個体のクラスタリングを考慮した因子分析の研究開発

三田村　倭

多変量データ解析は，個体（行）×変数（列）の多変量データ行列を分析対象とし，目的によって様々な手法が選択される．中でも［1］列の変数をより少数の因子（成分）に縮約する因子分析や主成分分析），［2］行の個体を少数の群に分類するクラスター分析はポピュラーな手法である．［1］と［2］を同時に達成する，つまり，次元縮約とクラスタリングを同時に行う手法は多く存在し，代表的なものとしてReduced K-means 法（De Soete & Carroll, 1994）が挙げられる．しかし，Reduced K-means 法（RKM）は［1］として主成分分析を選択しており，因子分析によって推定できる独自分散を推定できないという問題点がある．また，変数の因子（成分）を抽出した後にそれらをクラスタリングするという段階的な手法（タンデム分析法と呼ばれる）も存在するが，次元縮約によって得られた部分空間が必ずしも個体のクラスター構造に適切に寄与するとは限らないという非最適性が指摘されている．そこで本研究では，因子分析とクラスター分析を融合した手法を提案する．行列因子分析の目的関数に個体の因子得点をクラスター化するペナルティ関数を組み込むことで，負荷行列の推定に加え，独自分散の推定，そして個体のクラスタリングを同時に達成することが期待される．

提案手法の有用性を検証するために，2 つの実データを用いて RKM とタンデム分析法（因子分析＋K-means 法）との比較実験を行った．用いたデータは，各個体の所属する群が事前に判明しているラベル付きデータであり，分析後の分類結果と真の所属群を比較することで各手法のクラスタリング精度を確かめることができる．数値実験の結果，提案手法が RKM とタンデム分析法よりも正確な個体の分類を行ったことがわかった．また，2 つの実データのどちらにおいても，提案手法が RKM では推定できていない独自分散を推定できていることから，変数を解釈するうえで提案手法が優れていることがわかる．さらに，タンデム分析法の非最適性と結果のクラスタリング精度から，提案手法がタンデム分析法より優れていることも例証された．提案手法の正分類率が最も高かった理由として，次のことが考えられる．提案手法のベースとなっている因子分析では，各変数の変動は共通因子と対応する独自因子の変動に分解される．そして提案手法では共通因子の寄与を表す行列と，個体の分類を表す行列の差が最小になるように最適化されるため，共通因子の寄与が小さい変数は個体の分類にあまり関わらないことがわかる．つまり，独自分散が大きい変数は共通因子の寄与が小さく，個体の分類に寄与しないと考えらえる．このことは，各独自分散と対応する群間分離度（値が大きいほど群同士が分離していることを表す）が負の相関関係になっているという結果からも裏付けられ，独自因子の存在が個体の良い分類に寄与していることが示唆されている．変数間で異なる独自分散を推定しうる因子分析の効用が示されたと言えよう．

研究過程においていくつかの問題が明らかとなった．提案手法はそのモデルとアルゴリズムの性質上，共通因子と独自因子が無相関であるという制約を無視してる．それにもかかわらず，既存手法と似た負荷行列や独自分散を推定できたことは興味深い結果であるが，因子分析を用いた手法としては不完全である．本研究で用いた実データ以外のデータに対しても適切なパラメータ推定が可能かどうかの検証，そして，前述の制約を考慮したモデル・アルゴリズムの開発が必要である．さらに，提案手法の最小二乗基準に用いているチューニングパラメータの値の設定によって，各パラメータがどのように変化するかという，チューニングパラメータの最適化に関しても研究の余地が残されている．（行動統計科学）