

第4回 数学教材研究会

多変量解析の線形代数

(平成 24 年 3 月 22 日)

佐藤 尊文 (秋田高専 自然科学系)

1 多変量解析とは

数学辞典 (第 3 版) によると, **多変量解析** (multivariate analysis) は

各個体についての観測値が数個の値からなるとき, すなわち, p 次元 Euclid 空間の点として表されるようなデータを **多変量データ** (multivariate data) と呼び, 多変量データの統計的解析法を **多変量解析** という。

具体的には, 次のような解析法があり, これらの総称が多変量解析。

分析目的	量的データの分析	質的データの分析
初めの言葉	主成分分析 因子分析 クラスター分析	コレスポンデンス分析 数量化理論Ⅲ類 MDS
終りの言葉	重回帰分析 判別分析 正準相関分析	数量化理論Ⅰ類 数量化理論Ⅱ類 コンジョイント分析

1.1 主成分分析

主成分分析 (PCA, Principal Component Analysis) とは

ある現象に影響する多数の要因を 1 つ 1 つ独立に扱うのではなく, 小数の総合評価に集約する

という考えに基づく分析法。実際の現象は複雑で沢山の要因を持つが, 分析の目的は, それを利用すること。よって, 要因全てを考慮しても, 扱い難い分析結果が得られては意味がない。思い切ってまとめた総合指標を作ろうというもの。

1.2 重回帰分析

重回帰分析 (Multiple regression analysis) とは

ある現象を予測する為, これに影響する多数の要因について, それぞれがどの程度の影響を与えているか

を分析するもの。その現象を複数の要因から説明するモデルの構築。

2 多変量解析の線形代数

2.1 基本用語

説明変数 (explanatory variable) は、分析しようとする現象の要因となるものを表す変数。**基準変数** (criterion variable) は、結果を表す変数。説明変数 x_1 についての観測データは、次のように、ベクトルで表される。

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} \quad (1)$$

あるデータの成分の和を個数で割った

$$\bar{x}_1 = \frac{1}{n}(x_{11} + \cdots + x_{n1})$$

をデータ $\mathbf{x}_1$ の **平均** という。平均 $\bar{x}_1$ を n 個並べて得られる、次のベクトルを $\bar{\mathbf{x}}_1$ で表す。

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_1 \end{pmatrix}$$

データから平均を引いた差

$$\mathbf{x}_1 - \bar{\mathbf{x}}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 \\ \vdots \\ x_{n1} - \bar{x}_1 \end{pmatrix}$$

を **平均偏差 (ベクトル)** といい、平均偏差をとることを **平均偏差化** という。データの値が平均からどれだけズレているかを表す量。

あるデータの平均偏差の内積を個数で割った

$$s_{11} = \frac{1}{n}(\mathbf{x}_1 - \bar{\mathbf{x}}_1) \cdot (\mathbf{x}_1 - \bar{\mathbf{x}}_1) = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2$$

を **分散** (variance) という。データが平均からどれだけズレているかを表す量。 $\mathbf{x}$ の成分中に、平均からのズレが小さいものが沢山あるよりも、平均から大きくズレたものが少しでもある方によく反応する。

分散の正の平方根

$$\sigma_{11} = \sqrt{s_{11}}$$

を **標準偏差** (standard deviation) という。分散が2乗したままであるのに対し、標準偏差は平方根をとってるので、元のデータと単位が揃っているともいえる。また、データの標準化の際に必要な点でも重要。

平均偏差を標準偏差で割った

$$\hat{\mathbf{x}}_1 = \frac{1}{\sigma_{11}}(\mathbf{x}_1 - \bar{\mathbf{x}}_1)$$

をデータの **標準化** (normalization) という。標準化後のデータは、平均が0で、分散が1 (標準偏差1といってもよい) である。文字通り、データを標準化し、比較し易くしたもの。

2つのデータ $\mathbf{x}_1, \mathbf{x}_2$ をそれぞれ平均偏差化し、内積をとって個数で割った

$$s_{12} = \frac{1}{n}(\mathbf{x}_1 - \bar{\mathbf{x}}_1) \cdot (\mathbf{x}_2 - \bar{\mathbf{x}}_2)$$

を **共分散** (covariance) という。2つのデータの相関関係を表す。

2つのデータ $\mathbf{x}_1, \mathbf{x}_2$ をそれぞれ標準化し、内積をとって個数で割った

$$r_{12} = \frac{1}{n} \hat{\mathbf{x}}_1 \cdot \hat{\mathbf{x}}_2$$

を **相関係数** (correlation coefficient) という。2つのデータの相関関係、特に、直線的な関係の有無を表す。定義から、共分散と相関係数は

$$r_{12} = \frac{1}{\sigma_{11}\sigma_{22}} s_{12} \quad (2)$$

という関係にある。

対角成分に分散を、非対角成分に共分散を配置して得られる行列

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{nn} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

を **分散共分散行列** (variance-covariance matrix) という。分散共分散行列は n 次対称行列である。相関係数を成分として持つ行列

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{nn} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

を **相関行列** (correlation matrix) という。式 (2) より、相関行列も n 次対称行列である。

2.2 多変量解析の線形代数（あらすじ）

n 個の m 次元データ

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{m1} \end{pmatrix}, \cdots, \mathbf{x}_n = \begin{pmatrix} x_{1n} \\ \vdots \\ x_{mn} \end{pmatrix}$$

が与えられたとする。このとき

$$\mathbf{y} = \sum_{i=1}^n a_i \mathbf{x}_i$$

でベクトル $\mathbf{y}$ を定義する。 $\mathbf{y}$ の分散 $V(a_1, \cdots, a_n)$ を求める。

$$\begin{aligned} V(a_1, \cdots, a_n) &= \frac{1}{n} (\mathbf{y} - \bar{\mathbf{y}}) \cdot (\mathbf{y} - \bar{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^n a_i (\mathbf{x}_i - \bar{\mathbf{x}}_i) \cdot \sum_{j=1}^n a_j (\mathbf{x}_j - \bar{\mathbf{x}}_j) \\ &= \frac{1}{n} \sum_{i,j=1}^n a_i a_j (\mathbf{x}_i - \bar{\mathbf{x}}_i) \cdot (\mathbf{x}_j - \bar{\mathbf{x}}_j) \end{aligned} \quad (3)$$

そして、条件 $\sum a_i^2 = 1$ の下で、 $V(a_1, \dots, a_n)$ を最大とするような $(a_1, \dots, a_n)$ を求め、これを第1主成分の係数とする。

ここでベクトル $\mathbf{a}$ を

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

で定義し、データ $\mathbf{x}_1, \dots, \mathbf{x}_n$ の分散共分散行列を S とすると、式 (3) は

$$V(\mathbf{a}) = {}^t\mathbf{a}S\mathbf{a} \quad (4)$$

と表せる、すなわち、 $\mathbb{R}^n$ 内の2次形式を得る。

定理 1

対称行列 S について、 $\{{}^t\mathbf{a}S\mathbf{a} : |\mathbf{a}| = 1\}$ は閉区間であり、その両端は S の最小および最大の固有値に等しい。また、その両端の値をとるような $\mathbf{a}$ は、それぞれの固有値に対応する S の固有ベクトルである。

したがって、 S の固有値問題を解いて、最大固有値の固有ベクトルを $|\mathbf{a}| = 1$ という条件の下で求めればよい。こうして得られた $\mathbf{a}$ の成分を用いて、第1主成分 y を定義する。

さて、 n 次正方行列 S は、重複を許して n 個の固有値を持つ。これらを大きい順に並べて

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

とする。このとき、分散共分散行列 S の i 番目の固有値 λ_i に対応する（長さ1の）固有ベクトル

$$\mathbf{a}_i = \begin{pmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{pmatrix}$$

を用いて、第 i 主成分 y_i を定義する。