

主成分分析

1. 【重要】主成分分析が導出できる
2. 【重要】主成分分析が導出できる(多次元)
3. 因子負荷量が導出できる
4. 主成分方向の平方和と固有値が一致する理由がわかる
5. 主成分分析ができる(2次元)
6. 主成分分析ができる(3次元)
7. 主成分分析ができる(5次元)
8. 【注意】平方和・相関行列から求めた固有値・固有ベクトルは一致しない
9. 主成分分析ができる(3次元で重解がある場合)

【重要】主成分分析が導出できる

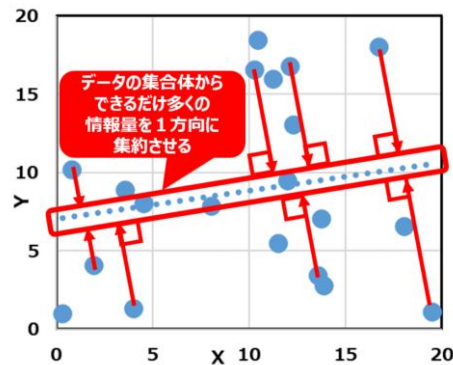
【1】【重要】主成分分析とは？

(1) 主成分分析はデータを要約するもの

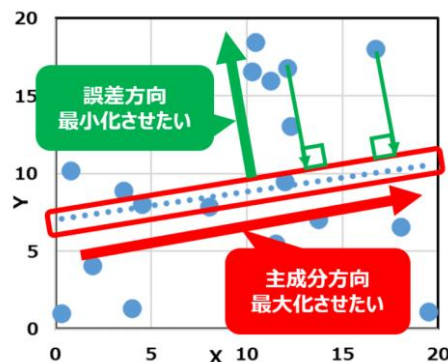
「主成分分析はデータを要約するもの
散布したデータの情報量が集まった方向を探す手法で
平方和を使って方向を探す」

ここはしっかり理解しましょう。

下図のように、青丸データが散布しており、その散布したデータからできるだけ多くの情報量が乗った方向を探し、散布したデータを集約させます。



その集約した方向が主成分方向であり、それと垂直な方向は誤差方向であり、主成分方向は情報量を最大化誤差方向は情報量を最小化させるのが主成分分析の目的です。



(2) 固有値から考える主成分分析がわからなくなる

主成分分析の解法を取得する時にどうしても固有値を求める固有方程式や固有ベクトルのイメージが強くなりがちです。

「主成分分析は固有値問題」を条件反射的にインプットするとかえって主成分分析が理解しにくくなる大事なのは、

平方和を使って情報量が多い方向を探す条件式を作ったら、結果的に固有方程式ができるだけの話。

(3) 重回帰分析との違いは何か？

よく同じデータが用意されると、「重回帰分析」と「主成分分析」は何が違うの？何で違う分析手法があるの？」と混乱します。

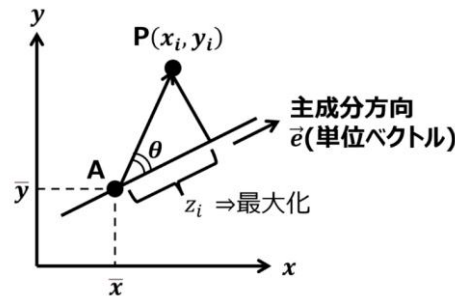
- 主成分分析は、平方和から散布したデータを要約する手法
- 重回帰分析は、平方和から散布したデータを誤差が最小になるように回帰直線を引く手法

【2】【重要】主成分分析は自分で導出できる

(1) 2次元でデータの情報を最大化する方向を立式

下図のように、点にある n 個のデータのうち、以下を定義します。

- 座標 $P(x, y)$
- 平均座標 $A(\bar{x}, \bar{y})$
- 主成分方向の単位ベクトル $\vec{e}$
- PA と主成分方向との角 θ
- 主成分方向の長さ z_i を情報量として評価する



まず、主成分方向の単位ベクトル $\vec{e}$ を $\vec{e} = \begin{pmatrix} a \\ b \end{pmatrix}$ と定義します。単位ベクトルなので、 $a^2 + b^2 = 1$ です。

①情報量 z_i の式を作る

次にまず、内積 $\overrightarrow{AP} \cdot \vec{e}$ を計算します。内積は高校数学の範囲ですね。

$$\overrightarrow{AP} \cdot \vec{e} = \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = a(x_i - \bar{x}) + b(y_i - \bar{y})$$

一方、内積 $\overrightarrow{AP} \cdot \vec{e}$ は $\overrightarrow{AP} \cdot \vec{e} = |\overrightarrow{AP}| |\vec{e}| \cos \theta$ です。ここで、単位ベクトルの大きさは $|\vec{e}| = 1$ なので、 $\overrightarrow{AP} \cdot \vec{e} =$

$$|\overrightarrow{AP}| |\vec{e}| \cos \theta = \overrightarrow{AP} \cdot \vec{e} = |\overrightarrow{AP}| \cos \theta = |z_i|$$

とうまくつながります。つまり、図と内積から

$$|z_i| = a(x_i - \bar{x}) + b(y_i - \bar{y})$$

です。これが散布したデータを要約した情報量の元となります。

$|z_i|$ は i 番目の情報量で全体の情報を最大化する条件式を作りたいので、
2乗和を取ります。何か平方和っぽいですが常套手段です。

②情報量 S の式を作る

情報量 S は $S = \sum_{i=1}^n |z_i|^2$ と書けるので、これを解いていきましょう。

$$S = \sum_{i=1}^n |z_i|^2 = \sum_{i=1}^n \{ a(x_i - \bar{x}) + b(y_i - \bar{y}) \}^2 = a^2 \sum_{i=1}^n (x_i - \bar{x})^2 + 2ab \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum_{i=1}^n (y_i - \bar{y})^2 \\ = a^2 S_{xx} + 2ab S_{xy} + b^2 S_{yy}$$

ここで、散布データは固定なので、

● S_{xx}, S_{xy} は定数

● a, b は変数

な点に注意ください。

まとめると、

$$\text{情報量 } S = a^2 S_{xx} + 2ab S_{xy} + b^2 S_{yy}$$

ただし、 $a^2 + b^2 = 1$

この条件で情報量 S を最大化（極値）をもつ条件式を求めたいので、

困ったときの「ラグランジュの未定係数法」を使って条件式を求めてみましょう。

(2) ラグランジュの未定係数法を使って条件式を決める

ラグランジュの未定係数は簡単にいうと

$g(a,b)=0$ のもとで、 $f(a,b)$ を最大化したいという等号制約付きの場合、関数 $L(a,b,\lambda)=f(a,b)-\lambda g(a,b)$ と置いて、

$$\bullet \frac{\partial L}{\partial a} = 0$$

$$\bullet \frac{\partial L}{\partial b} = 0$$

$$\bullet \frac{\partial L}{\partial \lambda} = 0$$

を同時に満たす解が最大化させる解となる。

情報量 $S(a,b)$ は

$g(a,b)=a^2+b^2-1=0$ のもと、 $S(a,b)=a^2S_{xx}+2abS_{xy}+b^2S_{yy}$ を最大化させたいので、

関数 $L(a,b,\lambda)=S(a,b)-\lambda g(a,b)$ と置いて、 $\frac{\partial L}{\partial a}=0$ 、 $\frac{\partial L}{\partial b}=0$ 、 $\frac{\partial L}{\partial \lambda}=0$ を同時に満たす方程式を解いてみましょう。

$$(L(a,b,\lambda) = (a^2S_{xx} + 2abS_{xy} + b^2S_{yy}) - \lambda(a^2 + b^2 - 1))$$

$$\bullet \frac{\partial L}{\partial a} = (2aS_{xx} + 2bS_{xy}) - \lambda 2a = 0$$

$$\bullet \frac{\partial L}{\partial b} = (2aS_{xy} + 2bS_{yy}) - \lambda 2b = 0$$

$$\bullet \frac{\partial L}{\partial \lambda} = -(a^2 + b^2 - 1) = 0$$

上の2つの式を整理して、行列表記します。

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

となり、よく見ると、2次元の固有方程式になっているのがわかりますね。

【3】「主成分分析＝固有値問題」と暗記するな！

(1) 情報量最大化条件が固有方程式になっただけ

情報量が最大化する条件をラグランジュの未定係数法から導出すると、

$$\bullet \begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

$$\bullet a^2 + b^2 = 1$$

となりました。つまり、

情報量の最大化の条件式が結果的に固有方程式になっただけで
主成分方向は固有値、固有ベクトルから話をスタートさせないこと！

(2) 固有値や固有ベクトルをイメージしても主成分分析は理解できない

固有値や固有ベクトルを可視化して、解法のイメージをわかりやすく解説する教科書やサイトが多いのですが、結局、その向きの意味がよくわからないから、主成分分析って何をやるものかがあいまいになってしまします。なので、固有方程式ではなく主成分分析の意図がより伝わる導出方法を本記事で解説しています。

情報量の最大化の条件式が結果的に固有方程式になっただけで
主成分方向は固有値、固有ベクトルから話をスタートさせないこと！

(3) 固有方程式は淡々と解けばいい

固有方程式は線形代数の代表選手なので、主成分分析よりインパクト強いんです。
だから「主成分分析」＝「固有値問題」と認識しがちです。

はっきり言って、

いろいろな場面で固有方程式が出て来るけど、固有値、固有ベクトルそのものにはあまり意味はないと
冷静に見た方が、各手法の本質が理解しやすい

固有方程式は淡々と解けばよいです。

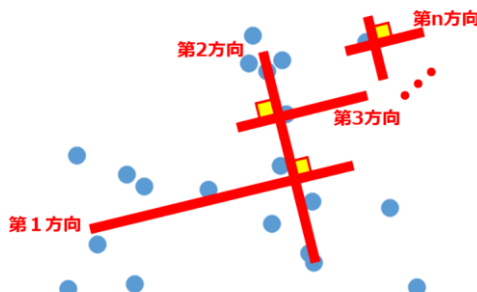
【4】主成分方向を理解する

(1) n 次元なら n 個の主成分方向がある

n 次元の固有方程式では、 n 個の固有値と固有ベクトルが計算できます。

なるべく少ない方向でたくさんの情報量が集めたいけど、数学的には、変数の数だけベクトルが解として出て来ます。

イメージは下図のとおりです。



n 個の方向が出て来る理由は、

第1方向である程度情報量を要約したら、その垂直な第2方向で情報量を要約する。
これを繰り返して、第 n 方向で情報量を要約して全データを要約するイメージ

と考えると理解しやすいです。

(2) 主成分方向は互いに直交する

それぞれの主成分方向はなぜか直交します。これは固有値解の特性でもあります。関連記事で具体例を挙げながら確かに直交することを確認しましょう。

(3) 主成分負荷量で方向の優先順位をつける

主成分方向は固有値解の個数だけあるので、優先順位をつけたいくなります。

寄与率や平方和、因子負荷量などの変数を使って評価していきます。これも関連記事で解説します。

以上、主成分分析の導出に必要なエッセンスを2次元の事例を使って解説しました。

主成分分析とは、散布データから、情報量を要約するもの。

情報量を最大限に要約する条件式が固有方程式になる。

主成分方向は固有ベクトルから求められるが、「主成分分析＝固有値問題」と暗記すると主成分分析の本質がぼやけるので要注意！

以上、「【重要】主成分分析が導出できる」を解説しました。

【重要】主成分分析が導出できる(多次元)

【1】主成分分析が導出(2次元)

(1) 主成分分析はデータを要約するもの

主成分分析の目的は、

「主成分分析はデータを要約するもの。散布したデータの情報が集まった方向を探す手法で平方和を使って方向を探す」

(2) 主成分分析が導出(2次元)のポイント

主成分分析とは、散布データから情報を要約するもの。

情報を最大限に要約する条件式が固有方程式になる。

主成分方向は固有ベクトルから求められるが、「主成分分析＝固有値問題」と暗記すると

主成分分析の本質がぼやけるので要注意！

【2】主成分分析が導出(3次元)

(1) 3次元でデータの情報を最大化する方向を立式

下図のように、点にある n 個のデータのうち、

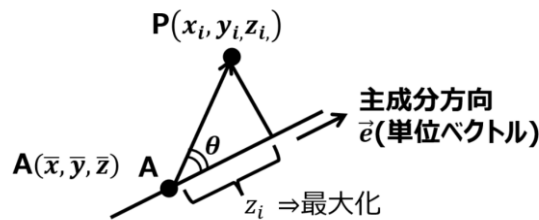
●座標 $P(x_i, y_i, z_i)$

●平均座標 $A(\bar{x}, \bar{y}, \bar{z})$

●主成分方向の単位ベクトル $\vec{e}$

●PA と主成分方向との角 θ

●主成分方向の長さ z_i を情報量として評価する
を定義します。



まず、主成分方向の単位ベクトル $\vec{e}$ を

$$\vec{e} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

と定義します。単位ベクトルなので、 $a^2 + b^2 + c^2 = 1$ です。この条件式はあとで何度も使います。

(2) 情報量 z_i の式を作る

次にまず、内積 $\vec{AP} \cdot \vec{e}$ を計算します。内積は高校数学の範囲ですね。

$$\begin{aligned} \text{内積 } \vec{AP} \cdot \vec{e} &= \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \\ z_i - \bar{z} \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = a(x_i - \bar{x}) + b(y_i - \bar{y}) + c(z_i - \bar{z}) \end{aligned}$$

また、内積 $\vec{AP} \cdot \vec{e} = |\vec{AP}| |\vec{e}| \cos \theta = |\vec{AP}| \cos \theta = |z_i|$ より、

$|z_i| = a(x_i - \bar{x}) + b(y_i - \bar{y}) + c(z_i - \bar{z})$
が散布したデータを要約した情報量の元となります。

(3) 情報量 S の式を作る

情報量 S は

$$\begin{aligned} S &= \sum_{i=1}^n |z_i|^2 = \sum_{i=1}^n \{a(x_i - \bar{x}) + b(y_i - \bar{y}) + c(z_i - \bar{z})\}^2 \\ &= a^2 \sum_{i=1}^n (x_i - \bar{x})^2 + b^2 \sum_{i=1}^n (y_i - \bar{y})^2 + c^2 \sum_{i=1}^n (z_i - \bar{z})^2 \\ &\quad + 2ab \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + 2ac \sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z}) + 2bc \sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z}) \\ &= a^2 S_{xx} + b^2 S_{yy} + c^2 S_{zz} + 2ab S_{xy} + 2ac S_{xz} + 2bc S_{yz} \end{aligned}$$

ここで、散布データは固定なので、

● $S_{xx}, S_{yy}, S_{zz}, S_{xy}, S_{xz}, S_{yz}$ は定数

● a, b, c は変数

な点に注意ください。まとめると、

$$\begin{aligned} \text{情報量 } S &= a^2 S_{xx} + b^2 S_{yy} + c^2 S_{zz} + 2ab S_{xy} + 2ac S_{xz} + 2bc S_{yz} \\ \text{ただし、} &a^2 + b^2 + c^2 = 1 \end{aligned}$$

この条件で情報量 S の最大化（極値）をもつ条件式を求めたいので、困ったときの「ラグランジュの未定係数法」を使って条件式を求めてみましょう。

(4) ラグランジュの未定係数法を使って条件式を決める

ラグランジュの未定係数は簡単にいうと

$$\begin{aligned} g(a, b, c) = 0 \text{ のもと、} f(a, b, c) \text{ を最大化したいという等号制約付きの場合、} \\ \text{関数 } L(a, b, c, \lambda) = f(a, b, c) - \lambda g(a, b, c) \text{ と置いて、} \\ \bullet \frac{\partial L}{\partial a} = 0 \\ \bullet \frac{\partial L}{\partial b} = 0 \\ \bullet \frac{\partial L}{\partial c} = 0 \\ \bullet \frac{\partial L}{\partial \lambda} = 0 \end{aligned}$$

を同時に満たす解が最大化させる解となる。

情報量 S は $g(a, b, c) = a^2 + b^2 + c^2 - 1 = 0$ のもと、

$S(a, b, c) = a^2 S_{xx} + b^2 S_{yy} + c^2 S_{zz} + 2ab S_{xy} + 2ac S_{xz} + 2bc S_{yz}$ を最大化させたいので、

関数 $L(a, b, c, \lambda) = f(a, b, c) - \lambda g(a, b, c)$ と置いて、最大化する方程式を解きます。

$$L(a, b, c, \lambda) = S(a, b, c) - \lambda g(a, b, c)$$

$$\begin{aligned} \bullet \frac{\partial L}{\partial a} &= (2a S_{xx} + 2b S_{xy} + 2c S_{xz}) - \lambda 2a = 0 \\ \bullet \frac{\partial L}{\partial b} &= (2a S_{xy} + 2b S_{yy} + 2c S_{yz}) - \lambda 2b = 0 \\ \bullet \frac{\partial L}{\partial c} &= (2a S_{xz} + 2b S_{yz} + 2c S_{zz}) - \lambda 2c = 0 \\ \bullet \frac{\partial L}{\partial \lambda} &= -(a^2 + b^2 + c^2 - 1) = 0 \end{aligned}$$

となります。この式を行列表記すると、

$$\begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

となり、よく見ると、

3次元の固有方程式になっているのがわかりますね。

(5)固有方程式を解けばよいけど

情報量の最大化の条件式が
結果的に固有方程式になっただけで
主成分方向は固有値、固有ベクトルから
話をスタートさせないこと！

3次元の場合の導出方法が理解できました！

2次元と全く同じ導出方法なので、手計算できますね。

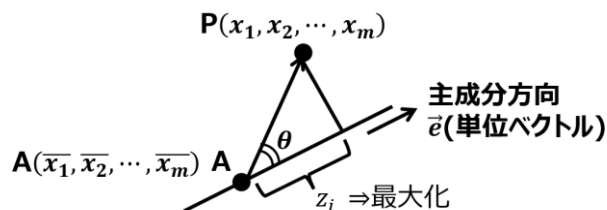
ただし、固有方程式の 3×3 逆行列はさすがに Excel 使ってください！

【3】主成分分析が導出(m次元)

(1) m次元でデータの情報を最大化する方向を立式

下図のように、点化する n 個のデータのうち、

- 座標 $P(x_{1i}, x_{2i}, \dots, x_{mi})$
- 平均座標 $A(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m)$
- 主成分方向の単位ベクトル $\vec{e}$
- PA と主成分方向との角 θ
- 主成分方向の長さ z_i を情報量として評価する
を定義します。



まず、主成分方向の単位ベクトル $\vec{e}$ を

$$\vec{e} = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_m \end{pmatrix}$$

と定義します。単位ベクトルなので、 $a_1^2 + a_2^2 + \dots + a_m^2 = 1$ です。この条件式はあとで何度も使います。
で、

(2) 情報量 z_i の式を作る

次にまず、内積 $\overrightarrow{AP} \cdot \vec{e}$ を計算します。内積は高校数学の範囲ですね。

$$\begin{aligned} & \text{内積 } \overrightarrow{AP} \cdot \vec{e} \\ &= \begin{pmatrix} x_{1i} - \bar{x}_1 \\ x_{2i} - \bar{x}_2 \\ \dots \\ x_{mi} - \bar{x}_m \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_m \end{pmatrix} \\ &= a_1(x_{1i} - \bar{x}_1) + a_2(x_{2i} - \bar{x}_2) + \dots + a_m(x_{mi} - \bar{x}_m) \end{aligned}$$

また、内積 $\overrightarrow{AP} \cdot \vec{e} = |\overrightarrow{AP}| |\vec{e}| \cos \theta = |\overrightarrow{AP}| \cos \theta = |z_i|$ より、

$|z_i| = a_1(x_{1i} - \bar{x}_1) + a_2(x_{2i} - \bar{x}_2) + \dots + a_m(x_{mi} - \bar{x}_m)$
が散布したデータを要約した情報量の元となります。

(3) 情報量 S の式を作る

情報量 S は

$$\begin{aligned}
 S &= \sum_{i=1}^n |z_i^2| \\
 &= \sum_{i=1}^n (a_1(x_{1i} - \bar{x}_1) + a_2(x_{2i} - \bar{x}_2) + \dots \\
 &\quad + a_m(x_{mi} - \bar{x}_m))^2 \\
 &= a_1^2 \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \\
 &\quad + a_2^2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 \\
 &\quad \dots \\
 &\quad + a_m^2 \sum_{i=1}^n (x_{mi} - \bar{x}_m)^2 \\
 &\quad + 2a_1a_2 \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \\
 &\quad + 2a_1a_3 \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{3i} - \bar{x}_3) \\
 &\quad \dots \\
 &\quad + 2a_{m-1}a_m \sum_{i=1}^n (x_{m-1,i} - \bar{x}_{m-1})(x_{mi} - \bar{x}_m)
 \end{aligned}$$

と展開できます。さらに Σ でまとめると、

$$\begin{aligned}
 &= \sum_{j=1}^m (a_j - 2S_{jj}) \\
 &\quad + 2 \sum_{j=1}^{m-1} \sum_{k=1, j \neq k}^m a_j a_k S_{jk} \\
 &\quad (S_{jk} = \sum_{i=1}^n (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k))
 \end{aligned}$$

ここで、散布データは固定なので、

● $S_{jk} (j = 1, \dots, m, k = 1, \dots, m)$ は定数

● $a_1, a_2, \dots, a_m$ は変数

な点に注意ください。まとめると、

$$\begin{aligned}
 \text{情報量 } S &= \sum_{j=1}^m (a_j - 2S_{jj}) + 2 \sum_{j=1}^{m-1} \sum_{k=1, j \neq k}^m a_j a_k S_{jk} \\
 \text{ただし、} & a_1^2 + a_2^2 + \dots + a_m^2 = 1
 \end{aligned}$$

この条件で情報量 S の最大化（極値）をもつ条件式を求めたいので、困ったときの「ラグランジュの未定係数法」を使って条件式を求めてみましょう。

(4) ラグランジュの未定係数法を使って条件式を決める

ラグランジュの未定係数は簡単にいうと

$g(a_1, a_2, \dots, a_m) = 0$ のもと、 $f(a_1, a_2, \dots, a_m)$ を最大化したいという等号制約付きの場合、関数 $L(a_1, a_2, \dots, a_m, \lambda) = f(a_1, a_2, \dots, a_m) - \lambda g(a_1, a_2, \dots, a_m)$ と置いて、

$$\bullet \frac{\partial L}{\partial a_1} = 0$$

$$\bullet \frac{\partial L}{\partial a_2} = 0$$

...

$$\bullet \frac{\partial L}{\partial a_m} = 0$$

$$\bullet \frac{\partial L}{\partial \lambda} = 0$$

を同時に満たす解が最大化させる解となる。

情報量 $S(a, b, c)$ は

$$g(a_1, a_2, \dots, a_m) = a_1^2 + a_2^2 + \dots + a_m^2 - 1 = 0 \text{ のもと、}$$

$$S(a_1, a_2, \dots, a_m) = \sum_{j=1}^m (a_j - 2S_{jj})$$

$$+ 2 \sum_{j=1}^{m-1} \sum_{k=1, j \neq k}^m a_j a_k S_{jk}$$

を最大化させたいので、

関数 $L(a, b, c, \lambda) = f(a_1, a_2, \dots, a_m) - \lambda g(a_1, a_2, \dots, a_m)$ と置いて、

条件式は結構複雑になりますが、

$$\begin{aligned} \bullet \frac{\partial L}{\partial a_1} &= (2a_1 S_{11} + 2a_2 S_{12} + \dots + 2a_m S_{1m}) - \lambda 2a_1 = 0 \\ \bullet \frac{\partial L}{\partial a_2} &= (2a_2 S_{22} + 2a_1 S_{12} + \dots + 2a_m S_{2m}) - \lambda 2a_2 = 0 \\ \bullet \frac{\partial L}{\partial a_m} &= (2a_m S_{mm} + 2a_1 S_{1m} + \dots + 2a_m S_{m-1,m}) - \lambda 2a_m = 0 \\ &\dots \\ \bullet \frac{\partial L}{\partial \lambda} &= -(a_1^2 + a_2^2 + \dots + a_m^2 - 1) = 0 \end{aligned}$$

となります。この式を行列表記すると、

$$\begin{pmatrix} S_{11} & S_{12} & \dots & S_{1m} \\ S_{12} & S_{22} & \dots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1m} & S_{2m} & \dots & S_{mm} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$$

となり、よく見ると、

m 次元の固有方程式になっているのがわかりますね。

(5)固有方程式を解けばよいけど

情報量の最大化の条件式が
結果的に固有方程式になっただけで
主成分方向は固有値、固有ベクトルから
話をスタートさせないこと！

n 次元の場合の導出方法が理解できました！

2,3 次元と全く同じ導出方法なので、手計算できますね。

ただし、固有方程式の **m × m** 逆行列はさすがに **Excel** 使ってください！

2 次元の導出方法が分かれば理解できます！

主成分分析の導出方法はあまり解説していないので、

2 次元、3 次元、n 次元としつこく同じ解法で導出方法を解説しました。

以上、「【重要】主成分分析が導出できる(多次元)」を解説しました。

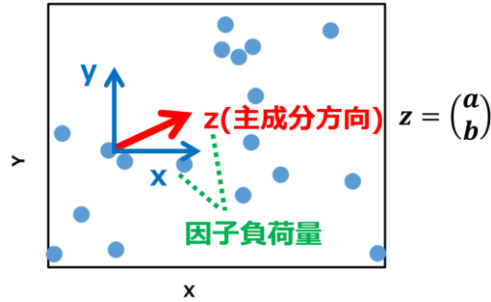
- 教科書の式は $r_{zx} = \sqrt{\lambda}\alpha$
- 実際解くと、 $r_{zx} = \frac{\sqrt{\lambda}\alpha}{S_{xx}}$
と式は異なるが、でも標準化して平方和 $S_{xx}=1$ の場合は一致

【1】因子負荷量とは

(1) 因子負荷量とは

因子負荷量は、いろいろ定義やイメージがありますが、シンプルに書くと

主成分方向と元の説明変数との相関係数



上図は2次元の場合なので、因子負荷量は2つあり、 r_{zx}, r_{zy} を計算します。

なお、上図では主成分方向 z はベクトル表記しますが、相関係数を計算するときは、

- $|z_i| = a(x_i - \bar{x}) + b(y_i - \bar{y})$

を使います。

$|z_i|$ からラグランジュの未定係数法を使って固有方程式を使って主成分方向を導出する過程は本冊子の【【重要】主成分分析が導出できる】でまとめています。

(2) 因子負荷量の式

教科書と比較すると QC プラネッツの式は少し違います。

- 教科書の式は $r_{zx} = \sqrt{\lambda}\alpha$
- 実際解くと、 $r_{zx} = \frac{\sqrt{\lambda}\alpha}{S_{xx}}$
と式は異なるが、でも標準化して平方和 $S_{xx}=1$ の場合は一致

教科書は他のサイトは、 $r_{zx} = \sqrt{\lambda}\alpha$ にするために、説明変数 x_i を標準化 $\frac{x_i - \bar{x}}{s}$ して平均0、分散1に改造してから解くことが多いです。

教科書の導出過程は長いし、複雑で理解しにくい! もっと短くシンプルな導出過程にならないのか?

という思いで、シンプルに導出します。ちょっと式が変わるけど、平方和 $S=1$ の場合は、教科書と同じ式になります。

【2】主成分分析の導出がスタート

2次元と3次元の場合で解説します。解き方は全く同じです。2回解き方を見れば、理解度は増しますよね!

本冊子、

【【重要】主成分分析が導出できる】、

【【重要】主成分分析が導出できる(多次元)】

で確認ください。

(2) 2次元の場合の大事なポイント

情報量 S を最大化するために、ラグランジュの未定係数法から方程式を作ると、固有方程式になります。

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

(3) 3次元の場合の大事なポイント

2次元と同様に解くと、

$$\begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

何次元であっても、

1. 情報量 S を定義
2. ラグランジュの未定係数法から立式
3. 固有方程式を作る

流れは同じです。以上を基礎において、因子負荷量を導出していきます。

【3】因子負荷量を導出(2次元)

(1) 因子負荷量を導出

2次元の場合なので、因子負荷量は2つあり、 r_{zx}, r_{zy} を計算します。

$$\begin{aligned} \bullet r_{zx} &= \frac{S_{zx}}{\sqrt{S_{zz}S_{xx}}} \\ &= \frac{\sum_{i=1}^n (a(x_i - \bar{x}) + b(y_i - \bar{y}))(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (a(x_i - \bar{x}) + b(y_i - \bar{y}))^2 \sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= \frac{\sum_{i=1}^n (a(x_i - \bar{x})^2 + b(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^n (a(x_i - \bar{x}) + b(y_i - \bar{y}))^2 \sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= \frac{aS_{xx} + bS_{xy}}{\sqrt{a^2S_{xx} + 2abS_{xy} + b^2S_{yy}} \sqrt{S_{xx}}} \end{aligned}$$

さて、固有方程式を再掲します。

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

式で表記すると

$$\bullet aS_{xx} + bS_{xy} = \lambda a$$

$$\bullet aS_{xy} + bS_{yy} = \lambda b$$

$$\bullet r_{zx} = \frac{aS_{xx} + bS_{xy}}{\sqrt{a^2S_{xx} + 2abS_{xy} + b^2S_{yy}} \sqrt{S_{xx}}}$$

の分子は上の式そのものですね。

$aS_{xx} + bS_{xy} = \lambda a$ を代入しましょう。

また、分母を

$$a^2 S_{xx} + 2ab S_{xy} + b^2 S_{yy}$$

$$= a(a S_{xx} + b S_{xy}) + b(a S_{xy} + b S_{yy})$$

と変形して、固有方程式を代入すると

$$= aa\lambda + bb\lambda$$

$$= (a^2 + b^2)\lambda$$

さらに a, b は単位ベクトルなので、

$$(a^2 + b^2) = 1 \text{より、}$$

$$= \lambda$$

となります。

まとめると

$$r_{zx} = \frac{\lambda a}{\sqrt{\lambda} \sqrt{S_{xx}}}$$

$$= \frac{\sqrt{\lambda} a}{\sqrt{S_{xx}}}$$

となります。一方、 S_{xx} は定数なので、このままです。標準化して $S_{xx}=1$ とすることが多いのはこの理由です。

$$r_{zx} = \frac{\sqrt{\lambda} a}{\sqrt{S_{xx}}}$$

$(S_{xx}=1 \text{の場合は、} r_{zx} = \sqrt{\lambda} a)$

r_{zy} も同様に計算ができて、

$$r_{zy} = \frac{\sqrt{\lambda} b}{\sqrt{S_{yy}}}$$

$(S_{yy}=1 \text{の場合は、} r_{zy} = \sqrt{\lambda} b)$

(2) 標準化してから因子負荷量を解く理由

標準化して因子負荷量を解く理由は、式を簡略したいためです。でも、 x, z を標準化する過程も必要なので、導出過程が面倒です。今回はあえて標準化しない場合で導出しました。

自分で解いてみると、教科書と一致しない式が出来たりしますが、その違いを理解することが大事です。

【4】因子負荷量を導出(3次元)

(1) 因子負荷量を導出

3次元の場合なので、因子負荷量は3つあり、変数が多いので、数字で区別します。

r_{z1} 、 r_{z2} 、 r_{z3} を計算します。

2次元の場合と全く同じ解法で導出できます。

$$\bullet r_{z1} = \frac{S_{z1}}{\sqrt{S_{z1} S_{11}}}$$

$$= \frac{\sum_{i=1}^n (a(x_{1i} - \bar{x}_1) + b(x_{2i} - \bar{x}_2) + c(x_{3i} - \bar{x}_3))(x_{1i} - \bar{x}_1)}{\sqrt{\sum_{i=1}^n (a(x_{1i} - \bar{x}_1) + b(x_{2i} - \bar{x}_2) + c(x_{3i} - \bar{x}_3))^2 \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}}$$

$$= \frac{aS_{11} + bS_{12} + cS_{13}}{\sqrt{a^2 S_{11} + b^2 S_{22} + c^2 S_{33} + 2abS_{12} + 2bcS_{23} + 2acS_{13}} \sqrt{S_{11}}}$$

さて、関連記事をベースに固有方程式を再掲します。

$$\begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{23} & S_{33} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

式で表記すると

$$\bullet aS_{11} + bS_{12} + cS_{13} = \lambda a$$

$$\bullet aS_{21} + bS_{22} + cS_{23} = \lambda b$$

$$\bullet aS_{31} + bS_{32} + cS_{33} = \lambda c$$

$$\bullet r_{zx} = \frac{aS_{11} + bS_{12} + cS_{13}}{\sqrt{a^2S_{11} + b^2S_{22} + c^2S_{33} + 2abS_{12} + 2bcS_{23} + 2acS_{13}} \sqrt{S_{11}}}$$

の分子は上の式そのものですね。

$aS_{11} + bS_{12} + cS_{13} = \lambda a$ を代入しましょう。

また、分母を

$$a^2S_{11} + b^2S_{22} + c^2S_{33} + 2abS_{12} + 2bcS_{23} + 2acS_{13}$$

$$= a(aS_{11} + bS_{12} + cS_{13})$$

$$+ b(aS_{21} + bS_{22} + cS_{23})$$

$$+ c(aS_{31} + bS_{32} + cS_{33})$$

と変形して、固有方程式を代入すると

$$= aa\lambda + bb\lambda + cc\lambda$$

$$= (a^2 + b^2 + c^2)\lambda$$

さらに a, b, c は単位ベクトルなので、

$$(a^2 + b^2 + c^2) = 1 \text{より、}$$

$$= \lambda$$

となります。

まとめると

$$r_{z1} = \frac{\lambda a}{\sqrt{\lambda} \sqrt{S_{11}}}$$

$$= \frac{\sqrt{\lambda} a}{\sqrt{S_{11}}}$$

となります。

一方、 S_{xx} は定数なので、このままです。標準化して $S_{xx}=1$ とすることが多いのはこの理由です。

r_{z2}, r_{z3} も同様に計算ができて、

$$r_{z2} = \frac{\sqrt{\lambda} b}{\sqrt{S_{22}}}$$

$$r_{z3} = \frac{\sqrt{\lambda} c}{\sqrt{S_{33}}}$$

$$(S_{22}=1, (S_{33}=1 \text{の場合は、} r_{z2}=\sqrt{\lambda}b), r_{z3}=\sqrt{\lambda}c)$$

と計算できます。

変数が増えても因子負荷量の公式の形が変化しないので面白いですね。

以上、「因子負荷量が導出できる」を解説しました。

【1】主成分の平方和を計算する

(1) 平方和を計算

主成分分析は点存在するデータから最大限の情報量を抽出する方法です。その情報量を平方和で求めますよね！ これは本冊子【【重要】主成分分析が導出できる】解説していますので、ご確認ください。

情報量 $\backslash(S\backslash)$ は

$$S = \sum_{i=1}^n |z_i|^2 \\ = a^2 S_{xx} + 2ab S_{xy} + b^2 S_{yy} \text{ (2次元)}$$

$$S = \sum_{i=1}^n |z_i|^2 \\ = a^2 S_{xx} + b^2 S_{yy} + c^2 S_{zz} \\ + 2ab S_{xy} + 2bc S_{yz} + 2ca S_{xz} \text{ (3次元)}$$

で表現できます。

(2) 主成分分析を導出する関係式

主成分分析を解く場合は、必ず次の関係式を使って解きます。1つの解法ですべて解けます！

情報量 S (平方和) から、ラグランジュの未定係数法を使って、固有方程式を作る。
固有ベクトルは単位ベクトルとする。

関係式を並べると

(i) 2次元の場合

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} \begin{pmatrix} a_j \\ b_j \end{pmatrix} = \lambda_j \begin{pmatrix} a_j \\ b_j \end{pmatrix}$$

ただし、

● $i=1, 2, \dots, n$ (データ数)

● $j=1, 2$ (固有値の種類)

$a_j^2 + b_j^2 = 1$ (単位ベクトル)

(ii) 3次元の場合

$$\begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{pmatrix} \begin{pmatrix} a_j \\ b_j \\ c_j \end{pmatrix} = \lambda_j \begin{pmatrix} a_j \\ b_j \\ c_j \end{pmatrix}$$

ただし、

● $i=1, 2, \dots, n$ (データ数)

● $j=1, 2, 3$ (固有値の種類)

$a_j^2 + b_j^2 + c_j^2 = 1$ (単位ベクトル)

証明方法は、本冊子【因子負荷量が導出できる】でも書いていますので、ご確認ください。

【2】主成分の平方和と固有値が一致する理由がわかる(2次元)

(1)今回証明したい式は

「主成分の平方和と固有値が一致する」

$$S_j = \lambda_j$$
$$S_j = \sum_{i=1}^n \left(a_j(x_i - \bar{x}) + b_j(y_i - \bar{y}) \right)^2$$

j 方向の平方和は j 番目の固有値に一致するということです。

証明します！

(2) 証明

まず、固有方程式を行列表記から式に直すと、

● $a_j S_{xx} + b_j S_{xy} = \lambda_j a_j$

● $a_j S_{xy} + b_j S_{yy} = \lambda_j b_j$

ですね。

次に j 番目の平方和(情報量)の式は

$$S_j = \sum_{i=1}^n |z_i|^2 = a_j^2 S_{xx} + 2a_j b_j S_{xy} + b_j^2 S_{yy} \quad (2 \text{ 次元})$$

また、

$$S_j = \sum_{i=1}^n |z_i|^2 = a_j^2 S_{xx} + 2a_j b_j S_{xy} + b_j^2 S_{yy} = a_j (a_j S_{xx} + b_j S_{xy}) + b_j (a_j S_{xy} + b_j S_{yy})$$

と変形すると、固有方程式を使って

$$= a_j (\lambda_j a_j) + b_j (\lambda_j b_j) = \lambda_j (a_j^2 + b_j^2) = \lambda_j \quad (a_j^2 + b_j^2 = 1)$$

よって、 $S_j = \lambda_j$

「主成分の平方和と固有値が一致する」

$$S_j = \lambda_j$$

結構シンプルに証明できました。

【3】主成分の平方和と固有値が一致する理由がわかる(3次元)

(1)今回証明したい式は

「主成分の平方和と固有値が一致する」

$$S_j = \lambda_j$$
$$S_j = \sum_{i=1}^n \left(a_j(x_i - \bar{x}) + b_j(y_i - \bar{y}) + c_j(z_i - \bar{z}) \right)^2$$

j 方向の平方和は j 番目の固有値に一致するということです。

証明します！

(2) 証明

まず、固有方程式を行列表記から式に直すと、

$$\bullet \quad a_j S_{xx} + b_j S_{xy} + c_j S_{xz} = \lambda_j a_j$$

$$\bullet \quad a_j S_{yx} + b_j S_{yy} + c_j S_{yz} = \lambda_j b_j$$

$$\bullet \quad a_j S_{zx} + b_j S_{zy} + c_j S_{zz} = \lambda_j c_j$$

ですね。

次に j 番目の平方和(情報量)の式は

$$S_j = \sum_{i=1}^n |z_i|^2 = a_j^2 S_{xx} + b_j^2 S_{yy} + c_j^2 S_{zz} + 2a_j b_j S_{xy} + 2a_j c_j S_{xz} + 2b_j c_j S_{yz} \quad (3 \text{ 次元})$$

また、

$$S_j = \sum_{i=1}^n |z_i|^2 = a_j^2 S_{xx} + b_j^2 S_{yy} + c_j^2 S_{zz} + 2a_j b_j S_{xy} + 2a_j c_j S_{xz} + 2b_j c_j S_{yz}$$
$$= a_j(a_j S_{xx} + b_j S_{xy} + c_j S_{xz}) + b_j(a_j S_{yx} + b_j S_{yy} + c_j S_{yz}) + c_j(a_j S_{zx} + b_j S_{zy} + c_j S_{zz})$$

と変形すると、固有方程式を使って

$$= a_j(\lambda_j a_j) + b_j(\lambda_j b_j) + c_j(\lambda_j c_j) = \lambda_j(a_j^2 + b_j^2 + c_j^2) = \lambda_j \quad (a_j^2 + b_j^2 + c_j^2 = 1)$$

よって、 $S_j = \lambda_j$

「主成分の平方和と固有値が一致する」

$$S_j = \lambda_j$$

結構シンプルに証明できました。

同様に、 n 次元の場合も同様に証明できます。シンプルに証明できました！

以上、「主成分方向の平方和と固有値が一致する理由がわかる」を解説しました。

主成分分析ができる(2次元)

【1】例題

(1) 主成分分析の本質

主成分分析はいろいろな値が計算できますが、本質をおさえることが最重要です。

本冊子【【重要】主成分分析が導出できる】で確認しましょう。

(2) 例題

2次元の主成分分析の例題は次の通りです。

【問題】以下のデータにおいて、主成分分析せよ。

(1) 相関係数行列 R

(2) 固有値 λ_i

(3) 固有ベクトル v_i

(4) 寄与率と累積寄与率

(5) 因子負荷量 (標準化)

(6) 主成分得点 (標準化)

データ(下表)

No	x	y
1	0	1
2	1	3
3	2	2
4	3	4
5	4	5
合計	10	15
平均	2	3

2次元の主成分分析なので、手計算で解いて主成分分析の理解を深めましょう。

(3) データの標準化

因子負荷量や主成分得点などをツールから計算するために、データを標準化しておきます。

データの標準化は $z_i = \frac{x_i - \bar{x}}{s}$ と変換して、平均 0、標準偏差 $s=1$ の変数 z_i に変換することです。

上のデータ表からは、標準偏差を計算し、データを標準化します。(標準偏差は $x:1.41, y:1.41$)

No	x	y
1	-1.414	-1.414
2	-0.707	0
3	0	-0.707
4	0.707	0.707
5	1.414	1.414
合計	0	0
平均	0	0
標準偏差	1	1

確かに、平均 0、標準偏差 1 に変換できていますね。

【2】 相関係数行列の計算

(1) 各平方和を先に計算

科目ごとの平方和を先に計算します。

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 10$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = 10$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 9$$

(2) 相関係数行列の計算

相関係数 r_{ij} は

$$\bullet r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

から計算し、まとめたら相関係数行列できます。平方和 S_{ij} から相関係数 r_{ij} を計算すると下表になります。

相関係数 r_{ij}	x	y
x	$r_{xx}=1$	$r_{xy}=0.9$
y	$r_{xy}=0.9$	$r_{yy}=1$

よって相関係数行 R は

$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

2次元なら、あえて相関係数行列は要らないですが、3次元、5次元と関連記事で解説していますので、流れは理解しておきましょう。

【3】 固有値、固有ベクトルの計算

(1) 固有値、固有ベクトルの解法

ツールを使った解法を解説する前に、主成分分析の本質を再確認しましょう。本記事のテーマではありませんが、主成分分析は必ず固有方程式を解きます。平方和でも相関係数でもどちらでも固有方程式は解けますが、今回は相関係数を使って解きます。

固有方程式

$$Rv = \lambda v$$

より

$$\begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

を満たす、固有値 λ と固有ベクトル v を解きます。

(2) 【重要】固有値の計算

2×2 行列における、固有方程式を書くと

$$Rv = \lambda v$$

$$(R - \lambda E)v = 0$$

より、行列式 $|(R - \lambda E)| = 0$ を満たす λ を計算します。 2×2 行列なので、 $ad - bc = 0$ の簡単な式になります。

$$(R - \lambda E) = \begin{pmatrix} r_{11} - \lambda & r_{12} \\ r_{21} & r_{22} - \lambda \end{pmatrix} = \begin{pmatrix} 1 - \lambda & 0.9 \\ 0.9 & 1 - \lambda \end{pmatrix}$$

よって、固有値は

● $\lambda_1=1.9$

● $\lambda_2=0.1$

となります。

(3) 【重要】固有ベクトルの計算

① $\lambda_1=1.9$ の場合

$$\begin{pmatrix} -0.9 & 0.9 \\ 0.9 & -0.9 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

から、関係式を作ると

● $-0.9v_{11}+0.9v_{12}=0$

● 単位ベクトル

を考えると

よって、固有ベクトル v_1 は

$$v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

② $\lambda_2=0.1$ の場合

同様に解くと、

$$v_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

(4) 固有値、固有ベクトルの計算結果を検算

固有値の計算の理解を深める方法の1つとして、実際に検算することをお勧めします。

実際、検算すると $Rv = \lambda v$ が成り立ちます。ここからは、個々の値を求めていきましょう。

単に数字があてばOKではなく、値の意味をまず関連記事で確認しましょう。

【関連記事】主成分分析が計算できる

<https://qcplanets.com/method/multivariate/parameters/>

【4】主成分の寄与率、累積寄与率

(1) 主成分の寄与率、累積寄与率

● 主成分の寄与率は、個々の固有値を自由度で割った値ですね。

● 累積寄与率は、寄与率の累積値です。

固有値はそれぞれ $\lambda_1=1.9$, $\lambda_2=0.1$ です。結果は下表のとおりです。

主成分寄与率	主成分1	主成分2
固有値	1.9	0.1
寄与率	0.95	0.05
累積寄与率	0.95	1

(2) 寄与率を平方和でなく固有値で考える理由

各主成分の平方和は固有値になります。これは関連記事で解説していますが、一番簡単にわかる証明方法があります。

固有方程式 $Sv = \lambda v$ から
 全ての変数をスカラー（数値）と考えると、
 両辺を v で割れば、
 (平方和) $S = \lambda$ (固有値)

主成分平方和は固有値として考えてよいということです。平方和と固有値は別物と思いがちですが、一致しています。意外ですね。</mark>

【5】因子負荷量

(1) 因子負荷量の公式

因子負荷量の公式を下表にまとめます。公式から計算しましょう。

因子負荷量	主成分1	主成分2
主成分1	$v_{11}\sqrt{\lambda_1}$	$v_{21}\sqrt{\lambda_2}$
主成分2	$v_{12}\sqrt{\lambda_1}$	$v_{22}\sqrt{\lambda_2}$

(2) 因子負荷量の計算結果

因子負荷量	主成分1	主成分2
主成分1	0.975	0.975
主成分2	-0.224	0.224

【6】主成分負荷量

「主成分負荷量」名前は立派ですが、簡単にいうと固有ベクトルの個々の値です。すでに表で結果が出ていますが、再掲します。

主成分負荷量	主成分1	主成分2
主成分1	0.7071	0.7071
主成分2	-0.7071	0.7071

【7】主成分得点

主成分得点の求め方は、

主成分得点は、標準化した各データに個々の主成分負荷量を掛け算して足し合わせたもの
 つまり、
 主成分得点 $z_i = \lambda_1 x_{1j} + \lambda_2 x_{2j}$

主成分得点の結果を下表にまとめます。

No	主成分1	主成分2
1.000	-2.000	1.000
2.000	-0.500	0.000
3.000	-0.500	0.500
4.000	1.000	-0.500
5.000	2.000	-1.000
合計	0.000	0.000

2次元主成分分析でしっかり解法の流れを理解しましょう。

主成分分析は解析のボリュームが多いですが、何を求めているのかを常に意識しましょう。

以上、「主成分分析ができる(2次元)」を解説しました。

主成分分析ができる(3次元)

【1】例題

(1) 主成分分析の本質

まず、本冊子【重要】主成分分析が導出できる】で復習しましょう。

(2) 例題

【問題】以下のデータにおいて、主成分分析せよ。

(1) 相関係数行列 R

(2) 固有値 λ_i

(3) 固有ベクトル v_i

(4) 寄与率と累積寄与率

(5) 因子負荷量 (標準化)

(6) 主成分得点 (標準化)

データ(下表)

No	x1	x2	x3
1	1	5	4
2	2	4	2
3	3	3	3
4	4	4	1
5	5	5	5
和	15	21	15
平均	3	4.2	3

3次元の主成分分析ですが、手計算で解いて主成分分析の理解を深めましょう。

(3) データの標準化

因子負荷量や主成分得点などをツールから計算するために、データを標準化しておきます。

データの標準化は $z_i = \frac{x_i - \bar{x}}{s}$ と変換して、平均 0、標準偏差 $s=1$ の変数 z_i に変換することです。

上のデータ表からは、標準偏差を計算すると、下表になります。

No	x1	x2	x3
標準偏差 s	1.41	0.75	1.41

なので、データを標準化します。結果は下表のとおりです。

No	x1	x2	x3
1	-1.414	1.069	0.707
2	-0.707	-0.267	-0.707
3	0	-1.604	0
4	0.707	-0.267	-1.414
5	1.414	1.069	1.414
和	0	0	0
平均	0	0	0
標準偏差	1	1	1

確かに、平均 0、標準偏差 1 に変換できていますね。

【2】相関係数行列の計算

(1) 各平方和を先に計算

科目ごとの平方和を先に計算します。

$$S_{ij} = \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

として、x1 と x2 の平方和は

$$S_{12} = \sum_{k=1}^5 (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

と計算します。

全部のパターンを計算した結果を下表にまとめます。

平方和S	x1	x2	x3
x1	$S_{11}=10$	$S_{12}=0$	$S_{13}=1$
x2	$S_{21}=0$	$S_{22}=2.8$	$S_{23}=3$
x3	$S_{31}=1$	$S_{32}=3$	$S_{33}=10$

(2) 相関係数行列の計算

相関係数 r_{ij} は

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

から計算できて、それをまとめたら相関係数行列が計算できますね。

まず、さっき求めた平方和 S_{ij} から相関係数 r_{ij} を計算すると下表になります。

相関係数 r_{ij}	x1	x2	x3
x1	$r_{11}=1$	$r_{12}=0$	$r_{13}=0.1$
x2	$r_{21}=0$	$r_{22}=1$	$r_{23}=0.567$
x3	$r_{31}=0.1$	$r_{32}=0.567$	$r_{33}=1$

よって相関係数行 R は

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0.1 \\ 0 & 1 & 0.567 \\ 0.1 & 0.567 & 1 \end{pmatrix}$$

ここで、相関係数 $r_{12}=0$ となっています。変な感じですが、このあとの行列式を簡単に解くためにあえて、相関係数 0 になるように、x2 のデータを作っています。x2-x1 のグラフをプロットすると 2 次関数っぽい感じになります。(プロットしてみてくださいね。)

【3】固有値、固有ベクトルの計算

(1) 固有値、固有ベクトルの解法

ツールを使った解法を解説する前に、主成分分析の本質を再確認しましょう。

本記事のテーマではありませんが、主成分分析は必ず固有方程式を解きます。

★固有方程式

平方和でも相関係数でもどちらでも固有方程式は解けますが、今回は相関係数を使って解きます。

固有方程式 $Rv = \lambda v$ より

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

を満たす、固有値 λ と固有ベクトル v を解きます。

3次元になりますが、行列式を解いて手計算で行きましょう。
 その後で、解析ツールに頼りましょう。
 手計算で何を解いているのかの理解を深めましょう。

(2) 【重要】固有値の計算

3×3行列における、固有方程式を書くと

$Rv = \lambda v$ 、 $(R - \lambda E)v = 0$ より、行列式 $|R - \lambda E| = 0$ を満たす λ を計算します。

$$R - \lambda E = \begin{pmatrix} r_{11} - \lambda & r_{12} & r_{13} \\ r_{21} & r_{22} - \lambda & r_{23} \\ r_{31} & r_{32} & r_{33} - \lambda \end{pmatrix}$$

$$|R - \lambda E| = \begin{vmatrix} r_{11} - \lambda & r_{12} & r_{13} \\ r_{21} & r_{22} - \lambda & r_{23} \\ r_{31} & r_{32} & r_{33} - \lambda \end{vmatrix} = \begin{vmatrix} 1 - \lambda & 0 & b \\ 0 & 1 - \lambda & a \\ b & a & 1 - \lambda \end{vmatrix}$$

($a = 0.567, b = 0.1$ とします。)

行列式の公式を紹介すると

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} - \lambda \end{vmatrix}$$

$$= A_{11}A_{22}A_{33} + A_{12}A_{23}A_{31} + A_{13}A_{21}A_{32} - A_{13}A_{22}A_{31} - A_{12}A_{21}A_{33} - A_{11}A_{23}A_{32}$$

となります。線形代数の教科書に書いていますので参考ください。</div>

行列式を解くと

$$\begin{vmatrix} 1 - \lambda & 0 & b \\ 0 & 1 - \lambda & a \\ b & a & 1 - \lambda \end{vmatrix}$$

$$= (1 - \lambda)^3 + 0 + 0 - b^2(1 - \lambda) - 0 - a^2(1 - \lambda)$$

$$= (1 - \lambda)((1 - \lambda)^2 - (a^2 + b^2)) = 0$$

となり、綺麗に因数分解できます。

一般的には、因数分解できない3次方程式が多いので、数値解析しますが、今回は勉強用に因数分解でできるデータを用意しました。

よって、固有値 λ は

$$\lambda = 1, 1 \pm \sqrt{a^2 + b^2} \text{ に } a = 0.567, b = 0.1 \text{ と}$$

$$\bullet \lambda_1=1.5757$$

$$\bullet \lambda_2=1$$

$$\bullet \lambda_3=0.4243$$

となります。 $c = \sqrt{a^2 + b^2}$ として、今後

$$\bullet \lambda_1=1+c$$

$$\bullet \lambda_2=1$$

$$\bullet \lambda_3=1-c$$

と簡略化して、最後に値を代入しましょう。

(3) 【重要】固有ベクトルの計算

① $\lambda_1=1+c$ の場合

$$\begin{pmatrix} -c & 0 & b \\ 0 & -c & a \\ b & a & -c \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \\ v_{13} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

から、関係式を作ると

$$\bullet -cv_{11} + bv_{13} = 0$$

$$\bullet -cv_{12} + av_{13} = 0$$

$$v_{13}=1 \text{ とすると } \quad \bullet v_{11} = \frac{b}{c} \quad \bullet v_{12} = \frac{a}{c}$$

固有ベクトル $v_1 = \begin{pmatrix} b \\ a \\ c \end{pmatrix}$ となり、大きさ $|v_1|^2 = a^2 + b^2 + c^2 = 2c^2$ より、 $|v_1| = \sqrt{2}c$ となります。よって、

$$v_1 = \frac{1}{\sqrt{2}c} \begin{pmatrix} 0.1228 \\ 0.6965 \\ 0.7072 \end{pmatrix}$$

② $\lambda_2=1$ の場合

同様に解くと、

$$\begin{pmatrix} 0 & 0 & b \\ 0 & 0 & a \\ b & a & 0 \end{pmatrix} \begin{pmatrix} v_{21} \\ v_{22} \\ v_{23} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

から、関係式を作ると

$$\bullet bv_{23} = 0$$

$$\bullet av_{23} = 0$$

$$\bullet bv_{21} + av_{22} = 0$$

を満たしつつ、単位ベクトルとなるように解くと、固有ベクトル v_2 は

$$v_2 = \begin{pmatrix} 0.9848 \\ -0.1737 \\ 0 \end{pmatrix}$$

③ $\lambda_3=1-c$ の場合

同様に解くと

$$v_3 = \begin{pmatrix} 0.1228 \\ 0.6964 \\ -0.7071 \end{pmatrix}$$

となります。

以上、固有値と固有ベクトルを下表にまとめます。

固有ベクトル	主成分1	主成分2	主成分3
x1	0.1228	0.9848	0.1228
x2	0.6964	-0.1737	0.6964
x3	0.7071	0	-0.7071
固有値	1.5757	1	0.4243

なお、関連記事でも確認しましょう。

【関連記事】主成分分析が計算できる

<https://qcplanets.com/method/multivariate/parameters/>

【4】主成分の寄与率、累積寄与率

(1) 主成分の寄与率、累積寄与率

●主成分の寄与率は、個々の固有値を自由度で割った値ですね。

●累積寄与率は、寄与率の累積値です。

結果は下表のとおりです。

主成分寄与率	主成分1	主成分2	主成分3
固有値	1.5757	1	0.4243
寄与率	0.525	0.333	0.142
累積寄与率	0.525	0.858	1

(2) 寄与率を平方和でなく固有値で考える理由

各主成分の平方和は固有値になります。関連記事で解説しますが、一番簡単にわかる証明方法があります。

固有方程式 $Sv = \lambda v$ から

全ての変数をスカラー（数値）と考えると、両辺を v で割れば、

(平方和) $S = \lambda$ (固有値)

確かに、(平方和) $S = \lambda$ (固有値) となりますよね。これを行列、ベクトル表記にして n 次元化していますが、考え方は同じなので、主成分平方和は固有値として考えてよいということです。平方和と固有値は別物と思いがちですが、一致しています。意外ですよね。

【5】因子負荷量

(1) 因子負荷量の公式

因子負荷量の公式を下表にまとめます。公式から計算しましょう。

因子負荷量	主成分1	主成分2	主成分3
主成分1	$v_{11}\sqrt{\lambda_1}$	$v_{21}\sqrt{\lambda_2}$	$v_{31}\sqrt{\lambda_3}$
主成分2	$v_{12}\sqrt{\lambda_1}$	$v_{22}\sqrt{\lambda_2}$	$v_{32}\sqrt{\lambda_3}$
主成分3	$v_{13}\sqrt{\lambda_1}$	$v_{23}\sqrt{\lambda_2}$	$v_{33}\sqrt{\lambda_3}$

(2) 因子負荷量の計算結果

因子負荷量	主成分1	主成分2	主成分3
主成分1	0.1542	0.8741	0.8876
主成分2	0.9848	-0.1737	0
主成分3	0.08	0.4536	-0.4606

【5】主成分負荷量

「主成分負荷量」は、簡単にいうと固有ベクトルの個々の値です。すでに表で結果が出ています。

主成分負荷量	主成分1	主成分2	主成分3
主成分1	0.1228	0.6964	0.7071
主成分2	0.9848	-0.1737	0
主成分3	0.1228	0.6964	-0.7071

【6】主成分得点

主成分得点の求め方は、

主成分得点は、標準化した各データに個々の主成分負荷量を掛け算して足し合わせたもの
つまり、主成分得点 $z_i = \lambda_1 x_{1j} + \lambda_2 x_{2j} + \lambda_3 x_{3j}$

主成分得点の結果を下表にまとめます。

主成分 得点	主成分1	主成分2	主成分3
1	-0.9577	-1.4118	0.0633
2	0.6914	-0.5813	0.2031
3	0.9988	0.2491	-0.9988
4	0.9832	0.6644	0.8056
5	-1.7156	1.0796	-0.0732

長い計算でしたが、3次元主成分分析ができましたね。

以上、「主成分分析ができる(3次元)」を解説しました。

主成分分析ができる(5次元)

【1】例題

(1) 主成分分析の本質

本冊子 【【重要】主成分分析が導出できる】で復習しましょう。

(2)QC 検定® 1 級受験に必須な教科書にある、5次元の主成分分析を実際に解いてみましょう。

【問題】30 人に対して 5 科目(英語、数学、国語、理科、社会)の試験を実施した点数結果である。これを主成分分析せよ。

- (1)相関係数行列 R (2)固有値 λ_i (3)固有ベクトル v_i
(4)寄与率と累積寄与率 (5)因子負荷量 (標準化) (6)主成分得点 (標準化)

データ(下表)

No	英語(1)	数学(2)	国語(3)	理科(4)	社会(5)
1	74	90	74	95	73
2	79	100	74	89	82
3	74	68	70	60	77
4	70	66	64	74	62
5	53	49	54	41	51
6	63	43	55	51	61
7	64	42	71	56	59
8	48	36	61	47	51
9	64	88	69	91	73
10	77	78	72	72	82

11	79	76	74	77	64
12	86	89	75	78	78
13	46	30	61	51	63
14	90	78	71	83	76
15	67	45	61	51	67
16	64	41	66	44	58
17	57	65	69	59	56
18	78	79	73	97	85
19	56	43	62	56	62
20	96	92	79	93	78

21	71	53	67	48	65
22	91	74	80	86	92
23	59	42	63	58	52
24	58	42	62	48	53
25	72	95	75	98	66
26	90	98	76	84	82
27	84	77	87	79	82
28	58	41	65	45	67
29	63	46	58	55	65
30	79	71	76	69	83
合計	2110	1937	2064	2035	2065
平均	70.33	64.57	68.8	67.83	68.83
標準偏差	13.03	21.23	7.63	18.14	11.28

ここで、各科目に数字を入れています。英語なら「1」、数学なら「2」として区別します。

(2) データの標準化

因子負荷量や主成分得点などをツールから計算するために、データを標準化しておきます。

データの標準化は $z_i = \frac{x_i - \bar{x}}{s}$ と変換して、平均 0、標準偏差 $s = 1$ の変数 z_i に変換することです。

結果は下表のとおりです。

No	英語(1)	数学(2)	国語(3)	理科(4)	社会(5)
1	0.281	1.198	0.682	1.498	0.369
2	0.665	1.669	0.682	1.167	1.167
3	0.281	0.162	0.157	-0.432	0.724
4	-0.026	0.067	-0.629	0.34	-0.606
5	-1.33	-0.733	-1.941	-1.48	-1.581
6	-0.563	-1.016	-1.81	-0.928	-0.694
7	-0.486	-1.063	0.288	-0.652	-0.872
8	-1.713	-1.345	-1.023	-1.149	-1.581
9	-0.486	1.104	0.026	1.277	0.369
10	0.511	0.633	0.42	0.23	1.167
11	0.665	0.538	0.682	0.505	-0.428
12	1.202	1.151	0.813	0.561	0.813
13	-1.867	-1.628	-1.023	-0.928	-0.517
14	1.509	0.633	0.288	0.836	0.635
15	-0.256	-0.921	-1.023	-0.928	-0.163
16	-0.486	-1.11	-0.367	-1.314	-0.96
17	-1.023	0.02	0.026	-0.487	-1.138
18	0.588	0.68	0.551	1.608	1.433
19	-1.1	-1.016	-0.892	-0.652	-0.606
20	1.969	1.292	1.337	1.388	0.813
21	0.051	-0.545	-0.236	-1.094	-0.34
22	1.586	0.444	1.469	1.002	2.054
23	-0.87	-1.063	-0.761	-0.542	-1.492
24	-0.946	-1.063	-0.892	-1.094	-1.403
25	0.128	1.433	0.813	1.663	-0.251
26	1.509	1.574	0.944	0.891	1.167
27	1.049	0.586	2.386	0.616	1.167
28	-0.946	-1.11	-0.498	-1.259	-0.163
29	-0.563	-0.874	-1.416	-0.708	-0.34
30	0.665	0.303	0.944	0.064	1.256
合計	0	0	0	0	0
平均	0	0	0	0	0
標準偏差	1	1	1	1	1

確かに、平均 0、標準偏差 1 に変換できていますね。

【2】相関係数行列の計算

(1) 各平方和を先に計算

科目ごとの平方和を先に計算します。

$$S_{ij} = \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

として、英語と数学の平方和は

$$S_{12} = \sum_{k=1}^{n=30} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

と計算します。全部のパターンを計算した結果を下表にまとめます。

平方和 S	英語(1)	数学(2)	国語(3)	理科(4)	社会(5)
英語(1)	30	23.72	23.99	21.99	24.53
数学(2)	23.72	30	22.73	27.29	21.5
国語(3)	23.99	22.73	30	22.41	22.63
理科(4)	21.99	27.29	22.41	30	21.19
社会(5)	24.53	21.5	22.63	21.19	30

(2) 相関係数行列の計算

相関係数 r_{ij} は

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

から相関係数行列が計算できますね。

平方和 S	英語(1)	数学(2)	国語(3)	理科(4)	社会(5)
英語(1)	1	0.791	0.8	0.733	0.818
数学(2)	0.791	1	0.758	0.91	0.717
国語(3)	0.8	0.758	1	0.747	0.754
理科(4)	0.733	0.91	0.747	1	0.706
社会(5)	0.818	0.717	0.754	0.706	1

よって相関係数行列 R は

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & r_{33} & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{34} & r_{44} & r_{45} \\ r_{15} & r_{25} & r_{35} & r_{45} & r_{55} \end{pmatrix} = \begin{pmatrix} 1 & 0.791 & 0.8 & 0.733 & 0.818 \\ 0.791 & 1 & 0.758 & 0.91 & 0.717 \\ 0.8 & 0.758 & 1 & 0.747 & 0.754 \\ 0.733 & 0.91 & 0.747 & 1 & 0.706 \\ 0.818 & 0.717 & 0.754 & 0.706 & 1 \end{pmatrix}$$

【3】固有値、固有ベクトルの計算

(1) 固有値、固有ベクトルの解法

ツールを使った解法を解説する前に、主成分分析の本質を再確認しましょう。

本記事のテーマではありませんが、主成分分析は必ず固有方程式を解きます。

平方和でも相関係数でもどちらでも固有方程式は解けますが、今回は相関係数を使って解きます。

固有方程式 $Rv = \lambda v$ より

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & r_{33} & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{34} & r_{44} & r_{45} \\ r_{15} & r_{25} & r_{35} & r_{45} & r_{55} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix}$$

を満たす、固有値 λ と固有ベクトル v を解きます。ただし、1つ問題があります。

5次元になると、固有方程式から5次方程式を解く必要があるため、手計算ではムリ！解析ツールを使って解きましょう。

(2)固有値、固有ベクトルの解法ツール

ツールは何でもいいですが、今回は、フリーソフトを使って計算しました。

<https://mam-mam.net/download/scatterplot/> mam 相関分析・主成分分析～（無料）フリーソフト

使い方はリンク先をご覧ください。

(3) 固有値、固有ベクトルの計算

実際に固有値と、固有ベクトルを解くと、

固有値	λ_1	λ_2	λ_3	λ_4	λ_5
固有値	4.094	0.4071	0.2398	0.1786	0.0805

固有ベクトル	主成分1	主成分2	主成分3	主成分4	主成分5
英語1	-0.4525	-0.3398	0.0753	0.7768	-0.2661
数学2	-0.4567	0.4857	0.1173	0.1796	0.7138
国語3	-0.4431	-0.2177	-0.8302	-0.2509	0.0642
理科4	-0.4478	0.5723	0.112	-0.2386	-0.6343
社会5	-0.4356	-0.5231	0.5281	-0.4945	0.1148

固有ベクトルは、最初に紹介した「QC 検定®受検テキスト 1 級」の結果と比較すると正負が入れ替わっています。ただし、固有方程式の両辺に固有ベクトルが付くので、正負±は両辺で消えますので、問題ありません。

(4) 固有値、固有ベクトルの計算結果を検算

固有値の計算の理解を深める方法の1つとして、実際に検算することをお勧めします。

実際、検算すると $Rv = \lambda v$ が成り立ちます。ここからは、個々の値を求めていきましょう。

単に数字があっていればOKではなく、値の意味をまず関連記事で確認しましょう。

【関連記事】主成分分析が計算できる

<https://qcplanets.com/method/multivariate/parameters/>

【4】主成分の寄与率、累積寄与率

(1) 主成分の寄与率、累積寄与率

●主成分の寄与率は、個々の固有値を自由度で割った値ですね。

●累積寄与率は、寄与率の累積値です。

結果は下表のとおりです。

寄与率	寄与率	累積寄与率
主成分1	$\frac{\lambda_1}{p=5} = \frac{4.094}{5} = 0.819$	0.819
主成分2	$\frac{\lambda_2}{p=5} = \frac{0.4071}{5} = 0.081$	0.9
主成分3	$\frac{\lambda_3}{p=5} = \frac{0.2398}{5} = 0.048$	0.948
主成分4	$\frac{\lambda_4}{p=5} = \frac{0.1786}{5} = 0.036$	0.984
主成分5	$\frac{\lambda_5}{p=5} = \frac{0.0805}{5} = 0.016$	1

(2) 寄与率を平方和でなく固有値で考える理由

各主成分の平方和は固有値になります。これは関連記事で解説していますが、一番簡単にわかる証明方法があります。

固有方程式 $Rv = \lambda v$ から全ての変数をスカラー（数値）と考えると、
両辺を v で割れば、(平方和) $S = \lambda$ (固有値)

確かに、(平方和) $S = \lambda$ (固有値) となりますよね。これを行列、ベクトル表記にして n 次元化していますが、考え方は同じなので、主成分平方和は固有値として考えてよいということです。平方和と固有値は別物と思いがちですが、一致しています。意外ですよね。

【5】因子負荷量

(1) 因子負荷量の公式

因子負荷量の公式を下表にまとめます。公式から計算しましょう。

因子負荷量	主成分1	主成分2	主成分3	主成分4	主成分5
英語	$v_{11}\sqrt{\lambda_1}$	$v_{21}\sqrt{\lambda_2}$	$v_{31}\sqrt{\lambda_3}$	$v_{41}\sqrt{\lambda_4}$	$v_{51}\sqrt{\lambda_5}$
数学	$v_{12}\sqrt{\lambda_1}$	$v_{22}\sqrt{\lambda_2}$	$v_{32}\sqrt{\lambda_3}$	$v_{42}\sqrt{\lambda_4}$	$v_{52}\sqrt{\lambda_5}$
国語	$v_{13}\sqrt{\lambda_1}$	$v_{23}\sqrt{\lambda_2}$	$v_{33}\sqrt{\lambda_3}$	$v_{43}\sqrt{\lambda_4}$	$v_{53}\sqrt{\lambda_5}$
理科	$v_{14}\sqrt{\lambda_1}$	$v_{24}\sqrt{\lambda_2}$	$v_{34}\sqrt{\lambda_3}$	$v_{44}\sqrt{\lambda_4}$	$v_{54}\sqrt{\lambda_5}$
社会	$v_{15}\sqrt{\lambda_1}$	$v_{25}\sqrt{\lambda_2}$	$v_{35}\sqrt{\lambda_3}$	$v_{45}\sqrt{\lambda_4}$	$v_{55}\sqrt{\lambda_5}$

(2) 因子負荷量の計算結果

因子負荷量	主成分1	主成分2	主成分3	主成分4	主成分5
英語	-0.916	-0.217	0.037	0.328	-0.075
数学	-0.924	0.31	0.057	0.076	0.203
国語	-0.897	-0.139	-0.407	-0.106	0.018
理科	-0.906	0.365	0.055	-0.101	-0.18
社会	-0.881	-0.334	0.259	-0.209	0.033

最初に紹介した「QC 検定® 受検テキスト 1 級」の結果と比較すると正負が入れ替わっています。固有ベクトルの正負が逆であることが理由です。特に問題はありませんが、気になるようなら、固有ベクトルの標準化対応して個々の値を計算すればよいでしょう。正負に自由度があるようなので、値が一致しないと焦る気持ちはよくわかりますが、数式的には問題ありません。

【5】主成分負荷量

「主成分負荷量」名前は立派ですが、簡単にいうと固有ベクトルの個々の値です。すでに表で結果が出ていますが、再掲します。

主成分負荷量	主成分1	主成分2	主成分3	主成分4	主成分5
英語1	-0.4525	-0.3398	0.0753	0.7768	-0.2661
数学2	-0.4567	0.4857	0.1173	0.1796	0.7138
国語3	-0.4431	-0.2177	-0.8302	-0.2509	0.0642
理科4	-0.4478	0.5723	0.112	-0.2386	-0.6343
社会5	-0.4356	-0.5231	0.5281	-0.4945	0.1148

【6】主成分得点

主成分得点の求め方は、

主成分得点は、標準化した各データに個々の主成分負荷量を掛け算して足し合わせたもの

つまり、主成分得点 $z_i = \lambda_1 x_{1j} + \lambda_2 x_{2j} + \lambda_3 x_{3j} + \lambda_4 x_{4j} + \lambda_5 x_{5j}$

計算は Excel でもいいでしょうね。具体的な計算の様子は下図のとおりです。主成分得点の結果を下表にまとめます。

No	主成分1	主成分2	主成分3	主成分4	主成分5
1	-1.808	1.002	-0.042	-0.277	-0.084
2	-2.396	0.493	0.427	-0.21	0.452
3	-0.393	-0.677	0.243	-0.047	0.408
4	0.371	0.69	0.247	0.369	-0.271
5	3.148	0.498	0.425	0.457	0.463
6	2.238	-0.076	0.87	0.399	-0.182
7	1.249	-0.331	-0.934	-0.054	-0.297
8	3.046	0.321	-0.401	-0.26	-0.023
9	-1.029	1.233	0.409	-0.673	0.151
10	-1.318	-0.437	0.406	-0.226	0.331

21	0.968	-0.679	-0.166	0.43	0.237
22	-2.914	-1.144	0.149	-0.311	-0.41
23	2.109	0.415	-0.407	0.192	-0.403
24	2.41	0.108	-0.319	0.253	-0.032
25	-1.708	1.559	-0.444	-0.12	-0.043
26	-2.728	-0.054	0.231	0.428	0.352
27	-2.583	-0.85	-1.148	-0.403	0.036
28	1.79	-0.745	-0.015	-0.429	0.208
29	1.746	-0.152	0.772	0.098	-0.156
30	-1.433	-0.904	-0.028	-0.302	0.203
合計	0	0	0	0	0

11	-0.889	0.4	-0.622	0.533	-0.119
12	-2.035	-0.131	0.042	0.401	0.291
13	2.682	-0.194	0.141	-1.009	-0.201
14	-1.751	-0.122	0.377	0.7	-0.389
15	1.476	-0.584	0.532	0.194	-0.085
16	1.896	-0.544	-0.516	0.304	0.037
17	1.156	0.668	-0.752	-0.119	0.467
18	-2.165	0.181	0.604	-0.652	-0.492
19	1.913	0.018	0.145	-0.358	-0.145
20	-3.049	0.036	-0.226	0.693	-0.303

長い計算でしたが、5次元主成分分析ができましたね。主成分分析は解析のボリュームが多いですが、何を求めているのかを常に意識しましょう。

以上、「主成分分析ができる(5次元)」を解説しました。

【1】平方和行列と相関係数行列は別物

(1) 平方和行列

主成分分析で固有値を求める際、固有方程式を作ります。

$$\begin{pmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{12} & S_{22} & \dots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1n} & S_{2n} & \dots & S_{nn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

となり、 $n \times n$ 行列における固有方程式ができます。この、

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{12} & S_{22} & \dots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1n} & S_{2n} & \dots & S_{nn} \end{pmatrix}$$

は平方和から計算した行列ですよね。

(2) 相関係数行列

一方、データを標準化してから、主成分分析する場合があります。その場合は、平方和ではなく、相関係数を使って固有方程式を作ります。

$$\begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{12} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \dots & r_{nn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

となり、 $n \times n$ 行列における固有方程式ができます。この、

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{12} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \dots & r_{nn} \end{pmatrix}$$

(3) 平方和行列と相関係数行列は別物

さて、固有値・固有ベクトルを算出する式が2つありますが、結果は同じなのかというと、異なります。なぜなら、

平方和行列と相関係数行列は別物だから

式でいうと、

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{12} & S_{22} & \dots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1n} & S_{2n} & \dots & S_{nn} \end{pmatrix} \neq \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{12} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \dots & r_{nn} \end{pmatrix} = R$$

なのです。

同じデータでも、固有方程式は

$$Sv = \lambda v$$

$$Rv = \lambda v$$

で $S \neq R$ なら、 λ, v は変わってきますよね！

【2】平方和行列と相関係数行列から固有値・固有ベクトルを算出

平方和行列と相関係数行列からそれぞれ固有値・固有ベクトルを算出して比較してみましょう。わかりやすく説明するために2次元データで考えます。

(1) データ事例

下表のデータを用意します。必要な数値も準備しておきます。

No	x	y	x' (標準化)	y' (標準化)
1	0	3	-1.414	0.267
2	1	2	-0.707	-1.069
3	2	2	0	-1.069
4	3	4	0.707	1.604
5	4	3	1.414	0.267
合計	10	14	0	0
平均	2	2.8	0	0
標準偏差	1.41	0.75	-	-

平方和と相関係数も計算しておきますが、是非、確かめてみてください。

x,y	値	標準化 x',y'	値
S_{xx}	10	$S_{x'x'}$	5
S_{xy}	2	$S_{x'y'}$	1.89
S_{yy}	2.8	$S_{y'y'}$	5
R	0.143	R	0.143

(2) 平方和行列から固有値・固有ベクトルを算出

平方和行列は

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} = \begin{pmatrix} 10 & 2 \\ 2 & 2.8 \end{pmatrix}$$

より、固有方程式を作ると、

$$Sv = \lambda v \quad \begin{pmatrix} 10 & 2 \\ 2 & 2.8 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

固有方程式から

$$(10 - \lambda)(2.8 - \lambda) - 2^2 = 0$$

$$\lambda = 10.518, 2.282$$

また単位ベクトルを気にせず、固有ベクトルを計算すると（実際やってみてください）

$$v_1 = \begin{pmatrix} 1 \\ 0.259 \end{pmatrix} \quad v_2 = \begin{pmatrix} 1 \\ -3.859 \end{pmatrix}$$

(3) 相関係数行列から固有値・固有ベクトルを算出

相関係数行列 R は

$$R = \begin{pmatrix} r_{xx} & r_{xy} \\ r_{xu} & r_{yu} \end{pmatrix} = \begin{pmatrix} 1 & 0.143 \\ 0.143 & 1 \end{pmatrix}$$

より、固有方程式を作ると、

$$Rv = \lambda v \quad \begin{pmatrix} 1 & 0.143 \\ 0.143 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

固有方程式から

$$(1 - \lambda)^2 - 0.143^2 = 0$$

$$\lambda = 1.143, 0.857$$

また単位ベクトルを気にせず、固有ベクトルを計算すると（実際やってみてください）

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

となります。

固有値と固有ベクトルを比較すると

	平方和から	相関係数行列から	一致、不一致
固有値1	10.518	1.143	×
固有ベクトル1	$\begin{pmatrix} 1 \\ 0.259 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	×
固有値2	2.282	0.857	×
固有ベクトル2	$\begin{pmatrix} 1 \\ -3.859 \end{pmatrix}$	$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$	×

と、固有値も固有ベクトルも一致していません。

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{12} & S_{22} & \dots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1n} & S_{2n} & \dots & S_{nn} \end{pmatrix} \neq \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{12} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \dots & r_{nn} \end{pmatrix} = R$$

から、固有値と固有ベクトルが一致しませんが、

両者に関係性があれば、固有値・固有ベクトルは一致する場合があります。

【3】平方和行列と相関係数行列は関係性がある場合がある
もともと相関係数は

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

です。基本的には、データはランダムデータですから、当然

$$S_{ii} \neq S_{jj}$$

これが、行列

$$S \neq R \text{ の理由です。}$$

(1) 平方和が等しい場合

$$S_{ii} = S_{jj}$$

の場合は、

$$\mathbf{S} = \text{定数} \times \mathbf{R}$$

の関係性ができます。

$S_{ii} = S_{jj} = T$ として、仮に $S_{ij} = kT$ とおくと

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{kT}{\sqrt{T^2}} = k \text{ となります。}$$

そうすると、平方和行列と相関係数行列に関係性が生まれ、

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1n} \\ S_{12} & S_{22} & \dots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1n} & S_{2n} & \dots & S_{nn} \end{pmatrix} = \begin{pmatrix} T & kT & \dots & kT \\ kT & T & \dots & kT \\ \vdots & \vdots & \ddots & \vdots \\ kT & kT & \dots & T \end{pmatrix} = T \begin{pmatrix} 1 & k & \dots & k \\ k & 1 & \dots & k \\ \vdots & \vdots & \ddots & \vdots \\ k & k & \dots & 1 \end{pmatrix} = TR$$

となり、 $\mathbf{S} = T \times \mathbf{R}$ と関係式ができます。

(2) 平方和が 1 の場合

さらに、平方和の値 $T=1$ まで特別なデータを用意すると、

$$\mathbf{S} = \mathbf{R}$$

となり、この場合、平方和でも相関係数でも固有値・固有ベクトルは一致します。

平方和行列と相関係数行列に関係性があるデータの場合をこれから、実際に主成分分析してみましょう。

【4】平方和が等しい場合

(1) データ事例

実データではほぼありえませんが、下表のように、 x, y 両方の平方和が等しい場合を想定します。

No	x	y
1	0	0
2	1	2
3	2	3
4	3	1
5	4	4
合計	10	10
平均	2	2
標準偏差	1.41	1.41

平方和と相関係数は

x,y	値
S_{xx}	10
S_{xy}	7
S_{yy}	10
R	0.143

平方和行列 $\mathbf{S}$ と相関係数行列 $\mathbf{R}$ はそれぞれ

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix} = \begin{pmatrix} 10 & 7 \\ 7 & 10 \end{pmatrix}$$
$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$

$\mathbf{S} = 10 \mathbf{R}$ の関係となることがわかりますね。

(2) 平方和行列から固有値・固有ベクトルを算出

固有方程式は $\mathbf{S}v = \lambda v$ より

$$\begin{pmatrix} 10 & 7 \\ 7 & 10 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

より、

$$(10 - \lambda)^2 = 7^2$$

よって、固有値 $\lambda = 17, 3$ 。固有ベクトルは単位ベクトルを気にせず求めると

$$v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

となります。

(3) 相関係数行列から固有値・固有ベクトルを算出

固有方程式は $\mathbf{R}v = \lambda v$

$$\begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

より、

$$(1 - \lambda)^2 = 0.7^2$$

よって、固有値 $\lambda = 1.7, 0.3$ 。

平方和行列の場合の $1/10$ の値になります。固有ベクトルは単位ベクトルを気にせず求めると

$$v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

と平方和行列の場合と同じ結果になります。

平方和行列と相関係数行列のそれぞれ結果をまとめると、

- | |
|--|
| 1. 行列は $\mathbf{S} = k\mathbf{R}$ の関係が成り立つ |
| 2. 固有値は k 倍異なる |
| 3. 固有値ベクトルは同じになる |

【5】平方和が 1 の場合

(1) データ事例

先のデータでは平方和 $S_{xx} = S_{yy} = 10$ でしたので、データを $1/\sqrt{10}$ 倍に変えてみましょう。

(2) データ事例

実データではほぼありませんが、下表のように、 x, y 両方の平方和が等しく、1 となる場合を想定します。

No	x	y
1	0	0
2	0.316	0.632
3	0.632	0.949
4	0.949	0.316
5	1.265	1.265
合計	3.16	3.16
平均	0.632	0.632
標準偏差	0.45	0.45

平方和と相関係数は

x,y	値
S_{xx}	1
S_{xy}	0.7
S_{yy}	1
R	0.143

平方和行列**S**と相関係数行列**R**はそれぞれ

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$
$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$

となり、

S = R の関係となることがわかりますね。

(3) 行列から固有値・固有ベクトルを算出

固有方程式は $Sv = \lambda v$

$$\begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

より、

$$(1 - \lambda)^2 = 0.7^2$$

よって、固有値 $\lambda = 1.7, 0.3$

固有ベクトルは単位ベクトルを気にせず求めると

$$v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

となります。

平方和行列と相関係数行列のそれぞれ結果をまとめると、

1. 行列は **S=R** の関係が成り立つ
2. 固有値も固有ベクトルも同じになる
3. でも、非常にレアなデータである点に注意

以上、「【注意】平方和・相関係数行列から求めた固有値・固有ベクトルは一致しないがわかる」を解説しました。

主成分分析ができる(3次元で重解がある場合)

【1】2次元データの場合、重解な固有値はない

(1) 主成分分析の本質

主成分分析はいろいろな値が計算できますが、本質をおさえることが最重要です。本冊子【重要】主成分分析が導出できる】で復習しましょう。

(2) 2次元データが重解な固有値を持つための条件

2次元データの場合の固有方程式は、固有方程式 $Rv = \lambda v$ より

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

から固有値を計算します。なお、主成分分析で固有方程式を解くときは、対称行列を使いますね。

固有方程式から、条件式

$$(a - \lambda)(c - \lambda) - b^2 = 0 \quad \text{から}$$

$$\lambda^2 - (a + c)\lambda + (ac - b^2) = 0$$

と2方程式を作り、重解をもつので、判別式 $D=0$ ですね。

$$\text{判別式 } D = (a + c)^2 - 4(ac - b^2) = 0 \quad \text{から}$$

$$(a - c)^2 + 4b^2 = 0$$

となります。等号成立条件は $a = c$ かつ $b = 0$ となり、これは単位行列しかないことになります。

2次元の主成分分析では、データの平方和や相関係数行列が

単位行列になることはないので、2次元データからは重解が出ることはありません。＜

では、3次元以上の場合も、主成分分析データから重解はないのかどうか、データを作ってみたら、3次元では重解になるデータがありました！というか見つけました！結構時間かかったけど！

では、3次元で重解となる主成分分析を解説します。

【2】例題

(1) 例題

3次元の主成分分析の例題は以下です。3次元の行列式を解く場面もあるので、行列式も練習しましょう。

【問題】

以下のデータにおいて、主成分分析せよ。

(1) 相関係数行列 R

(2) 固有値 λ_i

(3) 固有ベクトル v_i

(4) 寄与率と累積寄与率

(5) 因子負荷量（標準化）

(6) 主成分得点（標準化）

No	x1	x2	x3
1	1	1	3
2	2	4	2
3	3	2	1
4	4	5	5
5	5	3	4
和	15	15	15
平均	3	3	3

重解をもつデータとなるように、いろいろデータ値をいれて試した結果、運良く重解をもつデータが見つかりました！

3次元の主成分分析ですが、手計算で解いて主成分分析の理解を深めましょう。

(2) データの標準化

因子負荷量や主成分得点などをツールから計算するために、データを標準化しておきます。

データの標準化は $z_i = \frac{x_i - \bar{x}}{s}$ と変換して、平均 0、標準偏差 $s = 1$ の変数 z_i に変換することです。

結果は下表のとおりです。

No	x1	x2	x3
標準偏差 s	1.41	1.41	1.41

No	x1	x2	x3
1	-1.414	-1.414	0.000
2	-0.707	0.707	-0.707
3	0	-0.707	-1.414
4	0.707	1.414	1.414
5	1.414	0.000	0.707
和	0	0	0
平均	0	0	0
標準偏差	1	1	1

確かに、平均 0、標準偏差 1 に変換できていますね。

【3】 相関係数行列の計算

(1) 各平方和を先に計算

科目ごとの平方和を先に計算します。

$$S_{ij} = \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

として、x1とx2の平方和は

$$S_{12} = \sum_{k=1}^{n=5} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

と計算します。

全部のパターンを計算した結果を下表にまとめます。

平方和S	x1	x2	x3
x1	$S_{11} = 10$	$S_{12} = 5$	$S_{13} = 5$
x2	$S_{21} = 5$	$S_{22} = 10$	$S_{23} = 5$
x3	$S_{31} = 5$	$S_{32} = 5$	$S_{33} = 10$

(2) 相関係数行列の計算

相関係数 r_{ij} は

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}$$

相関係数 r_{ij}	x1	x2	x3
x1	$r_{11} = 1$	$r_{12} = 0.5$	$r_{13} = 0.5$
x2	$r_{21} = 0.5$	$r_{22} = 1$	$r_{23} = 0.5$
x3	$r_{31} = 0.5$	$r_{32} = 0.5$	$r_{33} = 1$

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$$

固有方程式が重解をもつための下準備が終わりました。
本題はここからです。

【4】固有値、固有ベクトルの計算

(1) 固有値、固有ベクトルの解法

ツールを使った解法を解説する前に、主成分分析の本質を再確認しましょう。

本記事のテーマではありませんが、主成分分析は必ず固有方程式を解きます。

平方和でも相関係数でもどちらでも固有方程式は解けますが、今回は相関係数を使って解きます。

固有方程式 $Rv = \lambda v$ より

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

を満たす、固有値 λ と固有ベクトル v を解きます。

3次元になりますが、行列式を解いて手計算で行きましょう。その後で、解析ツールに頼りましょう。
手計算で何を解いているのかの理解を深めましょう。

(2) 【重要】固有値の計算

3×3行列における、固有方程式を書くとき $Rv = \lambda v$ より、行列式 $|R - \lambda E| = 0$ を満たす λ を計算します。

$$|R - \lambda E| = \begin{vmatrix} r_{11} - \lambda & r_{12} & r_{13} \\ r_{21} & r_{22} - \lambda & r_{23} \\ r_{31} & r_{32} & r_{33} - \lambda \end{vmatrix} = \begin{vmatrix} 1 - \lambda & 0.5 & 0.5 \\ 0.5 & 1 - \lambda & 0.5 \\ 0.5 & 0.5 & 1 - \lambda \end{vmatrix}$$

行列式の公式を紹介すると

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} - \lambda \end{vmatrix}$$

となります。線形代数の教科書に書いていますので参考ください。

行列式を解くと

$$\begin{vmatrix} 1 - \lambda & 0.5 & 0.5 \\ 0.5 & 1 - \lambda & 0.5 \\ 0.5 & 0.5 & 1 - \lambda \end{vmatrix} = 0$$

$$(1 - \lambda)^3 + \frac{1}{8} + \frac{1}{8} - \frac{1}{4}(1 - \lambda) - \frac{1}{4}(1 - \lambda) - \frac{1}{4}(1 - \lambda) = 0$$

$$4\lambda^3 - 12\lambda^2 + 9\lambda - 2 = 0$$

となり、因数分解すると

$$(2\lambda - 1)^2(\lambda - 2) = 0$$

と因数分解でき、固有値は $\lambda = 0.5$ (重解), 2 が得られます。

よって、固有値 λ は

● $\lambda_1 = 2$

● $\lambda_2, \lambda_3 = 0.5$

となります。

(3) 【重要】固有ベクトルの計算

主成分分析で重解が出たら固有ベクトルはどう求めるの？

重解でない場合と同じですが、グラムシュミットの直交化法という処理が1つ必要です。

① $\lambda_1=2$ の場合

$$\begin{pmatrix} -1 & 0.5 & 0.5 \\ 0.5 & -1 & 0.5 \\ 0.5 & 0.5 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

から、関係式を作ると

$$\bullet -2a + b + c = 0$$

$$\bullet a - 2b + c = 0$$

$$\bullet a + b - 2c = 0$$

より、 $a = b = c$ の関係式ができるので、固有ベクトルは単位ベクトルに注意して

$$v_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

となります。

② $\lambda_2=0.5$ (重解)の場合

同様に解くと、

$$\begin{pmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

から、関係式を作ると

$$\bullet a + b + c = 0$$

の1つしか関係式ができません。これは重解だからですね。でも、焦る必要はなく、 $c = -(a + b)$ と変形して、ベクトル表記します。

$$v_2 = \begin{pmatrix} a \\ b \\ -(a+b) \end{pmatrix} = a \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + b \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

となり、2つのベクトル（単位ベクトル処理はあとにしますが）

$$v_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

が得られます。

ただし、注意点があり、内積 $v_2 \cdot v_3 = 1 \neq 0$ で直交しません。

なので、グラムシュミットの直交化法を使って処理が1つ増えます。

グラムシュミットの直交化法は他のサイトでも解説していますので、それに任せるとすると

$$u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

u_2 は単純に単位ベクトル化

$$a_3 = v_3 - (v_3 \cdot u_2)u_2 = \frac{1}{2} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

単位ベクトル化すると

$$u_3 = \frac{1}{2\sqrt{6}} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

となります。

以上、固有値と固有ベクトルを下表にまとめます。

固有ベクトル	主成分1	主成分2	主成分3
x1	0.5773	0.707	-0.204
x2	0.5773	0	0.408
x3	0.5773	-0.707	-0.204
固有値	2	0.5	0.5

【4】主成分の寄与率、累積寄与率

(1) 主成分の寄与率、累積寄与率

●主成分の寄与率は、個々の固有値を自由度で割った値ですね。

●累積寄与率は、寄与率の累積値です。

結果は下表のとおりです。

主成分寄与率	主成分1	主成分2	主成分3
固有値	2	0.5	0.5
寄与率	2/3	1/6	1/6
累積寄与率	2/3	5/6	1

(2) 寄与率を平方和でなく固有値で考える理由

各主成分の平方和は固有値になります。これは関連記事で解説していますが、一番簡単にわかる証明方法があります。

固有方程式 $Sv = \lambda v$ から全ての変数をスカラー（数値）と考えると、両辺を v で割れば、
 (平方和) $S = \lambda$ (固有値)

確かに、(平方和) $S = \lambda$ (固有値) となりますよね。これを行列、ベクトル表記にして n 次元化していますが、考え方は同じなので、主成分平方和は固有値として考えてよいということです。平方和と固有値は別物と思いがちですが、一致しています。意外ですよね。

以上、「主成分分析ができる(3次元で重解がある場合)」を解説しました。