主成分分析は因子分析ではない！

○狩野 裕
大阪大学 大学院人間科学研究科

0. はじめに　主成分分析 (Principal Compotent Analysis; PCA) と因子分析 (Factor Analysis; FA) との論争の歴史は長い．例えば，多変量実験心理学会の機関紙である Multivariate Behavioral Research が 1989 年に特集を組んでおり，その後も現在にいたるまで多くの論文が出版されている．

この議論を始める前に FA と PCA の定義を明確にしておく必要があろう．発表者は，何らかの方法で共通性を推定すれば因子分析，共通性を推定しない場合は回転を施したとしても主成分分析と呼ぶことにする．すなわち，S を標本共分散 (or 相関) 行列とするとき，因子負荷行列 Λ および主成分ベクトル (行列)L の推定値は

$$\text{FA:}\quad S \leftarrow \Lambda\Lambda' + \Psi \tag{1}$$

$$\text{PCA:}\quad S \leftarrow LL' \tag{2}$$

によって定められるとする．ここで Ψ は独自性を表す対角行列である．

1. 明確な違い　観測変数ベクトルを $\boldsymbol{X} = [X_1, \cdots, X_p]'$ とかく．PCA における主成分は，

$$PC_i = \ell_{i1}X_1 + \cdots \ell_{ip}X_p$$

であるから観測変数から作られる合成変数であって，観測変数が原因系で主成分が結果系である．一方，因子分析のモデルは

$$X_i = \lambda_{i1}F_1 + \cdots + \lambda_{ik}F_k + u_i$$

であるから，FA の因子は原因系であって観測変数が結果系である．そして，因子の変動が観測変数に伝わることで観測変数間の相関を説明しようとする．このように PCA と FA は因果の方向が逆だという明確な違いが存在する (図 1 参照)．同じ意味だが，観測変数は主成分の構成要素であるが，因子は観測変数の構成要素であると表現することもできる．

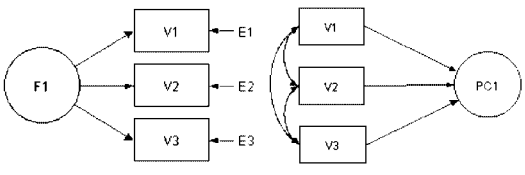

図 1．FA(左) と PCA(右) の違い

豊田 (1992, p.154; 東大出版) はこの点について次のような明解な解説を与えている：

> 因子分析は心理学における知能の研究にしばしば用いられるが，それは知能（構成概念）が高いことが原因でテスト得点（観測変数）が高くなると考える方が，テストの成績が高いことが原因で知能が高くなると考えるより自然だからである．主成分分析は経済学の各種指標にしばしば用いられるが，それはたとえば物価 (観測変数) が高いことが原因で，その結果として物価指数（構成概念）を高く設定すると考える方が，物価指数が高いことが原因で物価が高くなると考えるより自然だからである．

たとえ分析結果が似ていたとしても，原因と結果を逆転させた分析は失格だろう．単回帰分析において標準偏回帰係数が同じだからと言って，原因系変数を y に結果系変数を x に設定するだろうか．

PCA は FA において誤差がないモデルという考えも成り立つが，個体差が顕著で再現性が高くないデータを扱う社会科学においては,「誤差がない」という仮定にどれだけの正当性をもたせることができるだろうか．因子分析で考えれば，誤差 (+特殊性) が 50%を超えることにしばしば遭遇するのである．

主成分によって説明できなかった部分を誤差とする，すなわち，$\widehat{\text{Var}}(\boldsymbol{e}) = S - \hat{L}\hat{L}'$ を誤差と解釈するということも考えられる．しかし，この場合は，誤差間に相関を認めることになってしまう．

2. 類似点　前節で説明したような明確な違いがあるにもかかわらず，なぜ混乱が生じるのだろうか．それは，PCA は分散最大という軸を求める手法であるのだが，この基準が FA のように標本共分散行列を説明するということと同等であることから ((1) と (2) を参照)，ときに推定値が似通るからである (Sato, 1990(JJSS))．さらには，PC の方が推定が簡単で分析途中でトラブルに巻き込まれることがないこと，また，多くの汎用ソフトウェアが FA と PCA を同じモジュー

ルで扱いそしてPCAがデフォルトになっていることもユーザを混乱させる原因になっている．

Sの固有値行列と固有ベクトル行列をそれぞれ$V = [V_1, V_2]$, $D = \mathrm{diag}(D_1, D_2)$とすると，(2)の解は$\hat{L} = V_1 D_1^{1/2}$である．一方，もし，$\Psi = \sigma^2 I_p$であったならば (Anderson, 1963(AMS))，(1)の解は$\hat{\Lambda} = V_1(D_1 - \sigma^2 I_k)^{1/2}$と表現でき，両分析による解は実質的に同等になる．しかし，重要なのは，PCAでは，$\hat{L}$の長さがσ^2の分だけ長く推定されることである．すなわち，PCAでは因子負荷量が本来の値より大きく推定され，一見，分析結果がアピーリングになることがある．それは，誤差を主成分ベクトルに取り込んでいるからであり不当である．ときおり，PCAの方が良い結果がでるという感想を聞くことがあるが[1]，私は，その理由はここにあるのではないかと考えている．

1因子で$\Psi = \sigma^2 I_p$である場合を考えると，主成分ベクトル$\boldsymbol{l}$と因子負荷ベクトル$\boldsymbol{\lambda}$との間に

$$\boldsymbol{l} = \sqrt{1 + \sigma^2/||\boldsymbol{\lambda}||^2} \cdot \boldsymbol{\lambda} \qquad (3)$$

という関係を導くことができる．このことから，$||\boldsymbol{\lambda}||$が大きいときにFAとPCAの結果が近くなることが分かる．そして，一般に，因子負荷量が大きいか，観測変数が多い場合に$||\boldsymbol{\lambda}||$が大きいという仮定が満たされる．実は，この事実は$\Psi = \sigma^2 I_p$が成立していなくても正しく (Bentler-Kano, 1989(MBR))，また，多因子等の場合に拡張されている (e.g., Schneeweiss-Mathes, 1995(JMA); Ogasawara, 2000(Psychometrika))．

このような議論は理論的すぎると感じるかもしれないが，どのようなときにFAとPCAの結果が似るのかを理論的に示すことは非常に重要である．裏を返せば，どのようなときに無視できない違いが生じるのかを暗示してくれているのである．上記の議論は$\Psi = \sigma^2 I_p$の下で実質的にPCAとFAは同一ということであったが，このことから，独自性Ψの対角要素が等質でなければ，PCAをFAとの違いが大きくなることが予想される．それも，観測変数が少ないときに顕著になるということが理解されるのである．

3. 処方箋とまとめ　因子分析モデルのフィッティングには十分な経験を積みそのテクニックに習熟する必要がある．因子分析を適用すべき状況であったとしてもPCAが使われることがあるのは，PCAは分析が簡単でトラブル (e.g., 不適解) に巻き込まれることがないこと，そして，FAとあまり違わない分析結果が得られるという (誤った) 経験則があるからであろう．加えて，FAの方が推定すべきパラメータが多いのでやや不安定ということもある (小笠原, 2002(本セッション原稿))．

因子分析を実行するには，因子数の選定，観測変数の選択，outlierの同定など試行錯誤を避けて通れない．その過程で，不適解や反復の非収束などのトラブルが起こると面倒である．FAを使うべき研究であっても，分析の過程でPCAや反復しない主因子法を用いることは大きな問題ではないと思う．しかし，論文で報告する分析結果としては，最尤法（or 反復主因子法）による因子分析を採用したい．

最尤法による因子分析ではトラブルが生じるのにPCAではうまく分析できた，という話を聞くことがある．詳しく内容を調べてみると，観測変数に二値変数が含まれていたり，歪度や尖度が異常に大きい項目があったり，また，標本サイズが極めて小さかったり，ふたつの観測変数にしか効かない因子があったりと，因子分析を実行するためのいくつかの必要要件が満たされていない場合がある[2]．これらの問題を解決する手段がなくPCAや非反復主因子法でFAを解くという選択をせざるを得ない場合もあろう．しかし，それはあくまでもsecond bestな分析であることを了解しておく必要がある．

一方，因子分析で最尤解が求まるにもかかわらずPCAの解を報告している場合がある．因子負荷を大きく推定して立派な解であることを誇張したいとか，因子分析だとうまく分析できないような質の低いデータではないか，というふうに誤解される可能性がある．自ら論文の価値を低める必要はないのである．

因子分析は因子分析モデルによる分析であるから，いくつかの仮定がおかれている．例えば，線形性や誤差の独立性（局所独立性）である．因子が原因系の変数であったとしても，因子分析モデルの仮定が満たされない場合は使えない．しかし，このような場合であっても，FAを拡張した構造方程式モデリング (SEM) によって分析できる場合がある．安易なPCAの適用は控えるべきであろう．

[1] もちろん，逆の感想を聞くこともある．

[2] すべての場合がこれに当てはまるとは言わない．

狩野 裕（大阪大学 大学院人間科学研究科）

主成分分析は因子分析ではない！

主成分分析 (PCA) と因子分析 (FA) の基本的なコンセプト上の違いを再確認した．すなわち，PCA は主成分が結果変数であるのに対し，FA では因子は原因変数である．因子負荷量が大きいときや因子数に比して観測変数が多い場合は，両分析方法による推定結果が類似することがわかっている．このことを利用すれば，分析途中の試行錯誤の段階で PCA 等を用いることは大きな問題にならないだろう．しかし，論文として報告する最終解は，FA を用いるべき状況では FA の結果を報告すべきである．