2Q-8

計量MDSのストレスによる評価とグラフ表現の高度化

渡辺 洋平　三浦 孝夫
法政大学工学部電気電子工学科

1　前書き

多次元尺度構成 (MDS) は，一対の高次元データの非類似度を収集し，この間の距離を直接に低次元空間で表現する技法である。MDS の主たるねらいは，機械学習や統計・推定とは異なって，データオブジェクト間の関連性の把握にある。しかも大づかみな傾向を捕らえることが狙いであり，多くの場合，視覚的な表現を伴う [2]。各次元は解釈可能でなければならないため，結果の評価は”心理的”であり客観的な評価は難しい。反面，実用的な観点から広範囲な応用を有する [1]。

計量 MDS(Metric MDS) は，データ間の距離が数値として得られるときにその比を保ったままで表現する手法である。順位尺度データを扱う場合には非計量 MDS 手法があり，Kruskal 手法などが知られる。非計量データは，(非) 類似性を相関性などの統計値で与え，これをもとに類似度を推定するため，MDS による結果の”よさ”を客観的に評価するストレス (Stress) 関数が必要となる。

本研究では，計量 MDS (Torgerson 法) を用い，次元選択とカラー値 (RGB) による 6 次元までの低次元化を行う。評価関数 (ここでは単純にストレスを利用) と次元縮小を利用者が判断できるシステムを試作する。次元縮小によって最大 6 次元までの可視化を行う。次元は明示的に利用者が与えてもよいが，自動的に次元縮小を行える。これにより，縮小次元，ストレス，表示次元の選択を入力として可視化状況を制御することができる。

2　多次元尺度構成とストレス関数

非計量 MDS ではデータ間 i,j の類似度は δ_{ij} と表現されこれが次元縮小されて距離 d_{ij} をもつ 2 点に配置される。$F(\delta_{ij}) = d_{ij}$ を与える変換 F を多次元尺度構成 MDS と定式化する。しかし実際には d_{ij} ではなくて，誤差 e_{ij} を伴う。現実には，この誤差を小さくする変換 f を考えることになる: $f(\delta_{ij}) = d'_{ij} = F(\delta_{ij}) + e_{ij}$

すべての誤差の二乗和はデータと，空間上の距離の適合度を表すと考えられるのでストレス評価関数 S を次のように定義する：$S = \sqrt{\frac{\Sigma_{i<j}(d_{ij}-d'_{ij})^2}{\Sigma_{i<j}d_{ij}^2}}$

計量 MDS では，変換 f の計算は単純な線形代数計算に基づくものがある。このうち Torgerson 法は，すべてのデータオブジェクト間に距離 δ が与えられている (したがって対称関係) と仮定する [1]。n 次実対称正方行列 $A = (a_{ij})$ を次式により求める：

$$a_{ij} = \frac{1}{2}(\frac{1}{n}\Sigma_k\delta_{ik}^2 + \frac{1}{n}\Sigma_k\delta_{kj}^2 - \frac{1}{n^2}\Sigma_k\Sigma_h\delta_{kh}^2 - \delta_{ij}^2)$$

この要素は実は全オブジェクトの重心から i,j に至るベクトルの内積を意味し，固有値・固有ベクトルから因子分解して対応する点の座標を求めることができる。すなわち，A は直交行列 X を用いて固有値からなる対角行列 Λ に変換できる：$X^t \cdot A \cdot X = \Lambda$
このとき固有値は大きいものから並べ，直交行列 $X = (\mathbf{x_1}...\mathbf{x_n})$ は固有ベクトル列となる。すべての固有値が非負であるとき，Λ の各要素 λ_k を $\sqrt{\lambda_k}$ に置き換えた対角行列を Λ_2 と表せば，行列 $P = X \cdot \Lambda_2$ の第 i 行はデータオブジェクト i の座標を実ユークリッド空間で表現できる。

この因子分解の結果に対して縮小次元の決定が必要である。大きい固有値のみを取り出し (この個数が縮小する次元に対応する)，残りを無視することにより，次元縮小が可能となる。作為的に無視する固有値があるならば，縮小する次元に応じて”無視される”部分が増加し，類似性行列の近似的な復元が困難となる。この誤差は上述のストレス関数を用いて計算できる。

ストレスを計算するとき，直交行列 (固有ベクトル列) X で規定される座標系を用いる。本試作システムでは 通常の 3 次元表現に加え， RBG カラーを用いて最高 6 次元オブジェクトの可視化を行うが，視点の向きを変更する機能も有する。

本試作システムでは、固有値が正である限り表示次元 (1–6) を動的に選択しストレス最小を選ぶ機能を有する。あるいは、利用者の指定した閾値を下回るものがあればその時点で停止する。そのようなものがなけ

”Stress Function on Metric MDS for Advanced Visualization” : Yohei Watanabe, Takao Miura : Hosei University, Dept.of Elec. and Elec. Eng.
Kajino-cho 3-7-2, Koganei, Tokyo, JAPAN

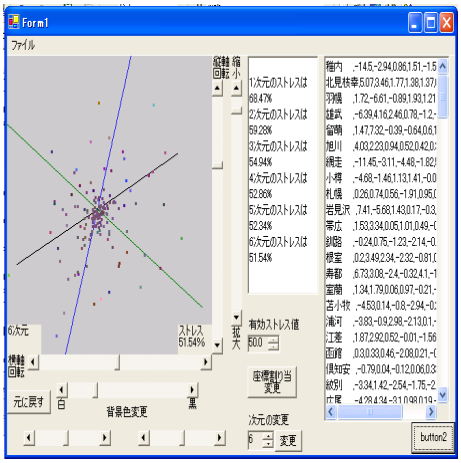
れば，エラーとする。

3　試作システム

本章では，前述の機能を実行する試作システムを構築し，その有効性を確認する。入力データはファイル形式で与えられ，Torgerson 法による計量 MDS 計算を行った結果を (最高 6 次元に) 可視化する。表示空間上で 3 つの直線が X，Y，Z 軸を表しデータオブジェクトが配置され，これらを射影変換した結果が表示する。表示次元は自動的に選定され，正の固有値により決定される。即ち，正の固有値が 2 つならば，2 次元の座標，6 つならば 3 次元座標と RGB カラー値に変換される。これを超える場合は，固有値の大きいほうから 6 個を選び，同様の処理を実行する。

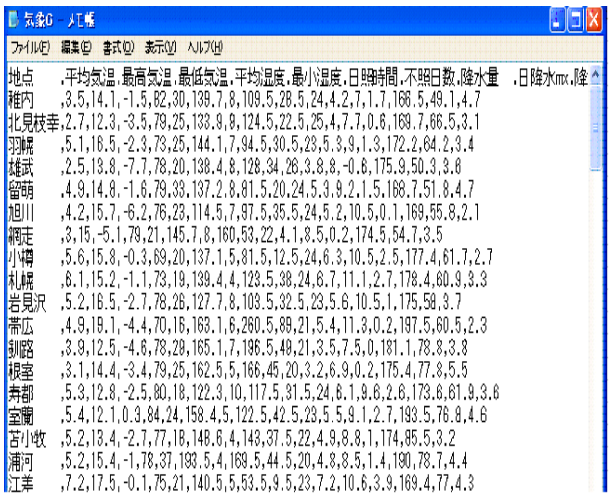
しかし，自動的に選定された次元値は，利用者が任意に変更できる。本試作システムでは，正の固有値が 3 個以下はそのまま，4 つ以上では大きいものから 3 つをそれぞれ X,Y,Z 座標に対応させ，それ以降を RGB 値に対応させる。各 RGB 値 は，0-256 に比例配分しあてはめる。

ストレス基準 (この評価値を上限とする) を設定することにより，計算結果がこれを満たさないときには，次善の方法をとる。例えば 4 次元でのストレス評価値が基準を満たさない場合，5 次元に設定しなおして再計算する。6 次元を超えた場合は，表示不能と判断する。次の例は 6 次元までのストレス計算をすべて実行し 51.54% の最小値になるまで繰り返している。

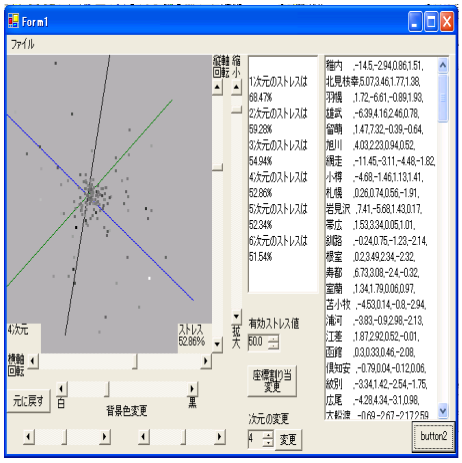


4　実験

以下では，試作システムを用いて 16 属性からなる全国の 6 月の気象データ 150 件を用いる。実験データ，得られたグラフを以下に示す。縮小次元は自動的に選定したものであり，ストレス基準 (この評価値を上限とする) を設定していないため，6 個以上の正の固有値により 6 次元で表示する。先の例はこれを表示したものである。



各データオブジェクトは観測点を示し，RGB の順で近接する。もとの数値と比べると，右側が北の地方に，左側が南に対応する。実際に計算した結果，4 次元空間を指定し表示させた例である。この結果では上で 52.86% のストレス値を得る。



5　結論

本稿では次元数を与え，これを満す解があるとして可視化・表示した。表示方法で，4 次元以上ではどのデータを選ぶかで印象が異なる。表示軸の設定を変更することで，(次元数やストレス値は同じであっても) 可視化結果の印象が変化する。しかし，”印象の計算”は，新たな評価の観点を必要とする。

参考文献

[1] 林知己夫，飽戸弘: 多次元尺度解析法，サイエンス者, 1976

[2] 細田博史:　高次元の利用による対話的グラフ配置法，WISS, 2003

[3] Young, F.W. and Hamer, R.M.: Multidimensional Scaling: History, Theory and Applications, Erlbaum, 1986