第8回メモ

**ハッシュタグ：#ybkstat**

主成分分析



参考：お化け煙突の謎解き（<http://www1.adachi.ne.jp/a.trm/13-6s.htm>）

お化け煙突：見る方角によって本数が変わる。

煙突の様子（4本あること）がもっともよくわかる方角はどこか？

主成分分析が解決する問題はこれに似ている。ただし、以下の点で異なる。

* 1方向だけから見るわけではない（この授業では2方向から見る）。
* 形をゆがませてもよい。
* 変数はたくさんある（煙突の場合は、xとyの2つだけ）。たくさんの変数を少ない変数に縮約してしまおう、というのが主成分分析の思想。

# Excelでの主成分分析

次のような4件のデータを使って主成分分析を試してみよう（かなり面倒）。

この表をExcelに貼り付ける。

|  |  |  |
| --- | --- | --- |
| **id** | **x** | **y** |
| 1 | 0 | 0 |
| 2 | 2 | 1 |
| 3 | 1 | 2 |
| 4 | 3 | 3 |

散布図を描いてみる。

データ分析の「共分散」を使って共分散行列を作る。



出力される共分散行列は下三角行列になっている。

|  |  |  |
| --- | --- | --- |
|  | x | y |
| x | 1.25 |  |
| y | 1 | 1.25 |

共分散行列の空白を埋める（行列が大きいときは、まず「値」だけ別の場所にコピーして、それを「行と列を入れ替え・空白を無視し」ながら元に貼り付ければよい。

|  |  |  |
| --- | --- | --- |
|  | x | y |
| x | 1.25 | 1 |
| y | 1 | 1.25 |

この行列の固有値と固有ベクトルを求める（線型代数大事）。この程度の大きさの行列なら、[Wolfram/Alpha](http://www.wolframalpha.com/)で計算できる。次のようなクエリを使えばよい。

Eigensystem[{{1.25,1},{1,1.25}}]

この固有ベクトルを**主成分**と呼ぶ。

|  |  |  |
| --- | --- | --- |
| 固有値 | 2.25 | 0.25 |
| 固有ベクトル | |  |
|  | PC1 | PC2 |
| x | 0.707107 | -0.70711 |
| y | 0.707107 | 0.707107 |

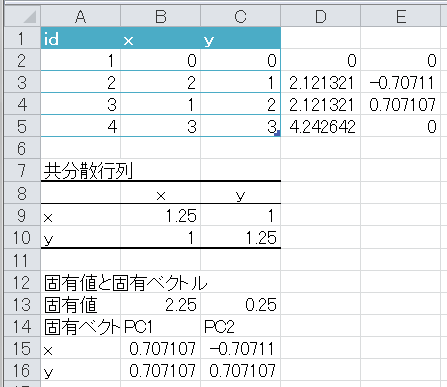
## 主成分の解釈

主成分を棒グラフで描くと下のようになる。**第1主成分**（PC1）は両方同じ程度だから、最初に描いた散布図の分布の分散が最大になる方角、**第2主成分**はそれと直交する方角であることがわかる。

## 主成分スコア

もとの行列（4行2列）と固有ベクトルを固有値の大きい順に並べた行列（2行2列）の積を求める。4行2列の領域を選択、次のような式を入力し、Ctrl+Shift+Enter

=MMULT(B2:C5,B15:C16)



行列の積の結果が**主成分スコア**（**主成分得点**）である。散布図を描くと下のようになる。

# Rでの主成分分析

## テストデータ

Excelで実験したデータを、Rで処理してみよう。

#テストデータの作成（CSVファイルを作って読み込んでもよい）

myData <- data.frame(x=c(0,2,1,3),y=c(0,1,2,3))

myData

#主成分分析（全体を括弧で囲むと結果が表示される）

myResult <- prcomp(myData)

#結果の概要

summary(myResult)

Importance of components:

PC1 PC2

Standard deviation 1.732 0.5774

Proportion of Variance 0.900 0.1000

Cumulative Proportion 0.900 1.0000

#Proportion of Variance: 寄与率（PC1は全体の9割を説明している）

#Cumulative Proportion: 累積寄与率（PC2まで見れば全体の10割、つまりすべてを説明している）

#結果に含まれている情報のリスト

attributes(myResult)

#主成分（固有ベクトル）（Excel+Wolfram/Alphaと同じ結果。正負は逆でもよい）

myResult$rotation

#主成分の図示（Excelの場合と違って、大きい順に並び替えている）

barplot(sort(myResult$rotation[,1])) #第1主成分

barplot(sort(myResult$rotation[,2])) #第2主成分

#主成分スコア

myResult$x

#主成分スコアの図示

biplot(myResult)

## 定食チェーン店アンケート結果

定食チェーン店でのアンケート結果には、10個もの変数がある。変数を2つくらいに縮約して、店舗の位置づけを把握したい。（先の実験では使わなかった機能を赤字にしてある。）

setwd("c:/cit")

myData <- read.csv("定食チェーン店アンケート結果.csv")

head(myData)

#biplotの結果を見やすくするために、行に名前（店名）を付けておく

row.names(myData) <- myData$店名

myData

#主成分分析（全体を括弧で囲むと結果が表示される）

myResult <-prcomp(myData[,-1]) #1列目（店名）は不要

#結果の概要

summary(myResult)

#主成分（固有ベクトル）

myResult$rotation

#主成分の図示

barplot(sort(myResult$rotation[,1]))

barplot(sort(myResult$rotation[,2]))

#主成分スコア

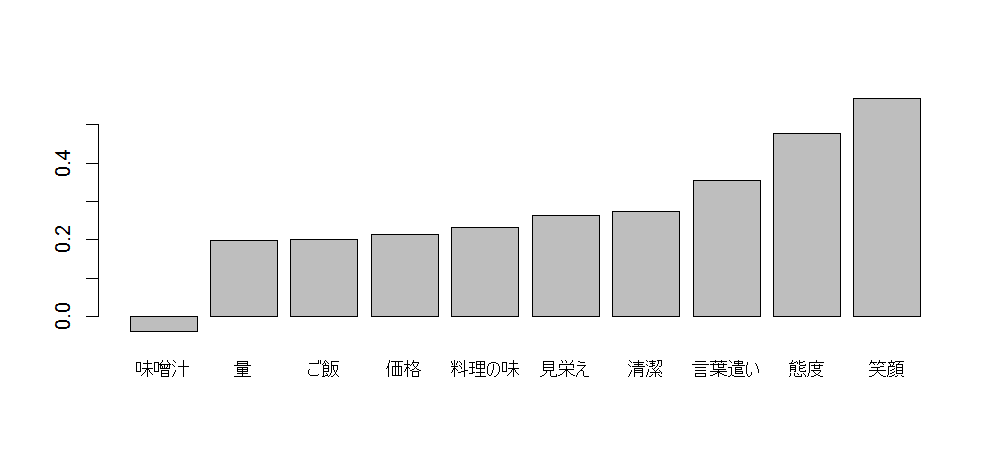
myResult$x

#主成分スコアの図示

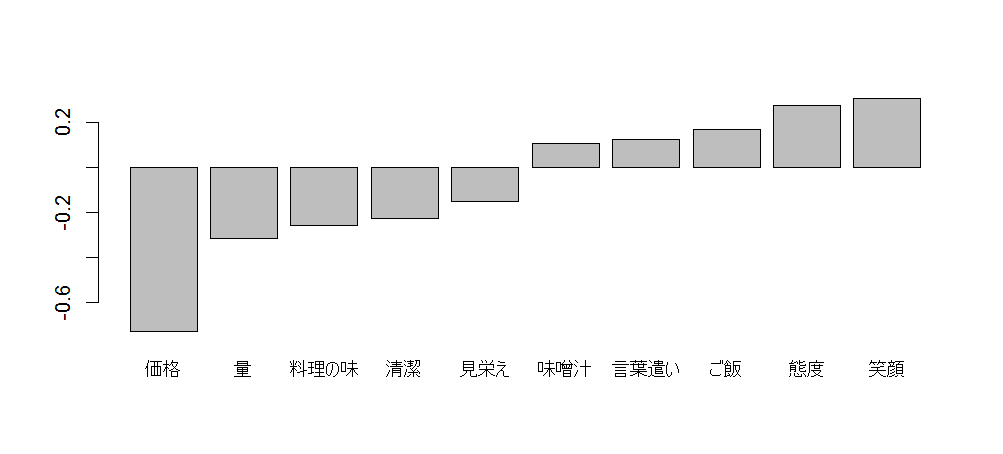
biplot(myResult)

### 主成分の解釈

第1主成分を図示すると次のようになる。ほとんどすべての要素が正だから、第1主成分は総合力を表す指標だと思われる。



第2主成分を図示すると次のようになる。第2主成分が大きいことは店の雰囲気がよいことを、ちいさいことは商品力が高いことを表していると思われる。



# レポート課題

「テスト結果」を主成分分析、主成分を図示し、解釈せよ。第1主成分と第2主成分は何を表しているか。ヒント：次のようなコマンドで主成分分析する（1列目（生徒No）は不要なため「[,-1]」としている。教科毎の分散を揃えるために「scale=TRUE」としている）。定食チェーン店アンケートの場合の「主成分の解釈」を参考にせよ。

(myResult <- prcomp(myData[,-1], scale=TRUE))