尺度混在データに対する次元縮約とクラスタリング

岡山理科大学　森　裕一・吉岡嵩紹・片山浩子・黒田正博

１．はじめに

次元縮約とクラスタリングを同時に行う Reduced k-means 法（RKM）を尺度混在データに適用することを考える。RKM は基本的に量的データに対する手法であるため，質的データを数量化すればよい。本報告では，RKM の次元縮約部分を非計量主成分分析で置き換えた手法の提案を行う。

２．方　法

$\mathbf{X}$ を n 個体 p 変数の中心化されたデータ行列，k をクラスター数，r（一般に，$k \geq r+1$）を次元数とする。通常の RKM は，$\mathbf{U}$ を $n \times k$ のメンバーシップ行列，$\mathbf{A}$ を $p \times r$ の負荷行列，$\mathbf{Z} = \mathbf{XA}$ を $n \times r$ の主成分得点行列，$\mathbf{F}$ を $k \times r$ のクラスター中心行列とすると，最小化する目的関数は，

$$f_{\mathrm{RKM}}(\mathbf{U}, \mathbf{F}, \mathbf{A}) = \left\| \mathbf{X} - \mathbf{UFA}^{\mathrm{T}} \right\|^2 \tag{1}$$

であり，この $\mathbf{U}$ と $\mathbf{F}$ を交互最小二乗法により求めるものである。いわゆる主成分分析（PCA）と k平均法を同時に行うものである（De Soete and Carroll, 1994）。

これに，質的データを最適尺度化して主成分分析を行う非計量主成分分析（MLPCA，Nonlinear PCA）を導入する。NLPCA は，$\mathbf{X}$ が質的変数で構成されているとき，最適に尺度変換された $\mathbf{X}^*$，すなわち，j 番目の変数を $\mathbf{x}_j^* = \mathbf{G}_j \mathbf{y}_j$ と表したときの $c_j \times r$ のカテゴリースコア $\mathbf{y}_j$ を主成分分析の文脈で求めるもので（$\mathbf{G}_j$ は $n \times c_j$ の指標行列，c_j は j 番目の変数のカテゴリー数），

$$\sigma(\mathbf{Z}, \mathbf{A}, \mathbf{X}^*) = \left\| \mathbf{X}^* - \mathbf{ZA}^{\mathrm{T}} \right\|^2 \tag{2}$$

を目的関数として，$\mathbf{A}$ と $\mathbf{X}^*$（すなわち，$\mathbf{G}_j$ と $\mathbf{y}_j$）を最小交互二乗法により求めるものである。このアルゴリズムには，PRINCIPALS（Young et al., 1978）や PRINCALS がある。数量化は変数ごとに行うので，$\mathbf{X}$ が尺度混在の場合，量的変数である $\mathbf{x}_j$ については数量化ステップは必要なく，また，$\mathbf{x}_j$ が順序尺度の場合，$\mathbf{y}_j$ を順序制約を満たすように推定する。

以上より，尺度混在データに対する RKM の目的関数は，

$$f_{\mathrm{RKM/NLPCA}}(\mathbf{U}, \mathbf{F}, \mathbf{A}) = \left\| \mathbf{X}^* - \mathbf{UFA}^{\mathrm{T}} \right\|^2 \tag{3}$$

となる。具体的な手順は，次のようになる。

[Step 1] 初期化: クラスター数 k，主成分数 r を決め，初期値として $\mathbf{U}$ に乱数を与える。

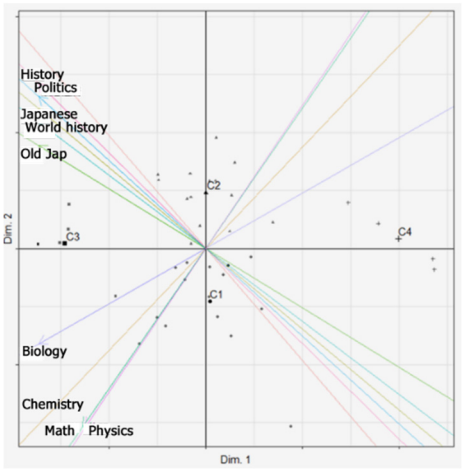
[Step 2] 数量化: PRICIPALS により $\mathbf{X}^{(t+1)}$ を数量化し，数量化行列 $\mathbf{X}^{*(t)}$ を求める。

[Step 3] クラスタリング: k平均法により，式(3)を最小にする $\mathbf{U}^{(t)}, \mathbf{F}^{(t)}, \mathbf{A}^{(t)}$を求める。

[Step 4] 終了判定: t番目と$t+1$番目の式(3)の値に差がなければ終了, さもなくば, $t=t+1$として Step2 へ戻る。

３．数値例

9 科目の成績を 5 段階評価したデータに本手法を適用したときの主成分空間におけるクラスター中心，および主成分得点，負荷量が右図である。元の素点データの RKM を基準としたとき，5 段階評価データを量的データとして RKM を実行した結果より付置の再現性などが高くなる。



参考文献

De Soete G, Carroll JD (1994). K-means Clustering in a Low-Dimensional Euclidean Space. In E Diday, Y Lechevallier, M Schader, P Bertrand, B Burts chy (eds.), New Approaches in Classification and Data Analysis, 212–219.

Young, F.W., Takane, Y., de Leeuw, J. (1978). Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. Psychometrika, 43, 279-281.