

Rと因子分析

1. 因子分析とは

因子分析(factor analysis)は、多くの変数により記述された量的データの分析方法として、1904 年にスピアーマン(Spearman)によって提案された。

因子分析で扱うデータの形式は主成分分析と基本的には同じであることから、同じ場面に利用されることが多いが、手法の開発の出発点は全く異なる。

主成分分析では、変数間の相関関係を用いて、無相関の合成変数を求めることで多くの変数を少ない変数に縮約するが、因子分析は、変数間の相関関係から共通因子を求めることで、多くの変数を少数個の共通因子にまとめて説明することを目的としている。

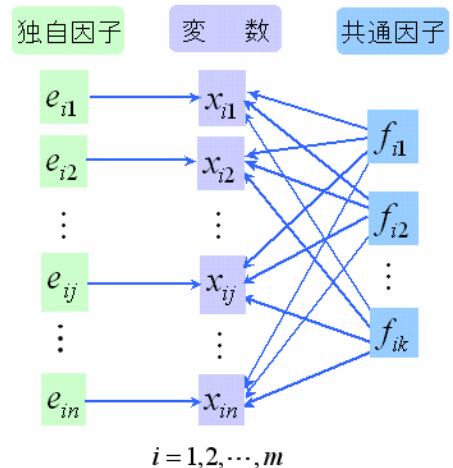
因子分析は、観測データにおける変数間の関連成分をまとめたものを共通因子(common factor)と呼び、他の変数と関係がなく、その変数のみ持っている成分を独自因子(unique factor)と呼ぶ。因子分析では、観測データはお互いに関連性を持っており、共通因子と独自因子に分解できることを前提としている。

例えば、表 1 のような観測データが図 1 に示すように k 個の共通因子と独自因子により構成されたとすると、式 1 のようなモデルで表現できる。

表 1 データ行列 $X_{m \times n}$

	x_1	x_2	...	x_j	...	x_n
個体 1	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}
個体 2	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
個体 i	x_{i1}	x_{i2}	⋮	x_{ij}	⋮	x_{in}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
個体 m	x_{m1}	x_{m2}	⋮	x_{mj}	⋮	x_{mn}

図 1 因子分析モデル



$$\begin{aligned}
 x_{i1} &= a_{11}f_{i1} + a_{12}f_{i2} + \dots + a_{1k}f_{ik} + e_{i1} \\
 x_{i2} &= a_{21}f_{i1} + a_{22}f_{i2} + \dots + a_{2k}f_{ik} + e_{i2} \\
 &\vdots \\
 x_{ij} &= a_{j1}f_{i1} + a_{j2}f_{i2} + \dots + a_{jk}f_{ik} + e_{ij} \quad (\text{式 1}) \\
 &\vdots \\
 x_{im} &= a_{m1}f_{i1} + a_{m2}f_{i2} + \dots + a_{mk}f_{ik} + e_{im}
 \end{aligned}$$

$i = 1, 2, \dots, m$

式の中の f_{ik} を共通因子 f_k の個体 i の因子得点(factor score)と呼び、共通因子の係数を因子負荷量(factor loading)と呼ぶ。表 2 に因子

負荷量、表 3 にそれに対応する因子得点を示す。

表 2 因子負荷量行列 $A_{n \times k}$

	因子負荷量			
	第 1	第 2	...	第 k
x_{i1}	a_{11}	a_{12}	...	a_{1k}
x_{i2}	a_{21}	a_{22}	...	a_{2k}
\vdots	\vdots	\vdots	\ddots	\vdots
x_{ij}	a_{j1}	a_{j2}	\vdots	a_{jk}
\vdots	\vdots	\vdots	\ddots	\vdots
x_{in}	a_{n1}	a_{n2}	...	a_{nk}

表 3 因子得点行列 $F_{m \times k}$

	因子得点			
	第 1	第 2	...	第 k
個体 1	f_{11}	f_{12}	...	f_{1k}
個体 2	f_{21}	f_{22}	...	f_{2k}
\vdots	\vdots	\vdots	\ddots	\vdots
個体 i	f_{j1}	f_{j2}	\vdots	f_{jk}
\vdots	\vdots	\vdots	\ddots	\vdots
個体 m	f_{n1}	f_{n2}	...	f_{nk}

因子分析では、式 1 から $A_{n \times k}$ 、 $F_{m \times k}$ を求めるのが主な課題である。式 1 の左側は観測・測定によって得られたデータで、右側は共通因子と因子負荷量の線形結合に独自因子を加えている。右側の各要素を求めることは比較的複雑であり、何らかの条件の仮定が必要である。異なる仮定のもとでいくつかのアルゴリズムが提案されている。その中で最も広く用いられているのは主因子法(反復法)と最尤法である。主因子法は安定した結果が得られるが、データが正規分布に従うときには最尤法を用いた方がよいと言われている。しかし、最尤法はセンシティブすぎて初心者にとっては主因子法より扱いにくい。

異なるアルゴリズムで計算された因子負荷量、共通因子は、異なる仮定と条件の下での推測値であるので、返された結果を断片的に比較することはナンセンスである。

因子分析では、分析の便利のため、因子軸を回転して用いる。回転方法は多数提案されてい

るが直交回転と斜交回転に分けられる。直交回転にはバリマックス回転、バイコーティマックス回転、コーティマックス回転、エクイマックス回転などがあり、斜交回転にはプロマックス回転、コバリミン回転、バイコーティミン回転、コーティミン回転などがある。

多く使用されている直交回転方法はバリマックス(varimax)回転で、斜交回転方法はプロマックス(promax)回転である。

因子分析における因子負荷量の推定、因子得点の推定、因子軸の回転は、それぞれ多くの方法が提案され、方法によって得られた結果がそれぞれ微妙に異なることが初心者困惑を与えている。

2. R の因子分析関数

R には最尤法による因子分析の関数 **factanal** が実装されている。

`factanal(x, factors, rotation, scores, ...)`

引数の x はデータセットで、 $factors$ は求める因子の数である。因子の数は解析を行う際に指定しなければならない。因子の数の値が不適切でエラーメッセージが返される際には、メッセージの内容に沿って因子の数を調整し、再実行すればよい。主成分分析で考察を踏まえて因子の数を決めるとエラーメッセージを回避することができる。

主成分分析の解を求めるアルゴリズムは最尤法による因子分析より柔軟性があるので、因子分析を行う前に主成分分析を試みるのが良い。これは主成分分析の結果と因子分析の結果を比較しながら分析を進める上でも意味がある。また、関数 `factanal` のように因子分析を始める初期段階に因子の数を決めなければならない場合

は、主成分の固有値が因子の数を定める1つの情報となる。通常因子分析の因子の数は、相関行列を用いた主成分分析と同じく固有値の値が1以上の数を用いる。

引数 rotation ではバリマックス(varimax)回転とプロマックス(promax)回転を指定することができる。デフォルトには、直交回転“varimax”になっている。

引数 scores では、因子得点を求める方法として、回帰方法(regression)とパートレット法(Bartlett)の中から1つ選択する。デフォルトには“none”になっているので、特別な方法を指定しないと、因子得点は求めない。

関数 factanal が返す主な値を次に示す。

\$loadings 因子負荷量

\$correlation 相関係数

\$factors 求めた因子数

\$STATISTIC カイ 2 乗値

\$dof カイ 2 乗決定の自由度

\$PVAL カイ 2 乗統計量の P 値

このカイ 2 乗検定統計量は、元のデータの分散と指定した共通因子のモデルに基づいて求めたデータの分散との間の有意差に関する検定統計量である。この検定統計量は、探索的に因子分析を行う際の因子の数を定める際の1つの参考材料となる

具体的なデータの例を用いてその使用法と読み方について説明することにする。

進路指導・職業指導に用いる職業適性検査方法として「厚生労働省編一般職業適性検査」がある。この適性検査は、アメリカ合衆国労働省が開発した GATB(General Aptitude Test Battery)を原案とし、昭和 27 年から改良を加えながら職業指導・進路指導のために用いている検査法である。進路指導・職業指導の検査は、

表 4 に示す項目により構成されている紙筆検査と、指先の器用さ・手腕の器用さを検査する器具検査に分かれている。その詳細については次のサイトで公開されている。

http://www.jil.go.jp/institute/seika/Top_page_000.htm

表 4 進路指導・職業指導の検査項目

検査 1	円打点検査 (円の中に点を打つ検査)
検査 2	記号記入検査 (記号を記入する検査)
検査 3	形態照合検査 (形と大きさの同じ図形を探しだす検査)
検査 4	名詞比較検査 (文字・数字の違いを見つけたる検査)
検査 5	図柄照合検査 (同じ図柄を見つけだす検査)
検査 6	平面図判断検査 (置き方をかえた図形を見つけだす検査)
検査 7	計算検査 (加減乗除の計算を行う検査)
検査 8	語意検査 (同意語かまたは反意語を見つけだす検査)
検査 9	立体図判断検査 (展開図で表された立体形を探しだす検査)
検査 10	文章完成検査 (文章を完成する検査)
検査 11	算数応用検査 (応用問題を解く検査)

本稿では、柳井ら(参考文献[1])が用いた表 4 の 11 項目に対する 50 人の被験者の紙筆検査データを借用する。データは次のサイトからダウンロードすることができる。

<http://www1.doshisha.ac.jp/~mjin/data/>

データは xls 形式になっているので、データシートのみを csv 形式に保存する手続きが必要となる。CATB50 として R に読み込み、次のようなコマンドで因子分析の関数を実行することができる。

```
>(CATB50.fa<-factanal(CATB50,factors=4,scores = "Bartlett"))
```

```
Call:
factanal(x = roudousyou, factors = 4, scores = "Bartlett")
```

Uniquenesses:											
	A	B	C	D	E	F	G	H	I	J	K
	0.575	0.560	0.853	0.429	0.467	0.340	0.500	0.354	0.278	0.277	0.400

Loadings:				
	Factor1	Factor2	Factor3	Factor4
A	0.146	0.180	0.303	0.649
B	0.333	0.332	-0.177	0.308
C	0.648	0.312	0.212	0.110
E	0.574	0.288	-0.321	0.137
F	0.280	0.722	0.280	0.280
G			0.882	0.187
H	0.747		0.242	0.144
I	0.211	0.802	0.185	
J	0.788	0.196	0.126	-0.226
K	0.276	0.351	0.816	-0.144

	Factor1	Factor2	Factor3	Factor4
SS loadings	2.247	1.853	1.230	1.038
Proportion Var	0.204	0.150	0.112	0.094
Cumulative Var	0.204	0.355	0.466	0.561

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 11.2 on 17 degrees of freedom.
The p-value is 0.846

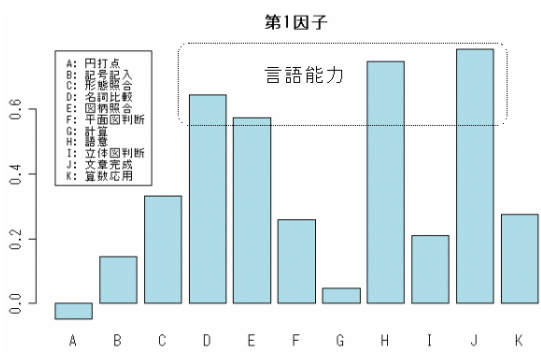
ここでは、因子の数を4にしているが、これは試行錯誤の結果である。ちなみに、因子の数を7以上にすると因子の数が多すぎるというエラーメッセージが返される。

返された結果の最下部には、元のデータの分散と因子分析モデルによる推測データの分散とのカイ2乗検定統計量のp値(0.846)がある。因子の数を2(factores=2)にすると、カイ2乗検定統計量のp値は0.05より小さく、5%有意水準でデータの分散と因子分析モデルによる推測データの分散が等しいという帰無仮説が棄却される。因子を3つにすると帰無仮説は棄却されないが、累積寄与率が50%にもならないので、ここでは4因子を用いている。

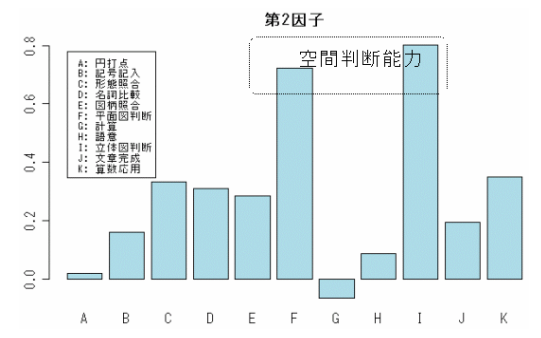
説明の便利のため、第1~4の因子負荷量の棒グラフを図2(a)~(d)に示す。次に図2(a)の作成コマンドのみを次に示す。

```
>barplot(CATB50.fa$loadings)
```

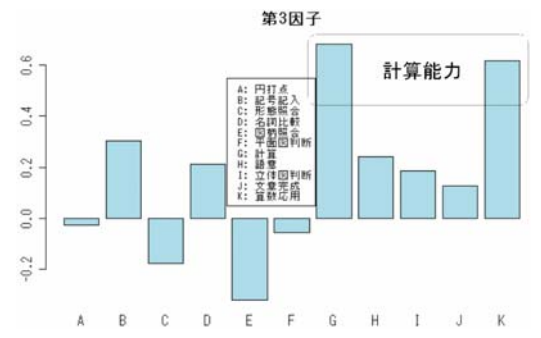
図2 因子負荷量の棒グラフ



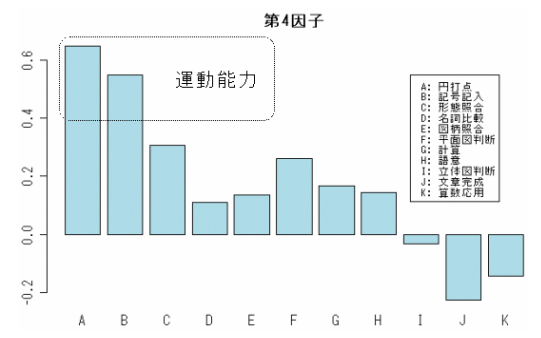
(a)



(b)



(c)



(d)

図2(a)で分かるように、第1因子の中で値が大きいのは「文章完成」、「語意」、「名詞比較」「図柄照合」の順である。前3者は言語能力と関わりのある項目であるので、この因子は「言

語能力」であると考えられる。しかし、「図柄照合」は「言語能力」と解釈できるかに関しては筆者の知識の範囲を超えている。

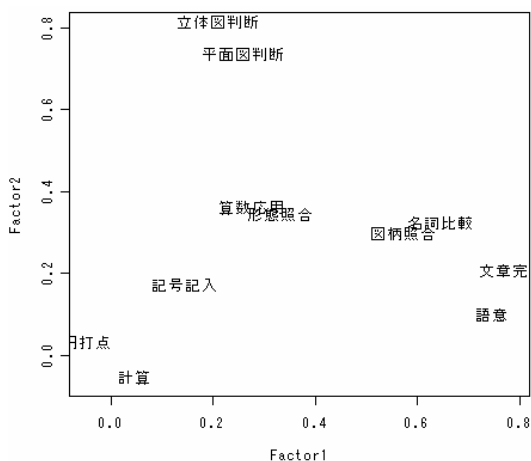
このように、第2因子は空間判断能力、第3因子は計算能力、第4因子は運動能力と解釈することができる。ただし、第4因子の運動能力は、「眼と手または指を共応させて、迅速かつ正確に作業を遂行する能力。眼で見ながら、手の迅速な運動を正しくコントロールする能力」を指す。

因子分析では、複数の因子負荷量を散布図に表して考察を行う。分析の便利のため因子回転を行うが、上記の因子分析のコマンドには回転の指定を行っていないので、返された結果はデフォルトで設定されているバリマックス回転の結果である。

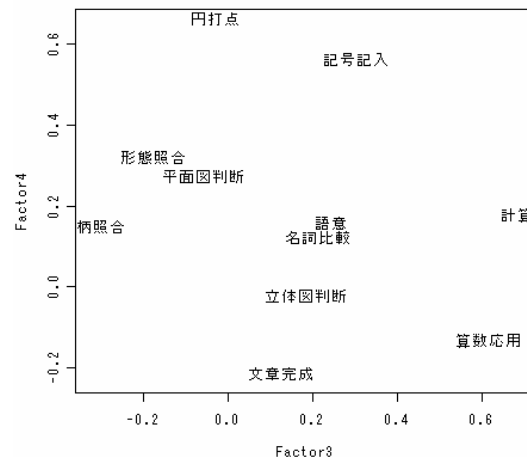
次に第1, 2因子負荷量の散布図のコマンドを示し、その結果を図3(a)に示す。同じの方法により作成した第3, 4因子負荷量の散布図を図3(b)に示す。

```
>plot(CATB50.fa$loadings[,1:2],type="n")
>text(CATB50.fa$loadings[,1:2],colnames(CATB50))
```

図3 因子負荷量の散布図



(a)



(b)

このような散布図では、各因子のみではなく、因子間の相互作用についても考察することができるので便利である。例えば、図2(c)では、「算数応用」は、「計算能力」という因子にまとめているが、図3(b)の散布図では「計算」と「文章完成」との中間に配置されていることから、この両変数に関連があると解釈される。

被験者がどのような能力を持っているかに関しては、因子得点で読み取れる。第1因子得点の値が大きい方は、言語能力が優れていると判断される。念のためにそうであるかどうかを、元のデータと照合して確かめることを勧める。因子得点も因子負荷量のように散布図を作成し、因子負荷量と対応させ、個体の特徴を分析することができる。次に第1, 2因子得点の散布図を作成するコマンドを示し、その結果を図4(a)に示す。第3, 4因子得点の散布図は紙面上の都合により省略する。

```
>plot(CATB50.fa$scores[,1:2],type="n")
>text(CATB50.fa$scores[,1:2],rownames(CATB50))
```

図4 因子得点の散布図

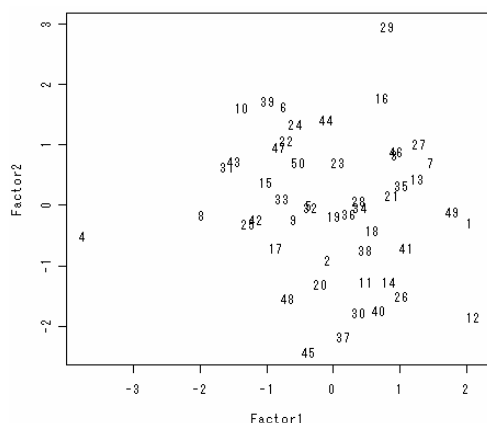


図4では、個体1、49が第1因子の右、かつ第2因子の中間に位置している。第1因子が「言語能力」であるので、この被験者は言語能力に優れていると判断することができる。

4つの因子でデータ CATB50 をモデル化した場合、元のデータが持つ情報のどのぐらいを説明できるかに関しては、累積寄与率と独自因子の推測値から読み取られる。この例における、累積寄与率は0.561(56.1%)である。これは、データが持つ情報の56.1%しか、モデルが説明できていないことを意味する。

返された独自因子(Uniquenesses)の値の平均値に累積寄与率を加えると近似的に1になる。よって、独自因子の推測値の分析で、どの変数に独自因子が強いかに関する分析が可能である。ちなみに、この問題では、変数A、B、Cの独自因子が0.5を超えているので、これらの変数は共通因子で説明できない情報が半分を超えていると言える。

因子分析は、人間の知能・能力・行動・心理・個性のような客観的に精密計測が困難な問題において、何らかの手段で収集した内因と外因に影響されやすいデータの分析に多く用いられている。因子分析は、若干粗いデータの中から妥

当と思われる情報をどう見つけ出すかという側面で発展し続けている。しかし、主成分分析などで、得られない知識が、因子分析によって驚くべき新しい発見ができることはほとんど期待できないであろう。

既に述べたとおり、因子分析は、多くのバリエーションがあり、かつどの方法でも解が決定的、不変的のものではない。例えば、本稿で用いたデータ CATB50 について、因子の数を換えるだけでも、返された結果が異なる。因子の推定方法、因子の回転方法が異なると返された結果が似ても似つかないケースもある。

よって、得られた結果について分析を行うときには、主観的な考えに解析の結果を恣意的に合わせるのではなく、探索的に様々なアルゴリズムによる因子分析を繰り返し、客観的に意味付けし、その解釈が多くの方々の納得と支持を得るものでなければならない。

因子分析の初心者にとっては、多義性が比較的少ない主成分分析や対応分析などの方法を兼用して、因子分析を行うことが賢明であろう。

Rに実装されている因子分析の関数 `factanal` は使い勝手が良くない。例えば、引数 `factors` を指定しないと、プログラムが実行されない。これは初心者にとっては非常に不便である。群馬大学青木繁伸教授は、引数 `factors` の指定を省略すると理論的に最も多い因子数が仮定され、結果を返すプログラム `factanal2` を次のサイトで公開している。

<http://aoki2.si.gunma-u.ac.jp/R/factanal2.html>

また、慶応義塾大学渡辺利夫教授が作成したS言語用の主因子法のプログラム(参考文献[2])が次のサイトで公開されている。このプログラムはRで問題なく作動する。プログラムは、因子負荷量を求める `factor1` と因子得点を求め

る fscore に分かれている。factor1 では、データの相関行列を用いるようになっている。

<http://web.sfc.keio.ac.jp/~watanabe/adstat10.htm#1>

返された結果の因子の回転は R に実装されている関数 **varimax**、**promax** を用いればよい。

[1] 柳井晴夫・高木廣文編著(1993):多変量データ解析ハンドブック、現代数学社

[2] 渡辺利夫(1994):使いながら学ぶS言語、オーム社