多変量解析

証券アナリスト

Takayuki Suzuki

This is institude of the author

多変量解析目標

線形代数（ベクトルと行列）では、「理解する」と「計算できる」が有る。

- 主成分分析の行列計算との対応
- 主成分分析の主成分ベクトルの意味が説明できる
- 主成分得点が計算できる
- 因子分析と主成分分析の違いが説明できる
- バリマックス回転？プロマックス回転？
- 判別分析、クラスター分析、マハラノビス距離などの用語を知っている

計算できる必要があるのは、２ｘ２の固有方程式から固有ベクトルを計算すること
および対角化、そして、関数を成分とするベクトルの微分表記、計算ができること。

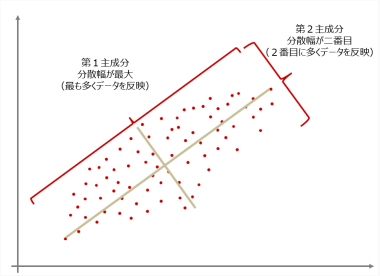
出る問題

結局、例の、水準（パラレルシフト）、傾き、曲率の問題に落ちる。
この問題について、主成分分析にフォーカスして解説する。例えば、水準はたいていの場合第一主成分だが、なぜ第一主成分なのか？など。

主成分分析

多変数からなるデータを分析する手法。
データを分析するとき、変数同士に相関が無い方が分析しやすい。
また、データの変動に影響が有る変数の順番が分かっていた方が分析しやすい。
しかし、一般的にデータは変数同士は相関があるし、どの変数が最も影響があるか、明らかではない。
主成分分析を行うと、１．互いに独立で、２．最も影響がある変数から順番に、新しい変数を作ることができる。

右は、主成分分析のイメージ。x,y の二つの変数のデータから、新しい軸（茶軸）を作る。
元のデータは、x,y で表現されると、相関がある。（斜めの傾向が見える）
しかし、新しい軸で見ると、相関が無いデータになる。
しかも、最も変動が大きい方向が第一主軸となっている。



=https://www.intage.co.jp/glossary/401/

主成分分析の計算方法

主成分分析は、データをよく表現する新しい変数を t 来ること。どうやって作る？
⇒ 元の変数の分散共分散行列の固有値よ固有ベクトルを求めればよい （覚える）
例えば、ある 2 変数 x_1, x_2 からなるデータが有って、それぞれの変数の分散が $\sigma_{x_1} = 4, \sigma_{x_2} = 7$, 共分散が $\mathrm{cov}(x_1, x_2) = 2, \mathrm{cov}(x_2, x_1) = 2$ であったとき (共分散は順番が変わっても同一であるから、分散共分散行列は常に対称行列である)、これを主成分分析するには、

分散共分散行列 $= \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}$ の固有値、固有ベクトルを求めればよい。前章の手順に従い、計算すると、固有値は $\lambda = 3, 8$, 対応する固有ベクトルは

$\boldsymbol{w}_1 = \begin{bmatrix} 2/\sqrt{3} \\ -1/\sqrt{3} \end{bmatrix}$ および $\boldsymbol{w}_2 \begin{bmatrix} 1/\sqrt{3} \\ 2/\sqrt{3} \end{bmatrix}$ であることが分かる（計算せよ）

絶対値が大きい固有値 8 に対応するベクトルが、第一主軸で、固有値 3 に対応するベクトルが、第二主軸である。

主成分得点

この主軸はどのように使うか？
あるサンプル A（サンプルというのは、データの中の一点のこと）を考える。
$A \sim \boldsymbol{x}_A = (x_1, x_2) = (3, 2)$ というデータが有ったとする。このデータは、新しい軸では

$$
\begin{aligned}
A \sim \boldsymbol{x}'_A = (x'_1, x'_2) &= (\text{A の第一主軸への射影成分}, \text{A の第二主軸への射影成分}) \\
&= (w_1 \cdot \boldsymbol{x}_a,\ \ w_2 \cdot \boldsymbol{x}_a) \\
&= \boldsymbol{P}\boldsymbol{x}_A \\
&= \begin{bmatrix} \frac{8}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}
\end{aligned}
$$

ただし、$\boldsymbol{P}$ は固有ベクトルを並べた行列

つまり、固有ベクトルとの内積を取れば求められる。内積は正射影ということもある。（行列 P を掛けるともいえる）
こうして求めた新しい軸での値 (x'_1, x'_2) のことを、主成分得点という。
主成分得点から、元のデータを再現する方法は、計算できるようにしておく。

寄与率

主成分のは、それぞれ寄与率というものがある。
データの分散のどの程度が、どの主成分で計算できるか、を示す。
それぞれの主軸の寄与率は、各主軸に対応する固有値を、すべての固有値の和で割ることで求められる。