＜教育講演（第34回年次学術集会より）＞

臨床検査データを合成・縮約するための主成分分析
第3回　主成分分析の仕組みと解析実例

山西 八郎

Synthesis and Reduction of the Coefficient of Variation by Principal Component Analysis
Part 3: Principal Component Analysis Mechanism and Analysis Examples

Hachiro Yamanishi

Summary　Examples of mathematical and analytical examples of principal component analysis (PCA) were explained. PCA is an analytical method that summarizes the information possessed by multiple observed variables and synthesizes it into a smaller number of variables (principal components). In particular, the principal component score, a measure of the strength of the principal components of the individual giving the variable, is an important statistic. As an example of the analysis, the relationship between obesity and lifestyle was described. In PCA, it is important to understand the number of principal components to be synthesized and how to interpret them. PCA is a useful analysis method not only in the social sciences, but also in medicine.

Ⅰ．はじめに

主成分分析とは、複数の変数の有する情報を要約し、それをより少数の変数（主成分）に組み替える（合成する）分析法である。そしてその仕組みには、第2回で解説した主成分回帰の考え方が取り入れられている[1)2)]。ただし、新しい変数を合成できたとしても、それは解析のOutput（結果）であり、これを統計的に2次・3次加工してOutcome（成果）を導く必要がある。Outcomeは「終わりの言葉」と言い換えることもできる。なお、主成分分析の仕組みをグラフ上（平面上）で解説するためには2変量が限界であることをご了承ください。

Ⅱ．主成分分析の仕組み

Fig. 1に10人の生徒の数学と国語の試験成績（10点満点）と散布図を示す。回帰直線は主成分回帰により求めたものである。両科目の平均（破線）の交点が散布図の重心となる。主成分回帰による回帰直線は、各プロットから回帰直

天理大学 医療学部
〒632-0018　奈良県天理市別所町80-1
天理大学別所キャンパス

連絡先：山西 八郎
Tel: +81-743-63-7811
E-mail: yamaha@sta.tenri-u.ac.jp

Tenri University Department of Clinical Laboratory Science,
80-1 Bessho-cho, Tenri, Nara 632-0018, Japan

生徒	A	B	C	D	E	F	G	H	I	J	平均
数学	2	1	2	3	5	4	8	6	7	4	4.2
国語	3	4	2	2	4	4	5	3	6	9	3.8

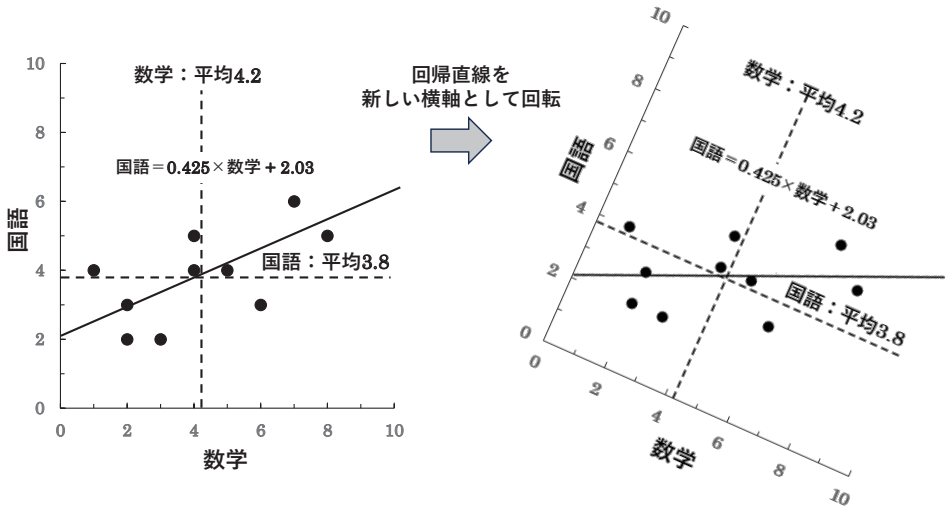

Fig. 1　Rotation with the regression line as the horizontal axis

線に下した垂線の2乗和が最小となるように決定されているため[2]、Fig. 1右のグラフのように回帰直線を新しい横軸､重心を原点と考えると、重心より右に位置する生徒ほど国語と数学の成績が良好で、左に位置する生徒では両科目の成績が低い傾向にある（Fig. 2）。主成分分析では、横軸と考えた回帰直線を第1主成分軸と定義する。したがって、少々乱暴な解釈となるが、第1主成分軸は数学と国語の成績が合成された｢総合学力｣を意味していると考えることができる。合成された主成分の解釈とネーミングは解析者に委ねられる。

一方、重心から各プロットまでの第1主成分軸方向の距離が、各生徒の有する「総合学力」の大きさを表す主成分スコアとなる。また、重心を原点とすると、生徒Iの主成分スコアの符号は（+)、生徒Bは（－）となる。

次に、原点を通り第1主成分軸に垂直な軸を第2主成分と定義する（Fig. 3）。この場合、原点よりも上に位置する生徒は数学の成績が良好で、原点よりも下にある生徒は国語の成績が良い傾向にある。したがって、これも乱暴な解釈となるが、第2主成分は理系学力と文系学力の「分野別学力」を表していると解釈できる。

第1主成分スコアと同様に、原点からプロットまでの第2主成分軸方向の距離が「分野別学力」の大きさ：第2主成分スコアを表しており、生徒Iは（+）の､生徒Hは（－）のスコアとなる。ただし、第1主成分スコアの正負は「総合学力」の優劣を表しているが、第2主成分スコアの正負は「理系学力」と「文系学力」の区別を意味している。また、第1、2主成分スコアが0に近い生徒は、その集団の中での平均的な学力を有していると解釈する。

Ⅲ．主成分分析の数理[1) 2)]

1．固有値と固有ベクトルの算出

以上までは、散布図を利用して主成分分析の仕組みを解説したが、数理的に主成分分析の目的は、p個の観測変数をより少ないm個の主成分：Zmに要約、合成することにある。これを次元削除という｡具体的には次式に示すように、観測変数に重み（係数）を付けた線形な一次式よりZmを合成することにある。係数a_{11}～a_{mp}を固有ベクトルという。

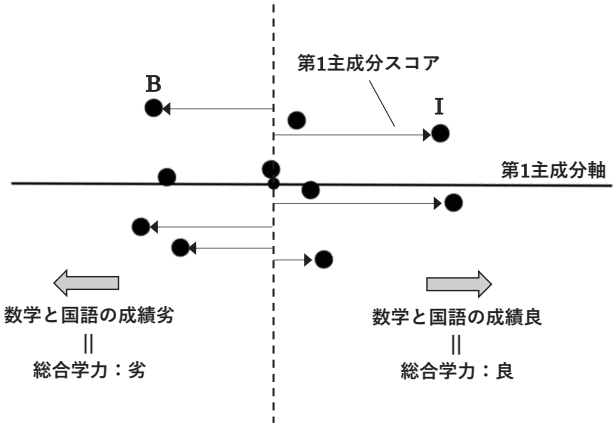

Fig. 2　First Principal Component and Principal Component Score

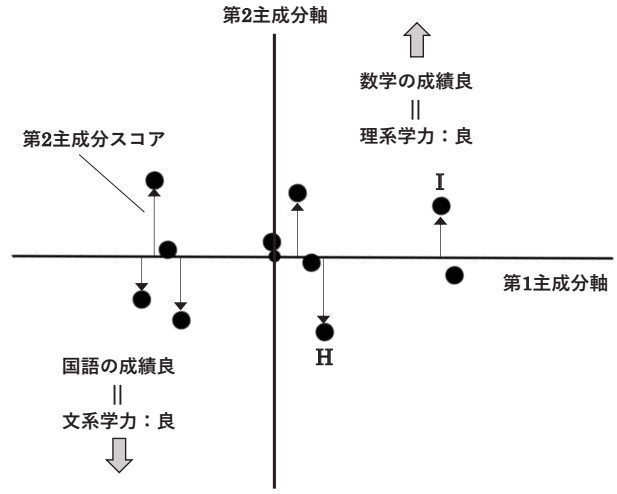

Fig. 3　Second Principal Component and Principal Component Scores

$$Z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$

$$Z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$$

$$\vdots$$

$$Z_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mp}x_p \ (m < p)$$

説明を簡略化するために測定変数をx_1、x_2の2変数とし固有ベクトルの算出プロセスについて解説すると、まずx_1、x_2の分散をV_{11}、V_{22}、共分散をV_{12}、V_{21}とすると、分散・共分散行列より、それぞれの主成分スコアの分散に相当する固有値：λを次の行列式より求める。

$$\begin{vmatrix} V_{11}-\lambda & V_{12} \\ V_{21} & V_{22}-\lambda \end{vmatrix} = 0 \quad \text{より}$$

$$(V_{11}-\lambda)(V_{22}-\lambda) - V_{12} \times V_{21} = 0$$

2つの主成分を合成するとして、第1および第2主成分の固有ベクトルをそれぞれa_1、a_2とすると、次の連立方程式より固有ベクトルを求める。

$$V_{11}a_1 + V_{12}a_2 - \lambda a_1 = 0$$

$$V_{21}a_1 + V_{22}a_2 - \lambda a_2 = 0$$

$${a_1}^2 + {a_2}^2 = 1$$

以上のプロセスでFig. 1の試験成績から固有値

Table 1　Principal Component Analysis Results

観測変数	Z_1: 固有ベクトル	Z_2: 固有ベクトル
数学	0.9215	−0.3884
国語	0.3884	0.9215
固有値	6.0569	0.9653
寄与率	0.8625	0.1375
累積寄与率	0.8625	1.0000

（分散・共分散行列より算出）

と固有ベクトルを求めるとTable 1に示す結果が得られた。

固有値と固有ベクトルは観測変数の相関行列からも算出することができる。計算のプロセスは分散・共分散行列と同様である。その使い分けは、観測変数の単位が異なる場合は相関行列から、単位が同じ場合は共分散行列より算出する。近年では単位が同じでも相関行列から算出される場合が多い。Table 1の結果より、主成分は次式で定義される。

$$Z_1 = 0.9215 \times \text{数学} + 0.3884 \times \text{国語}$$
$$Z_2 = -0.3884 \times \text{数学} + 0.9215 \times \text{国語}$$

2. 主成分スコアの算出

得られた主成分式に各観測変数の値を代入することにより主成分スコアが算出される。例として、生徒A（数学2点、国語3点）の主成分スコアは、

総合学力：

$$Z_1 = 0.9215 \times 2 + 0.3884 \times 3 = 3.0082$$

分野別学力：

$$Z_2 = -0.3884 \times 2 + 0.9215 \times 3 = 1.9877$$

と計算される。

ここで、Fig. 4に示すように、主成分スコア間の相関係数は数学的に0となる。これは、主成分軸は互いに直交している、言い換えると主成分ベクトルの内積が0となるように軸が定義されているからである。なお、上式で算出した値から、各主成分における平均を差し引いた値、偏差を主成分スコアとする場合もある。いずれにせよ、解析結果から「終わりの言葉」を導くうえにおいて、主成分スコアは極めて有用な統計量となる。

生徒	Z_1スコア	Z_2スコア
A	3.0082	1.9877
B	2.4751	3.2976
C	2.6198	1.0662
D	3.5413	0.6778
E	6.1611	1.7440
F	5.2396	2.1324
G	9.3140	1.5003
H	6.6942	0.4341
I	8.7809	2.8102
J	5.6280	3.0539
平均	5.3462	1.8704

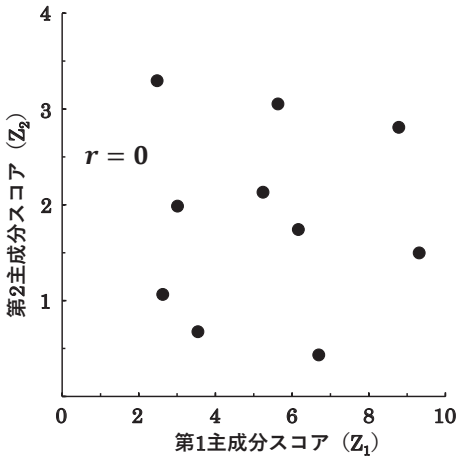

Fig. 4　Correlation of principal component scores

Table 2　Results of Principal Component Analysis with Lifestyle as a Variable

	第1主成分：Z_1	第2主成分：Z_2	第3主成分：Z_3
飲酒度	**0.329**	0.176	− 0.264
喫煙度	**0.356**	0.202	− 0.204
食速	0.186	− 0.070	**0.395**
満腹度	0.233	− 0.175	**0.399**
塩気	**0.366**	0.073	0.214
香辛	**0.402**	0.077	0.099
脂身	**0.360**	0.120	− 0.080
不規則度	0.205	**− 0.352**	− 0.299
睡眠	0.039	**0.324**	0.178
充実度	− 0.078	**0.518**	0.033
憂鬱度	− 0.010	**− 0.483**	0.215
固有値	**2.77**	**1.64**	**1.32**
寄与率	**0.173**	**0.102**	**0.082**
累積寄与率	**0.173**	**0.275**	**0.357**

3. 固有値

Ⅲ-1でも述べたが、固有値は主成分スコアの分散に相当する。したがって、固有値の大きさはその主成分軸方向への変数を与える固体間の相違を反映している。つまり、固有値の大きな主成分ほど、より多くの情報を含有していると考えることができる。また、第1主成分での固有値が最大となり、第2主成分以降、その固有値は低下していく。

4. 寄与率

観測データ全体が有している情報量を1としたときの、各主成分が保持している情報量の比率を意味する。固有値の総和を$\sum\lambda$、k番目の主成分の固有値をλ_kとすると、その寄与率は、

$$\frac{\lambda_k}{\sum\lambda}$$

で計算される。また、第1主成分からの寄与率を加算した値を累積寄与率という。また、相関行列より主成分分析を行った場合は、固有値の総和は観測変数の数:pに等しくなるため、上式は、

$$\frac{\lambda_k}{p}$$

となる。

5. 主成分負荷量

標準化した観測データと主成分スコアとの相関係数。この値が大きいほど、主成分を良く説明する観測変数と解釈できる。i番目の主成分の固有値をλ_i、固有ベクトルをa_iとすると次式より算出される。

$$a_i \times \sqrt{\lambda_i}$$

6. 合成する主成分の数

主成分分析において合成する主成分の数は解析者が指定する。基準となるルールはないが、個人的には、累積寄与率が0.6以上まで、相関係数行列による主成分分析では、固有値が1.0以上まで、あるいは上記の条件を満たしていても主成分の解釈が可能な主成分までとしている。

Ⅳ. 解析実例[2)]

1. 主成分の解釈とネーミング

Table 2は、健常日本人成人504名（女性182人、男性322人、平均年齢40.9歳）について調査された生活習慣についてのアンケート結果（1 ～ 5の5段階で回答）を観測変数とした主成分分析の結果である。また、アンケート項目の身長と体重からBody Mass Index：BMIを算出し、BMI

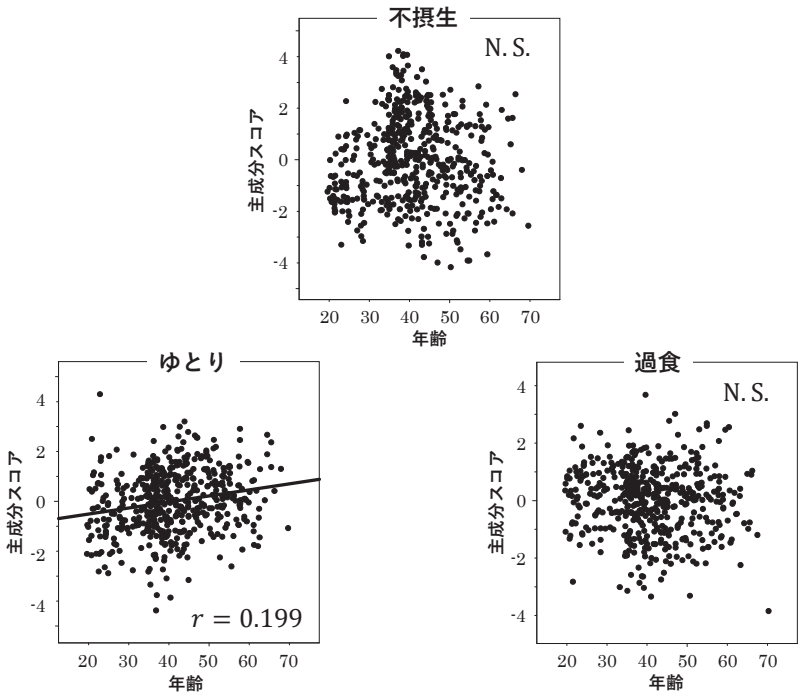

Fig. 5　Correlation between principal component score and age

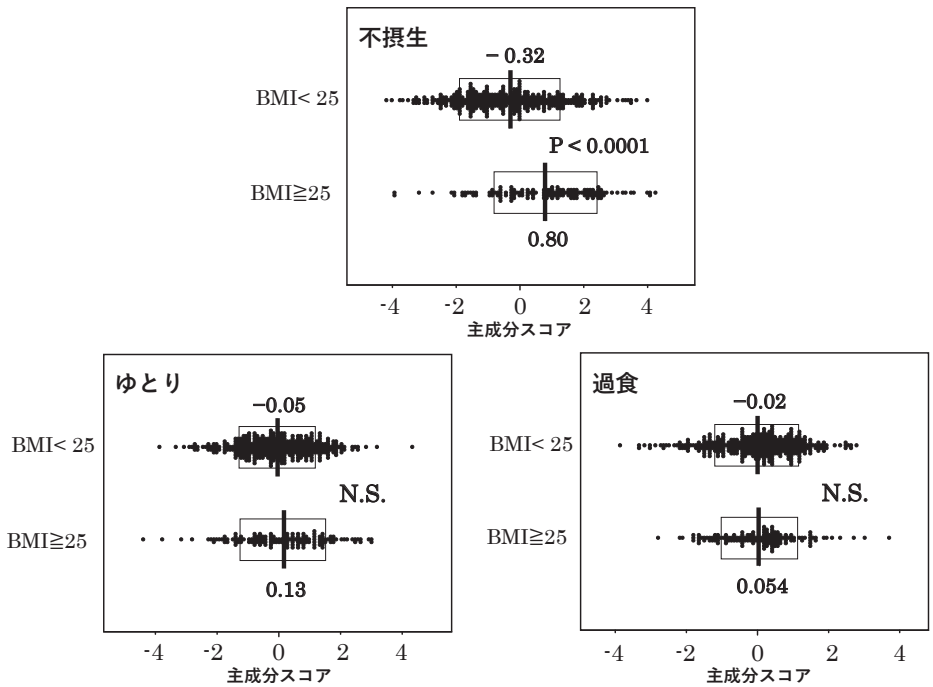

Fig. 6　Distribution of principal component scores grouped by BMI

Table 3　Multiple regression analysis of principal component scores with BMI as the objective variable

変数	β	SE	std β	t 値	df	P
（初期値）	23.8	0.229				
性別	**-1.940**	0.475	-0.253	4.090	499	**0.0001**
不摂生	**0.357**	0.128	0.161	2.794	499	**0.0054**
ゆとり	0.114	0.123	0.039	0.925	499	0.3551
過食	**0.387**	0.146	0.121	2.654	499	**0.0082**

β: 偏回帰係数　SE: 標準誤差　stdβ:標準偏回帰係数　df: 自由度　P: 有意確率

≧25を「肥満」とした。表中の数値は相関行列より求めた固有ベクトルを示しており、その絶対値が0.3以上の観測変数の組合せに注目して各主成分の解釈とネーミングについて解説する。

まず、Z_1では、「飲酒度」「喫煙度」と嗜好を表す「塩気」「香辛」「脂身」に0.3以上の固有ベクトルを示していることから、この主成分は「不摂生」と解釈、ネーミングすることができる。次にZ_2では「睡眠」が十分で、生活の「充実度」も高く、逆に「不規則度」「憂鬱度」に対しては負のベクトルを示していることから、「ゆとり」とネーミングすることができる。さらにZ_3は食べる速度が速く、満腹になるまで食事を取ると解釈できることから「過食」とネーミングした。主成分のネーミングは解析者の自由であるので。例えば「不摂生」を「不健康度」としてもかまわない。なお、第3主成分までの累積寄与率は0.357であるが、第4主成分まで合成すると、その主成分の解釈が困難となるために主成分数を3で打ち切っている。

以上のように3つの主成分が合成されネーミングすることができたが、これは解析のOutput：「結果」あるいは「はじめの言葉」でしかなく、これを使ってOutcome：「終わりの言葉」につなげる必要がある。そのツールとなるのが主成分スコアである。

2. 主成分スコアによる考察

Fig. 5に年齢と主成分スコアの相関関係を示す。加齢とともに、特に食生活に対する関心が高まるために、「不摂生スコア」との間には負の相関性を予想していたが、両者間に有意な関係は認められなかった。一方、「ゆとりスコア」と年齢の間には5％水準で有意な正の相関関係が認められたが、標本数が500以上であることを考え合わせると、実質的な相関性はないと考えられる。

BMI≧25を肥満群としてBMI＜25群（正常群）と主成分スコアを比較した結果、肥満群の不摂生スコアが有意に高値を呈した（Fig. 6）。これに対して、肥満の原因の一つとして考えられる食習慣に関連した過食スコアにおいては、2群間に有意差は認められなかった。そこで、BMIを目的変数、不摂生、ゆとり、過食スコアおよび性別（女性＝1、男性＝0）を説明変数とした重回帰モデルを検証すると、不摂生、過食スコアが有意な説明変数であるとともに、性別も高度に有意な変数であった（Table 3）。この場合、性別の標準偏回帰係数が負の値であることは、男性よりも女性のBMIが低値傾向であることを意味している。またその絶対値の大きさより、不摂生、過食スコアよりも性別のBMIに対する影響が大きいと解釈された。以上の結果より、肥満群と正常群で過食スコアに有意差が認められなかったのは、性別が交絡することにより有意差が隠されていたものと考えられる。

Ⅴ．終わりに

以上、主成分分析の数理と解析実例について解説した。固有値や固有ベクトルの算出は統計ソフトで簡単に対応できるので、合成すべき主成分数とその解釈に対する理解が最も重要なポ

イントとなる。主成分分析は、特に社会科学やマーケティングの分野で広く用いられるが[3)4)5)]、臨床検査の分野においても、生活習慣と検査成績、疾患の有無・程度を解析するうえにおいて極めて有用な分析法である。また個別の検査項目の基準範囲だけでなく、「肝機能」「腎機能」「脂質代謝」といったより幅の広い病態概念を合成し、その基準範囲を設定することも可能である。同時に、予防医学を実践するうえにおいても主成分分析は強力な解析法となるものと考えている。

本論文内容に関連する著者の利益相反：なし

文献

1） 管 民郎：主成分分析．多変量解析の実践（上），128-158，現代数学社，京都（2005）

2） 山西 八郎：主成分分析．寺子屋統計教室．92-100，情報機構、東京（2020）

3） Zajicek JL, Tillitt DE, Schwartz TR, Schmitt CJ and Harrison RO. Chemosphere, 40:539-548, 2000.

4） Torri GM, Torri J, Gulian JM, Vion-Dury J, Viout P and Cozzone PJ. Magnetic resonance spectroscopy of serum and acute-phase proteins revisited: a multiparametoric statistical analysis of metabolite variations in inflammatory, infectious and miscellaneous disease, Clinica Chimica Acta，279, 77-96,1999.

5） Castro IA, Barroso LP and Sinnecker P. Functional foods for coronary heart disease risk reduction: a meta-analysis using a multivariate approach. Am J Clin Nutr, 40, 32-40, 2005.