7-1. 主成分分析

1　主成分分析の問題意識

以下は10人の生徒の期末試験結果である。

生徒NO.	国語	英語	数学	理科
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

- それぞれの生徒の特徴を抽出できないか
  全体的に成績がよい悪い、理系が得意不得意など

2　行列の対角化

固有値，固有ベクトル

(n, n)型の正方行列Aに対して，

$$A\boldsymbol{x} = \alpha\boldsymbol{x}, \quad \alpha \in \mathbb{R} \tag{1}$$

となるn項列ベクトル$\boldsymbol{x}$であって，ゼロベクトルでないものを，Aの固有ベクトルといい，αを固有値という[*1].
固有ベクトルは，スカラー倍（ゼロを除く）しても固有ベクトルなので，この性質を用いて，固有ベクトルはそれ自体との内積が1となるように（つまり絶対値が1となるように）規格化できる．以下では，固有ベクトルはこのように規格化してあると約束する．

[*1] 固有値，固有ベクトルの求め方については，線型代数の教科書を参照のこと．一般的に，行列を扱えるプログラム言語（Rなど）には，固有値，固有ベクトルを求める関数が用意されている．

行列の対角化

ある(n,n)型の正方行列Aに対して，n個の線型独立な固有ベクトルが存在するとする[*2]．Aのn個の固有値$\alpha_1, \alpha_2, \ldots, \alpha_n$を対角線上に並べ他の成分をゼロとした$(n,n)$型の正方行列を$V$とする．すなわち，

$$V = \begin{bmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \alpha_n \end{bmatrix} \tag{2}$$

[*2] (n,n)型の正方行列には，重解も含めてn個の固有値が存在する．固有値は虚数となる場合があり，(1)式のαは虚数でもよい．この重解も含めたn個の固有値に対して，n個の固有ベクトルが互いに線型独立になるように上手く選べたとここでは仮定する．

とする[*3]．また，対応する固有ベクトル$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$を横に並べて作った$(n, n)$型の正方行列を$S$とする．すなわち，

$$S = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \end{bmatrix}. \tag{3}$$

このとき，

$$\begin{bmatrix} A\boldsymbol{x}_1 & A\boldsymbol{x}_2 & \ldots & A\boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} \alpha_1\boldsymbol{x}_1 & \alpha_2\boldsymbol{x}_2 & \ldots & \alpha_n\boldsymbol{x}_n \end{bmatrix}$$

$$\Leftrightarrow \; A \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \alpha_n \end{bmatrix}$$

$$\Leftrightarrow \; AS = SV \tag{4}$$

である．(4)式の両辺に左からS^{-1}を掛けると，

$$S^{-1}AS = V \tag{5}$$

[*3] このとき，どの順序で固有値を対角線上に並べるかには任意性がある．

という表現が得られる．これを行列の対角化という．

対称行列のスペクトル分解

対称行列とは、$A = A^\top$ である対称行列をいう。対称行列の固有ベクトルは互いに直交する。したがって、固有ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ を横に並べて作った S は $S^\top S = I$ という性質を有し、$S^{-1} = S^\top$ であることがわかる。このような性質を持つ行列を直交行列とよぶ。
固有値 $\alpha_1, \alpha_2, \ldots, \alpha_n$ を対角線上に並べた V と上記の S から (4) 式を作り、両辺に右から $S^\top$ を掛けると、

$$A = SVS^\top \tag{6}$$

という分解ができるが、固有ベクトルは互いに直交するという性質を用いると

$$A = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{7}$$

と表せる。これを対称行列のスペクトル分解という。

3　主成分分析の方法

主成分分析はデータの特徴量を取り出すことを目的とする。2ページのデータ（生徒の期末試験結果）を例に、Rでの計算方法とあわせて主成分分析の方法を説明する。

　まず、行と列を変えずデータを$X0$という行列として定義する：

```
X0 <- matrix(c(86, 79, 67, 68,
               71, 75, 78, 84,
               42, 43, 39, 44,
               62, 58, 98, 95,
               96, 97, 61, 63,
               39, 33, 45, 50,
               50, 53, 64, 72,
               78, 66, 52, 47,
               51, 44, 76, 72,
               89, 92, 93, 91), ncol = 4, byrow = T)
```

このままだと扱いづらい点があるため、各列の平均が0、標準偏差が1となるように標準化を行う。

```
X <- matrix(NA, 10, 4)
for(i in 1:4)
  for(j in 1:10)
    X[j, i] <- (X0[j, i]-mean(X0[, i]))/sqrt(var(X0[, i]))
> X
             [,1]        [,2]        [,3]        [,4]
 [1,]  0.9540628  0.69585780 -0.01548941 -0.03337686
 [2,]  0.2239127  0.51029572  0.55245560  0.85667266
 [3,] -1.1877109 -0.97420092 -1.46116762 -1.36845113
 [4,] -0.2141774 -0.27834312  1.58508289  1.46858170
 [5,]  1.4408296  1.53088716 -0.32527760 -0.31151733
 [6,] -1.3337409 -1.43810612 -1.15137943 -1.03468256
 [7,] -0.7982975 -0.51029572 -0.17038350  0.18913552
 [8,]  0.5646494  0.09278104 -0.78995988 -1.20156685
 [9,] -0.7496208 -0.92781040  0.44919287  0.18913552
[10,]  1.1000929  1.29893456  1.32692607  1.24606933
```

ここで、$n = 10$（行数＝生徒数）として$X^\top X/(n-1)$を計算すると（分散共分散行列という）、

```
> n <- 10
> t(X) %*% X / (n-1)
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.9669841 0.3760929 0.3112756
[2,] 0.9669841 1.0000000 0.4146367 0.3982751
[3,] 0.3760929 0.4146367 1.0000000 0.9721285
[4,] 0.3112756 0.3982751 0.9721285 1.0000000
```

となり、対角成分が全て1となるので、Xの各列の分散および標準偏差が1となるように標準化されているのがわかる。非対角成分は、Xのi列目をx_iとすると、$X^\top X/(n-1)$のj行k列はXのj列とk列の共分散（各列の分散が1のため相関係数に等しい）を表す。

$\Sigma \equiv X^\top X$から(4)式のような方程式を作る：

$$\Sigma S = SV \tag{8}$$

$$\Leftrightarrow \; X^\top X S = SV \tag{9}$$

$$\Leftrightarrow \; X^\top Z = SV \tag{10}$$

途中$Z = XS$と定義した。
　Rでは、

```
> EG <- eigen(t(X) %*% X)
> v <- EG$values
> S <- -EG$vectors # 解釈を容易にするため正負を反転させてSを定義
> Z <- X %*% S
>
> v
[1] 24.48659669 10.99619582  0.47170308  0.04550441
> S
          [,1]       [,2]       [,3]       [,4]
[1,] 0.4872727 -0.5273374  0.4989741 -0.4852890
[2,] 0.5105362 -0.4739968 -0.5386722  0.4738271
[3,] 0.5083186  0.4807477  0.5041136  0.5063234
[4,] 0.4934879  0.5158720 -0.4546720 -0.5325590
```

とする。Sの各列が固有ベクトルである（固有ベクトルの定数倍も固有ベクトルであることに注意)。また、

```
> Z
```

```
             [,1]        [,2]          [,3]         [,4]
 [1,]  0.7958047 -0.85761206  0.1085805662 -0.123347365
 [2,]  1.0732121  0.34756917 -0.2741601262 -0.043377951
 [3,] -2.4941567 -0.32030618 -0.1822601660  0.103737793
 [4,]  1.2839874  1.76450267  0.1744055831 -0.007490484
 [5,]  1.1645755 -1.80251845 -0.1280482286  0.027362884
 [6,] -2.4799717  0.29770231 -0.0008193416 -0.066104775
 [7,] -0.6427860  0.67850890 -0.2953349640 -0.041381976
 [8,] -0.6720037 -1.34136483  0.3798561779  0.009874113
 [9,] -0.5172813  1.14860039  0.2661938363  0.050871985
[10,]  2.4886197  0.08491807 -0.0484133371  0.089855777
```

である。Sの各列をXの主成分、Zの各列を主成分得点とよぶ。
$X^\top Z = SV$は

```
> V <- diag(v)
> t(X) %*% Z - S %*% V
              [,1]          [,2]          [,3]          [,4]
[1,] 3.552714e-15 -2.664535e-15 -1.221245e-15  7.806256e-16
[2,] 1.776357e-15  8.881784e-16  1.665335e-16 -1.190020e-15
```

```
[3,] 2.131628e-14  1.953993e-14  1.249001e-15 -1.318390e-16
[4,] 2.131628e-14  2.042810e-14 -1.942890e-16 -5.932754e-16
```

とすると、（ほぼ）ゼロ行列が得られるので確かめられる。

Σは対称行列で正定値という性質をもつため、その固有値v_iは全て正の実数である。固有値を対角線上に並べたVを作る際、どの順序で固有値を対角線上に並べるかには任意性があるが、固有値が大きい順に左上から並べたとき（上記のRの例ではこの順で並んでいる）、Sのi列を第i主成分、Zのi列を第i主成分得点とよぶ。s_iを第i主成分、z_iを第i主成分得点とするとき、

$$z_i = X s_i \tag{11}$$

である。

単位行列のs_iによるスペクトル分解

$$I = \sum_{i=1}^{n} s_i s_i^\top \tag{12}$$

に左からXを掛けると

$$X = \sum_{i=1}^{n} X s_i s_i^\top = \sum_{i=1}^{n} z_i s_i^\top \tag{13}$$

であるから、Xは$z_i s_i^\top$の和に分解できることがわかる。

```
> Z[, 1]%*%t(S[, 1]) + Z[, 2]%*%t(S[, 2]) +
+ Z[, 3]%*%t(S[, 3]) + Z[, 4]%*%t(S[, 4])
```

がXに一致することで確かめられる（出力は省略）。

以下、主成分の番号jを固定して考える。$\Sigma = X^\top X$のスペクトル分解

$$X^\top X = \sum_{i=1}^{n} v_i s_i s_i^\top \tag{14}$$

の右からs_jをかけると$s_j^\top s_j = 1$、$j \neq i$のとき$s_i^\top s_j = 0$であるため、

$$X^\top X s_j = X^\top z_j = v_j s_j \tag{15}$$

である。$X^\top z_j/(n-1)$ の k 番目の要素は X の k 列目と z_j の相関係数を表し、“相関係数ベクトル” $X^\top z_j/(n-1)$ の大きさ $|X^\top z_j/(n-1)|$ は $s_j^\top s_j = 1$ を使うと $v_j/(n-1)$ である。
固有値の比 $\lambda_j = v_j/\sum v_i$ を寄与率といい、z_j で X をどの程度説明できるか、を意味する。例えば、v_1 が最も大きい固有値であるため、第1主成分得点 z_1 が X を最もよく説明する。上記の例では

```
> v/sum(v)
[1] 0.680183241 0.305449884 0.013102863 0.001264011
```

したがって、第1主成分の寄与率は0.680、第2主成分の寄与率は0.305である。

　第1主成分から第 k 主成分得点で全体を1として X を $\lambda_1 + \lambda_2 + \cdots + \lambda_k$ だけ説明できると解釈できる。上記の例では、第1主成分と第2主成分で全体を1として $0.680 + 0.305 = 0.986$ と、ほぼ全てを説明できることがわかる。

　冒頭で主成分分析はデータの特徴量を取り出すことを目的とすると述べた。ここで、各主成分得点はデータ（生徒の期末試験結果）のどのような特徴を表すか考える。第1主成分得点とXの各列との相関係数は

```
> cor(Z[, 1], X)
          [,1]      [,2]      [,3]      [,4]
[1,] 0.8037389 0.8421113 0.8384534 0.8139908
```

である。第1主成分得点はXのどの列とも相関が高いので、生徒の総合的な学力（大きいほど学力が高い）を表すと解釈できる。第2主成分得点とXの各列との相関係数は

```
> cor(Z[,2], X)
           [,1]       [,2]      [,3]      [,4]
[1,] -0.5828926 -0.5239325 0.5313947 0.5702194
```

である。第2主成分得点はXの1列目と2列目（国語と英語）と負の相関関係を有し、3列目と4列目（数学と理科）と正の相関関係を有する。したがって、生徒の理系文系のタイプを表す（第2主成分得点が大きいと理系、小さいと文系）第1主成分と第2主成分でほぼ全てを説明できるので、第3主成分以下の解釈は行わない。

　改めてデータ（X）と第1主成分得点(z1)と第2主成分得点(z2)を表にすると以下の通りとなる。

ID	X1	X2	X3	X4	Z1	Z2
1	86	79	67	68	0.80	-0.86
2	71	75	78	84	1.07	0.35
3	42	43	39	44	-2.49	-0.32
4	62	58	98	95	1.28	1.76
5	96	97	61	63	1.16	-1.80
6	39	33	45	50	-2.48	0.30
7	50	53	64	72	-0.64	0.68
8	78	66	52	47	-0.67	-1.34
9	51	44	76	72	-0.52	1.15
10	89	92	93	91	2.49	0.08

主成分得点から、生徒を以下のように分類できる：

高学力・理系タイプ：生徒2、生徒4、生徒10

高学力・文系タイプ：生徒1、生徒5

低学力・理系タイプ：生徒6、生徒7、生徒9
低学力・文系タイプ：生徒3、生徒8
主成分分析によりデータ（生徒の期末試験結果）の特徴量を取り出したことで、生徒の学力のタイプの分類ができた。

練習問題8-1（景気動向指数）

政府（内閣府）は景気の現状把握及び将来予測に資するため、毎月景気動向指数を公表している[*4]。生産、雇用など様々な経済活動での重要かつ景気に敏感に反応する指標の動きを統合することによって計算され、先行指数、一致指数、遅行指数の3種類が存在する。

1. 一致指数の構成系列（9の指標からなる）を内閣府のウェブサイトからダウンロードし、その主成分を計算しなさい。
2. 計算した第1主成分得点と景気動向指数・一致指数をそれぞれグラフに描き、相関係数を計算しなさい。
3. 第1主成分は何を表していると考えられるか、簡潔に答えなさい。

[*4] https://www.esri.cao.go.jp/jp/stat/di/menu_di.html