PCA, FA, SEM and ICA

大阪大学 大学院基礎工学研究科　狩野　裕

故丘本正先生は多変量解析の分野においても多くの影響ある貢献をなされた．本発表では，同先生の 1960 年代の主成分分析 (PCA) の研究と 1980 年代の因子分析 (FA) の研究にふれ，構造方程式モデリング (SEM) などその後の発展と現在の同分野の研究へのつながりを考慮しつつ，独立成分分析 (ICA) における separability についての研究を紹介した．

1．モデル

$\boldsymbol{x}$, $\boldsymbol{s}$, $\boldsymbol{n}$ をそれぞれ p 次元観測ベクトル，m 次元潜在変数ベクトル，p 次元誤差ベクトルとし，次のモデルを考える．

$$\boldsymbol{x} = A\boldsymbol{s} + \boldsymbol{n}$$

ここで，A は $p \times m$ の因子負荷行列 (混合行列)，$\boldsymbol{s}$ と $\boldsymbol{n}$ は独立で，$V[\boldsymbol{s}] = I_m$, $V[\boldsymbol{n}] = \Psi$ を仮定すると，

$$V[\boldsymbol{x}] = AA^T + \Psi$$

を得る．このモデルは，PCA, FA と ICA，そして，SEM の中の重要なコンポーネントである検証的因子分析モデル (CFA) を含む．たとえば，$\Psi = O$ とすると PCA ないしは ICA のモデルとなる．m は主成分の数，因子数またはソースシグナルの次元と呼ばれている．$\boldsymbol{n}$ については正規性が仮定される．

(noisy)ICA では $\boldsymbol{s}$ の各要素は独立・非正規であることが必須であるが，他のモデルでは $\boldsymbol{s}$ は無相関であればよい．これらのモデルは誤差の有無と回転の方法の違いによって特徴付けられる．表 1 を参照されたい．

表 1: いくつかのモデル

モデル	**構造方程式**	**共分散構造**	**回転の考え方**
PCA	$\boldsymbol{x} = A\boldsymbol{s}$	$\Sigma = AA^T$	$\hat{\boldsymbol{s}}$が主成分になるよう回転
FA	$\boldsymbol{x} = A\boldsymbol{s} + \boldsymbol{n}$	$\Sigma = AA^T + \Psi$	A が sparse になるよう回転
ICA	$\boldsymbol{x} = A\boldsymbol{s}$	$\Sigma = AA^T$	$\hat{\boldsymbol{s}}$ が最も独立になるよう回転
noisy ICA	$\boldsymbol{x} = A\boldsymbol{s} + \boldsymbol{n}$	$\Sigma = AA^T + \Psi$	$\hat{\boldsymbol{s}}$ が最も独立になるよう回転

† noisy ICA は IFA と呼ばれることがある

ICA は PCA や FA の亜流と見られることもあるようだが，たとえば，FA では因子数 m は観測変数の次元 p よりかなり小さくなければならないし Ψ は基本的に対角行列である．一方，noisy ICA では $m = p$ でよく，また，Ψ は飽和モデルでもよい．これらから，両モデルは，構造方程式は同じだがモデルの本質はまったく異なることが理解されよう．

2．Identifiability, Uniqueness, Separability

さて，$\Psi = O$ とした ICA モデルにおける識別性の問題を考える．ICA モデルが identifiable

であるとは，観測変数ベクトル $\boldsymbol{x}$ の情報から混合行列 A が縦ベクトルのスケールと順番を除いて一意的に定まることを言う．unique とは $\boldsymbol{x}$ の情報から $\boldsymbol{s}$ の分布が一意的に決まることを言う．separable とは $\boldsymbol{s}$ が $\boldsymbol{x}$ の関数として表現されることを言う．一般に，

$$\text{separable} \implies \text{unique} \implies \text{identifiable}$$

である．ICA は歴史的には blind source separation の問題の解法として提示されたことを考えれば，信号が完全に復元できることを保証する separability が最も重要であることが理解される．ICA モデルが separable であるための条件は，i) A が full column rank; ii) $\boldsymbol{s}$ は degenerate せず，高々１つの正規変量を含む，である (Comon, 1994; Eriksson and Koivunen, 2003)．一方，誤差のある noisy ICA は separability を満たさないことが分かっており，ICA の分野では大きな問題とされている．この問題は伝統的な FA における因子不確定性 (factor score indeterminacy) と酷似しており，FA ではこの問題に対して観測変数の数を増やすという方策をとるのが一般的である．

ICA の推定においては，ノイズのない ICA では FastICA や JADE など de facto standard が開発されているが，noisy ICA の推定プログラムについては広く用いられているものがなく，たとえば，誤差項を無視視して ICA の解法で解くか，Ψ を既知として FastICA を用いるなどの方法が取られる．そこで，ノイズがある場合に，それを無視した ICA の解法を用いたときに，それが妥当であるためにはどのような条件があればよいかという問題を考える．これは，伝統的な多変量解析においては，FA を用いる場面で PCA を代用する問題に対応し，典型的な処方箋は観測変数の数を増やすことである．

そこで，ノイズのある ICA モデルが成立しているときに，ノイズを無視して ICA の解法を用い，そして separability を求めるにはどのようにすればよいかという実際的な問題を考え，伝統的な多変量解析の議論と同様に，観測変数の数を増やすことが有効な方法であるかを理論的な観点とシミュレーションを用いて検討した．

理論的な条件として以下が導かれた．混合行列等には，それらが次元の関数であることを明記するため添え字 p を付与し，また，$A_p = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]$, $D_p = \mathrm{diag}(||\boldsymbol{a}_1||, \cdots, ||\boldsymbol{a}_m||)$ とおく．

(i) $\Gamma' = (A_p^T A_p)^{-1} A_p^T \Psi_p A_p (A_p^T A_p)^{-1} \to O \quad (p \to \infty)$

(ii) $\lim_{p\to\infty} \min_{1\le i\le m} ||\boldsymbol{a}_i|| = \infty$

(iii) $\sup_p \lambda_{max}(\Psi_p) < \infty$

(iv) $\inf_p \det\left(A_p^T D_p^{-2} A_p\right) > 0$

(i) は不確定性を減少させるための条件であり，(ii)-(iv) は FA を PCA で近似するための条件で Schineeweiss and Mathes (1995) によって与えられたものである．

Eriksson and Koivunen (2003). Identifiability and separability of linear ICA models revisited. ICA2003.

Hyvarinen, Karhunen and Oja (2001). Independent Component Analysis. Wiley.

Williams (1978). A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy. Psychometrika.

Schneeweiss and Mathes (1995) Factor analysis and principal components. JMA.

村山香織・狩野裕 (2006). ノイズのある独立成分分析の一意性問題．統計関連学会連合大会報告．