MDS 法における最適次元の推定

伊藤　里江　(指導教官 : 吉田 裕亮)

1　はじめに

対象間の類似度のデータが与えられたとき, 類似したものどうしを近くに, そうでないものどうしを遠くに布置する手法に多次元尺度法 (以下 MDS 法) がある. 本研究では距離行列が与えられた際, それら対象を布置する MDS 法における最適次元の推定のためのひとつの手法を提案する.

データ点の布置結果の当てはまりの良さを測る尺度としてストレス値が既にあるが, これだけに着目すると, 最適次元よりも高次元が選ばれてしまう場合がある. 本研究では, この問題を回避するために情報量基準 (AIC) を援用した手法を提案し, その有効性を考察する.

2　多次元尺度法 (MDS)

MDS 法は複数の対象間の非類似度すなわち距離が対象間の Euclid 距離として推定されている場合, 対象を Euclid 空間の点として位置付ける方法である.

n 個の対象 $O_1, O_2, \ldots, O_n$ のうち. いま仮に O_n を原点にとり, 残りの $n-1$ 個の対象を終点とする列ベクトルを $x_1, x_2, \ldots, x_{n-1}$ とし, 対象 O_i の第 j 軸の座標値を x_{ij} とする. 行列 X は $n-1$ 個の対象の埋め込まれる空間の次元数分の座標値を対象ごと各行に並べたものである.

$n-1$ 個の対象間の内積を要素とする行列 B を,

$$B = \left(b_{ij}\right) = XX^t$$

とおくと, ベクトル $v_{ij} = x_j - x_i$ の長さ, つまり対象 O_i と O_j 間の Euclid 距離 d_{ij} は

$$d_{ij}^2 = ||v_{ij}||^2 = v_{ij}^t v_{ij} = ||x_i||^2 + ||x_j||^2 - 2b_{ij}$$

とかける. また, 対象間距離の 2 乗 d_{ij}^2 を (i,j) 要素とする行列を $D^{(2)}$ で表し, 中心化行列を

$$J_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$$

と定義する. ただし, $\mathbf{I}_n$ は n 次の単位行列で,$\mathbf{1}_n$ はすべての成分が 1 の列ベクトルである. この J_n を用いて $D^{(2)}$ に Young - Householder 変換 (両側から中心化行列をかける演算) を施すことにより, 重心を原点とする内積行列 B_c

$$B_c = -\frac{1}{2}J_n D^{(2)} J_n$$

が得られる. この B_c を

$$B_c = X_c X_c^t$$

と固有値分解することで布置座標を求める.

3　固有値分解

Young - Householder 変換により得られた内積行列 B_c の固有値分解 (スペクトル分解) を行う数値計算法として, 本研究では, 絶対値の大きな固有値とそれに付随する固有ベクトルを逐次求めていく累乗 (パワー) 法を用いた. その方法は以下の通り.

1. 内積行列 $B_1 = B_c$ の最大固有値 λ_1 とそれに付随する固有ベクトル μ_1 (ただし, $||\mu_1|| = 1$) を求める. これは適当な単位ベクトルからはじめて B_1 を施して正規化し, 再び B_1 を施すという操作を予め定めた収束基準を満たすまで繰り返すことにより求められる.
2. $P_1 = \mu_1\mu_1^t$ とおく. この P_1 は固有値 λ_1 に対応するスペクトル射影になるので $B_2 = B_1 - \lambda_1 P_1$ とする.
3. B_2 の最大固有値 λ_2 とそれに付随する固有ベクトル μ_2 を 1 と同様に求め, $P_2 = \mu_2\mu_2^t$ とおく.
4. 必要な回数 1〜3 の操作を繰り返すことにより, 大きい順に固有値とそれに付随する固有ベクトルを求めることになる.

4　最適次元の選択指標

次元選択の指標としてストレスという値が知られている. ストレスとは

$$S = \sqrt{\frac{\sum\sum_{j<k}(d_{jk} - \widehat{d_{jk}})^2}{\sum\sum_{j<k} d_{jk}^2}}$$

で定義され, この量 S を最小にするような布置により, 当てはまりの良さを判断するものである. これは, ある種の最小 2 乗による最適化である.

そこで, 本研究では, $\widehat{d_{jk}}$ を最小 2 乗推定された対象間の距離で置き換え, 最適な可視化次元の選択問題を回帰モデルでのパラメータ選択および係数推定の問題として捉えることにした.

最適なパラメータ数を選択する基準としてモデル選択基準である情報量基準 AIC を援用する. AIC は値が最小となるモデルが良いモデルである, と判断される.

本研究では, 回帰モデルにおける AIC の式より, 以下の量を導入することにする.

$$C = n \log S^2 + 2\big(\ell\, n + 1\big)$$

ここで, モデルの自由パラメータ数は, 可視化次元を ℓ とした場合, n 次元の固有ベクトルと固有値の組が ℓ 個とする. この C が小さいものを選択するようなものが最適な布置される空間の次元と考える.

5　有効性の検証実験

5.1　シミュレーションデータの場合

5.1.1　実験の概要

前項で導入した, 判定基準の有効性を見るために, 第 1, 2 軸平面上に相関があり, 第 3 軸方向に正規乱数によるノイズを入れたデータを用意し, 今回提案する手法により 2 次元であると判断可能かを調べた.

5.1.2　実験結果

結果は, ストレス値では 2 次元の場合, $S = 0.7113$, 3 次元になると $S = 0.7071$ となり, 3 次元を示しているが, 本研究で提案する手法の C 値においては, 2 次元では $C = 45.154$, 3 次元では $C = 70.989$ となり, 2 次元を選ぶべきであると判断された.

5.2　実データの場合

5.2.1　近鉄路線

MDS 法の実データへの応用例として, 鉄道運賃表から地点の相対的位置関係の復元を試みた.

近畿日本鉄道の駅から 20 駅を選び, 駅間の料金表を元に MDS 法で復元してみると 図 1 のような結果を得た. いくつかの相対位置がずれた駅は距離と比例せずに料金が過度に加算されているものと考えられる. 奈良以西の 12 駅間で再び MDS を適用させて復元したものが 図 2 である.

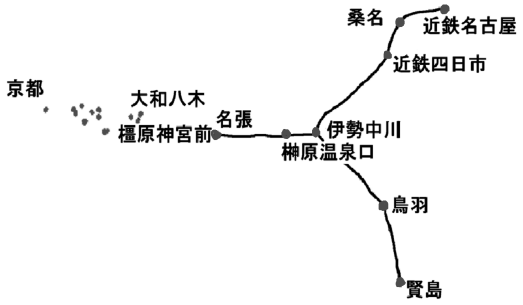

図 1:

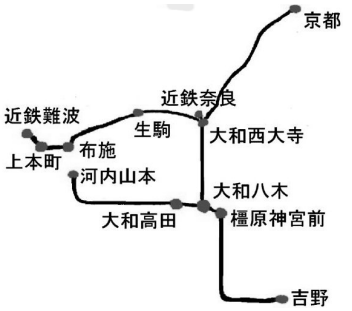

図 2:

5.2.2　復元結果

このとき 2 次元のストレス値は $S = 0.1259$, 3 次元のストレス値は $S = 0.1021$ になった. もちろん, これは高低差の比較的小さい平面の鉄道路線なので, 2 次元であるべきであろう. 本研究で提案した C 値を用いると 2 次元のときは $C = 0.2713$, 3 次元のときは $C = 19.229$ になり, 2 次元であると判断される.

6　まとめ

計量 MDS 法で最適次元を定めるのに, ストレス値だけでは, 次元を推定するのは難しいが, 本研究で提案した簡便な基準により推定することが可能といえる. 本来の AIC はパラメータの個数が多くなると, 不安定になりパラメータの大きいモデルが選ばれる傾向にある. そのため, 最適次元が比較的高次の場合に適応可能であるかの検討は今後の課題としたい.

なお, 本研究の一部は 2007 年 12 月 20 日に開催された情報処理学会第 63 回「 MPS 研究会」で発表済みである.

参考文献

1. 斉藤堯幸, 多次元尺度構成法, 朝倉書店 (1980).
2. 坂元慶行, 石黒真木夫, 北川源四郎, 情報量統計学, 共立出版株式会社 (1983).