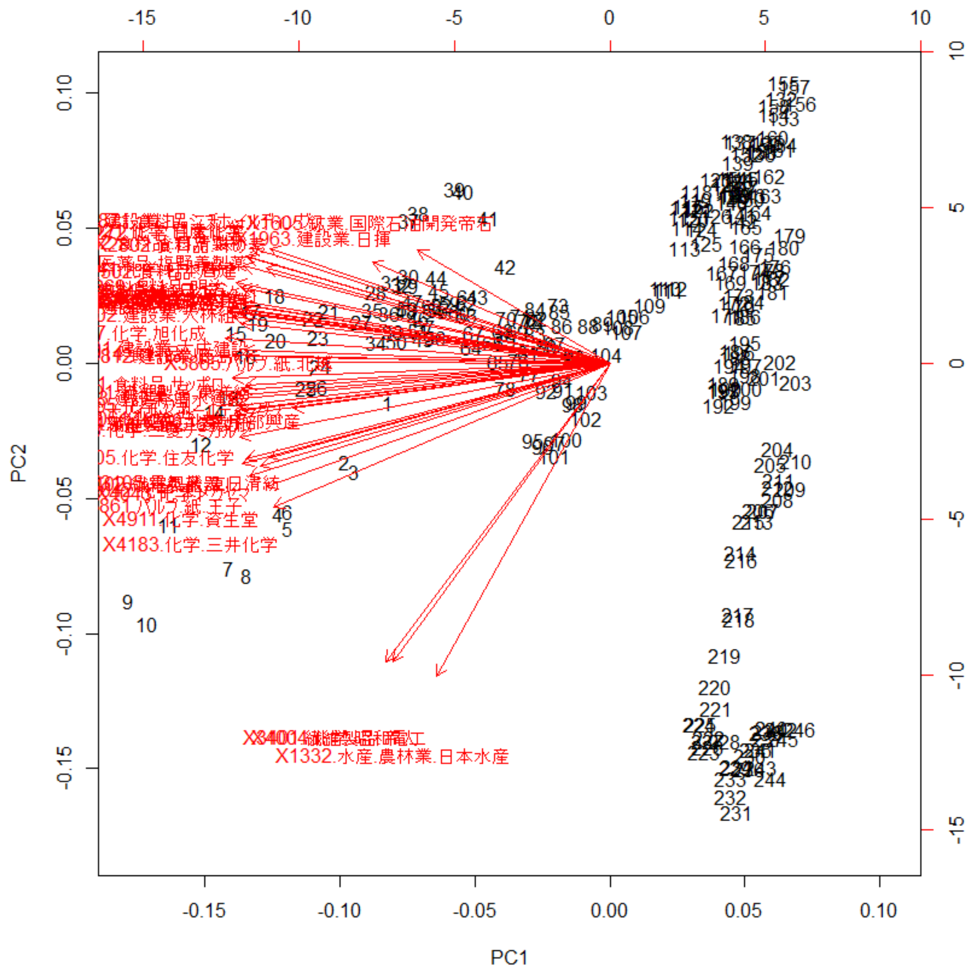
★ 右特異ベクトルの値，主成分スコアを同時に図にしてくれる biplot が便利

```
> biplot(res)
```

★ 図は次のスライド参照

- ★ 横軸は第１主成分，縦軸は第２主成分
- ★ 赤い矢印は，各企業の第１主成分，第２主成分に関する特異ベクトルの値（定数倍無視すれば因子負荷量）
- ★ 黒い数値は，その日付における第１主成分・第２主成分のスコア

★ 実行結果


PC1
PC2
-15
-10
-5
0
5
10
0.10
0.05
0.00
-0.05
-0.10
-0.15
X4183.化学.三井化学
X4911.化学.資生堂
X1332.水産.農林業.日本水産
X1963.建設業.日揮
05.化学.住友化学

読み取る

- ★ データ番号120ぐらいから10ぐらいまでにかけて第1主成分のスコアが増加→全体的に株価が上昇
  - ★ データ番号は小さい方が新しい，古いほうが大きい
- ★ データ番号230ぐらいから160らいまでにかけて第1主成分のスコアはあまり変わらず，第2主成分のスコアが上昇
  - ★ 全体的な株価はあまり変わらなかったが，建築業などが好調になったか？

- ★ などと主成分を使って，全体的な株価や，各業種の様子を可視化することができる（かも）

主成分回帰 - 演習

概要

★ 主成分を用いて回帰分析を行う主成分回帰を行ってみましょう

★ 回帰分析のときに行ったビールの売上を例題に行ってみます.

★ 主成分分析の利用方法としては,「面白い軸を見つける」ではなく,「次元の縮小を行う」の方に近い内容です

データ

- ★ 各月のビールの売上，東京の気温，京都の気温，12月かどうか，のデータ
  - ★ 詳しい概要は回帰分析の演習の資料を確認してください

- ★ PandA にある `beer.csv` を使用します

ファイルに関して

- ★ beer.csv
  - ★ http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/beer.csv

- ★ 京都大学の授業中であれば，PandA の授業のサイトよりダウンロードすることもできます

データの読み込み

★ まずは，データを読み込み，dat に格納し，中身を確認します

```
> dat <- read.csv('beer.csv', row.name='X')
> head(dat)
       beer Tokyo Kyoto December
2011/1  475   5.1   2.8        0
2011/2  625   7.0   6.3        0
2011/3  800   8.1   6.8        0
2011/4  960  14.5  12.5        0
2011/5  730  18.5  19.0        0
2011/6  980  22.8  24.1        0
```

通常の回帰分析の確認

★ R を用いて回帰分析をする場合は `lm` 関数を利用します

★ データフレーム `dat` の全てのデータを利用する場合

- ★ `lm(dat)`

★ 回帰式 $y = a_1x_1 + a_2x_2 + a_3x_3 + b$ において，対応するデータが入っている R のオブジェクトが `y`, `x1`, `x2`, `x3` の場合

- ★ `lm(y ~ x1 + x2 + x3)`

★ 結果の詳細を見る場合は，`lm` 関数が返すオブジェクトを `summary()` や `plot()` にわたします

★ 以下の場合について通常の回帰分析を行って結果を確認してみましょう

- ★ 説明変数：東京の気温
- ★ 説明変数：東京の気温と 12 月かどうか
- ★ 説明変数：全て（東京の気温と京都の気温と 12 月かどうか）

説明変数：東京の気温

★ 単回帰分析を行います

★ まずはlm関数が返すオブジェクトをそのまま表示してみます

```
> (res <- lm(dat$beer ~ dat$Tokyo))

Call:
lm(formula = dat$beer ~ dat$Tokyo)

Coefficients:
(Intercept)    dat$Tokyo
     656.93        12.99
```

説明変数：東京の気温

★ summary() にわたした結果を見ます（表示を一部省略しています）

```
> summary(res)
Residuals:
    Min      1Q  Median      3Q     Max 
-291.28 -125.89  -23.06   39.23  631.03 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  656.931     56.521  11.623   <2e-16 ***
dat$Tokyo     12.992      3.159   4.113    1e-04 ***
-
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 212.3 on 74 degrees of freedom
Multiple R-squared:  0.1861, Adjusted R-squared:  0.1751 
F-statistic: 16.92 on 1 and 74 DF,  p-value: 0.0001002
```

説明変数：東京の気温・12月の気温

★ summary() で結果を見ます

```
> summary(lm(dat$beer ~ dat$Tokyo + dat$December))
Residuals:
     Min        1Q    Median        3Q       Max
-244.605   -63.713     2.525    67.220   272.553

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   479.715     32.077   14.96   <2e-16 ***
dat$Tokyo      20.551      1.734   11.85   <2e-16 ***
dat$December  698.542     49.580   14.09   <2e-16 ***
-
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘ ’   1

Residual standard error: 110.8 on 73 degrees of freedom
Multiple R-squared:  0.7812, Adjusted R-squared:  0.7752
F-statistic: 130.3 on 2 and 73 DF,  p-value: < 2.2e-16
```

説明変数：全て

★ summary() で結果を見ます

```
> summary(lm(dat))
Residuals:
    Min      1Q  Median      3Q     Max
-237.32  -61.18   -1.22   68.11  278.10

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  495.648     41.805  11.856   <2e-16 ***
Tokyo         10.502     16.890   0.622    0.536
Kyoto          9.226     15.424   0.598    0.552
December     700.110     49.869  14.039   <2e-16 ***
-
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘ ’   1

Residual standard error: 111.3 on 72 degrees of freedom
Multiple R-squared:  0.7822, Adjusted R-squared:  0.7732
F-statistic: 86.21 on 3 and 72 DF,  p-value: < 2.2e-16
```

結果の確認

- ★ 東京の気温の係数は，それぞれのモデルによって異なる
  - ★ 説明変数間で相関関係があるため
- ★ 相関関係が非常に強い場合は推定結果が不安定になる（多重共線性）
  - ★ 東京の気温と京都の気温の間でそのような様子が観察される
- ★ 主成分同士は無相関になるからそのようなことは起こらない

主成分分析

★ 説明変数を主成分分析してみよう

★ まずは被説明変数と説明変数を分離する

```
> y <- dat[,1]
> tmp_x <- dat[,-1]
```

★ 説明変数の主成分分析を行う

```
> (pr <- prcomp(tmp_x, scale=TRUE))
Standard deviations (1, .., p=3):
[1] 1.47030392 0.91290230 0.06939582

Rotation (n x k) = (3 x 3):
                PC1        PC2          PC3
Tokyo     0.6609969 -0.2523500 -0.706684198
Kyoto     0.6615571 -0.2484996  0.707524012
December -0.3541544 -0.9351831  0.002686445
```

主成分分析の内容の確認

★ だいたい以下のような意味合いがあるように見える

- ★ 第1主成分は全国的な気温の高さ
- ★ 第2主成分は12月ではない（大きいと12月でない，小さいと12月）
- ★ 第3主成分は関東に比べて関西の気温の高さ

主成分スコアを用いて回帰分析を行う

★ 主成分回帰を行う

```
> summary(lm(y ~ pr$x))
Residuals:
    Min      1Q  Median      3Q     Max 
-237.32  -61.18   -1.22   68.11  278.10 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  866.711     12.770  67.871  < 2e-16 ***
pr$xPC1       38.500      8.743   4.404 3.63e-05 ***
pr$xPC2     -217.800     14.081 -15.467  < 2e-16 ***
pr$xPC3       -1.551    185.238  -0.008    0.993    
-
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘ ’   1

Residual standard error: 111.3 on 72 degrees of freedom
Multiple R-squared:  0.7822, Adjusted R-squared:  0.7732 
F-statistic: 86.21 on 3 and 72 DF,  p-value: < 2.2e-16
```

主成分スコアを用いて回帰分析を行う

★ 一部の主成分のみを説明変数に用いても係数が変わらないことが確認できる

```
> summary(lm(y ~ pr$x[,1] + pr$x[,2]))
Residuals:
     Min       1Q   Median       3Q      Max
-237.371  -61.132   -1.307   68.094  278.168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  866.711     12.682  68.341  < 2e-16 ***
pr$x[, 1]     38.500      8.683   4.434  3.2e-05 ***
pr$x[, 2]   -217.800     13.984 -15.574  < 2e-16 ***
-
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘ ’   1

Residual standard error: 110.6 on 73 degrees of freedom
Multiple R-squared:  0.7822, Adjusted R-squared:  0.7763
F-statistic: 131.1 on 2 and 73 DF,  p-value: < 2.2e-16
```

相関係数を用いたモデル選択

- ★ 主成分同士は無相関のため，以下のような簡易的なモデル選択を行うことがある
  - ★ 説明変数と主成分の相関係数を調べ，相関係数の絶対値が大きい方から説明変数採用する

- ★ 実際に相関係数を調べる

```
> cor(y,pr$x)
           PC1        PC2           PC3
[1,] 0.2421743 -0.8506388 -0.0004603593
```

- ★ この結果より，以下のモデルを検討する
  - ★ 説明変数は第２主成分のみ
  - ★ 説明変数は第２主成分・第１主成分
  - ★ 説明変数は第２主成分・第１主成分・第３主成分

補足：AIC

★ R の場合は，AIC なども簡単に確認できる

```
> AIC(lm(y ~pr$x[,2]))
[1] 951.988
> AIC(lm(y ~pr$x[,2] + pr$x[,1]))
[1] 935.8635
> AIC(lm(y ~pr$x[,2] + pr$x[,1] + pr$x[,3]))
[1] 937.8634
```

★ この結果からは，第 2 主成分と第 1 主成分を説明変数とするのが良さそう，と判断できる