

スパース三相主成分分析法の研究開発

池本 大樹

1. はじめに

人間科学では、「刺激×反応×実験参加者」のように 3 種類の変数群からなるデータがしばしば得られる。こうしたデータは 3 相 3 元データと呼ばれ、心理統計学では 3 相 3 元データのための様々な分析法が提案されてきた。三相主成分分析法は、変数群から少数の主成分を抽出しデータを要約する分析法として Tucker (1966)によって提案された。三相主成分分析法は 3 相のうち主成分を抽出する相の数に応じて、Tucker1 モデル、Tucker2 モデル、Tucker3 モデルに大別される。本研究では Tucker2 モデルのみを取り扱う。Tucker2 モデルを示すにあたり、3 相 3 元データを行列で次のように表す。はじめに、 $I \times J \times K$ の 3 相 3 元データを $\mathbf{X}$ と定義し、 $I \times J$ の行列を $\mathbf{X}_k (k=1, 2, \dots, K)$ とする。そして、 $\mathbf{X}_k$ を並べた $I \times JK$ のブロック行列を $\mathbf{X}_A = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K]$ とする。Tucker2 モデルでは $\mathbf{X}_A$ を次のように近似する。

$$\mathbf{X}_A \cong \mathbf{A} \mathbf{H}_A (\mathbf{I}_K \otimes \mathbf{B}^T)$$

ここで、 $\mathbf{A}$ 、 $\mathbf{B}$ はそれぞれ $I \times P (P < I)$ 、 $J \times Q (Q < J)$ の主成分得点行列である。 $\mathbf{H}_A = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_K]$ は、 $P \times Q$ 行列 $\mathbf{H}_k (k=1, 2, \dots, K)$ からなる核配列と呼ばれるブロック行列である。核配列の要素は、 $\mathbf{A}$ と $\mathbf{B}$ の主成分間の交互作用の強度を表す。 $\otimes$ はクロネッカー積を表す。Adachi (2011) は図 1 のように、Tucker2 モデルがネット図で表現されることを示した。

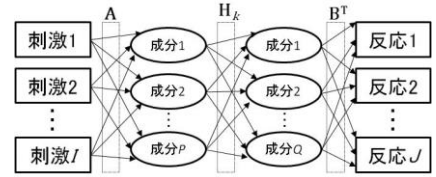


図1: Tucker2モデルのネット図表現

三相主成分分析法は、3 相 3 元データの構造を理解する上で欠かせない分析法といえるが、実際の研究では利用される機会は少ない。この理由として、豊田(2005)は、三相主成分分析法は異なる相から抽出された主成分間の交互作用をモデル化しているため、モデルが複雑になりやすいことを指摘している。複雑なモデルはその意味を解釈することがしばしば困難であるため、実際の研究では利用が避けられるのである。本研究では、そうした問題の解決を目指し、交互作用の個数を分析者が予め定めておく Tucker2 モデルを提案する。交互作用の個数を限定することで、通常よりも簡潔なモデルを得ることができる。

2. 提案手法

提案手法では、次の目的関数

$$f(\mathbf{A}, \mathbf{B}, \mathbf{H}_A) = \|\mathbf{X}_A - \mathbf{A} \mathbf{H}_A (\mathbf{I}_K \otimes \mathbf{B}^T)\|^2$$

を最小化する $\mathbf{A}$ 、 $\mathbf{B}$ 、 $\mathbf{H}_A$ を推定する。ただし、制約条件として、

$$\mathbf{A}^T \mathbf{A} = \mathbf{I}_P, \quad \mathbf{B}^T \mathbf{B} = \mathbf{I}_Q, \quad SP(\mathbf{H}_A) = m$$

を設ける。ここで、 $SP(\cdot)$ は行列に含まれる零要素の数を表す。 $\mathbf{H}_A$ の零要素数 m は、 $0 < m \leq PQK$ の整数から分析者が自由に定めることのできるチューニング・パラメータである。 $\mathbf{H}_A$ の零要素は、それに対応する交互作用の強度が 0 であること、すなわち交互作用が存在しないことを意味する。 m の値を大きくとるほど交互作用の個数が減り、より簡潔なモデルが得られる。このように、いくつかの要素を正確に 0 として推定する分析法はスパース推定法と呼ばれる。提案手法は、Adachi & Trendafilov (2013) が提案した Penalty-free Sparse PCA を 3 相 3 元データへ拡張したものとして位置づけられる。

m の値は分析者の主観だけでなく、赤池情報量規準(AIC)とベイズ情報量規準(BIC)を利用して定める

ことができる. AIC と BIC は値が小さいほど適合が良いことを表す. m の値を変化させて AIC・BIC が最小になるモデルを探すことで, モデル選択を行なうことができる.

3. シミュレーション

Tucker2 モデルに基づいて発生させたデータを用いて, 提案手法の真値の再現度を評価した. 実験条件として, データ行列の二乗ノルムの 20%が誤差である低誤差条件と, 40%が誤差である高誤差条件を設定した. 真値の再現度の評価には, 3 種類の指標を使用した. ひとつ目は真の核配列の零要素のうち, モデルがその位置を誤って推定した個数である. 真の零要素の位置を正しく推定すれば, 下限 0 をとる. ふたつ目は, モデルの零要素数と真の零要素数の差である. 推定されたモデルの零要素数の方が多い場合は正の値, 少ない場合は負の値をとる. 最後は, 真値との差の絶対値の平均(AAD)である. 例えば, $\mathbf{A}=\{a_{ip}\}$ を真の行列, $\mathbf{A}'=\{a'_{ip}\}$ を推定された行列とすると, $AAD(\mathbf{A})=(IP)^{-1} \sum_i \sum_p |a_{ip} - a'_{ip}|$ である. AAD は $\mathbf{A}$, $\mathbf{B}$, $\mathbf{H}_A$ について求めた. AIC によって選ばれたモデルと, BIC によって選ばれたモデルからこれらの指標を算出し, 比較を行なった. その結果, BIC の方が AIC よりも零要素数が多いモデルを選びやすいことが明らかとなった. 低誤差条件では AIC, BIC ともに零要素の個数をおよそ 1 つ誤る程度であったが, 高誤差条件では BIC は AIC よりも零要素の個数を多く誤った. AAD は高誤差条件では低誤差条件よりも大きくなった.

4. 実データ解析

実データへ適用を通して, 分析法の有効性を検証した. 用いたデータは, Osgood & Luria (1954)によって測定された “Multiple Personality” データであった. 彼らは, 解離性同一性障害により三種類の人格 (Jane, Eve Black, Eve White)をもつ女性を実験参加者とした. 実験参加者の各人格は 2 回ずつ, 15 の概念へ 10 の形容語がどれくらい当てはまるかを 7 段階で評定した. したがって, データは 15 概念×10 形容語×6 測定 の 3 相 3 元データである. 前処理として, 評定値の中心化, 及び, 測定ごとの分散の規準化を行なった. 通常の Tucker2 モデル, AIC でモデル選択した提案手法, BIC でモデル選択した提案手法の結果を比較した. 主成分の解釈は検討した 3 モデルで概ね一致していたが, モデル間で負荷の大きな変数が異なる主成分もいくつか抽出された. また, BIC で選ばれたモデルは, 通常の Tucker2 モデルと比べると約 69%の核配列の要素が 0 となった. そのため, 核配列に見られる共通性や相違が把握しやすくなり, 解釈が容易となった. 注目すべき点として, 通常の Tucker2 モデルで大きな強度をもつ交互作用であっても, 提案手法では 0 として推定されることがある点が挙げられる.

5. 結論

三相主成分分析法は, 3 相 3 元データの解釈に重要な役割を持つ分析法である. しかし, 異なる相の主成分間に交互作用を許容するためモデルが複雑になりやすく, 実際の研究では利用が避けられる傾向があった. 本研究では, そうした問題を解決するため, 交互作用の数を予め指定する新たなモデルを提案した. 提案手法は, Tucker2 モデルにおける核配列の推定にスパース推定法をとり入れることで, 上記の目的を達成した. 提案手法には核配列の零要素の数がチューニング・パラメータとして含まれ, その選択には AIC と BIC を利用した. シミュレーションでは, 人工的に発生させたデータを用いて, AIC と BIC によるモデル選択の挙動を検証した. “Multiple Personality” データへの適用結果から, 提案手法は核配列の解釈を容易にすることが確認できた. 今後の課題として, より大規模なデータへの適用を目指し, 局所解の発生頻度を減らす工夫が必要となる(行動統計科学).