多変量解析を理解するために(4)

主成分分析の基礎

高木 廣文 | Takagi Hirofumi

東邦大学医学部看護学科国際保健看護学研究室教授

■1950年生まれ。74年東京大学医学部保健学科卒。79年同大大学院博士課程修了。同年米国国立衛生院奨励研究員として米国国立環境保健学研究所勤務。81年聖路加看護大学講師、82年同大助教授。89年文部省統計数理研究所助教授。99年新潟大学医療技術短期大学部看護学科教授、同年同大医学部保健学科教授。2006年4月より現職。著書に『ナースのための統計学 — データのとり方生かし方』(医学書院)、『多変量解析ハンドブック』(現代数学社)、『健康科学とコンピュータ』(共立出版)など。

1. はじめに

大学などの入学試験では、英語、数学、国語の3科目の得点を合計し、合計得点により受験生の順位を付けることがよく行われる。この場合、合計得点は3つのデータが持つ情報を統合して、1つの情報にまとめたものである。このように多数のデータの情報を、より少数の合成得点に集約させるような操作は「情報の圧縮」と呼ばれている。

情報を圧縮する場合、どのような操作を行えば、オリジナルのデータの持つ情報を代表させることができるかという問題がある。例えば、入学試験の場合、単純にすべての科目の得点を合計するだけでよいのだろうか。理科系の入試ならば数学の得点を2倍にしたり、文化系ならば英語の得点を2倍にする必要はないのだろうか。すなわち、目的に応じて各科目の得点に重み付けをする必要はないのかという問題がある。

臨床検査のように多くの検査項目を調べた場合、検査項目間の関係を基に、肝機能、脂質代謝、糖代謝など、各検査項目が示す情報に関して合成得点を求め、臨床診断に応用したい場合もあるだろう。

分析する変数の中に基準変数がない場合、変数相互間の相関構造から新たな合成得点を求める方法として、「**主成分分析** principal component analysis」がある。

主成分分析の目的は、

①情報の圧縮

②合成変数の持つ意味から、変数間の関係を検討

の2つに大きく分けて考えることができる。今回は、主成分分析の理論と計算方法について、簡単に解説する。

2. 主成分分析の理論と用語

n人についてp個の多変量データ$X=(x_1, x_2, \cdots, x_p)$があるものとする。また、計算に便利なように、ここでは各データはその変数の平均値を引いてあるものとする。従って各変数の平均は0になるようにしてある。

合成変数ベクトルfは、各変数ベクトルx_iに重み

$w_i=(i=1,2,\cdots,p)$を掛けて合計する。すなわち、

$$f=w_1x_1+w_2x_2+\cdots+w_px_p$$

により求められる。重みをすべて1にすれば、fはすべての変数の単純な合計となる。

主成分分析では、合成得点によって全変数の変動をなるべく多く説明するように、重みベクトルを定めるようにする。計算上は、

$$C^2_{XX}w=\lambda C_{XX}w$$

を解けばよいことが導かれる。ここで、C_{XX}はp個の変数間の分散共分散行列であり、λは定数である。今、

$$a=C_{XX}w$$

と置くと、ベクトルaは、合成変数fと各変数の共分散を示し、「**構造ベクトル** structure vector」と呼ばれている。結局、上式より、

$$C_{XX}a=\lambda a$$

を求めればよいことが分かる。すなわち、分散共分散行列C_{XX}の固有値λとその固有値ベクトルaを求めればよいことが分かる。

上記の説明では、変数間の分散共分散行列を用いる場合について解説したが、通常は変数間の単位の違いによる影響を除くために、相関係数行列R_{XX}を用いて同様の計算を行う。すなわち、

$$R_{XX}a=\lambda a$$

を満たすような固有値と固有ベクトルを求めればよい。なお、構造ベクトルaは主成分と各変数の相関係数を示し、「**主成分負荷量** principal loading」と呼ばれている。

ところで、主成分は1つとは限らず、最大で変数の個数分存在する。p個の変数の主成分分析で、正の固有値λ_1, λ_2,..., $\lambda_m(m\leq p)$に対応して、m個の主成分があるものとする。

このとき、各主成分の間には、

①異なる主成分間の相関は0。

②第i主成分f_iの各変数の主成分負荷量の2乗和は、固有値λ_iに等しい。

③重みベクトルw_iの各要素の2乗和を1になるように設定すると、主成分f_iの分散はλ_iとなる。

という関係がある。

求めた主成分がどの程度、分析に用いた変数が持っている情報を説明しているのかということを表すための指標が必要である。全部でm個の主成分があり、全固有値の和をTとする。すなわち、

$$T=\lambda_1+\lambda_2+\cdots+\lambda_m$$

とする。第i主成分の固有値をTで除したもの、すなわち、

$$c_i=\lambda_i/T,(i=1,2,...,m)$$

は、第i主成分の「**寄与率** contribution rate」と呼ばれ、その主成分が説明できる割合を示す指標となる。また、第1主成分から第i主成分までのi個の主成分の寄与率の合計は「**累積寄与率** cumulative contribution rate」と呼ばれ、そこまでの主成分で説明できる割合を示す指標となる。これは、

$$cc_i=(\lambda_1+\lambda_2+\cdots+\lambda_i)/T$$

により求められる。

実際の分析では、なるべく少数の主成分で、全体の変数の持つ変動を説明できる方がよい。しかし、意味のある主成分の個数を定める決定的な方法はない。便宜的には、以下のような方

法を用いることが多い。

①累積寄与率がある程度大きくなること。通常は、大体70～80％が目安である。

②相関係数行列に基づく主成分分析の場合、固有値が１以上の主成分のみを採用する。

これらの方法と実質的な主成分の意味を考慮して、主成分の個数を恣意的に決めるのが普通である。

3. 主成分分析の計算例

変数として、X_1が身長、X_2が体重、X_3が最高血圧、そしてX_4が最低血圧である５人のデータ行列（素データ行列）X_Rを考えよう。そのようなデータ行列は、以下のようになる。

$$X_R = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 \end{matrix} \\ \left[\begin{matrix} 170 & 65 & 120 & 80 \\ 165 & 57 & 122 & 78 \\ 180 & 72 & 115 & 75 \\ 168 & 55 & 124 & 60 \\ 162 & 50 & 130 & 82 \end{matrix}\right] \end{array} \begin{matrix} \\ \text{1番目のケース} \\ \text{2番目のケース} \\ \text{3番目のケース} \\ \text{4番目のケース} \\ \text{5番目のケース} \end{matrix}$$

この素データ行列X_Rの各データから、各変数の平均値を引き算すると、偏差データ行列Xを求めることができる。各変数の平均値は、身長169、体重59.8、最高血圧122.2、最低血圧75、なので、これらの値を引くと、

$$X = \begin{bmatrix} 170-169 & 65-59.8 & 120-122.2 & 80-75 \\ 165-169 & 57-59.8 & 122-122.2 & 78-75 \\ 180-169 & 72-59.8 & 115-122.2 & 75-75 \\ 168-169 & 55-59.8 & 124-122.2 & 60-75 \\ 162-169 & 50-59.8 & 130-122.2 & 82-75 \end{bmatrix}$$

となり、結果は、

$$X = \begin{bmatrix} 1 & 5.2 & -2.2 & 5 \\ -4 & -2.8 & -0.2 & 3 \\ 11 & 12.2 & -7.2 & 0 \\ -1 & -4.8 & 1.8 & -15 \\ -7 & -9.8 & 7.8 & 7 \end{bmatrix}$$

となる。

分散共分散行列C_{XX}は上記の偏差データ行列Xの積$X'X$を標本数５で割ることで求められる。すなわち、

$$C_{XX} = X'X/5$$

である。実際に計算すると、

$$C_{XX} = \begin{bmatrix} 1 & -4 & 11 & -1 & -7 \\ 5.2 & -2.8 & 12.2 & -4.8 & -9.8 \\ -2.2 & -0.2 & -7.2 & 1.8 & 7.8 \\ 5 & 3 & 0 & -15 & 7 \end{bmatrix} \begin{bmatrix} 1 & 5.2 & -2.2 & 5 \\ -4 & -2.8 & -0.2 & 3 \\ 11 & 12.2 & -7.2 & 0 \\ -1 & -4.8 & 1.8 & -15 \\ -7 & -9.8 & 7.8 & 7 \end{bmatrix} \times \frac{1}{5}$$

となり、

$$C_{XX} = \begin{bmatrix} 37.6 & 44.8 & -27.4 & -8.2 \\ 44.8 & 60.6 & -36.8 & 4.2 \\ -27.4 & -36.8 & 24.2 & 3.2 \\ -8.2 & 4.2 & 3.2 & 61.6 \end{bmatrix}$$

と求められる（計算は小数点以下第２位四捨五入）。

分散共分散行列C_{XX}の対角成分は、各変数の分散を示している。例えば、２番目の変数である体重の分散は60.6であることが分かる。また、縦と横の変数同士の組み合わせは、それらの変数間の共分散を示している。例えば、２番目の変数（体重）と３番目の変数（最高血圧）の共分散は－36.8である。なお、分散共分散行列は、対角成分を挟んで右上の三角部分の成分と左下の三角部分の成分は対称になる（対称行列）。

分散共分散行列C_{XX}の各成分を、対応する２変数の標準偏差で割ることで相関係数が求められるので、相関係数行列R_{XX}を簡単に求めることができる。その結果は、

$$R_{XX}=\begin{bmatrix} 1.000 & 0.939 & -0.909 & -0.170 \\ 0.939 & 1.000 & -0.961 & 0.068 \\ -0.909 & -0.961 & 1.000 & 0.082 \\ -0.170 & 0.068 & 0.082 & 1.000 \end{bmatrix}$$

である（小数点以下第4位四捨五入）。

主成分分析では、相関係数行列の固有値、固有ベクトルを求めればよい。上記の相関係数行列の固有値と固有ベクトルを求めると、表1のようになる。

表1で示したように、固有値・固有ベクトルは4つ存在する。なお、表1では各固有ベクトルの要素の2乗和は1になるように計算されている。主成分負荷量を求めるには、固有値の平方根を各変数の固有ベクトルに掛ければよい。表2に各変数の主成分負荷量を示した。

表2で示したように、第1主成分の寄与率は約72%、第2主成分が25.6%であり、この2つで97.6%の寄与率である。従って、4つの変数の情報は、この2つの主成分に集約されていると言ってもよいだろう。また、変数ごとの主成分負荷量は第3主成分以降では、0.2程度しかないことも分かり、主成分と各変数との相関も小さい。

4. おわりに

今回は、主成分分析の概略を解説するとともに、データ行列からの計算例を示した。実際には、これらの計算を電卓で求めたりすることは、絶対にあり得ないし、やっても大抵はどこかで計算間違いするのが普通である。多変量解析では、データから相関係数行列、固有値、固有ベクトルなどを求める計算は、すべてパソコンを用いて行う。

我々が行うことは、研究のためにデータを集めること、正しい分析方法の選択、そして得られた結果の正しい解釈ということになる。

表1　相関係数行列の固有ベクトルと固有値

	固有ベクトル			
変数名	第1	第2	第3	第4
1) 身長	0.574	0.082	0.717	−0.387
2) 体重	0.579	−0.162	−0.014	0.799
3) 最高血圧	0.576	0.015	0.693	0.433
4) 最低血圧	0.056	−0.983	0.072	−0.157
固有値	2.879	1.024	0.087	0.010

表2　主成分負荷量と寄与率など

	主成分			
変数名	第1	第2	第3	第4
1) 身長	−0.973	0.083	0.211	−0.039
2) 体重	−0.983	−0.165	−0.004	0.081
3) 最高血圧	0.978	0.014	0.204	0.044
4) 最低血圧	0.095	−0.995	0.021	−0.016
固有値	2.879	1.024	0.087	0.010
累積固有値	2.879	3.903	3.990	4.000
寄与率(%)	71.968	25.611	2.167	0.254
累積寄与率(%)	71.968	97.579	99.746	100.000

＊参考文献

[1] 高木廣文(2000)：多変量解析の基礎知識：超音波検査技術, 25(5), pp.344-350.

[2] 柳井晴夫・高木廣文編著(1986)：多変量解析ハンドブック：現代数学社, 京都.

[3] 高木廣文・柳井晴夫編著(1995)：HALBAUによる多変量解析の実践：現代数学社, 京都.

[4] 高木廣文(2006)：HALBAU7によるデータ解析：シミック㈱.