2010 年度前期　応用統計学　第５回

主成分分析 (1) ー ２変量の主成分分析

主成分分析は，多変量データが与えられたとき，各変量を組み合わせて，各データ間の違いをより際立たせて表現できるような新しい変量を作る方法です．今回は，主成分分析の基本的考え方を説明するために，もっとも基本的な２変量の場合の主成分分析を説明します．

総合得点をいかに評価するか？

次のような状況を考えてみましょう．

> クラスで数学と国語（各々 100 点満点）の試験を行いました．数学を x_1 軸，国語を x_2 軸として，各人の成績を散布図に描くと図 1 のようになりました．どのように総合得点を求めて各学生を評価すればよいでしょうか？

２科目の試験を行なった場合，ひとりの成績は２つの数の組で表されていますから，受験者どうしを比較することができません．そこで，ひとりの成績が 1 つの数で表されるように変換し，受験者間の比較ができるようにしたものが「総合得点」です．

散布図のうえでは，総合得点を求めることは，散布図上に「ものさし」を置き，散布図上の各点からものさしに垂線をおろして，ものさし上のその場所の目盛りを総合得点とすることに相当します（図 2）．

一番簡単な２科目の総合得点とは，２科目の得点の合計で，上の例題でいえば $z = x_1 + x_2$ を求めることです．この式を変形すると $x_2 = -x_1 + z$ で，この式は「散布図上で，総合得点 z が同じである受験者を表す点は，直線 $x_2 = -x_1 + z$ の上にある」ことを示しています．このことを表したのが図 3 です．さまざまな z に対応する直線 $x_2 = -x_1 + z$ が，点線で表されています．ひとつの点線上のある受験者は，総合得点が皆同じ z ですから，これが上の説明でいう「ものさしへの垂線」にあたります．したがって，この場合の「ものさし」は，原点を通り右上 45 度方向に直線ということになります．

ここで，「ものさしに垂線をおろして，目盛りを読む」というのは，z を新しい座標軸と考えれば，単に「座標軸上の位置を読む」ということです．したがって，$z = x_1 + x_2$ という計算は，x_1 と x_2 という２つの座標軸から z という新しい座標軸を作ったことに相当します．

はたして，この総合得点は本当に「一番よい」のでしょうか．そこで，「よい総合得点」とは何かを考えてみましょう．総合得点は，２つの数の組で表された各人の成績を，１つの数で表して評価するのが目的です．ですから，各人の成績を総合得点に変換したとき，なるべく総合得点に差がつくほうが，各人を比較して評価しやすくなります．この例では，数学の成績は各人の差が大きく，国語の成績の差は小さくなっています．ですから，数学の成績を重視して総合成績を求めた方が，各人の差がつきやすくなります．これに対して，次ページの図 4 のように数学の得点を無視して，国語の得点，すなわち $x2$ 軸上の位置を総合得点とすると，この総合得点では各人の成績に差がつかず，総合得点の意味をなさなくなります．

つまり，「よい総合得点」とは，図 5 のように，その総合得点に変換すると各人の成績にもっとも大きく差がつくような総合得点であるといえます．言い換えれば，軸を移動したとき，軸上での点数の分散が最大になるような軸を求めればよいことになります．軸上に置き換えた点数を第 1 主成分といいます．

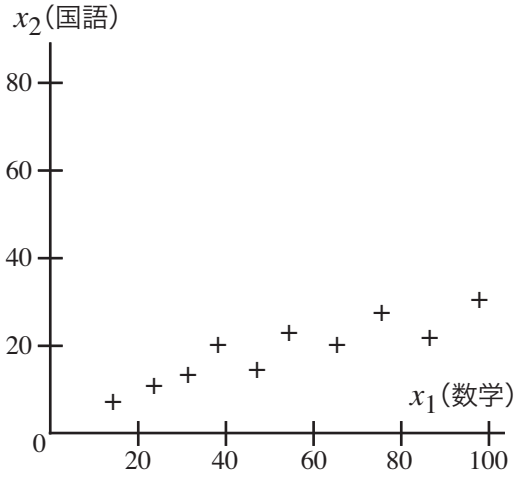


図 1: 数学と国語の散布図

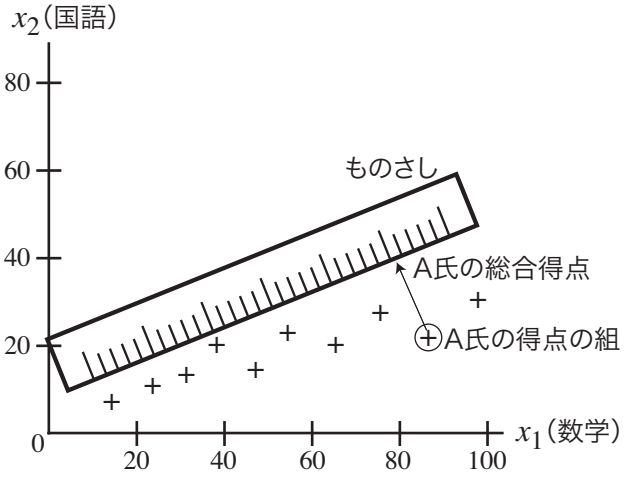


図 2: 総合得点

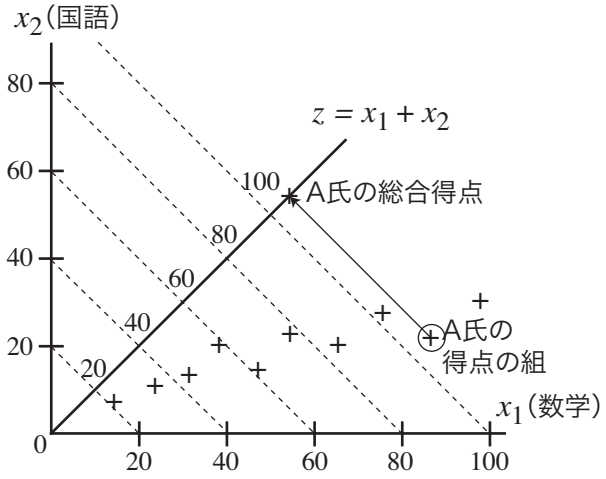


図 3: 総合得点（数学と国語の合計）

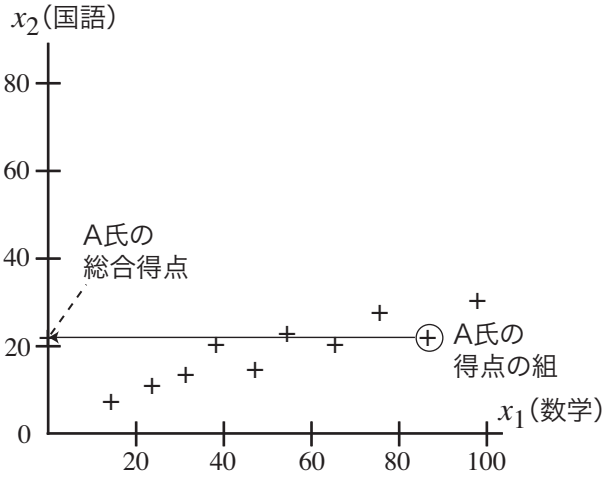


図 4: 総合得点（国語のみ）

主成分を求める

では，このような軸を求める方法を考えてみましょう．ここで，x_1, x_2 軸を回転させて $z_{(1)}, z_{(2)}$ 軸[1]とし，$z_{(1)}$ 軸上での分散が最大になるようにします．この状態が図 6 です．散布図をこの $z_{(1)}, z_{(2)}$ 軸から見てみると，図から直観的に読み取れるとおり，$z_{(1)}$ と $z_{(2)}$ の相関が 0 になっています．

相関が 0 になるような軸を求めるため，次の**分散共分散行列**を考えます．元の変量 x_1, x_2 軸でみたとき，x_1 の分散を s_{11}，x_2 の分散を s_{22}，x_1 と x_2 の共分散を s_{12} とするとき，分散共分散行列は次のように定義されます．

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \tag{1}$$

さて，散布図上の点を $z_{(1)}, z_{(2)}$ 軸から見たときに相関が 0 ということは，共分散が 0 ということですか

[1]以下，添え字の「1, 2」は元の座標軸 (x) の番号，「(1), (2)」は主成分 (z) の番号を表します．

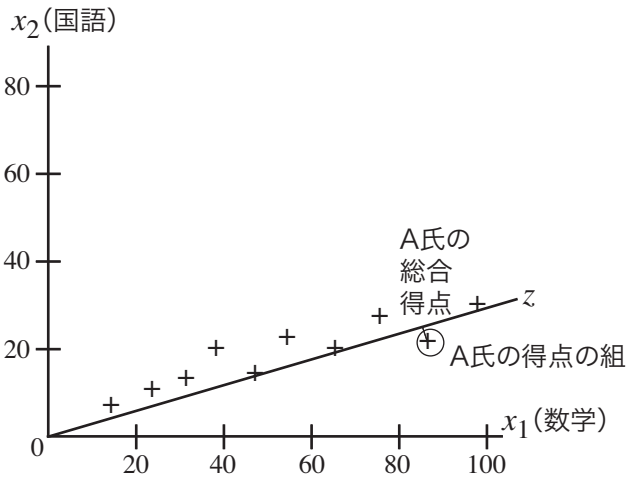


図 5: もっともよい総合得点

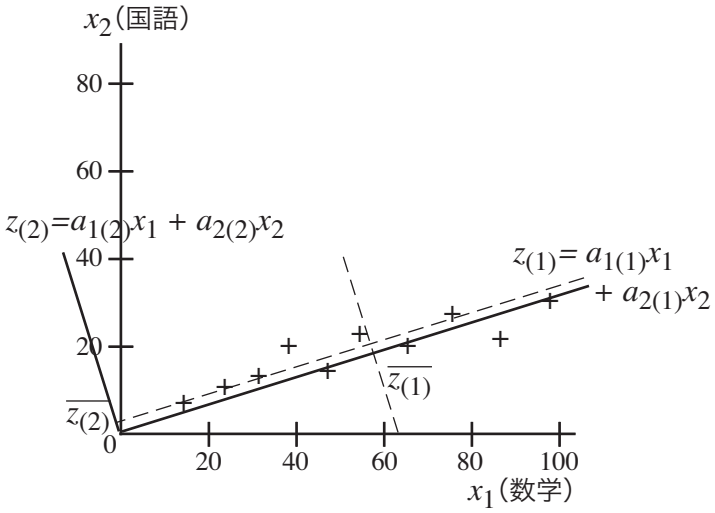


図 6: 軸の回転

ら，$z_{(1)}, z_{(2)}$ 軸から見たときの分散共分散行列は，$z_{(1)}, z_{(2)}$ 軸の分散をそれぞれ $\lambda_{(1)}, \lambda_{(2)}$ とすると

$$\begin{pmatrix} \lambda_{(1)} & 0 \\ 0 & \lambda_{(2)} \end{pmatrix} \tag{2}$$

と表すことができます．このような $z_{(1)}, z_{(2)}$ 軸を求める操作は，行列の**対角化**と呼ばれています．

対角化を行うため，変量 x_1, x_2 を $z_{(1)}, z_{(2)}$ に変換する式を

$$z_{(1)} = a_{1(1)}x_1 + a_{2(1)}x_2, \quad z_{(2)} = a_{1(2)}x_1 + a_{2(2)}x_2 \tag{3}$$

とします．このとき，a_1, a_2 と λ（いずれも，$_{(1)}$ と $_{(2)}$ の２つずつある）は

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{4}$$

の２つの解となることが知られています．これは**固有値問題**とよばれ，これを満たす λ は**固有値**，$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ は**固有ベクトル**と呼ばれています．解は２組得られるので，２つの λ のうち大きいほうを $\lambda_{(1)}$ とし，そのときの固有ベクトルを $\begin{pmatrix} a_{1(1)} \\ a_{2(1)} \end{pmatrix}$，もう一方を $\lambda_{(2)}$ と $\begin{pmatrix} a_{1(2)} \\ a_{2(2)} \end{pmatrix}$ とします．

ここで，$\lambda_{(1)}, \lambda_{(2)}$ は新しい軸 $z_{(1)}, z_{(2)}$ での分散を表していますから，その大きいほうである $\lambda_{(1)}$ が第 1 主成分の分散になります．そして，それに対応する固有ベクトル $\begin{pmatrix} a_{1(1)} \\ a_{2(1)} \end{pmatrix}$ から (3) 式で得られる $z_{(1)}$ が第 1 主成分となります．

固有値問題を解く

(4) 式は

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$\text{すなわち} \begin{pmatrix} s_{11}-\lambda & s_{12} \\ s_{12} & s_{22}-\lambda \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5}$$

であり，これが $a_1 = a_2 = 0$ 以外の解を持つためには，左辺の行列の行列式が 0，すなわち

$$\begin{vmatrix} s_{11}-\lambda & s_{12} \\ s_{12} & s_{22}-\lambda \end{vmatrix} = 0 \tag{6}$$

でなければなりません（これを特性方程式といいます）．この方程式は λ の 2 次方程式で，その解は

$$\lambda = \frac{(s_{11}+s_{22}) \pm \sqrt{(s_{11}-s_{22})^2 + 4s_{12}^2}}{2} \tag{7}$$

となります．

次に固有値に対応する固有ベクトルを求めます．固有値 λ のうち大きいほうを $\lambda_{(1)}$ として，$s_{12} \neq 0$ のとき (4) 式から

$$a_{(1)2} = \frac{\lambda_{(1)} - s_{11}}{s_{12}} a_{(1)1} \tag{8}$$

です．ここで，主成分は「軸」ですから，これを表す固有ベクトルは方向さえ決まればよく大きさは関係ないので，

$$a_{(1)1}^2 + a_{(1)2}^2 = 1 \tag{9}$$

と決めておいて，これに (8) 式を代入すると

$$\begin{aligned} a_{(1)1} &= \frac{s_{12}}{\sqrt{s_{12}^2 + (\lambda_{(1)} - s_{11})^2}} \\ a_{(1)2} &= \frac{\lambda_{(1)} - s_{11}}{\sqrt{s_{12}^2 + (\lambda_{(1)} - s_{11})^2}} \end{aligned} \tag{10}$$

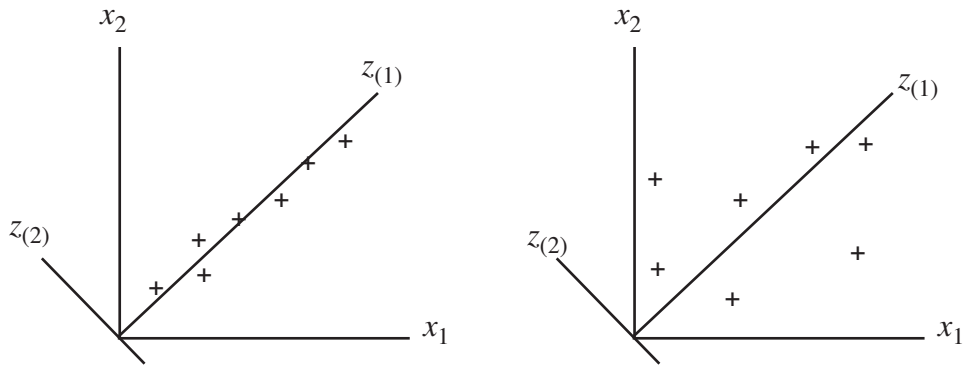
となって（複号同順），固有ベクトルが求まります．(10) 式の複号が + でも − でも，(3) 式で求められる $z_{(1)}$ 軸は向きが正反対なだけで同じものですから，以後は + のほうだけを固有ベクトルとします．

このようにして求められる $a_{(1)1}, a_{(1)2}$ で決まる新しい軸，すなわち変量 $z_{(1)}$ が第 1 主成分であり．また，各データ x_1, x_2 から得られる新しい変量の値 $z_{(1)} = a_{(1)1}x_1 + a_{(1)2}x_2$ を**主成分得点**といいます．

第２主成分と寄与率

さて，もう一方の固有値を $\lambda_{(2)}$ とすると，これに対応する固有ベクトルによってもう 1 つの軸が求められます．この軸はさきほどの軸に直交する軸となります．$\lambda_{(2)}$ から求められる $a_{(2)1}, a_{(2)2}$ で決まる変量 $z_{(2)}$ を第２主成分といいます．

(2) 式のように，$\lambda_{(1)}$ と $\lambda_{(2)}$ はそれぞれ $z_{(1)}, z_{(2)}$ 軸上での分散です．$\lambda_{(1)}$ や $\lambda_{(2)}$ の，分散の和に占める割合を，各主成分の寄与率といいます．第 1 主成分の寄与率が大きい時は，元のデータの各々の違いを第 1 主成分でほとんど表せていることになります．一方，寄与率が小さいときは元のデータの各々の違いは第 1 主成分だけでは十分に表すことができず，２つの軸で表現する必要があることを示しています（図 7）．



第１主成分の寄与率・大
（$z_{(2)}$方向の分散はほとんどない）

第１主成分の寄与率・小
（$z_{(2)}$方向の分散がある程度ある）

図 7: 寄与率