2010 年度前期　応用統計学　第６回

主成分分析 (2) ー 主成分の導出と意味

今回は，２変量の場合の主成分分析について，前回は省略した「なぜ固有値問題になるのか」の説明，および，主成分分析における「主成分による分散の表現」の考え方と回帰分析の考え方との違い，変量の標準化と相関行列について説明します．

なぜ固有値問題になるのか

前回は，「主成分とは，元の変量を組み合わせて新しい変量を作ったとき，その分散が最大になるもの」とし，それが分散共分散行列を対角化する固有値問題となると述べましたが，なぜそうなるのかは説明しませんでした．ここでは，なぜ固有値問題になるのかを説明します．

元の変量 x_1, x_2 に対して，第 1 主成分 $z_{(1)}$ が

$$z_{(1)} = a_1 x_1 + a_2 x_2 \tag{1}$$

で表されるものとします．i 番目のデータの変量 x_1, x_2 の値を x_{1i}, x_{2i} とすると，変量 x_1, x_2 の平均 $\bar{x_1}, \bar{x_2}$，分散 s_{11}, s_{22}，共分散 $s_{12} = s_{21}$ は，データ数を n として

$$\begin{aligned} \bar{x_1} &= \frac{1}{n}\sum_{i=1}^{n} x_{1i}, \quad \bar{x_2} = \frac{1}{n}\sum_{i=1}^{n} x_{2i} \\ s_{11} &= \frac{1}{n}\sum_{i=1}^{n} (x_{1i} - \bar{x_1})^2, \quad s_{22} = \frac{1}{n}\sum_{i=1}^{n} (x_{2i} - \bar{x_2})^2 \\ s_{12} &= s_{21} = \frac{1}{n}\sum_{i=1}^{n} (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2}) \end{aligned} \tag{2}$$

で表されます．

このとき，主成分の分散 $V(z_{(1)})$ は，(1)(2) 式から次のように表されます．

$$\begin{aligned} V(z_{(1)}) &= \frac{1}{n}\sum_{i=1}^{n} (z_{1i} - \bar{z_1})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \{(a_1 x_{1i} + a_2 x_{2i}) - (a_1 \bar{x_1} + a_2 \bar{x_2})\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \{(a_1 (x_{1i} - \bar{x_1}) + a_2 (x_{2i} - \bar{x_2})\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{(a_1^2 (x_{1i} - \bar{x_1})^2 + 2a_1 a_2 (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2}) + a_2^2 (x_{2i} - \bar{x_2})^2\right\} \\ &= a_1^2 s_{11} + 2a_1 a_2 s_{12} + a_2^2 s_{22}. \end{aligned} \tag{3}$$

したがって，この $V(z_{(1)})$ を最大にする a_1, a_2 を求めればよいわけです．ここで，

$$a_1 = \cos\theta_1, \quad a_2 = \cos\theta_2 \tag{4}$$

とおくと，θ_1, θ_2 は z_1 軸が x_1, x_2 軸となす角となり，(a_1, a_2) は新しい座標軸 $z_{(1)}$ の方向余弦ということになります．このとき，a_1, a_2 は

$$a_1^2 + a_2^2 = 1 \tag{5}$$

を満たします．したがって，問題は，(5) 式の条件のもとでの，(3) 式の $V(z_{(1)})$ の最大化ということになります．

このような，制約条件付き最大化問題は，Lagrange の未定乗数法によって解くことができます．これによれば，この問題は未定乗数を λ とおいて

$$F(a_1, a_2, \lambda) = a_1^2 s_{11} + 2a_1 a_2 s_{12} + a_2^2 s_{22} - \lambda(a_1^2 + a_2^2 - 1) \tag{6}$$

を最大化する制約条件なしの最大化問題に帰着されます．これを解くため，F を a_1, a_2, λ でそれぞれ偏微分して，それらが 0 に等しいとおくと

$$\begin{aligned} \frac{\partial F}{\partial a_1} &= 2a_1 s_{11} + 2a_2 s_{12} - 2a_1\lambda = 0 \\ \frac{\partial F}{\partial a_2} &= 2a_2 s_{22} + 2a_1 s_{12} - 2a_2\lambda = 0 \\ \frac{\partial F}{\partial \lambda} &= -\lambda(a_1^2 + a_2^2 - 1) = 0 \end{aligned} \tag{7}$$

となります．ここで，第 3 式は (5) 式と同じですから，すでに満たされています．残りの式からは，

$$\begin{aligned} a_1 s_{11} + a_2 s_{12} &= a_1\lambda \\ a_2 s_{22} + a_1 s_{12} &= a_2\lambda \end{aligned} \tag{8}$$

という関係が得られます．これを行列を使って書くと

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{9}$$

となり，前回の講義で示した固有値問題が導かれます．

第 1 主成分と回帰直線

主成分を表す (1) 式を，第 1 主成分 $z(1)$，第 2 主成分 $z(2)$ の両方について組み合わせると，主成分とは，直交座標 x_1，x_2 による散布図上の 2 変量データを

$$\begin{pmatrix} z_{(1)} \\ z_{(2)} \end{pmatrix} \begin{pmatrix} a_{1(1)} & a_{1(1)} \\ a_{1(2)} & a_{2(2)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{10}$$

という式で新たな直交座標 $z_{(1)}$，$z_{(2)}$ に変換し，$z_{(1)}$ 上での分散が最大になるようにしたもの，ということができます（図 1）．各々の a は (4) 式を満たすので，実は，(10) 式の変換は，座標系の角度 θ_1 の回転となります．ここで，新しい座標軸 z_1，z_2 を平行移動しても，それぞれの軸上での分散は変わりませんから，今回は，２つの座標軸を，直交座標 x_1，x_2 での平均 $\bar{x_1}, \bar{x_2}$ を通るように移動して考えます（図 2）．

さて，第 1 主成分を表す軸 z_1 を引きます．データの分散とは，おのおののデータと平均とのへだたりの２乗の合計（をデータ数で割ったもの）です．ここで，図 3 のように，あるデータ $(x_{1i},\ x_{2i})$ について，

「平均とのへだたりの２乗」を考えてみましょう．あるデータが平均からへだたっていればいるほど，そのデータは「目立っている」，あるいは「価値・特色のあるデータ」であると考えることができます．これは，図４のように平均とのへだたりがまったくないデータ群は，たくさんデータを用意しなくても平均をあらわす１つで十分で，「全く価値がない」ことからもわかります．

$(x_{1i},\ x_{2i})$ とそれを第１主成分に投影した z_{1i} とのへだたりは，元の $(x_{1i},\ x_{2i})$ と平均 $\bar{x_1}, \bar{x_2}$ とのへだたりに比べて，小さくなっています．このことは，第１主成分上では，２次元のデータを１次元で表したために，元のデータに比べて，そのデータの価値が失われていることを示しています．その損失の量は，図３からピタゴラスの定理によって求められます．

さて，第１主成分は「軸上でのデータの分散が最大」になる軸，すなわち「各データを軸上に投影したときの，データの平均からのへだたりの２乗が最大」になる軸ですから，言い換えれば「軸上での価値の合計が最大」になる軸と言えます．一方，もとの各データの平均からのへだたりはどの座標軸でも同じ，すなわち各データの価値の合計はどの座標においても同じです．したがって第１主成分は，「各データの軸上に投影したときの価値の損失の合計が最小」である軸ということになります．

前回の講義で「分散が最大の軸」という表現で第１主成分を表現しましたが，もっと明確にいえば「多次元データを，それらがもつ価値をなるべく損なわないように，１次元の値（主成分得点）に縮約したもの」ということができます．また，図３から明らかなように，第１主成分に投影することによって生じた価値の損失は，第２主成分によって表現されていることがわかります．

ところで，この「各データと，その軸上への投影とのへだたりの２乗の合計が最小」という考え方は，回帰直線による直線のあてはめとよく似ています．回帰分析では，被説明変数の分散を，回帰直線からみた分散で何パーセント表現できるかを，回帰直線の当てはまり具合と考え，決定係数で表しました．

しかし，両者の間には大きく違うところがあります（図5）．図３からわかるように，主成分分析では，各データから第１主成分軸に垂線を下ろしたときの，各垂線の長さの２乗を最小にしています．

これに対して，回帰分析では，各データから回帰直線に対して x_1 軸に垂直に線分をひき，その長さの２乗を最小にしています．回帰分析では「x_1 で x_2 を説明する」あるいは「x_1 から x_2 を予想する」という考えにもとづいており，説明変数と被説明変数の区別があるので，x_2 軸に沿ったへだたりを計算しています．

主成分分析ではこのような区別はなく，どちらでどちらを説明しているわけでもないので，軸と各データとの「距離」を最小化の対象としています．

変量の標準化と相関行列

主成分は，データの分散・共分散からなる分散共分散行列の固有値・固有ベクトルとして求められます．ここで，２つの変量のうち一方の単位だけが，例えば cm → mm のように，違う倍率の単位に変わったとしましょう．この場合，図６のように，散布図が一方向だけ（この場合は x_1 軸方向だけ）に引き伸ばされることになります[1]．このとき，x_2 方向の分散 s_{22} は変わらず，他の分散 s_{11} は 100 倍，共分散 s_{12} は 10 倍というように，偏って変化しますから，得られる固有値・固有ベクトルも違ってきます．

しかし，これは困ったことです．図６のように，どちらも長さという同じ次元の単位であれば，単位

[1]本当は 10 倍に引き伸ばされますが，紙に入らないので 2 倍に引き伸ばして描いてあります．

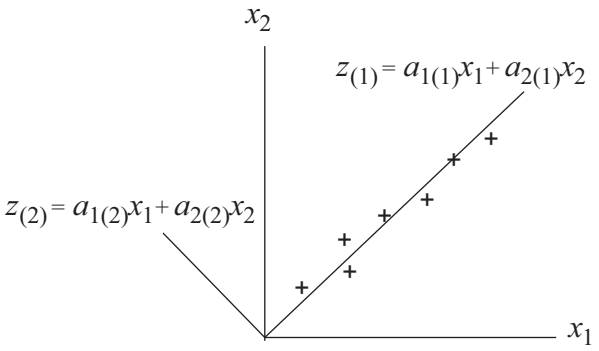

図 1: ２変量の主成分分析

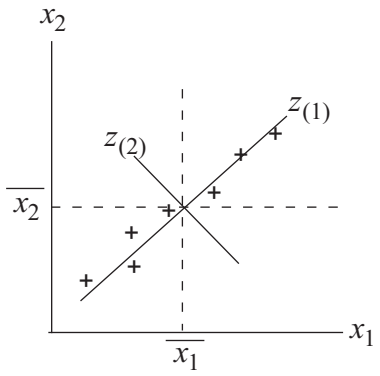

図 2: 主成分軸の移動

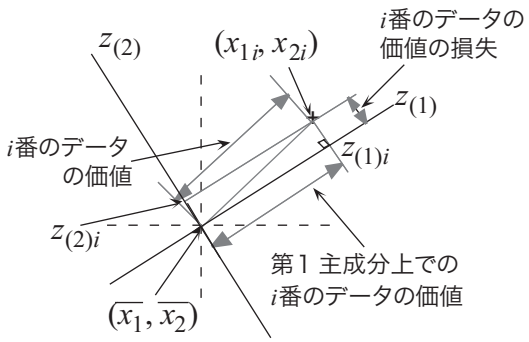

図 3: あるデータの価値の損失

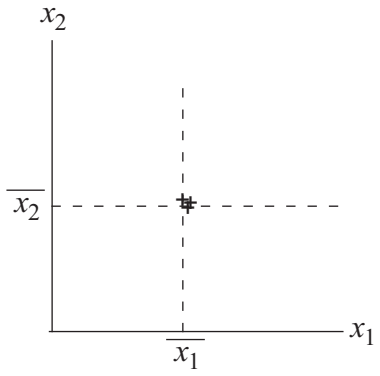

図 4: 価値のほとんどないデータ群

をそろえることができますが，片方が重さで片方が長さというように違っている場合は，g と m がよいのか，kg と cm がよいのか判断できません．

そこで，各変量を平均 0，分散 1 のいわゆる標準得点に変換し，無名数にしてしまうという方法があります．こうすればもとのデータがどんな単位で測定されていても，常に同じように主成分分析をすることができます．

もとのデータを x_{1i}, x_{2i} とし，標準得点を X_{1i}, X_{2i} とすると

$$X_{1i} = \frac{x_{1i} - \bar{x_1}}{\sqrt{s_{11}}},\ X_{2i} = \frac{x_{2i} - \bar{x_2}}{\sqrt{s_{22}}} \tag{11}$$

の関係があります（s_{11}, s_{22} は，それぞれ x_1, x_2 の分散）．このとき，変換後の両変量の分散は 1 で，ま

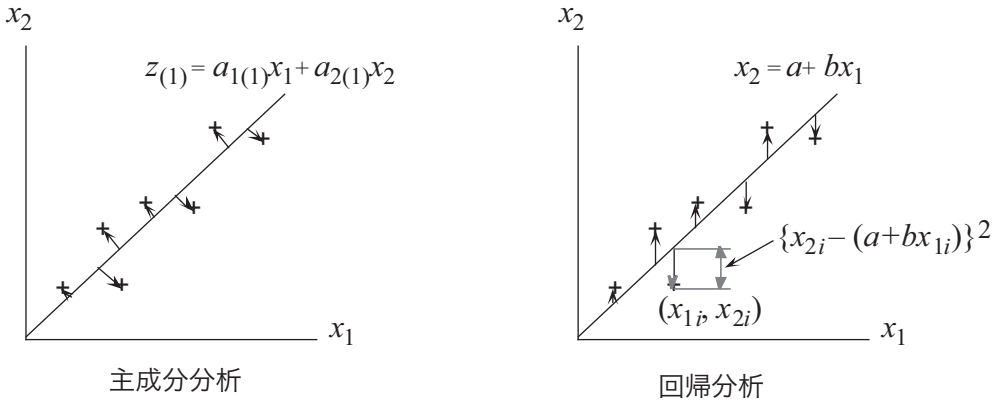

図 5: 主成分分析と回帰分析の違い

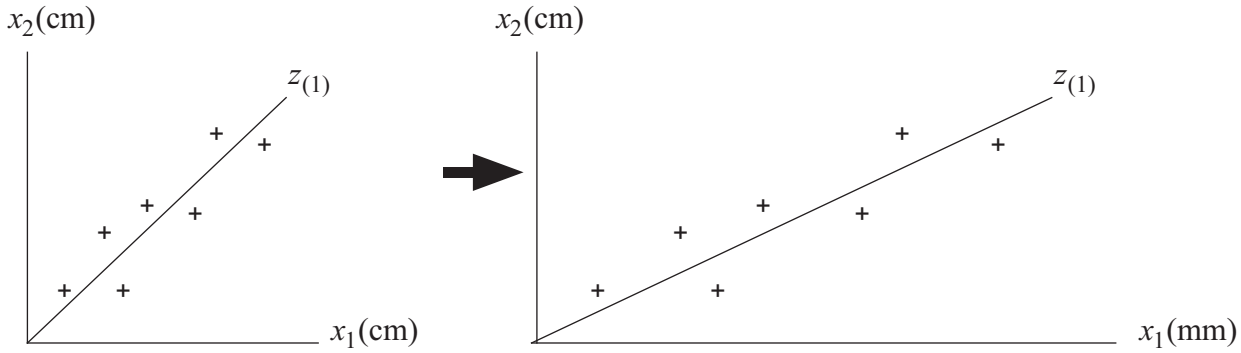

図 6: 単位が違うとどうなるか

た変換後の共分散を S_{12} とすると

$$
\begin{aligned}
S_{12} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_{1i} - \bar{X_1}\right)\left(X_{2i} - \bar{X_2}\right) \\
\bar{X_1} = 0, \bar{X_2} = 0 \text{ だから} \quad &= \frac{1}{n}\sum_{i=1}^{n} X_{1i}X_{2i} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{1i} - \bar{x_1}}{\sqrt{s_{11}}}\right)\left(\frac{x_{2i} - \bar{x_2}}{\sqrt{s_{22}}}\right) \\
&= \frac{1}{\sqrt{s_{11}}\sqrt{s_{22}}}\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{n} \\
&= \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} \qquad (12)
\end{aligned}
$$

となり，これはすなわち x_1 と x_2 との相関係数 r_{12} となります．したがって，標準化した変量による主成分分析とは，元のデータについて

$$
\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad (13)
$$

という固有値問題を解くことになります．(13) 式の左辺の行列を**相関行列**といい，この形式の主成分分析を「相関行列による主成分分析」といいます．これに対して，最初に説明した形式の主成分分析を「分散共分散行列による主成分分析」といいます．相関行列による主成分分析は，いわば「偏差値」による主成分分析ということになります．