応用統計学 2017 第3回 主成分分析

2017年10月11日 (水)
清 智也 sei@mist.i.u-tokyo.ac.jp
講義用ページ http://ur0.pw/yTzt

- 主成分分析：特異値分解，個体と変量，中心化，標準化，主成分得点，寄与率，因子負荷量，バイプロット[1]。

演習問題

データ行列 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ の各列は変量を表し，各行は個体を表すものとする。また

$$\boldsymbol{X} = (x_{ti}) = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p) = \begin{pmatrix} \boldsymbol{x}^{(1)\prime} \\ \vdots \\ \boldsymbol{x}^{(n)\prime} \end{pmatrix}$$

のようにベクトル $\boldsymbol{x}_i$, $\boldsymbol{x}^{(t)}$ を定義する[2]。

問題 3-1. 次の行列を特異値分解せよ。ただし $h \geq 0$ は定数とする。

$$\boldsymbol{X} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ -1 & -1 & h \\ -1 & -1 & -h \end{pmatrix}.$$

h の値によって結果はどのように変化するか。

問題 3-2. 授業のアンケート結果のうち，通学時間，睡眠時間，小説，期待値，1円玉枚数の5変数について，計算機を用いて主成分分析を行ってみよ。具体的には，データを標準化した後，主成分得点，回転行列，寄与率を求め，バイプロットを描いてみよ。また可能ならば結果を解釈せよ。

問題 3-3. 実行列の分解のうち，統計学でよく用いられるものとして

スペクトル分解，特異値分解，Cholesky 分解，QR 分解

がある。これらを適用できる行列のクラス（任意の行列，正方行列，対称行列，正定値行列）を答えよ。ついでに Jordan 標準形，Schur 標準形，LU 分解，Sylvester 標準形についても答えよ[3]。

---

[1] Principal component analysis: singular value decomposition, individual & variate, centering, standardization, principal component score, proportion of variance, factor loading, biplot.
[2] 柴田里程「データ分析とデータサイエンス」近代科学社。
[3] 伊理正夫「一般線形代数」岩波書店。

問題 3-4. $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}' \in \mathbb{R}^{n\times n}$ とおく。$\boldsymbol{X}$ の主成分得点は $\boldsymbol{K}$ だけから求まることを示せ。この考えを拡張して，個体間の類似度を表す半正定値行列 $\boldsymbol{K}$ を直接定義し，主成分分析を行う手法をカーネル主成分分析という[4]。関連する手法として，非類似度に対応する距離行列を出発点とする多次元尺度構成法がある[5]。

問題 3-5 (Eckart-Young の定理；やや難しい)**.** $n \times p$ 行列 $\boldsymbol{X}$ の特異値分解を $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ とし，特異値は大きい順に並んでいるものとする。また $1 \le k \le \min(n, p)$ を 1 つ固定する。このとき

$$\begin{aligned} &\text{Minimize} \quad \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}^2 \\ &\text{subject to} \quad \operatorname{rank}(\boldsymbol{Y}) \le k \end{aligned}$$

の最適解は，$\boldsymbol{Y} = \boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{V}_k'$ で与えられることを示せ。ただし $\boldsymbol{U}_k$ と $\boldsymbol{V}_k$ はそれぞれ $\boldsymbol{U}$ と $\boldsymbol{V}$ の最初の k 列を表し，$\boldsymbol{D}_k$ は $\boldsymbol{D}$ の左上 $k \times k$ 部分行列を表す。また $\|\boldsymbol{A}\|_{\mathrm{F}} = (\operatorname{tr}(\boldsymbol{A}'\boldsymbol{A}))^{1/2}$ は Frobenius ノルムと呼ばれる。

宿題 3

問題 3- 6. Times Higher Education の World University Rankings のサイトにアクセスし，世界の上位 50 大学の “performance breakdown” （Teaching, International Outlook, Research, Citations, Industry Income）を変量として主成分分析をおこなってみよ。

先週の宿題に関する補足　R 言語では，データの要約は関数 `summary` で得られる：

```
> X = read.csv("questionnaire2017.csv")
> summary(X)
```

出力結果は各自で確認されたい。得られたオブジェクト `X` は `data.frame` と呼ばれるクラスに属す。`data.frame` クラスは行列とリストの両方の扱いが可能である：

```
> class(X)      # class of X
> X[,3]         # 3rd column of X as a matrix
> X[[3]]        # 3rd element of X as a list
> lapply(X, class)   # class of each element of X
```

`X` の第 1 列は `factor` というクラスのオブジェクトになっている。しかし第 1 列はタイムスタンプ（アンケートの入力時刻）を表しており，時刻として扱いたい。そのような場合は，`POSIXct` と呼ばれるクラスに変換するとよい：

```
> timeStamp = as.POSIXct(X[,1], format="%Y/%m/%d %H:%M:%S")  # convert
> summary(timeStamp)
> plot(timeStamp, X[,10], type="h")   # time profile of 1-yen data
```

なお `factor` クラスは離散的な変量を扱うためのクラスであり，分割表や，ダミー変数を用いた回帰式を求める際に便利である：

```
> table(X[,c(6,7)])   # contingency table
> lm(X[,3] ~ X[,4])   # regression
```

[4]福水健次「カーネル法入門」朝倉書店。
[5]宮川「統計技法」5.4 節。