統計講座（第11回）

生理学的データの主成分分析

高木　廣文

はじめに

大学の入学試験で英語，数学，国語の3科目の個人の得点を合計し，総合得点の順に順位を付けることがある．このような場合，総合得点とは3つのデータがもつ情報を合計して，1つの情報にまとめたものと考えることができるだろう．一般に多数のデータがもつ情報を，より少数の合成得点により代表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の問題として，入学試験などですべての科目の得点を単純に加えればよいのだろうか．例えば，理科系ならば数学を2倍にしたり，文化系ならば英語を2倍にする必要はないのだろうか．単純にすべての科目の合計得点を求めるのではなく，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，多種多様な検査項目を調べた場合，その検査項目間の関係を分析し，肝機能，脂質代謝，糖代謝などそれぞれの項目が代表する情報に関する合成得点を求めて，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く，変数相互間の構造から，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)合成変数のもつ意味から，変数間の関係を検討

と大きく2つに分けて考えることができる．

Hirofumi TAKAGI：新潟大学医学部保健学科
〒951-8518　新潟県新潟市旭町通2番町746番地
e-mail：takagi@clg.niigata-u.ac.jp
http://www.clg.niigata-u.ac.jp/~takagi/

主成分分析の理論

いま，n人についてp個の多変量データ$X=(x_1, x_2, \ldots, x_p)$があるものとしよう．ここでは計算に便利なように，あらかじめ各データはその変数の平均値を引いてあるものとする．すなわち，各変数の平均が0になるようにあらかじめ変換してあるものとする．

一般に，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, ..., p) を掛けて加えたもの，

$$f = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

と書くことができる．ちなみに，重みw_i (i=1, 2, ..., p) 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f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wを定めることにする．最終的には，

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

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を示しており，λは後述するようにある定数となっている．いま，

$$a = C_{XX} w$$

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すものとなる．ベクトルaは「構造ベクトル structure vector」と呼ばれるものであり，各変数の合成変数への寄与の程度を示す．結局，

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

となる．定数λとベクトルaは，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になっている.

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λとaを求めればよい.この場合，構造ベクトルaは主成分と各変数の相関係数を示しており，「主成分負荷量 principal loading」とも呼ばれている.

実際には，主成分は1つとは限らず，最大で分析に用いた変数の個数だけ存在する．いまp個の変数の主成分分析で，正の固有値$\lambda_1 > \lambda_2 > \cdots > \lambda_m$に対応して，m個の主成分があるものとする$(m \leqq p)$.

その場合，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ろう．いま，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，

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

とすると，第j主成分の固有値をTで除したものは，

$$c_j = \lambda_j / T \quad (j=1, 2, \ldots, m)$$

は，第j主成分の「寄与率 contribution rate」と呼ばれ，全変数の全情報に対して，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j主成分までのj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.

$$cc_j = (\lambda_1 + \lambda_2 + \cdots + \lambda_j) / T$$

により求められる.

実際の分析では，あまり小さな固有値に対応する主成分を求めることが，意味をもつとは考えられない．なぜならば，主成分分析では，なるべく少数の主成分で，全体の変数のもつ変動を説明することを意図しているからである.

しかし，意味のある主成分の個数を定めるための決定的な方法はない．そのため，便宜的に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.ただし，絶対的な基準はない．通常は，だいたい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以上の主成分のみ採用する.

実際には，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ことが多いようである.

臨床検査データへの応用例

ここでは，重回帰分析の応用例でも使用した約1,200名の生理学的な臨床検査データ[1]をもとに，主成分分析の実際例を示してみよう.

表1に主成分分析に用いた変数間の相関係数行列を示した.分析に用いたのは，①年齢，②身長，③体重，④BMI（Body Mass Index，体重／身長の2乗（Kg/m^2)），⑤EC腎臓周囲脂肪厚，⑥EC皮下脂肪厚，⑦TG，⑧TC，⑨HDL，⑩GOT，⑪GPT，⑫γ-GTP，⑬UA，⑭FBS，の14変数である．なお，**表1**では掲載スペースを少なくするために，相関係数は小数点以下第3位で四捨五入して表示しているが，実際の計算ではより大きな桁数を使用している.

表1をみると，GOTとGPTの間には0.82という高い相関があることが分かる．さらに，GOTとγ-GTPには0.43，GPTとγ-GTPには0.36の相関があることが分かる．同様に，体重，BMI，EC腎臓周囲脂肪厚，EC皮下脂肪厚の間には相互に正の相関関係があることが分かるだろう．その他にも，いくつかの変数間にひとつのまとまりになるような相関関係をみつけられるかもしれない.

表1　主成分分析に用いる変数間の相関係数行列

変数名	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)
1) 年齢	1.00	−0.34	−0.29	−0.10	0.07	−0.17	−0.08	0.10	0.06	−0.02	−0.19	−0.09	−0.11	0.05
2) 身長	−0.34	1.00	0.58	−0.02	0.24	−0.25	0.07	−0.14	−0.13	−0.02	0.06	0.12	0.32	−0.01
3) 体重	−0.29	0.58	1.00	0.79	0.56	0.33	0.26	−0.00	−0.34	0.19	0.35	0.18	0.36	0.14
4) BMI	−0.10	−0.02	0.79	1.00	0.51	0.59	0.25	0.10	−0.33	0.25	0.38	0.13	0.21	0.18
5) EC腎臓周囲脂肪厚	0.07	0.24	0.56	0.51	1.00	0.16	0.23	0.13	−0.27	0.20	0.28	0.23	0.32	0.15
6) EC皮下脂肪厚	−0.17	−0.25	0.33	0.59	0.16	1.00	0.10	0.16	−0.13	0.18	0.27	0.03	−0.03	0.06
7) TG	−0.08	0.07	0.26	0.25	0.23	0.10	1.00	0.29	−0.38	0.16	0.22	0.25	0.21	0.20
8) TC	0.10	−0.14	−0.00	0.10	0.13	0.16	0.29	1.00	0.09	0.11	0.11	0.08	0.07	0.10
9) HDL	0.06	−0.13	−0.34	−0.33	−0.27	−0.13	−0.38	0.09	1.00	−0.07	−0.20	−0.01	−0.20	−0.14
10) GOT	−0.02	−0.02	0.19	0.25	0.20	0.18	0.16	0.11	−0.07	1.00	0.82	0.43	0.14	0.12
11) GPT	−0.19	0.06	0.35	0.38	0.28	0.27	0.22	0.11	−0.20	0.82	1.00	0.36	0.17	0.16
12) γ-GTP	−0.09	0.12	0.18	0.13	0.23	0.03	0.25	0.08	−0.01	0.43	0.36	1.00	0.25	0.13
13) UA	−0.11	0.32	0.36	0.21	0.32	−0.03	0.21	0.07	−0.20	0.14	0.17	0.25	1.00	−0.03
14) FBS	0.05	−0.01	0.14	0.18	0.15	0.06	0.20	0.10	−0.14	0.12	0.16	0.13	−0.03	1.00

(953例)

表2　各変数の主成分負荷量

変数名	主成分1	主成分2	主成分3	主成分4	主成分5	主成分6	主成分7
1) 年齢	−0.260	0.361	−0.104	0.518	0.312	−0.547	−0.083
2) 身長	0.325	−0.774	0.293	−0.008	−0.008	0.046	0.220
3) 体重	0.821	−0.391	−0.220	−0.118	0.097	−0.037	0.169
4) BMI	0.770	0.089	−0.490	−0.136	0.133	−0.091	0.046
5) EC腎臓周囲脂肪厚	0.652	−0.121	−0.127	0.264	0.330	−0.287	0.113
6) EC皮下脂肪厚	0.436	0.372	−0.537	−0.391	0.067	0.208	−0.012
7) TG	0.487	0.083	0.004	0.504	−0.318	0.381	−0.233
8) TC	0.174	0.419	−0.036	0.420	0.298	0.568	0.167
9) HDL	−0.467	0.180	0.249	−0.242	0.460	0.177	0.496
10) GOT	0.550	0.482	0.499	−0.237	−0.002	−0.179	−0.104
11) GPT	0.679	0.371	0.370	−0.285	−0.104	−0.107	−0.091
12) γ-GTP	0.447	0.161	0.563	0.074	0.055	0.048	0.084
13) UA	0.459	−0.347	0.237	0.241	0.345	0.083	−0.228
14) FBS	0.267	0.211	−0.047	0.323	−0.509	−0.187	0.625
固有値	3.787	1.818	1.512	1.317	1.031	1.023	0.895
累積固有値	3.787	5.605	7.117	8.435	9.466	10.488	11.383
寄与率 (%)	27.05	12.99	10.80	9.41	7.37	7.30	6.39
累積寄与率 (%)	27.05	40.04	50.84	60.25	67.61	74.92	81.31

しかし，相関係数行列から，このような関係をみつけるのは，変数の個数が増加するにつれ，極めて困難になることは，容易に理解できるだろう．

表2は，主成分分析により，相関係数行列から各変数の主成分負荷量を求めたものを示したものである．

実際には分析に用いた変数が14個あるので，主成分も14個抽出されたが，ここでは，累積寄与率がほぼ80%となる第7主成分まで示した（**表2**では，第1主成分を「主成分1」のように示している）．このうち，第7主成分は固有値が0.895と1より小さいので，1つの変数がもつ情報よりも小さい情報しか持たないことになる．情報の圧縮という点からは，第6主成分までが主成分として有用なものと考えられる．

すでに説明したように，主成分負荷量は各主成分と変数との相関係数を示すので，主成分負荷量の大きな変数同士もまたその主成分に関して関係

があることになる．主成分分析は，多くの変数のもつ情報を圧縮して，少数の主成分によって表わすことを第一目的としているが，このような相関関係を利用して，主成分のもつ意味を考察し，変数間の相関構造を検討することにも用いられている．この点から，第1主成分の各変数の主成分負荷量をみると，体重とBMIの値が正に大きいことが分かる．また，EC腎臓周囲脂肪厚，GPT，GOTも比較的大きな負荷量を持っている．逆に，HDLでは負の負荷量となっている．年齢も負の負荷量になっている点から，おそらくは，体脂肪に関係する主成分と考えられるが，EC皮下脂肪厚の負荷量が0.436と他の変数に比べて，それほど大きくはない．

第2主成分は，身長の負荷量が負に最も大きく，体重とUAも同様に負の負荷量であった．年齢，TC，GOT，GPTなどが弱い正の負荷量であった．この主成分の正の方向は，身長の低い傾向と関係している点に注意が必要である．主成分負荷量の値の正負は，単に相関の方向を示しているだけである．第2主成分のもつ意味の解釈は困難である．

ところで，EC皮下脂肪厚についてみると，第3主成分の負荷量が－0.537と負の方向に他の主成分より大きくなっている．第3主成分はγ-GTP，GOT，GPTが正に大きく，体重は負に大きいことから，EC皮下脂肪厚は体重が多いと厚くなるが，肝機能が悪いと薄くなるような傾向を示す主成分であることが分かる．また，第3主成分ではEC腎臓周囲脂肪厚の負荷量が－0.127と小さいことから，ほとんど関係がないことが分かる．EC腎臓周囲脂肪厚はすでにみたように，第1主成分で極めて負荷量が高く，肝機能検査に関する項目の値が大きいと厚くなる傾向があり，この成分は同時にEC皮下脂肪厚とも正に関係していた．このように，脂肪厚に関する主成分は，1つだけとは限らないようである．

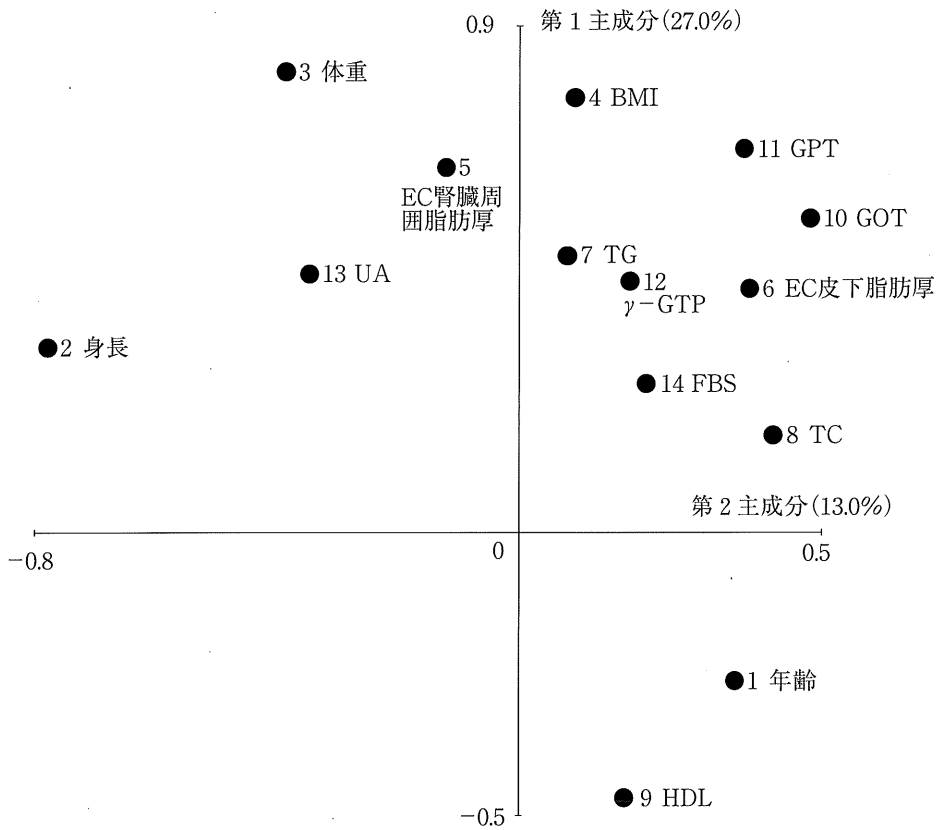

図1　第1主成分と第2主成分による変数の平面配置図

表3　各ケースの主成分得点

番号	主成分1	主成分2	主成分3	主成分4	主成分5	主成分6	主成分7
1	---	---	---	---	---	---	---
2	2.252	3.523	0.398	−2.130	−0.592	−1.002	0.082
3	−0.795	1.283	−0.944	−2.120	−0.353	0.127	−0.074
4	0.567	−0.613	−0.955	−0.052	−0.214	−0.352	0.532
5	−0.800	0.898	1.427	−1.659	0.106	0.421	0.675
6	−0.257	0.521	1.582	−1.009	0.967	0.252	0.091
7	−0.583	1.164	−1.924	−0.892	−0.286	0.514	0.138
8	0.305	−0.762	−0.423	−0.752	0.630	−1.282	0.004
9	−0.468	0.632	−0.978	−1.463	0.769	−0.865	−0.750
10	−0.259	1.507	−0.460	−1.225	1.075	0.926	−0.355
11	−0.206	0.614	−0.791	−0.617	0.707	0.200	0.707
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
1235	−1.322	0.957	−0.096	0.059	0.275	0.024	−0.410
1236	−0.901	0.555	0.505	−1.087	−0.577	−0.664	0.188
1237	−0.936	0.570	−1.145	−0.705	0.211	1.033	−0.031
1238	−1.391	1.194	−0.310	−0.586	0.918	1.638	0.386
1239	−0.585	0.032	−0.169	1.730	0.282	0.953	−1.049
1240	−0.325	0.640	0.692	0.670	−0.159	0.142	−0.190
1241	1.036	1.863	0.498	−1.005	−0.357	−1.477	0.530
1242	0.066	1.637	−1.935	0.580	−0.415	0.481	−1.649
1243	−0.009	−0.928	−1.021	−0.577	−0.189	−2.475	0.314
1244	1.260	1.782	0.000	−0.854	0.214	−2.401	2.038

(注) ---は欠損データ。

図1は，縦軸に第1主成分を，横軸に第2主成分をとり，各変数の主成分負荷量を座標として，平面上にプロットしたものである．このような図を作成することで，2次元平面上ではあるが，変数間の関係および各主成分のもつ意味を解釈するうえで，直観的な理解の助けになるだろう．例えば，身長と年齢，HDLは対極の位置にあるようにみえるだろう．また，各変数の主成分負荷量について，上ですでに説明した文章を図をみながら読むと，より簡単に内容が理解できることも分かると思う．

第4主成分以降については，ここでは説明を省略するが，正負に関わらず負荷量の大きな変数に注目することが重要である．

表3は各個人（ケース）の主成分得点を示したものである．主成分得点は，平均が0，分散が1になるように標準化されている．したがって，主成分得点が標準正規分布に従って分布しているものと仮定すると，各ケースのその主成分に関する傾向が判断できる．すなわち，主成分得点が0の近くならば，平均的であるが，得点が2以上ならば，非常にその主成分が代表する傾向が強い（上位2.5%未満に入る）ことになる．逆に，得点が−2以下ならば，正方向とは逆の傾向が極めて強い（負の方向に2.5%未満に入る）ことになる．

例えば，ケース番号2の第1主成分得点は2.252となっているので，このケースは，体重とBMIの値が大きく，EC腎臓周囲脂肪厚，GPT，GOTなの数値も大きい傾向にあり，逆に1238番のケースは得点が−1.391なので，これらの項目の数値は低い傾向にあることが分かるだろう．なお，1番目のケースは，今回使用した14変数のデータのうち，1つ以上に欠損があるために，分析に使用できなかったため，主成分得点が計算されていないことを示している．

主成分分析の事後分析

主成分分析の結果に，主成分分析に使用したなかった変数を加えて，解析後にさらに分析を行うことがよくある．例えば，重回帰分析で説明変数

表4　性別による主成分得点の差

変数名	平均値（SD）男（706例）	平均値（SD）女（247例）	Welchのt値	p値
1）主成分1	0.134（0.988）	−0.382（0.935）	7.338	0.000
2）主成分2	−0.363（0.783）	1.038（0.807）	23.620	0.000
3）主成分3	0.208（0.940）	−0.593（0.927）	11.622	0.000
4）主成分4	0.132（0.911）	−0.378（1.136）	6.363	0.000
5）主成分5	0.041（0.953）	−0.116（1.115）	1.968	0.050
6）主成分6	−0.078（0.953）	0.224（1.092）	3.859	0.000
7）主成分7	0.048（0.901）	−0.138（1.230）	2.173	0.031

間の相関が高く多重共線性の疑いがある場合には，疑いのあるいくつかの説明変数を用いて主成分分析を行い，求めたいくつかの主成分をそれらの変数の代わりにとして，重回帰分析を行うこともある[1]．

一般によく行う分析方法は，主成分分析に使用しなかった変数のうち，ケースをサブグループに分類するような項目（性，喫煙や飲酒の有無，患者か健常者か，学年，など）を用いて，各グループでの（母）平均値の差の検定を行うというものである．

表4は，主成分分析の結果得られたケースの主成分得点について，男女の差があるかを検討するために，2群の母平均値の差の検定を行ったものである．なお，ここでは2グループの主成分得点に等分散性を仮定しないで，Welch（ウェルチ）の検定（t検定）を行った[2]．

表4をみると分かるように，検定結果の有意性を示すp値（有意確率）はすべて0.05（5％）以下である．もっとも，統計学的な有意差が男女で認められるといっても，第5主成分や第7主成分での実質的な平均値の差は，それほど大きなものではない．第2主成分のt値が23.6と最も大きく，平均値は男が−0.363，女が1.038とかなりの差があった．これは，第2主成分の負の方向が，身長の高さと相関があり，主成分の正の方向はEC皮下脂肪厚と相関があることを考えれば，もっともな結果といえよう．

第1主成分では，平均値は男が0.134，女が−0.382となっていた．この結果は，体重，BMI，EC腎臓周囲脂肪厚，GPT，GOTなどに関係する総合的な値が，男で女に比べて高いことを示すものといえよう．

上記のように，主成分得点のサブグループごとの平均値を比較することで，各グループの総合的な特性を比較することが可能となる．

おわりに

今回は，基準変数がない場合の多変量解析法として，主成分分析を説明した．主成分分析は，情報圧縮（次元縮小）の方法として知られているが，応用上は分析に使用した変数間の相関構造を検討するために用いることが多い．また，事後解析として，患者と健常者での主成分得点の平均値を比較することで，特定の健康事象に関する因果推論に役立つこともある．

実際の計算は，変数間の相関係数行列を求め，その固有値と固有行列を求めるという，極めて単純な手法であるが，不慣れな読者には数式に惑わされる可能性が高いかもしれない．ただし，実際の応用はそれほど難しいものではないので，適当なデータがある場合には，変数間の相互関係を調べるために試してみるのもよいだろう．

分析には，今回もHALWIN[3]を使用したが，試作版は私のホームページからダウンロードできるので，興味ある読者は使ってみて欲しい．

参考文献

1）高木廣文：生理学的データの重回帰分析，超音波検査技術，2001；26：28-34.
2）高木廣文：2グループの平均値を比べる，超音波検査技術，1998；23：400-405.
3）高木廣文：HALWINによるデータ解析，現代数学社（京都），1998.