

多次元尺度構成法を用いた大量データの表示

宿久 洋 (鹿児島大学 理学部), 橋口 博樹, 岡 隆一 (新情報処理開発機構)

1 はじめに

通常の多次元尺度構成法 (MDS) では, 対象をユークリッド空間上の点で表し, 対象間の関係を空間上の位置関係によって表現しようとするものである. ところが対象の数が多くなってくると結果の解釈が難しくなり, データマイニング等で扱う規模のデータになれば實際上, 適用困難となる.

そこで, 複数のデータを 1 つの塊として, ユークリッド空間上に表すことを考える. 最も単純な方法は, 元データを何らかの規準でクラスター分けし, それらの平均等の代表値をもってそのクラスターのデータすることである. このことによりデータの個数は対象の数からクラスターの個数へと減少し, クラスター数を適当に選択することにより, MDS の現実的な適用が可能となる. しかしながら, 代表値のみでそのクラスターを表現することは情報の損失をとまなうことは言うまでもなく, なるべく多くの情報を表現するためには何らかの工夫が必要である.

そこで, 従来の MDS のように対象 (クラスター) を点のみで表すのではなく, ユークリッド空間内の領域で表現することを考える. 本報告では, クラスター分析, Procrustes 変換と MDS を併用する方法を提案する.

2 クラスター分析, Procrustes 変換と MDS の併用

N 個のデータが k-means 法等の何らかのクラスタリング法により n 個のクラスター C_i ($i = 1, \dots, n$) に分類されているとする. もちろん, 外的な情報により分類されていてもかまわない. この時, クラスター C_i の代表点の座標 y_r ($r = 1, \dots, n$) が適当に定められているものとする (通常は, 各クラスターに属する対象の座標の平均であろう). クラスター C_i と C_j の間の非類似性 o_{ij} が対応する代表点間の非類似性に基づき求められていると仮定する. これらの代表点を新たに付置を求めるべき対象と考えると付置の個数は $n \ll N$ となり, 適当な MDS を用いて求めることができる.

しかしながら, このままでは表示されるのはあくまでも代表点のみであり, 各クラスター内での対象の関係などは無視されてしまう. そこで, 代表点を求められた付置座標に対応させる Procrustes 変換 (Hurley and Cattell, 1962) を利用し, すべてのデータに基づく超楕円体を 1 つのクラスターに対して 1 つ描画することを考える. ここでは, 通常の Procrustes 変換による方法と最小射影誤差に基づく Procrustes 変換による方法の 2 つの方法について述べる.

2.1 通常の Procrustes 変換による方法

Procrustes 変換とは, 2 つの座標に対して, 拡大縮小, 回転, 平行移動を許して, 一方を他方にできるだけ近づける変換である. この言葉は Hurley and Cattell (1962) により初めて用いられたが,

この問題自体は他にも Green (1952), Schönemann (1966) などによっても解決されている。

代表点に対応する q 次元空間における n 個の座標を y_r ($r = 1, \dots, n$), MDS により求められた $p(\leq q)$ 次元空間における付置座標を x_r ($r = 1, \dots, n$) で表す。次元数の違いを無くすために、各 x_r ベクトルの最後に $q - p$ 個の 0 を加えた q 次元ベクトルを新たに x_r ($r = 1, \dots, n$) と考える。

このとき、これら 2 つの座標間の 2 乗距離の和は

$$R^2 = \sum_{r=1}^n (x_r - y_r)^T (x_r - y_r) \quad (1)$$

で表され、これを最小化する変換を y_r に施すこと考える。

y_r に回転, 拡大縮小, 平行移動を施した座標 y'_r を

$$y'_r = \rho A^T y_r + b \quad (2)$$

で表す。ここで、 A は回転を表す直交行列, ρ は拡大縮小を表す実数, b は平行移動を表すベクトルである。このとき、最小化すべき 2 乗距離の和は

$$R^2 = \sum_{r=1}^n (x_r - \rho A^T y_r - b)^T (x_r - \rho A^T y_r - b) \quad (3)$$

で表される。

R^2 を最小化する A, ρ, b は、 $X^T Y$ の特異値分解を

$$X^T Y = U \Lambda V^T \quad (4)$$

ただし、 U と V は直交行列, Λ は対角行列としたとき、それぞれ以下のように求められる (Cox and Cox, 2001 および Gower and Hand, 1996 参照)。

$$A = V U^T \quad (5)$$

$$\rho = \text{tr}(A X^T Y) / \text{tr}(Y^T Y) \quad (6)$$

$$b = x_0 - \rho A^T y_0 \quad (7)$$

ここで、 $x_0 = \sum_{r=1}^n x_r / n$, $y_0 = \sum_{r=1}^n y_r / n$ である。

もし、 $X^T Y$ が非特異であれば A, ρ は更に以下のように書くことができる。

$$A = (Y^T X X^T Y)^{1/2} (X^T Y)^{-1} \quad (8)$$

$$\rho = \text{tr}(Y^T X X^T Y)^{1/2} / \text{tr}(Y^T Y) \quad (9)$$

$$(10)$$

ここで注意しておきたいのは、通常の Procrustes 変換では $p \geq q$ が仮定されていることである。そのため、上では 0 を加えて次元を揃える操作を行っている。

このようにして求めた A, ρ, b を用いて N 個の生データの座標すべてを Procrustes 変換した座標を z_{ri} ($r = 1, \dots, n; i = 1, \dots, N_r$) (ただし、 $\sum_{r=1}^n N_r = N$) とする。

このとき、

$$(x_r - z)^T \text{Cov}_r^{-1} (x_r - z) \leq c \quad (11)$$

を満たす z の領域を対象が付置されている空間に描画する。ここで、 Cov_r^{-1} は C_r に属する N_r 個の変換された座標 z_{ri} から計算した分散共分散行列の逆行列を表し、 c は適当な定数を表している。

これは、各クラスターの代表点 x_r が付置されている空間に各クラスターに属するすべての対象の座標を変換したものから導かれた超楕円体を描画していることになる。このことにより、代表点の持つ情報のみではなく、クラスター内の対象の関係の情報も同一の空間に表現できる。

2.2 最小射影誤差に基づく Procrustes による方法

Gower and Hand (1996) は、 $p < q$ の場合に、回転行列ではなく射影行列を用いた変換で 2 組の対象の座標のマッチングさせることを提案している。ここでは、彼らの方法に基づき最適変換を求め、超楕円体を描画する方法を述べる。

2.1 と同様に、代表点に対応する q 次元空間における n 個の座標 y_r ($r = 1, \dots, n$) と MDS により求められた $p(< q)$ 次元空間における付置座標を x_r ($r = 1, \dots, n$) が与えられているとする。ここでは、次元を揃えていないことを注意しておく。

このとき、

$$r^2 = \sum_{r=1}^n (x_r - A^T y_r)^T (x_r - A^T y_r) \quad (12)$$

を最小化する $q \times p$ の射影行列 A を以下のステップで求める。

1. 各 x_r ベクトルの最後に $q - p$ 個の 0 を加える。
2. 通常の Procrustes 変換と同様に $R^2 = \sum_{r=1}^n (x_r - A^{*T} y_r)^T (x_r - A^{*T} y_r)$ を最小化する $q \times q$ の回転行列 A^* を求め、 $A^{*T} y_r$ を新たに各 y_r とする。
3. x_r ベクトルの最後の $q - p$ 個の成分を $A^{*T} y_r$ の対応する成分と入れ替えたものを新たに x_r とし、 R^2 を計算する。
4. R^2 が収束していれば、ステップ 5 へ、そうでなければ、この段階の x_r, y_r で再度ステップ 2 から繰り返す。
5. A^* の最初の p 列を射影行列 A とする。

このようにして求めた A を用いて前節と同じ操作を行い、 z_{ri} ($r = 1, \dots, n; i = 1, \dots, N_r$) (ただし、 $\sum_{r=1}^n N_r = N$) を求め、超楕円体を描画する。

2.3 超楕円体の解釈

与えられたデータを通常の MDS によって解析した場合、結果はデータ数と同数の点の座標として与えられる。結果の解釈はこれらの点間の関係に基づき行われる。今回、我々が提案した方法では、データから決定されたクラスターが点として付置され、その周りに上述の方法によって求められた超楕円体が描画される。この楕円体は各クラスターに属する対象間の付置次元における関係（散らばり）を表していると考えられる。クラスターを表す各点間の関係のみではなく楕円体間の距離関係を合わせて解釈することにより、元データの持つ情報をより捉えた解釈が可能になると考えている。

参考文献

- [1] Borg, I and Groenen, P (1997). *Modern Multidimensional Scaling*, Springer-Verlag, New York.
- [2] Cox, T. F. and Cox, M. A. (2001). *Multidimensional Scaling, Second Edition*, Chapman & Hall, London.
- [3] Commandeur, J. J. F. (1991). *Matching configurations*, DSWO Press, Leiden.
- [4] Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis, *Psychometrika*, **17**, 429–440.
- [5] Gower, J. C. and Hand, D. J. (1996). *Biplots*, Chapman & Hall, London.
- [6] Hurley, J. R. and Cattell, R.B. (1962). The Procrustes program: producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, **7**, 258–262.
- [7] Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem, *Psychometrika*, **31**, 1–10.