クレジット：

Mathematics and Informatics Center 文科系のための線形代数・解析II
2020 藤堂 眞治・松尾 泰・藤原 毅夫

主成分分析 (Principal Component Analysis)

多くのデータが与えられ，それらが互いに相関のある多数の変数からなっているとき，互いに相関の無い少数の変数（主成分）で全体のデータを記述することが望まれる．そのような成分を探す方法の1つが主成分分析である．

標準化されたデータ

N個のサンプルについて其々P個の変数（データ）を持つデータ群が与えられていて，サンプル i (i=1~ N) についての生のデータを $t_{i\alpha}\ (\alpha = 1 \sim P)$ とする．これに関して（N 行 P 列 の） データ行列 X にまとめられているとする：

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NP} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} \quad , \quad \mathbf{x}_i = (x_{i1},\ x_{i2},\ \cdots,\ x_{iP}) \quad .$$

各サンプルの特徴（データ）が，P次元変数空間におけるベクトル $\mathbf{x}_i\ (i = 1 \sim N)$ としてあらわさられている．

簡単のために、このデータは，データの中心を0と定め，値はそれぞれの標準偏差で規格化されているとする（標準化）：

$X = \{x_{ij}\}$ ， $i = 1 \sim N$， $j = 1 \sim P$ ，平均値 $\frac{1}{N}\sum_{m=1}^{N} x_{ms} = 0$ ， 標準偏差 $\frac{1}{N-1}\sum_{m=1}^{N} x_{ms}^2 = 1$

共分散行列

共分散 $\sigma_{x_{s_1}x_{s_2}} = \frac{1}{N-1}\sum_{m=1}^{N} x_{ms_1}x_{ms_2}$ ， $\sigma_{x_{s_1}x_{s_1}} = \sigma_{x_{s_1}}^2 = 1$ を各成分とする $P \times P$ 行列

$$Q = \frac{1}{N-1}X^T X = \begin{pmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_P} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & \cdots & \sigma_{x_2x_P} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_Px_1} & \sigma_{x_Px_2} & \cdots & \sigma_{x_Px_P} \end{pmatrix} = \begin{pmatrix} 1 & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_P} \\ \sigma_{x_2x_1} & 1 & \cdots & \sigma_{x_2x_P} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_Px_1} & \sigma_{x_Px_2} & \cdots & 1 \end{pmatrix}$$

を 共分散行列 と呼ぶ．標準化したデータの共分散は相関係数と一致する．（共分散行列の対角化）．

共分散行列は（定数を別にすれば）データ行列のグラム行列である．

共分散行列の固有値

正規直交行列

$$V = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_P) = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1P} \\ v_{21} & v_{22} & \cdots & v_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ v_{P1} & v_{P2} & \cdots & v_{PP} \end{pmatrix}$$

を用いて

$$Z = XV \quad , \quad (Z)_{m\alpha} = (XV)_{m\alpha} = \sum_{\beta=1}^{P} x_{m,\beta} v_{\beta,\alpha}$$

という変換を行おう．これは各データを特徴づける $1 \sim P$ の変数の線形結合で新しい変数 $Z_{m\alpha}$ を定義したことになる．

V は共分散行列 Q を対角にするように，

$$V^T Q V = V^T \ \frac{X^T X}{N-1} \ V = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_P^2 \end{pmatrix} = \Sigma$$

と決める．そうすれば，上の式は

$$QV = V\Sigma \quad , \quad Q\mathbf{v}_\alpha = \sigma_\alpha^2 \mathbf{v}_\alpha \ , \quad \sum_{\alpha=1}^{P} \sigma_\alpha^2 = P$$

となり，V を決める問題は Q の固有値問題であることが分かる．つまり主成分分析ではまず共分散行列の固有空間への分解を行う．

主成分分析の目的

与えられた N個の各データが P 個の独立な変数で特徴付けられている．

${}^tXX$ の特異値を順次大きい順番で考えると，大きい固有値が行列 ${}^tXX$ の特徴を決めることが分かる．

すなわち小さい固有値を無視すれば，少しの変数でデータの本質を捉えられると期待できる．

データを主成分分析の各成分で表す

各サンプル（k-サンプルを考える）の元データが

$$\mathbf{x}_k = (x_{k1}, x_{k2}, \cdots, x_{kP})$$

と与えられた．これを各主成分で表してみよう．$X = U\Lambda V^T$ である（下のまとめ（５））から，

$$x_{kl} = \sum_{\nu} U_{k\nu} \lambda_\nu V_{l\nu} = \sum_{\nu} \lambda_\nu u_{k\nu} v_{l\nu}$$

が得られる．勿論、上の議論を総て正確に実行すればこの等式は成り立つ．一方，主成分分析の主目的に立ち返って最初の数個（大きな方から2個としよう）の主成分で表現することにするなら，

$$\hat{x}_{kl} = \lambda_1 u_{k1} v_{l1} + \lambda_2 u_{k2} v_{l2}$$

であり，残差（このように表した値と正確な値の差）の l 成分は $x_{kl} - \hat{x}_{kl}$ である．

あるいは各データ $\mathbf{x}_i \ (i = 1 \sim N)$ が作る行列（データ行列）は

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = ZV^{-1} = Z^t V = Z \begin{pmatrix} {}^t\mathbf{v}_1 \\ {}^t\mathbf{v}_2 \\ \vdots \\ {}^t\mathbf{v}_P \end{pmatrix} = \begin{pmatrix} \sum_\alpha Z_{1\alpha} {}^t\mathbf{v}_\alpha \\ \sum_\alpha Z_{2\alpha} {}^t\mathbf{v}_\alpha \\ \vdots \\ \sum_\alpha Z_{N\alpha} {}^t\mathbf{v}_\alpha \end{pmatrix} = \begin{pmatrix} \sum_\alpha Z_{1\alpha} v_{1\alpha} & \cdots & \sum_\alpha Z_{1\alpha} v_{P\alpha} \\ \sum_\alpha Z_{2\alpha} v_{1\alpha} & \cdots & \sum_\alpha Z_{2\alpha} v_{P\alpha} \\ \vdots & \ddots & \vdots \\ \sum_\alpha Z_{N\alpha} v_{1\alpha} & \cdots & \sum_\alpha Z_{N\alpha} v_{P\alpha} \end{pmatrix},$$

$$\mathbf{x}_l = \sum_\alpha Z_{l\alpha} {}^t\mathbf{v}_\alpha$$

となる．ここで係数 $Z_{l\alpha}$ は既に与えたように，$(Z)_{m\alpha} = \sum_{\beta=1}^{P} x_{m,\beta} v_{\beta,\alpha}$ である．これは $\mathbf{x}_m$ から $\mathbf{v}_\alpha$ への射影成分の大きさである．

最後の式は元のデータを主成分分析で得られたベクトル空間のベクトル成分で表したものである．また主成分のベクトルをもとの性質 $1 \sim P$ の空間で表せば次のようになる：

$${}^t\mathbf{v}_\alpha = (v_{1\alpha}, v_{2\alpha}, \cdots, v_{P\alpha}) \ .$$

まとめ

（1）Qの固有値 $\sigma_\alpha^2 \ (\alpha = 1 \sim P)$ を対角に並べた行列 $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_P)$ は

共分散行列 Q を対角化（固有空間に分解）したものである．

（2）行列 ${}^tXX$ の固有値問題；${}^tXX\mathbf{v}_\alpha = \lambda_\alpha^2 \mathbf{v}_\alpha$ の固有値 $\{\lambda_\alpha^2\}$ が大きな順に $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P > 0$ と並べられているとする．$(\lambda_\alpha, \mathbf{v}_\alpha)$ を 第α主成分}という．

（3）上の諸量をまとめて，

$$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_P)$$

$$\mathbf{u}_\alpha = \frac{1}{\lambda_\alpha} X\mathbf{v}_\alpha$$

$$U = (\mathbf{u}_1, \cdots, \mathbf{u}_P)$$

を定義する（$\lambda_\alpha^2 = (N-1)\sigma_\alpha^2$）．

$X\mathbf{v}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$

${}^tX\mathbf{u}_\alpha = \lambda_\alpha \mathbf{v}_\alpha$

などは定義から直ぐに証明できる．下の式の証明も難しくない．

（4）$\mathbf{v}_\alpha$ は正規直交ベクトルだから，$\mathbf{u}_\alpha$ も正規直交ベクトル（$U^TU = E$ ：$P \times P$ 単位行列）．

（5）上から以下の式が成り立つことが分かる：

$$XV = (X\mathbf{v}_1, \cdots, X\mathbf{v}_P) = (\lambda_1\mathbf{u}_1, \cdots, \lambda_P\mathbf{u}_P) = (\mathbf{u}_1, \mathbf{u}_2, \cdots)\Lambda = U\Lambda$$.

$$X = U\Lambda\,{}^tV$$

（6）また行列 ${}^tXX$、$X{}^tX$ のスペクトル分解；

$${}^tX\ X = \sum_{\alpha=1}^{P} \lambda_\alpha^2 \mathbf{v}\ {}^t\mathbf{v}\ ,$$

$$X\ {}^tX = \sum_{\alpha=1}^{P} \lambda_\alpha^2 \mathbf{u}\ {}^t\mathbf{u}$$

が成り立つ．

課題 1

上の「まとめ」の諸式を証明せよ．

主成分分析の目的

与えられたN個の各データがP個の変数で特徴付けられている．${}^tXX$ の特異値（固有値）を順次大きい順番で考えると，大きい固有値が行列X^TXの特徴を決めることが分かる（スペクトル分解）．小さい固有値を無視すれば，少しの変数でデータの本質を捉えられると期待できる．

主成分分析の例

MATLABが提供しているデータ hald （セメントの発熱と混合成分）を例にとろう．

13 種類のセメント組成に対して，ingredients 部分（4つの列）には4 種類のセメント原料の組成率が与えられている．この部分をデータ行列 T とする．

```
T=[7 26 6 60;1 29 15 52;11 56 8 20;11 31 8 47;7 52 6 33;11 55 9 22;3 71 17 6; 1 31 22 44;2 54 1
  21 47 4 26;1 40 23 34;11 66 9 12;10 68 8 12]
```

```
T = 13×4
     7    26     6    60
     1    29    15    52
    11    56     8    20
    11    31     8    47
     7    52     6    33
    11    55     9    22
```

```
     3    71    17     6
     1    31    22    44
     2    54    18    22
    21    47     4    26
     ⋮
     ⋮
```

● 各列について要素の平均値 ts および標準偏差 sigts を計算しよう．

平均値は mean, 分散はvarが与える．

```
ts=mean(T)
```

```
ts = 1×4
    7.4615   48.1538   11.7692   30.0000
```

```
sigts=sqrt(var(T))
```

```
sigts = 1×4
    5.8824   15.5609    6.4051   16.7382
```

● 用意されているコマンド zscore を使い，平均値，標準偏差，標準化されたデータ行列を計算する．

mean, varを用いた計算と比較して上の結果を確認してみよう．

```
 [X, ts, sigts]=zscore(T)
```

```
X = 13×4
   -0.0785   -1.4237   -0.9007    1.7923
   -1.0985   -1.2309    0.5044    1.3144
    0.6015    0.5042   -0.5885   -0.5974
    0.6015   -1.1024   -0.5885    1.0156
   -0.0785    0.2472   -0.9007    0.1792
    0.6015    0.4400   -0.4323   -0.4779
   -0.7585    1.4682    0.8167   -1.4338
   -1.0985   -1.1024    1.5973    0.8364
   -0.9285    0.3757    0.9728   -0.4779
    2.3015   -0.0742   -1.2130   -0.2390
     ⋮
     ⋮
ts = 1×4
    7.4615   48.1538   11.7692   30.0000
sigts = 1×4
    5.8824   15.5609    6.4051   16.7382
```

X の共分散行列 Q は（ここでは N=13 だから）

```
 Q=X'*X/12
```

```
Q = 4×4
    1.0000    0.2286   -0.8241   -0.2454
    0.2286    1.0000   -0.1392   -0.9730
   -0.8241   -0.1392    1.0000    0.0295
   -0.2454   -0.9730    0.0295    1.0000
```

となる．

共分散行列の固有値 σ_α^2 を対角に並べた行列 SIGMA2，固有ベクトル（主成分 $\alpha = 1 \sim P$)を各列成分（もとの ingredients=1~ P

を行成分に）並べた行列 V を計算する：

```
[V, SIGMA2]=eig(Q)
```

```
V = 4×4
    0.2411    0.6755    0.5090   -0.4760
    0.6418   -0.3144   -0.4139   -0.5639
    0.2685    0.6377   -0.6050    0.3941
    0.6767   -0.1954    0.4512    0.5479
SIGMA2 = 4×4
    0.0016         0         0         0
         0    0.1866         0         0
         0         0    1.5761         0
         0         0         0    2.2357
```

SIGMA2 は Q の固有値行列であるから，主成分分散 σ_α^2 を対角に並べたになる． 列は σ_α の昇順である．

- さらに固有ベクトルをサンプルの成分を行成分（j=1 ～ N)で書き表した行列 Z=XV を計算してみよう．

```
[U1,Lambda1,V1]=svd(X)
```

```
U1 = 13×13
   -0.2833   -0.4376   -0.3542    0.2760   -0.3714    0.1485   -0.0471    0.0030 ···
   -0.4124   -0.0548   -0.1939   -0.2137   -0.1192    0.0637    0.6052   -0.0588
    0.2181   -0.0423   -0.0072   -0.6713   -0.2152   -0.0880   -0.0060    0.4188
   -0.1274   -0.3626    0.1198   -0.2372    0.4403   -0.0384    0.0071   -0.1882
    0.0693   -0.1112   -0.4946    0.1375    0.7167   -0.0190    0.0461    0.2612
    0.1866   -0.0391    0.0573   -0.0872    0.0560    0.9713   -0.0054    0.0173
    0.1797    0.4909   -0.1156    0.0594    0.0466    0.0268    0.6277   -0.0633
   -0.4309    0.1590    0.3072    0.1619    0.0876    0.0735    0.0038    0.7661
   -0.0679    0.3293   -0.0211   -0.3223    0.1816    0.0367   -0.3001   -0.1699
    0.3210   -0.4204    0.5688    0.1421    0.1088   -0.0791    0.3592    0.0180
      ⋮
      ⋮
Lambda1 = 13×4
    5.1796         0         0         0
         0    4.3489         0         0
         0         0    1.4964         0
         0         0         0    0.1396
         0         0         0         0
         0         0         0         0
         0         0         0         0
         0         0         0         0
         0         0         0         0
         0         0         0         0
      ⋮
      ⋮
V1 = 4×4
    0.4760   -0.5090    0.6755    0.2411
    0.5639    0.4139   -0.3144    0.6418
   -0.3941    0.6050    0.6377    0.2685
   -0.5479   -0.4512   -0.1954    0.6767
```

ここの $V1$ は前項の V と同じもの（列の並べ方は降順），Lambda1 は X の特異値 λ_α を並べた行列.

$\lambda_\alpha^2 = (N-1)\sigma_\alpha^2$ より，各成分に対して $\mathrm{Lambda1} = \sqrt{12 \times \mathrm{SIGMA2}}$ であることが確かめられる．SVD=特異値分解

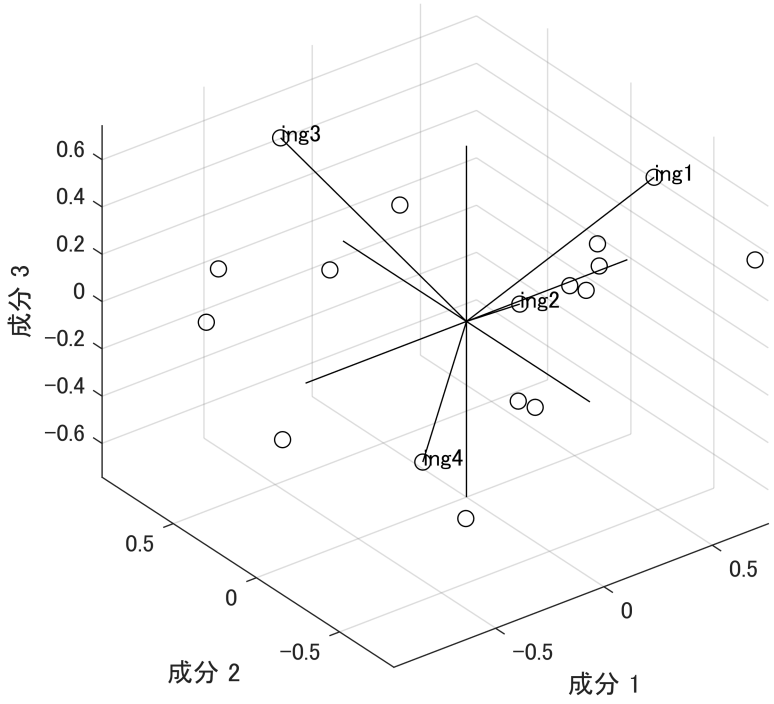
- これまでの計算（主成分分析）を一度に行うコマンドがpcaである：

```
[COEFF,SCORE,LATENT] = pca(X)
```

```
COEFF = 4×4
    0.4760   -0.5090    0.6755    0.2411
    0.5639    0.4139   -0.3144    0.6418
   -0.3941    0.6050    0.6377    0.2685
   -0.5479   -0.4512   -0.1954    0.6767
SCORE = 13×4
   -1.4672   -1.9030   -0.5300    0.0385
   -2.1358   -0.2384   -0.2902   -0.0298
    1.1299   -0.1839   -0.0107   -0.0937
   -0.6599   -1.5768    0.1792   -0.0331
    0.3588   -0.4835   -0.7401    0.0192
    0.9666   -0.1699    0.0857   -0.0122
    0.9307    2.1348   -0.1730    0.0083
   -2.2321    0.6917    0.4597    0.0226
   -0.3515    1.4322   -0.0316   -0.0450
    1.6625   -1.8281    0.8512    0.0198
     ⋮
LATENT = 4×1
    2.2357
    1.5761
    0.1866
    0.0016
```

これらの返り値の列は主成分分散 σ_α^2 の大きい方からの順（降順）に並ぶ．COEFF は $N \times p$（標準化された）データ行列 X の主成分係数（負荷量）V1を，SCORE は主成分スコア $Z = X * V$ を，LATENT は主成分分散 σ_α^2 （SIGMA2）を返す．

```
vbls={'ing1','ing2','ing3','ing4'};
biplot(COEFF(:,1:3),'Score',SCORE(:,1:3),'Color','k','Marker','o','VarLabels',vbls)
daspect([1 1 1])
```



MATLABには各COEFFおよびSCOREを描くためのbiplotというプログラムが用意されている.

ただし同じ図に納めるためにSCOREの値はスケーリングされている.

(実際の操作は，各Scoreを全てのScoreの最大絶対値で割り，coeffの最大係数をかけるというもの．)

課題1

上の問題で得られた結果から、元のデータにもどって、元のデータの性質について「何が分かったのか」，元のデータ（13のサンプルと1~4のingredients）について、説明せよ.

課題2

前回示したデータ　Jap-Math-Sci.xlsx　を用いて主成分分析を行ってみよ.