第7回 主成分分析（後半）

Code

1. 前回の復習

前回は、主成分分析とその仕組みについて解説しました。一度、復習しておきましょう。

第1主成分 : 主成分分析は、d 個の変数 $x_1, \cdots, x_d$ からなるデータを、より少ない $d' < d$ 個の変数 $z_1, \cdots, z_{d'}$ で表現する（いわゆる次元削減とよばれる）手法のひとつでした。z_1 は $x_1, \cdots, x_d$ の重みつき和

$$z_1 = a_1 x_1 + \cdots + a_d x_d$$

です。係数 $a_1, \cdots, a_d$ は $a_1^2 + \cdots + a_d^2 = 1$ をみたすもののなかで変数 z_1 の分散が最も大きくなるように決めるのでした。変数 z_1 を第1主成分といいます。実は、この係数 $a_1, \cdots, a_d$ は分散共分散行列の大きさ1の固有ベクトルであることがわかるのでした。

主成分と寄与率 : 第1主成分と直交する一次式 $z_2 = b_1 x_1 + \cdots + b_d x_d$ のうち分散が最大になるようなものを第2主成分といいます。第3主成分以降も同様の考え方で定義することができます。実は、次の事実が成り立ちます。

1. 主成分は、データに含まれる変数の個数の分だけ作ることができます。
2. 主成分の係数は分散共分散行列の大きさ1の固有ベクトルとして求めることができます。
3. 固有値は主成分の分散に等しくなります。

なお、主成分の分散の和はデータの変数の分散の和に等しくなることから、第1主成分の寄与率
を 第1主成分の分散/データの変数の分散の和 と定義するのでした。これは、第2主成分以降でも同様です。

今回はこの話に基づいて、実際のデータに主成分分析を計算してみましょう。

2. eigen関数を用いた主成分分析

A. ラーメンデータの紹介

Iさんは、ラーメン春木屋の店員です。彼は、春木屋のラーメンの味が周辺にあるラーメン店と比較してどんな特徴があるのか調べてみようと思いました。そこで、麺・具・スープの味を5段階（1:低〜5:高）で評価したデータをとり、以下のような結果を得ました。

Hide

```
# データの読み込み
dat <- read.csv("./data/ramen.csv", fileEncoding = "cp932", row.names = "店名")
dat
```

	麺 <int>	具 <int>	スープ <int>
春樹屋	2	4	5
もりまる	1	5	1
ラーメン宇宙	5	3	4
ラーメン三郎	2	2	3
大山屋	3	5	5

	麺 <int>	具 <int>	スープ <int>
ばつね	4	3	2
ラーメン十七番	4	4	3
ぐうのね	1	2	1
きらきら星	3	3	2
麺屋・小次郎	5	5	3

1-10 of 10 rows

このデータに対して、まず固有値問題を解く eigen 関数によって主成分分析を行い、主成分分析を計算する prcomp 関数を紹介して結果が同じになっていることを確認します。

B. 前処理（データの標準化）

最初に主成分分析を行うにあたっての注意点を一つ紹介します。主成分分析では、データをまずは標準化することが一般的です。

問題 : 主成分分析の前にデータを標準化する目的を考えてみましょう。

解答 : 大きな分散を持つ変数がデータに含まれていたとき、出来る主成分がその変数に大きく依存してしまうため。■

標準化したデータの分散共分散行列は相関行列に等しいことが知られています。以下のスクリプトでこのことを確認してみましょう。

Hide

```
# 標準化したデータの分散共分散行列
X_scaled <- scale(dat)
covmat_scaled <- cov(X_scaled)
covmat_scaled
```

```
           麺        具     スープ
麺     1.0000000 0.1905002 0.3600411
具     0.1905002 1.0000000 0.3004804
スープ 0.3600411 0.3004804 1.0000000
```

Hide

```
# 相関行列
cormat <- cor(dat)
cormat
```

```
           麺        具     スープ
麺     1.0000000 0.1905002 0.3600411
具     0.1905002 1.0000000 0.3004804
スープ 0.3600411 0.3004804 1.0000000
```

C. 主成分と寄与率を求める

R 言語では eigen 関数を用いることで、大きさ1の固有ベクトルとその固有値を求めることができます。つまり、B節で求めた相関行列 cormat を eigen 関数に渡すと、主成分の係数と寄与率を求めることができるはずです。

Hide

```
# 相関行列の固有値分解
result <- eigen(cormat)
result
```

```
eigen() decomposition
$values
[1] 1.5728539 0.8140083 0.6131378

$vectors
           [,1]       [,2]       [,3]
[1,] 0.5715110  0.6044710  0.5549685
[2,] 0.5221161 -0.7896069  0.3223595
[3,] 0.6330639  0.1055260 -0.7668731
```

この結果から第1主成分の式は

$$\text{第1主成分} = 0.57 \times \text{麺の}z\text{得点} + 0.52 \times \text{具の}z\text{得点} + 0.63 \times \text{スープの}z\text{得点}$$

で、第1主成分の分散は1.57であることがわかります。同様に第2主成分の式は

$$\text{第2主成分} = 0.60 \times \text{麺の}z\text{得点} - 0.79 \times \text{具の}z\text{得点} + 0.11 \times \text{スープの}z\text{得点}$$

で、第2主成分の分散は0.81であることがわかります。第3主成分も同様にして得ることができます。

Remark : 主成分を求めるときデータを標準化したので、1次式は麺・具・スープのz得点に対して作られることに注意してください。■

また第1主成分・第2主成分・第3主成分の寄与率も次のように求めることができます。

Hide

```
# 寄与率の計算
result$values / sum(result$values)
```

```
[1] 0.5242846 0.2713361 0.2043793
```

ところで主成分分析では、主成分の式を求めたあと、その式に分析者が解釈を与えることが大切です。

問題 : 第1主成分・第2主成分の解釈を与えてください。

解答 : 第1主成分の係数はどれもほぼ等しいので「全体的なうまさ」を表していて、正の値に大きくなるほどうまいことがわかります。第2主成分の係数は麺と具の符号が逆になっていて、スープの係数がほぼ0に近いので「麺重視か具重視か」を表しています。そして、正の値に大きくなるほど麺重視、負の値に大きくなるほど具重視であることがわかります。■

D. 主成分得点を求める

主成分が各データ点についていくらであるかを求めたものを**主成分得点**といいます。例えば「春木屋の全体的なうまさ（第1主成分得点）と麺重視か具重視か（第2主成分得点）」に興味があるわけです。

これも行列の掛け算を使って計算することができます。次の通りです。

Hide

```
# 主成分得点
X_scaled <- scale(dat)
X_scaled %*% result$vectors
```

```
                  [,1]       [,2]         [,3]
春樹屋          0.7119408 -0.5216497 -1.373736133
もりまる       -0.9740499 -1.8911205  0.645382316
ラーメン宇宙    0.9804158  1.2947047 -0.002322692
ラーメン三郎   -1.0513965  0.6781104 -0.864614382
大山屋          1.5401350 -0.7888582 -0.726820118
ばつね         -0.2766766  0.7435735  0.683778524
ラーメン十七番  0.6049920  0.1436935  0.429217649
ぐうのね       -2.3084890  0.1269792 -0.178513165
きらきら星     -0.6600579  0.3380821  0.311494336
麺屋・小次郎    1.4331863 -0.1235150  1.076133664
```

この結果から、春木屋の全体的なうまさ（第1主成分得点）は 0.71 、麺重視か具重視か（第2主成分得点）は -0.52 です。春木屋は全体的なうまさの得点が10店舗中4番目であることから、味は「そこそこ美味しい」くらいなんだろうなということがわかります。また、麺重視か具重視かの得点が負であることから、具重視の店であることがわかります。

E. 主成分分析が適切でない場合

問題 : 全ての変数の間の相関係数が 0 の場合、主成分分析を行うとどのような結果になるでしょうか。考えてみてください。

解答 : 例えば、3変数のデータを考えます。全ての変数の間で相関係数が 0 のとき、相関行列は次のようになります。

Hide

```
# 相関行列
cormat <- diag(c(1, 1, 1))
cormat
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

この行列を固有値分解すると、次のような結果が得られます。

Hide

```
# 固有値分解
eigen(cormat)
```

```
eigen() decomposition
$values
[1] 1 1 1

$vectors
     [,1] [,2] [,3]
[1,]    0    0    1
[2,]    0    1    0
[3,]    1    0    0
```

得られた固有ベクトルは、データの変数がそのまま主成分になっていることを表しています。また固有値が全て等しく、各主成分の寄与率が等しく $1/3$ になっています。このことから、各変数間の相関係数が 0 のとき、主成分分析によってデータを要約するような主成分を得ようとしても失敗することがわかりました。■

このようなことは、変数間の相関係数がぴったり 0 でなくても成り立ちます。主成分分析を行う前に、しっかりデータの変数間の関係の強さを散布図や相関係数を使って確認し、主成分分析によって得られる結果を事前に想定するようにしましょう。

3. prcomp関数を用いた主成分分析

A. 主成分・寄与率・主成分得点を求める

prcomp 関数は、これまでの主成分分析の計算を一気に行うことができる R 言語の関数です。

Hide

```
# prcompによる主成分分析
result <- prcomp(dat, scale. = TRUE)
result
```

result の結果が固有値分解で得られた結果と同じになっていることが確認できるでしょう。細かい違いを列挙しておきます。 * prcomp の Standard deviation には固有値のルート（主成分の標準偏差）が出力されている。 * 第2主成分の解釈が eigen の結果と逆になっている。 また、主成分得点は次のようにして得られます。

Hide

```
# 主成分得点
result$x
```

B. 主成分負荷量

データの各変数と主成分得点との相関係数を**主成分負荷量**といいます。主成分負荷量は、 係数*主成分の標準偏差 に等しくなることが知られています。

問題 : 第1主成分の麺・具・スープに対する主成分負荷量を求めてください。

解答 : 第1主成分の麺に対する主成分負荷量は $0.57 \times 1.25 = 0.71$、具に対する主成分負荷量は 0.65、スープに対する主成分負荷量は 0.79 とわかります。■

この事実は次のように、麺と主成分の相関係数を直接計算することで確認することができます。

Hide

```
# 麺と主成分の相関係数
cor(dat, result$x)
```

1列めに第1主成分の麺・具・スープに対する主成分得点の計算結果が示されていますが、問題の解答と値が一致していることがわかります。主成分負荷量は主成分の係数のかわりに、主成分の意味を解釈するヒントとして用いられます。

C. biplot

主成分得点と主成分負荷量を示したものに、biplotがあります。biplotはとても便利な図なので、主成分分析を行った際にはぜひかいてみることをお勧めします。さまざまなことをこのグラフから読み取ることができますが、例えば次のようなことがわかります。

- 第2主成分は、正の値に大きいほど具重視、負の値に大きいほど麺重視。
- うまい店トップ3は大山屋、麺屋・小次郎、ラーメン宇宙。

Hide

```
par(family = "ヒラギノ角ゴシック W3")    # Macのみ
biplot(result)
```

4. mtcarsデータセットによる演習

問題： mtcars_numeric.csv は、1974年の雑誌 “Motor Trend US” から32台の自動車モデルについて、

- mpg ：燃費
- cyl ：シリンダーの数
- disp ：排気量
- hp ：馬力
- drat ：rear axle 比
- wt ：重量
- qsec ：ドラッグレースのタイム
- gear ：ギア数
- carb ：キャブレーターの数

を抜粋し、記録したデータです。主成分分析を用いて、各自動車モデルの特徴や自動車モデルの間の類似性を把握してみてください。

解答：まずはデータの各変数間の相関関係を確認します。相関係数を確認する上で、ヒートマップを用いると便利です。

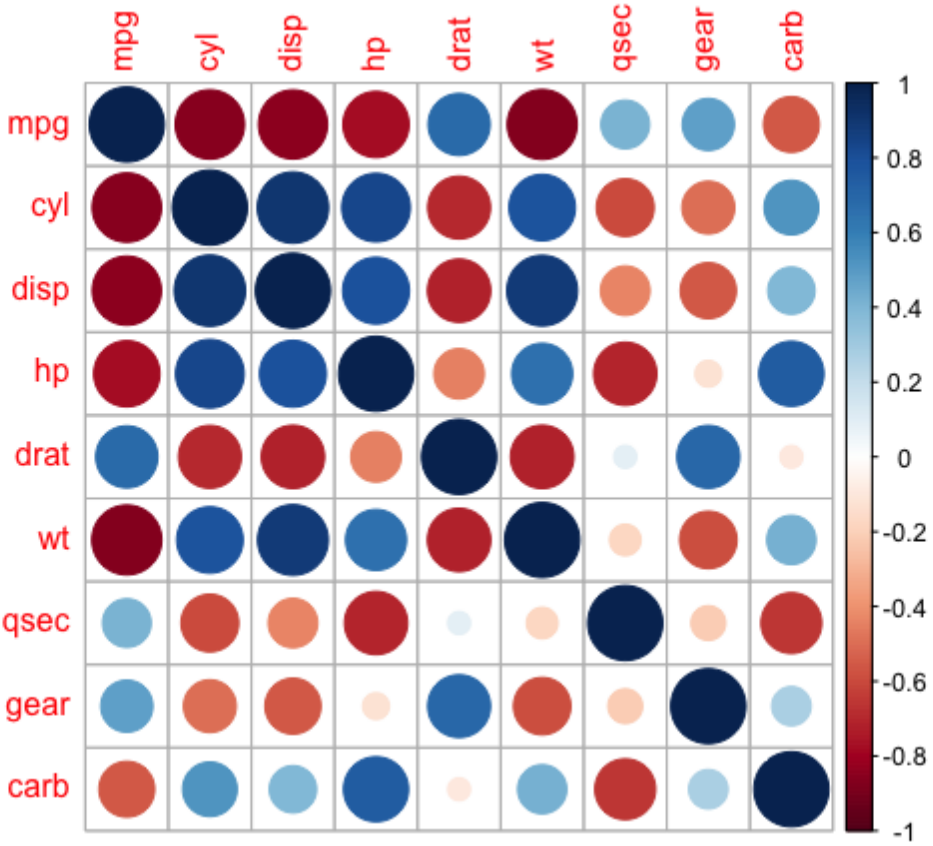
Hide

```
# ヒートマップをかくためのパッケージの読み込み
# install.packages("corrplot")
library(corrplot)
```

```
corrplot 0.88 loaded
```

Hide

```
dat <- read.csv("./data/mtcars_numeric.csv",
                fileEncoding = "utf-8",
                row.names = "model")
cormat <- cor(dat)    # 相関行列の計算
corrplot(cormat)    # ヒートマップの作成
```



prcomp 関数を用いて 主成分分析を行った結果が以下になります。なお、第2主成分までの累積寄与率は、86.0%でした。

Hide

```
result <- prcomp(dat, scale. = TRUE, rank. = 2)    # rankは使う主成分の数
result
```

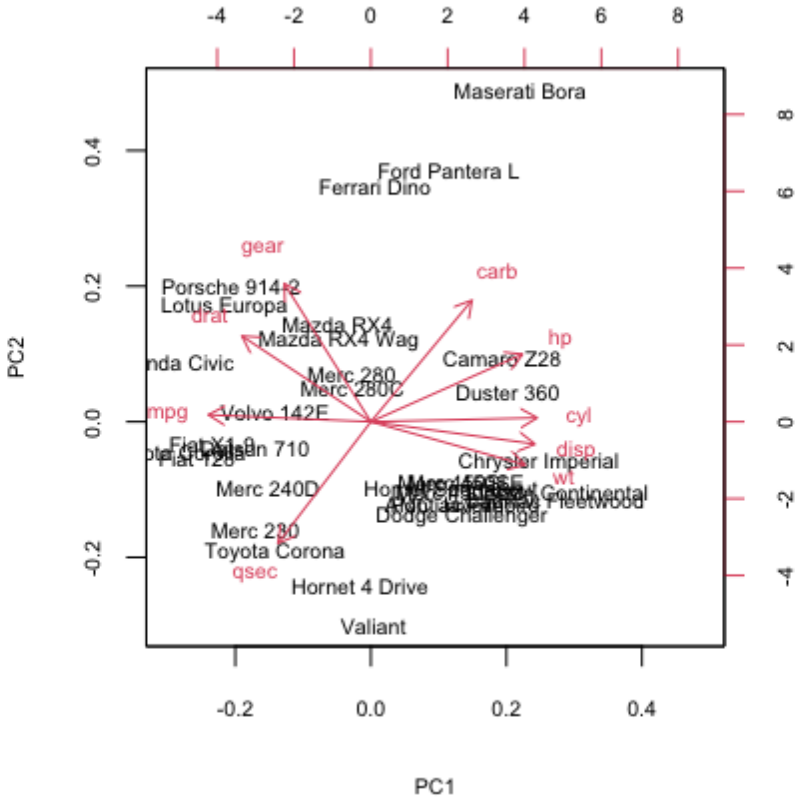
```
Standard deviations (1, .., p=9):
[1] 2.3782219 1.4429485 0.7100809 0.5148082 0.4279704 0.3518426 0.3241326
[8] 0.2418962 0.1489644

Rotation (n x k) = (9 x 2):
            PC1         PC2
mpg  -0.3931477  0.02753861
cyl   0.4025537  0.01570975
disp  0.3973528 -0.08888469
hp    0.3670814  0.26941371
drat -0.3118165  0.34165268
wt    0.3734771 -0.17194306
qsec -0.2243508 -0.48404435
gear -0.2094749  0.55078264
carb  0.2445807  0.48431310
```

主成分得点の解釈を与えます。 mpg , cyl , disp , wt などの主成分負荷量から、第1主成分得点は燃費を表し、負の値に大きいほど燃費が良く、正の値に大きいほど燃費が悪い自動車モデルであると解釈できます。また、 carb , gear , qsec などの主成分負荷量から、第2主成分得点はスポーツカーらしさを表し、正の値に大きいほどスポーツカーらしく、負の値に大きいほどスポーツカーらしくないことに対応していることがわかります。

最後にbiplotをかいてみます。

Hide

```
par(ps = 8)
biplot(result)
```



例えば、 Maserati Bora という自動車モデルは、biplotの一番右上の点に対応しています。このことから、 Maserati Bora は燃費の悪いスポーツカーであることがわかります。■