

データ解析

第九回「主成分分析」

鈴木 大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

休講情報

6/24 は休講

今日の講義内容

- 主成分分析

構成

① 主成分分析の概要

② 実データ解析

③ レポート

主成分分析の目的

主成分分析: PCA (Principal Component Analysis) と呼ばれる。

使いどころ: 多変量データを少ない変数に要約したい。

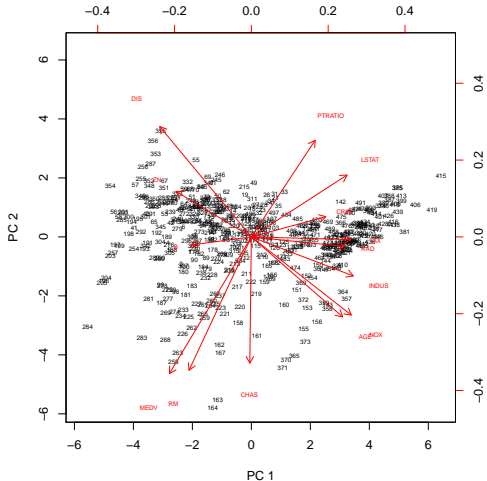
→ データの視覚化。

→ 線形回帰の前処理。

データを低い次元に落とすことを「次元削減」と言う。

主成分分析はデータ解析において「とりあえずやってみること」の一つ。

主成分分析で何が得られる？



- CRIM 各町の一人あたりの犯罪率
- ZN 宅地割合
- INDUS 非商用地の割合
- CHAS チャールズ川沿いかどうか
- NOX 一酸化窒素濃度
- RM 住居の平均部屋数
- AGE 1940 年より古くに建てられた住居の割合
- DIS ボストンのビジネス街からの距離
- RAD ハイウェイへのアクセスの良さ
- TAX 固定資産税
- PTRATIO 教師人口の割合
- B アフリカ系アメリカ人の割合を B_k としたときの $1000(B_k - 0.63)^2$
- LSTAT 低所得者層の割合
- MEDV 持ち家価格の中央値

多変量データを二次元に射影してデータを要約することができる。

主成分分析の流れ

- 1 データの標準化：中心化，分散の基準化
- 2 分散共分散行列の計算
- 3 分散共分散行列を固有値固有ベクトル分解
- 4 固有値の大きい方からいくつかの固有値固有ベクトルを取ってくる
→主成分！
- 5 主成分にデータを射影して視覚化および回帰などの処理を続行

データの形式

$$X = \left(\begin{array}{cccc} X_{1,1} & X_{1,2} & \dots & X_{1,d} \\ X_{2,1} & X_{2,2} & \dots & X_{2,d} \\ \vdots & \vdots & & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,d} \end{array} \right) \left. \vphantom{\begin{array}{cccc} X_{1,1} & X_{1,2} & \dots & X_{1,d} \\ X_{2,1} & X_{2,2} & \dots & X_{2,d} \\ \vdots & \vdots & & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,d} \end{array}} \right\} n \text{ サンプル}$$

d 次元

$$= \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix}$$

データの標準化

データの標準化

- **中心化** 元データから平均を引いて平均を 0 にする。
- **分散の基準化** 中心化したデータを標準偏差で割って、分散を 1 に基準化。

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_{ij} : \text{平均値 (の推定量)}$$

$$\hat{\sigma}_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \hat{\mu}_j)^2} : \text{標準偏差 (の推定量)}$$

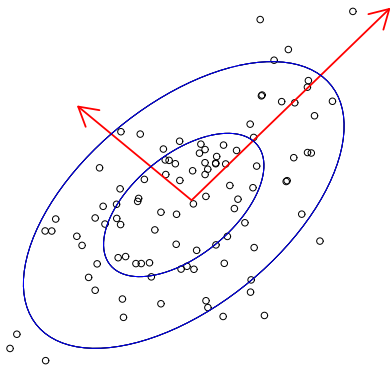
標準化：中心化して分散を 1 に基準化

$$X_{ij} \leftarrow \frac{X_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

→ 各成分は平均 0 分散 1 になる。

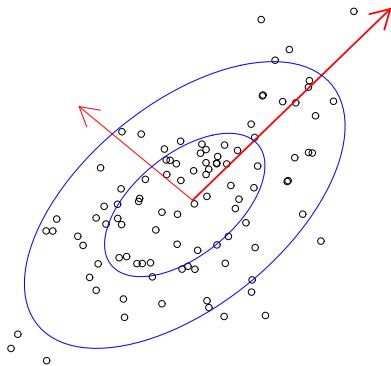
※ 主成分分析においては分散は 1 に揃えない場合も多い。

バラツキ（分散）が最大の方



第一主成分とは、バラツキが一番大きい方向である。
分散が大きい→そのデータの特徴付ける方向→データの要約

バラツキ（分散）が最大の方角

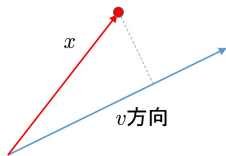


第一主成分とは、バラツキが一番大きい方角である。
分散が大きい→そのデータの特徴付ける方角→データの要約

バラツキ（分散）が最大の方向の計算

ある方向ベクトルを v ($\|v\| = 1$) とおく．この方向への x の長さは

$$v^T x$$



で求まる．

よって $v^T x_i$ の分散は

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n [v^T (x_i - \hat{\mu})]^2 &= v^T \underbrace{\left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T \right)}_{\text{分散共分散行列}} v \\ &=: v^T \Sigma v, \end{aligned}$$

である．これを最大にする方向 v を求める：

$$\max_{v: \|v\|=1} v^T \Sigma v.$$

→ 最大固有値に対応する固有ベクトルにほかならない．

分散共分散行列の固有値

Σ は (実対称) 半正定値行列 (チェックせよ)

一般に半正定値行列は直交行列で対角化可能 (固有値固有ベクトル分解):

$$\Sigma v_j = \lambda_j v_j \quad (j = 1, \dots, d),$$

ただし, v_j らは互いに直交 ($\langle v_j, v_{j'} \rangle = 0$ ($j \neq j'$)) し, $\lambda_1 \geq \dots \geq \lambda_d \geq 0$.

行列表現

$$V = [v_1, \dots, v_d], \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$$

に対して,

$$\Sigma V = V \Lambda.$$

V は直交行列なので,

$$V^T \Sigma V = \Lambda, \quad \Sigma = V \Lambda V^T$$

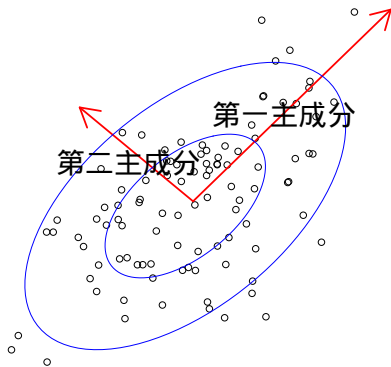
でもある.

最大固有値

$$\begin{aligned}\max_{v: \|v\|=1} v^T \Sigma v &= \max_{v: \|v\|=1} v^T V \Lambda V^T v \\ &= \max_{v: \|v\|=1} v^T \Lambda v \quad (\because \|Vv\| = 1 \Leftrightarrow \|v\| = 1) \\ &= \max_{v: \|v\|=1} \sum_{j=1}^d v_j^2 \lambda_j \\ &= \lambda_1,\end{aligned}$$

であり、最大化元は v_1 ($v_1^T \Sigma v_1 = \lambda_1$).

第二第三の主成分



v_1 (第一主成分) に直交した成分で、バラツキの一番大きな成分:

$$\max_{v: v_1 \perp v, \|v\|=1} v^T \Sigma v.$$

第二第三主成分の計算

$$\begin{aligned}\max_{v: v_1 \perp v, \|v\|=1} v^\top \Sigma v &= \max_{v: v_1 \perp v, \|v\|=1} v^\top V \Lambda V^\top v \\ &= \max_{v: v_1 \perp v, \|v\|=1} v^\top [v_1 v_2 \dots v_d] \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} [v_1 v_2 \dots v_d]^\top v \\ &= \max_{v: v_1 \perp v, \|v\|=1} v^\top [v_2 \dots v_d] \begin{pmatrix} \lambda_2 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} [v_2 \dots v_d]^\top v \\ &= \lambda_2,\end{aligned}$$

最適解は v_2 .

以下同様に第 j 主成分は j 番目の固有ベクトル v_j である。

まとめ

分散共分散行列を固有値分解して上から必要な数分だけ取ってくれば良い。

$$\Sigma = V \Lambda V^T = [v_1 \dots v_d] \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} [v_1 \dots v_d]^T.$$

固有値固有ベクトル分解

第 j 主成分: v_j

第 j 主成分スコア: $v_j^T x$ (サンプル x が第 j 主成分をどれだけ含んでいるか)

第 j 主成分の寄与率: $\frac{\lambda_j}{\sum_j \lambda_j} (> 0)$

寄与率はその主成分方向がデータの何割を表現しているかを表している。
寄与率の大きい成分から順に取ってくることでデータの良い要約を得る。
それが主成分分析。

構成

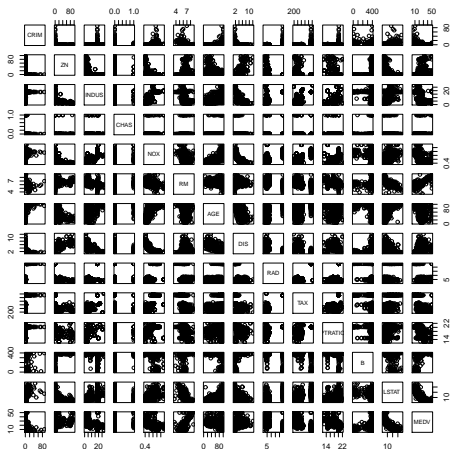
① 主成分分析の概要

② 実データ解析

③ レポート

ボストンハウジングデータ

```
x <- read.table("housing_table.data", header=T)
plot(x)
```



変数の意味

- CRIM 各町の一人あたりの犯罪率
- ZN 宅地割合
- INDUS 非商用地の割合
- CHAS チャールズ川沿いかどうか
- NOX 一酸化窒素濃度
- RM 住居の平均部屋数
- AGE 1940年より古くに建てられた住居の割合
- DIS ボストンのビジネス街からの距離
- RAD ハイウェイへのアクセスの良さ
- TAX 固定資産税
- PTRATIO 教師人口の割合
- B アフリカ系アメリカ人の割合を B_k としたときの $1000(B_k - 0.63)^2$
- LSTAT 低所得者層の割合
- MEDV 持ち家価格の中央値

変数の標準化

scale 関数で標準化可能

```
> x <- scale(x)      #標準化
> colMeans(x)       #全変数の平均 0
```

	CRIM	ZN	INDUS	CHAS	NOX	RM
	-6.899468e-18	2.298337e-17	1.516683e-17	-3.510587e-18	-2.149412e-16	-1.058524e-16
	AGE	DIS	RAD	TAX	PTRATIO	B
	-1.645039e-16	1.144506e-16	4.651527e-17	1.906139e-17	-3.931034e-16	-1.155991e-16
	LSTAT	MEDV				
	-7.012260e-17	-1.379311e-16				

```
> diag(var(x))      #全変数の分散 1
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO
	1	1	1	1	1	1	1	1	1	1	1
	B	LSTAT	MEDV								
	1	1	1								

分散共分散行列の計算

```
> Sigma = cov(x) #分散共分散行列を計算.  
> Sigma[1:5,1:5]
```

	CRIM	ZN	INDUS	CHAS	NOX
CRIM	1.00000000	-0.20046922	0.40658341	-0.05589158	0.42097171
ZN	-0.20046922	1.00000000	-0.53382819	-0.04269672	-0.51660371
INDUS	0.40658341	-0.53382819	1.00000000	0.06293803	0.76365145
CHAS	-0.05589158	-0.04269672	0.06293803	1.00000000	0.09120281
NOX	0.42097171	-0.51660371	0.76365145	0.09120281	1.00000000

分散共分散行列の固有値固有ベクトル分解

```
> res <- eigen(Sigma) #分散共分散行列を固有値固有ベクトル変換
```

```
> res
```

```
$values
```

```
[1] 6.54598958 1.64953191 1.34890592 0.88653987 0.85089944 0.66001077 0.53  
[13] 0.13400970 0.06032666
```

```
$vectors
```

	[,1]	[,2]	[,3]	[,4]	[,5]	
[1,]	0.242284451	0.065873108	0.395077419	0.100366211	0.004957659	-0.2
[2,]	-0.245435005	0.148002653	0.394545713	0.342958421	0.114495002	-0.3
[3,]	0.331859746	-0.127075668	-0.066081913	-0.009626936	-0.022583692	-0.0
[4,]	-0.005027133	-0.410668763	-0.125305293	0.700406497	-0.535197817	0.1

`res$values` は固有値. 降順に並んでいる.

`res$vectors` は固有ベクトルを並べた行列 (V のこと).

固有値・固有ベクトルのチェック

固有値・固有ベクトルの性質をチェック

```
> norm(Sigma - res$eigenvectors %*% diag(res$eigenvalues) %*% t(res$eigenvectors)) #確認
```

```
[1] 2.668005e-14
```

```
> tmp <- (res$eigenvectors %*% t(res$eigenvectors)); tmp[1:5,1:5]
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.000000e+00	-5.551115e-17	1.040834e-16	2.602085e-18	-3.729655e-17
[2,]	-5.551115e-17	1.000000e+00	-1.942890e-16	-1.170938e-16	2.064321e-16
[3,]	1.040834e-16	-1.942890e-16	1.000000e+00	-8.239937e-17	-6.834810e-16
[4,]	2.602085e-18	-1.170938e-16	-8.239937e-17	1.000000e+00	-6.591949e-17
[5,]	-3.729655e-17	2.064321e-16	-6.834810e-16	-6.591949e-17	1.000000e+00

```
> tmp <- (t(res$eigenvectors) %*% res$eigenvectors); tmp[1:5,1:5]
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.000000e+00	1.387779e-16	3.191891e-16	4.857226e-17	2.081668e-17
[2,]	1.387779e-16	1.000000e+00	-4.024558e-16	6.938894e-17	-3.122502e-17
[3,]	3.191891e-16	-4.024558e-16	1.000000e+00	-5.551115e-17	-5.204170e-17
[4,]	4.857226e-17	6.938894e-17	-5.551115e-17	1.000000e+00	-3.070461e-16
[5,]	2.081668e-17	-3.122502e-17	-5.204170e-17	-3.070461e-16	1.000000e+00

※ 対角行列の場合、固有ベクトルは直交行列をなす。

固有値・固有ベクトルのチェック 2

固有値・固有ベクトルの性質をチェック

```
> norm(Sigma %*% res$vector - res$vector %*% diag(res$value)) #確認  
[1] 2.207262e-14
```

$$\begin{aligned} \Sigma v_j &= \lambda_j v_j \quad (\forall 1 \leq j \leq d) \\ \Rightarrow \Sigma V &= V\Lambda, \end{aligned}$$

$$\text{ただし } V = [v_1, \dots, v_d], \quad \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}.$$

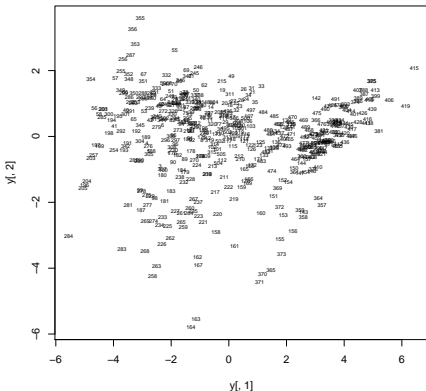
主成分分析

```
> #主成分分析  
> Lam <- res$values  
> V <- res$vectors  
> y = x %*% V #主成分スコアの計算  
> dim(y) # n × d  
[1] 506 14
```

$$y_i^T = x_i^T V = [x_i^T v_1, \dots, x_i^T v_d],$$
$$y = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix}.$$

第一， 第二主成分スコアのプロット

- > plot(y[,1],y[,2],type='n') #第一， 第二主成分スコアをプロット
- > text(y[,1],y[,2],seq(length=nrow(y)),cex=0.5)



主成分スコア (Y の各列) は互いに無相関.

> round(cov(y),10) #y は無相関

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	6.54599	0.000000	0.000000	0.000000	0.000000
[2,]	0.00000	1.649532	0.000000	0.000000	0.000000
[3,]	0.00000	0.000000	1.348906	0.000000	0.000000
[4,]	0.00000	0.000000	0.000000	0.8865399	0.000000
[5,]	0.00000	0.000000	0.000000	0.000000	0.8508994

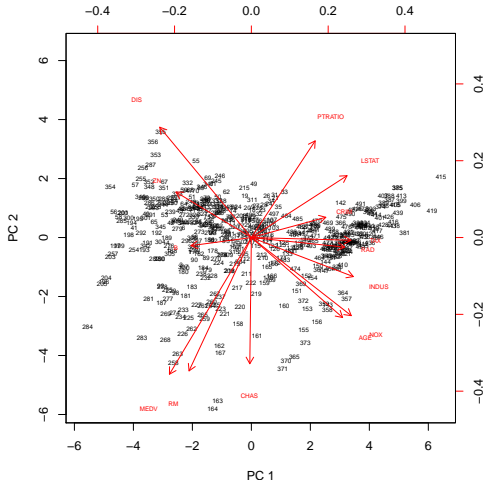
$$y = XV \in \mathbb{R}^{n \times d}$$

$$\Rightarrow y^T y = V^T X^T X V = V^T (\Sigma) V = V^T (V \Lambda V^T) V = (V V^T) \Lambda (V^T V) = \Lambda.$$

主成分スコアと軸のプロット

```
biplot(y[,c(1,2)],V[,c(1,2)],cex=0.5) #第一，第二主成分スコア
```

第一・第二主成分スコア



$$\text{各点} : y_i^T = [y_{i,1}, y_{i,2}] = [x_i^T v_1, x_i^T v_2]$$

$$\text{矢印の方向} : u_j = [v_{1,j} v_{2,j}] = e_j^T [v_1, v_2].$$

矢印は各変数が主成分の上でどの方向を向いているかを示している。

- CRIM 各町の一人あたりの犯罪率
- ZN 宅地割合
- INDUS 非商用地の割合
- CHAS チャールズ川沿いかどうか
- NOX 一酸化窒素濃度
- RM 住居の平均部屋数
- AGE 1940年より古くに建てられた住居の割合
- DIS ボストンのビジネス街からの距離
- RAD ハイウェイへのアクセスの良さ
- TAX 固定資産税
- PTRATIO 教師人口の割合
- B アフリカ系アメリカ人の割合を B_k としたときの $1000(B_k - 0.63)^2$
- LSTAT 低所得者層の割合
- MEDV 持ち家価格の中央値

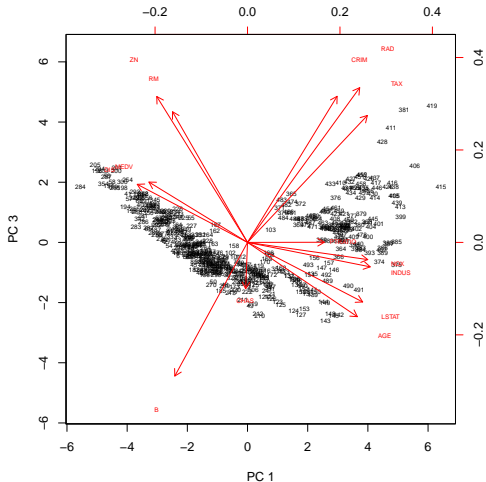
第一・第二主成分スコアの考察

第一主成分は、TAX や RAD, INDUS が大きく寄与していて、主に住環境に関する情報が乗っていると考えられる。

第一主成分が大きいほど、産業地域のようなあまり居住に適さない住環境。

第二主成分は CHAS や MEDV, RM の寄与が大きく、住宅の質の良さ（「いい家」かどうか）を表している。

第一・第三主成分スコア



各点: $y_i^T = [y_{i,1}, y_{i,2}] = [x_i^T v_1, x_i^T v_2]$

矢印の方向: $u_j = [v_{1,j}, v_{2,j}] = e_j^T [v_1, v_2]$.

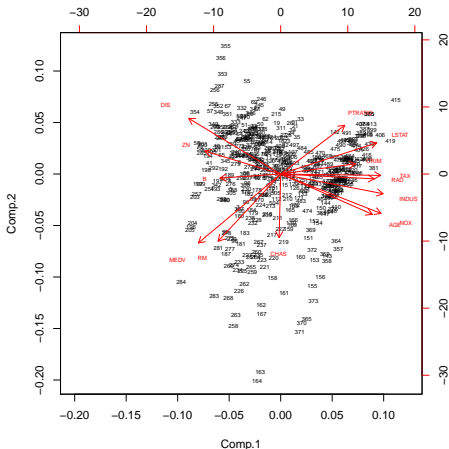
矢印は各変数が主成分の上でどの方向を向いているかを示している。

- CRIM 各町の一人あたりの犯罪率
- ZN 宅地割合
- INDUS 非商用地の割合
- CHAS チャールズ川沿いかどうか
- NOX 一酸化窒素濃度
- RM 住居の平均部屋数
- AGE 1940年より古くに建てられた住居の割合
- DIS ボストンのビジネス街からの距離
- RAD ハイウェイへのアクセスの良さ
- TAX 固定資産税
- PTRATIO 教師人口の割合
- B アフリカ系アメリカ人の割合を B_k としたときの $1000(B_k - 0.63)^2$
- LSTAT 低所得者層の割合
- MEDV 持ち家価格の中央値

Rの関数で主成分分析

princomp で主成分分析, prcomp もほぼ同じ. eigen か svd のどちらを使うかの違い.

- > PCA <- princomp(x,cor=TRUE) %相関行列を用いて PCA
- > biplot(PCA)



さっきと表示が違う... ?

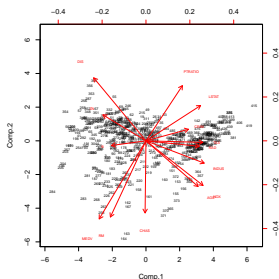
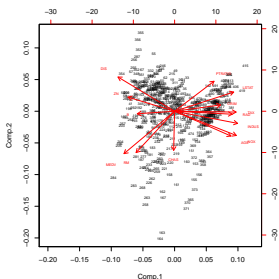
主成分スコアのスケージング

主成分スコアの分散を基準化：

$$z_{i,j} = y_{i,j} / \sqrt{\lambda_j}$$

第 j 主成分スコアの分散：

$$\text{var}(y_{:,j}) = y_{:,j}^\top y_{:,j} = V_{:,j}^\top X X^\top V_{:,j} = \sqrt{\Lambda_{j,j}} = \sqrt{\lambda_j}.$$



`biplot(PCA,scale=1,cex=0.5)` `biplot(PCA,scale=0,cex=0.5)`

`princomp` オブジェクトに対しては `scale=1` がデフォルト。

画像認識：固有顔



顔画像データから上位20個の主成分を生成.

<http://www.kixor.net/school/2008spring/comp776/assn3/>

構成

① 主成分分析の概要

② 実データ解析

③ レポート

レポート問題 (第四回)

- 1 自分の興味のあるデータで講義第八回に紹介した検定手法を当てはめてみよう。もしデータが見つからなければ「autopoll.csv」で分散分析を行ってみよう。
- 2 データ x を受け取って第 k 番目までの主成分スコア, 対応する主成分負荷 (λ_j) および主成分のリストを返す関数を作成せよ。
 - Y_k : k 番目までの主成分スコア
 - L_k : k 番目までの主成分負荷 ($\lambda_1, \dots, \lambda_k$)
 - V_k : k 番目までの主成分ベクトルを並べた行列
- 3 上で得た自作の関数と R の組み込み関数 `princomp` の戻り値とを比較せよ。
- 4 「ken-c-kakou.csv」にあるデータを用いて主成分分析を行え。なお、このデータは総務省統計局「統計でみる都道府県のすがた。経済基盤」から抜粋した：
<http://www.stat.go.jp/data/k-sugata/index.htm>
<http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001052235&cycode=0>
データの読み込みは `x <- read.csv("ken-c-kakou.csv", header=TRUE)` で可能。
- 5 (optional) 自分が興味のあるデータで主成分分析を行え。

レポートの提出方法

- 私宛にメールにて提出。
- 件名に 必ず 「データ解析第 n 回レポート」と明記し、R のソースコードと結果をまとめたレポートを送付のこと。
- 氏名と学籍番号も忘れず明記すること。
- レポートは本文に載せても良いが、pdf などの電子ファイルにレポートを出力して添付ファイルとして送付することが望ましい (これを期に tex の使い方を覚えることを推奨します)。
- 提出期限は講義最終回まで。

※相談はしても良いですが、コピペは厳禁です。

講義情報ページ

<http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html>