Chapter 5

主成分分析とその応用

5.1 はじめに

行列の固有値は様々な応用分野があり、その一つである主成分分析はデータを解析する際に重要な道具となる。主成分分析は、与えられたベクトルの組の類似度を見つけ出す手法である。例えば、n 個の標本に対し、それぞれ m 個のパラメータに関して測定を行ったとする。それぞれの標本の個体のパラメータ間にどのような関係が現れるかを調べるものである。

標本は $A_1, A_2, ...A_n$ ただし $(A_k = [a_k^1, a_k^2, ..., a_k^m])$ というベクトルの列として与えられる。これら n 本のベクトルの類似性を見つけ出すのが主成分分析の目的である。

例えば、以下のような場合に主成分分析が有効である。

例 1 n 人が m 科目の試験を受け、総合的な得点を求めたいとする。つまり、各個人について m 次元のベクトルがあるが、m 科目でなく 1 ないし小数のスコアによって評価したい。

例 2 ある一つの機械に取り付けたセンサーからの多くのデータのベクトルがある。これから、この機械の動きについて実際の自由度がどれくらいあるか評価したい。

例 3 画像から特徴を抽出してパターン認識を行いたい。

つまり、主成分分析とは、多次元のデータから、特徴的な指標を得るための方法である。主成分 Z は、例えば 2 成分の場合以下のように表現される。

$$Z = w_1 X_1 + w_2 X_2 \tag{5.1}$$

係数 w_1、w_2 は計算で求まる係数である (後述)。この形は回帰分析と似ているが、実際にその応用も重なりをもっている。ただし、回帰分析の場合は、片方の変数 (例えば X_1) によりもう一つの変数を説明する、という非対称性がある。後述するが、主成分分析の場合は、双方の変数を公平に扱う。

5.2 主成分分析

主成分分析は、データ $X_1, X_2, ...X_n$ ただし ($X_k = [x_k^1, x_k^2, \ldots, x_k^m]$) から各軸間の関係を圧縮し、Z の分散が最大になるような

$$Z = w_1 x^1 + w_2 x^2 + \cdots + w_m x^m \tag{5.2}$$

という関係を見いだすことである。ただし、$\sum w_j^2 = 1$ とする。また、X は平均 0、分散 1 となるように標準化されているとする。最も分散が大きい方向というのは、その方向に垂直な方向の分散が小さいということでもある。別の言葉でいうと「空間の中から、平面を切り出して、その平面の上での座標で空間全体の評価が近似できるという、次元を圧縮する作業」である。(pp. 8 図 5.1 を参照するとわかりやすい)

Z の分散を評価するのであるから、まず Z の分散 σ_Z^2, X_j の分散 σ_j^2 とすると、以下のように表される。

$$\begin{aligned} \sigma_Z^2 &= \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_i)^2 && (5.3) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(w_j x_i^j - w_j \bar{x}^j) && (5.4) \end{aligned}$$

簡単のため、m=2 とすると

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n}(w_1 x_i^1 + w_2 x^2 - (w_1 \bar{x^1} + w_2 \bar{x^2}))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(w_1 (x_i^1 - \bar{x^1}) + w_2 (x_i^2 - \bar{x^2}))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}((w_1)^2 (x_i^1 - \bar{x^1})^2 + 2w_1 w_2 (x_i^1 - \bar{x^1})(x_i^2 - \bar{x^2}) + (w_2)^2 (x_i^2 - \bar{x^2})^2) \\ &= \frac{(w_1)^2}{n}\left(\sum_{i=1}^{n}(x_i^1 - \bar{x^1})^2\right) + 2w_1 w_2 \frac{1}{n}\sum_{i=1}^{n}(x_i^1 - \bar{x^1})(x_i^2 - \bar{x^2}) + \frac{(w_2)^2}{n}\left(\sum_{i=1}^{n}(x_i^2 - \bar{x^2})^2\right) \\ &= (w_1)^2\sigma_{11} + 2w_1 w_2 \sigma_{12} + (w_2)^2 \sigma_{22} && (5.5) \end{aligned}$$

となる。ここで、σ_{ij} は共分散と呼ばれる量で、

$$\sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_k^i - \bar{x^i})(x_k^j - \bar{x^j}) \tag{5.6}$$

と定義される。直感的には、2 つのベクトル X_i、X_j の「内積」と考えるとわかりやすい。$i = j$ の場合は分散である。(例:$\sigma_{11} = \sigma_1^2$)

ここでは m=2 の場合に限ったが、

$$(\sum_i x_i)^2 = \sum_i x_i^2 + 2\sum_{i\neq j} x_i x_j \tag{5.7}$$

であることを考えると、一般に

$$\sigma_Z^2 = \sum_i w_i^2\sigma_i^2 + 2\sum_{i\neq j}\sigma_{ij} \tag{5.8}$$

が成立することがわかる。

さて、やや回り道をしたが式 5.5 を最小化する w_i の組を求めたいというのが目的である。共分散 (断りがない限り、以降では分散を含む) は定数であるので、2 変数 w_1、w_2 に関する問題である。

これは、Lagrange の未定乗数法によって以下のように関数 F としてかける。束縛条件は $w_1^2 + w_2^2 = c$ とする。

$$F(w_1, w_2, \lambda) = w_1^2\sigma_{11} + 2w_1w_2\sigma_{12} + w_2^2\sigma_{22} - \lambda(w_1^2 + w_2^2 - c) \tag{5.9}$$

これを w_1,w_2,λ それぞれについて偏微分し、0 と等しいとおく (極値を求めるためであるから)。

$$\frac{\partial F}{\partial w_1} = 2w_1\sigma_{11} + 2w_2\sigma_{12} - 2w_1\lambda = 0 \tag{5.10}$$

$$\frac{\partial F}{\partial w_2} = 2w_2\sigma_{22} + 2w_1\sigma_{12} - 2w_2\lambda = 0 \tag{5.11}$$

$$\frac{\partial F}{\partial \lambda} = -\lambda(w_1^2 + w_2^2 - c) = 0 \tag{5.12}$$

ただし c は定数である。第 3 式は定数の選び方できまるので、上の 2 式のみ考える。(あるいは、冒頭で係数の自乗和が 1 であると定義しているので、すでに満たされていると考えてもよい)。これは、

$$\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}\begin{pmatrix}w_1\\ w_2\end{pmatrix} = \lambda\begin{pmatrix}w_1\\ w_2\end{pmatrix} \tag{5.13}$$

という、固有値を求める方程式に帰着する。つまり、主成分分析は、共分散行列の対角化によって求められる。Z は、最大固有値に対応する固有ベクトル w_1、w_2 によって決められる。これが第一主成分と呼ばれるもので、以降固有値の数だけ主成分が存在する。

5.3 主成分分析の手法:2 変数の場合

では、実際に主成分分析を行ってみよう。以下に、20 人のクラスで 2 科目のテストを行った結果 X を示す。行は科目、各列は生徒を表すとする。(例は上田尚一「主成分分析」朝倉書店 2003 より引用)

```
X=[22, 38;
   24, 51;
   33, 45;
   35, 45;
   38, 46;
   40, 48;
   41, 52;
   41, 46;
   46, 52;
   46, 49;
   50, 50;
   51, 48;
   56, 51;
   56, 47;
   58, 57;
   58, 42;
   59, 39;
   61, 51;
   65, 61;
   68, 68;]
```

まず、X を標準化する。つまり、$X_S = \frac{X-\mu}{\sigma}$ と平均 μ と標準偏差 σ により変換する。

各列ごとの平均からの差を出力する関数 center() という関数を使うと分子が簡単に計算できる。

```
octave:69> X2=center(X)
X2 =

  -25.40000  -11.30000
  -23.40000    1.70000
  -14.40000   -4.30000
  -12.40000   -4.30000
   -9.40000   -3.30000
```

```
   -7.40000   -1.30000
   -6.40000    2.70000
   -6.40000   -3.30000
   -1.40000    2.70000
   -1.40000   -0.30000
    2.60000    0.70000
    3.60000   -1.30000
    8.60000    1.70000
    8.60000   -2.30000
   10.60000    7.70000
   10.60000   -7.30000
   11.60000  -10.30000
   13.60000    1.70000
   17.60000   11.70000
   20.60000   18.70000
```

分母は、まず標準偏差を求める。

```
octave:68> sigma=std(X)
sigma =

  13.0360   6.9744
```

により 行ベクトル sigma に標準偏差の値を格納する。 Kronecker 積を用いて、この行を列の数だけ (ここでは 20 回) 繰り返す。

```
octave:72> den=kron(std(X),ones(20,1))
den =

  13.0360   6.9744
  13.0360   6.9744
  13.0360   6.9744
   ....      ....
  13.0360   6.9744
  13.0360   6.9744
```

これで準備が整った。X_s は、行列の成分ごとの和を求める演算子 ./ を用いて

```
octave:73> XS = center(X)./den
XS =
```

```
  -1.948453  -1.620214
  -1.795032   0.243749
  -1.104635  -0.616541
     .....      ....
   1.350109   1.677566
   1.580241   2.681238
```

となる。

共分散行列は corrcoef() により求めることができる。

```
octave:75> R1=corrcoef(XS)
R1 =

  1.00000  0.51788
  0.51788  1.00000
```

ここまで求められたら、固有値を求めることができる。

```
octave:76> [v,l]=eig(R1)
v =

  -0.70711   0.70711
   0.70711   0.70711

l =

  0.48212  0.00000
  0.00000  1.51788
```

ここから、固有値は 1.52 と 0.48 で、固有ベクトルは [0.707;0.707],[-0.707;0.707] であることが求められた。

総合スコアは

$$Z_1 = 0.707X_1 + 0.707X_2 \tag{5.14}$$

$$Z_2 = -0.707X_1 + 0.707X_2 \tag{5.15}$$

である。ただし、これは標準化されたスコアでの上の話である。

また、固有値を正規化 (和を 1 にする) したものが寄与度と呼ばれる。この例では、

```
octave:77> sum(diag(l))
```

```
ans = 2
octave:78> l/sum(diag(l))
ans =

  0.24106  0.00000
  0.00000  0.75894
```

からわかるように、それぞれ 0.76, 0.24 である。

次に、XS と第一主成分の積をとろう。

```
octave:81> XS*v(:,2)
ans =

  -2.523428
  -1.096922
  -1.217055
    ......
   2.140890
   3.013321
```

これが、主成分分析によって求められた総合スコアである。平均値を原点にとっているので、スコアがマイナスになっているものもある。(図 5.1 参照)

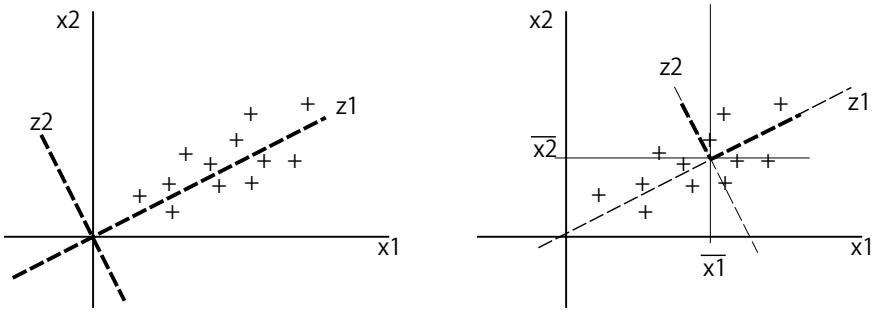

Figure 5.1: PCA によって作られた座標系

5.4 主成分分析の手法 II:3 変数の場合

では、せっかくの多変量解析なので、3 変数の場合を考えてみたい。

```
Octave:> Xadd=[61;62;56;64;66; 60;46;43;70;49; 54;52;62;68;68; 39;49;79;69;58];
```

このベクトルを、X と結合して、X3 という行列を作る。

```
octave:91> X3=[X,Xadd]
X3 =

  22  38  61
  24  51  62
  .........
  65  61  69
  68  68  58
```

前章で行ったことをまとめたのが以下のプログラム doPCA.m である。

```
#doPCA.m
function [v,l,Xs]=doPCA(X)
  [rows,cols]=size(X);
  Xs= center(X) ./ kron(std(X),ones(rows,1));
  R = corrcoef(Xs);
  [v,l]=eig(R);
endfunction
```

実行例は以下の通りである。

```
octave:96> source("doPCA.m")
octave:97> [v,l,xs]=doPCA(X3);
octave:105> v
v =
   0.607759   0.522738   0.597808
  -0.728409   0.067114   0.681848
   0.316307  -0.849848   0.421556
octave:106> l
l =
```

```
  0.42677  0.00000  0.00000
  0.00000  0.91825  0.00000
  0.00000  0.00000  1.65498
octave:107> xs
xs =

  -1.948453  -1.620214   0.219978
  -1.795032   0.243749   0.317745
  -1.104635  -0.616541  -0.268861
     ....       ....       ....
   1.043266   0.243749   1.979798
   1.350109   1.677566   1.002120
   1.580241   2.681238  -0.073326
```

練習問題

上の例で、第一主成分に基づいたスコア Z1 を求めよ。可能な限りプログラムで実行せよ。

この例では、主成分は 3 つあり、

$$Z_1 = 0.60x_1 + 0.68x_2 + 0.42x_3 \tag{5.16}$$
$$Z_2 = 0.52x_1 + 0.07x_2 - 0.85x_3 \tag{5.17}$$
$$Z_3 = 0.61x_1 - 0.73x_2 + 0.32x_3 \tag{5.18}$$

ただし、それぞれの寄与度は、 0.55, 0.31, 0.14 であるから、第 2 主成分まで考慮すればいいことがわかる (86%の寄与度である)。

では、主成分を解釈してみる。第一主成分はすべて符号が正であるから、総合的なスコアを表すものと考えることが出来る。第二主成分は、x_1 と x_2 に関しては正だが、x_3 に関しては負の係数をもつ。このことは、科目により、得意/不得意科目がグループ化されていることを示唆している。

5.5 主成分分析の手法 III:時系列解析

時系列解析にも主成分分析は応用できる。まず単純な場合として、以下の力学系の軌跡を考える。(この例は Daffertshofer e.al., Clin. Biomech. 19 (2004), pp.415-418 より引用)

$$x_1 = sin2\pi t \tag{5.19}$$
$$x_2 = 0.5sin2\pi t \tag{5.20}$$
$$x_3 = cos2\pi t \tag{5.21}$$

これは、空間中に、傾いた円を描いたものである。Octave で作成するためには、以下のプログラム PCAdat1.m を実行する。

```
#PCAdat1.m
t=(linspace(0,10,200))';
xc=[sin(2*pi*t),0.5*sin(2*pi*t),cos(2*pi*t)];
```

linspace(a,b,c) とは、[a,b] 区間に等間隔に c 個の数列をつくる関数である。列ベクトルの形にするために転地している。

3 次元でプロットするには、gsplot コマンドを使う。

```
octave:80> source("PCAdat1.m")
octave:82> gset parametric
octave:83> gsplot xc
```

3 次元空間に傾いた円が出現したはずである。

```
octave:84> Rxc=corrcoef(xc)
Rxc =

   1.0000e+00   1.0000e+00   5.1159e-16
   1.0000e+00   1.0000e+00   5.1159e-16
   5.1159e-16   5.1159e-16   1.0000e+00
octave:86> [vc,lc]=eig(Rxc)
vc =

    7.0711e-01    5.1159e-16    7.0711e-01
   -7.0711e-01    5.1159e-16    7.0711e-01
    1.6065e-31   -1.0000e+00    7.2349e-16

lc =

  -0.00000   0.00000   0.00000
   0.00000   1.00000   0.00000
   0.00000   0.00000   2.00000
```

第 3 固有値が 0 になっていることに注意。つまり、これは本質的に 2 次元平面内の運動であることがわかる。

上の例では次元が落ちていることは自明であったが、次に、ノイズがのった場合を考える。PCAdat2.m を実行すると sin カーブにノイズがのった時系列データが得られる。

$$x_1 = sin(2\pi t + noize_1) \tag{5.22}$$
$$x_2 = 0.5 * sin(2\pi t + noize_2) \tag{5.23}$$
$$x_3 = noize_3 \tag{5.24}$$

noize は正規乱数、係数 0.05 としたものが PCAdat2.m である。

```
#PCAdat2.m
n=1000;
na=0.05;
t=(linspace(0,10,n))';
Noize1=randn(n,3)*na;
```

```
t1 = [t,t,t]+Noize1;
Noize2=randn(n,3)*na;
xc2=[sin(2*pi*t1(:,1)),0.5*sin(2*pi*t1(:,2)),zeros(n,1)]+Noize2;
```

同様に、実行する。

```
octave:170> source("PCAdat2.m")
octave:171> plot(xc2)
octave:172> r2=corrcoef(xc2)
r2 =

  1.000000  0.884643  0.012551
  0.884643  1.000000  0.020536
  0.012551  0.020536  1.000000

octave:173> [vc2,lc2]=eig(r2)
vc2 =

   0.7070327   0.0231963   0.7068003
  -0.7071521   0.0141676   0.7069195
   0.0063842  -0.9996305   0.0264202

lc2 =

  0.11532  0.00000  0.00000
  0.00000  0.99942  0.00000
  0.00000  0.00000  1.88526

octave:174> rx1=xc2*vc2(:,3);
octave:175> rx2=xc2*vc2(:,2);
octave:176> rx3=xc2*vc2(:,1);
octave:177> plot(rx1)
octave:178> plot(rx2)
octave:179> plot(rx3)
```

最後の 3 行は、固有ベクトルと時系列データとの内積である。これは、固有ベクトル方向の成分を取り出していることになる。これにより、次元を落とした形で時系列をみることが出来るようになる。