社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (14)
2007－BIO－11 (14)
2007／12／20

# MDS 法における最適次元の推定

伊藤 里江*, 吉田 裕亮**
*お茶の水女子大学 理学部 情報科学科
**お茶の水女子大学 大学院人間文化創成研究科

本研究では多次元尺度法 (MDS:Multi-Dimensional Scaling) において, 対象を布置する空間の最適次元を推定する基準を提案する. すなわち, 通常 Kruscal のストレス値をもって布置の最適性に用いられることが多いが, これは必ずしも最適な布置空間の次元を求めることはできない. そこで, このストレス値に情報量基準と同様の手法で次元の増大に伴うペナルティを与えることにより, 最適な次元を選択する基準を導入する. また MDS 法の応用として鉄道の運賃表から各都市の位置関係の復元を例示する.

# An estimation of the optimal dimension in the MDS method

Rie Itoh*, Hiroaki Yoshida**
* Department of Information Sciences, Ochanomizu University
** Graduate School of Humanities and Sciences, Ochanomizu University

In this study, we introduce a criterion for an estimation of the optimal dimension of the relocation subspace in the Multi-Dimensional Scaling method. Generally, a value of Kruscal's stress can be used as a measure of goodness for an estimation in the MDS method but unfortunately, it does not always lead the optimal dimension well. Thus we shall modify the stressvalue by giving penalty on the increasing of the dimension, which is the same notion as in the information criteria like the AIC. As an application, we also give a reconstruction of the map from the fare table of the railway.

## 1 はじめに

対象間に距離を定義し, 距離が近い対象どうしをグループにまとめる作業は一般的にクラスタリングと呼ばれている. 1 次元クラスタリングでは距離行列が与えられた際, デンドログラムの横軸によって対象間の距離を表現することは可能であるが, 個体間の位置を視覚的に把握することは難しい. それを可能にし, 視覚的に対象間の相対的な位置関係を考察する手法として多次元尺度法 (MDS) がある. MDS 法とは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で, 距離データを低次元に配置する計量 MDS 法と順序尺度のデータの類似度あるいは変換可能な親近性データを低次元に配置する非計量 MDS 法とに分けられる. 計量 MDS においては, Young - Householder 変換によって, 距離行列から内積行列を構成し, 位置関係を復元することが可能になる. 本研究では計量 MDS を利用し, 距離行列を与えられたとき, それら対象をある可視化次元で空間に布置する場合における最適次元の決定に関する簡便な手法を提案したい.

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, ストレス値だけに着目すると, 本当の最適次元よりも高次元が選ばれてしまう場合があり, いわゆるオーバーフィッテング問題に陥る. この問題の回避のために情報量基準 (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)

まずはじめに, MDS 法について簡単に紹介しよう.[1] MDS 法は Richardson のアイディアを Young - House-

holder や Torgerson らが発展させたものであり，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，Young - Householder の定理をもとに対象を Euclid 空間の点として位置付ける方法である．

Young - Householder 定理とは複数の対象間の Euclid 距離が既知の場合，対象が埋め込まれる空間に関する定理であり，MDS 法の基礎定理としてみることができる．

$n$ 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．

また，$n-1$ 個の対象間の内積を要素とする行列 (内積行列)$B$ を，

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

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．これより，

$$b_{ij} = \frac{1}{2}(d_{in}^2 + d_{jn}^2 - d_{ij}^2). \qquad (*)$$

を得る．この式から対称間の距離が既知ならば，対称間の内積が求まることがわかる．また，点間距離 $d_{ij}(= d_{ji})$ の組が本当に Euclid 空間の点の集合間の相互距離であるための必要十分条件として，$(*)$ 式で定義される要素からなる内積行列 $B$ が正定値であることもわかる．この場合には $B$ を固有値分解することにより，各対象の布置される座標を定めることが可能である．

上の内積行列 $B$ は対象 $O_n$ を原点とし場合の内積行列であり，どの対象を原点に取るかで当然異なるが，対象間の距離に誤差が含まれない場合には，布置は原点の平行移動と軸の回転を除き不変である．しかし，現実のデータの場合は，対象間距離の推定に誤差が含まれるので，特定の対象を原点に取ることは好ましくない．これを回避するひとつの方法が，対象の重心を原点に選ぶ手法であり，これに基づく内積行列は，以下のように与えられる．

対象間距離の 2 乗 $d_{ij}^2$ を $(i,j)$ 要素とする行列を $D^{(2)}$ で表す．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,1 はすべての成分が 1 の列ベクトルである．したがって，行列 $\mathbf{1}_n\mathbf{1}_n^t$ はすべての成分が 1 である $n \times n$ 行列になる．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により，布置座標を求めることになる．この場合も対象間距離に全くの誤差がない場合には，行列 $B_c$ の階数 $r$ が $X_c$ の階数に等しく，$n$ 個の対象を原点とする各対象の $r$ 次元 Euclid 空間の座標値を要素とする行列になっている．

また，$r$ 次の直交行列を $T$ とし，$X' = XT$ と定義すると，

$$X'(X')^t = XTT^tX^t = XX^t$$

であるので，Euclid 距離，すなわち対象の布置は回転に関して不変であることもわかる．

さて，重心を原点とするようにして，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) を行う数値計算法は幾つかある．本研究では比較的小さいランクの場合が扱う対象であるため，累乗 (パワー) 法が有効の手法であると考えられる．すなわち，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方法である．

つまり，まず内積行列 $B_1 = B_c$ の最大固有値 $\lambda_1$ とそれに付随する固有ベクトル $\mu_1$ (ただし，$||\mu_1|| = 1$) を反復により求める．このとき，

$$P_1 = \mu_1\mu_1^t$$

が固有値 $\lambda_1$ スペクトル射影になるので, このスペクトル成分を除いた行列

$$B_2 = B_1 - \lambda_1 P_1$$

を求め, 再びこの行列 $B_2$ の最大固有値 $\lambda_2$ とそれに付随する固有ベクトル $\mu_2$ 求める. この固有値, 固有ベクトルは元の行列 $B_1$ の第 2 固有値とそれに付随する固有ベクトルになっている. したがって,

$$P_2 = \mu_2 \mu_2^t$$

が $B_c$ の固有値 $\lambda_2$ スペクトル射影である. 以下, 必要な回数

$$B_{n+1} = B_n - \lambda_n P_n$$

という操作を繰り返すことにより, 第 $2, 3, \cdots$ 最大固有値およびそれらに付随する固有ベクトルを求めることになる.

## 3 最適次元の選択指標

まず Kruscal により導入された次元選択の指標であるストレスを再考しよう. 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のもである. ここで, $\widehat{d_{jk}}$ はこの値が観測された対象 $j$ から対象 $k$ への非類似度 $\delta_{jk}$ と同順位になるという条件下で, $S$ を最小にする数値である. 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, $\widehat{d_{jk}}$ を布置された空間における対象間の Euclid 距離, すなわち最小 2 乗推定された対象間の距離で置き換え, このストレスとの類似量を推定誤差と考える. すなわち, 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する.

よく知られているように, 回帰モデルではモデルのパラメータを増やすことにより, いくらでも推定誤差は小さくすることが可能ではある. しかし, 一般にはオーバーフィッティングにより必ずしも推定されるべき良いモデルとは限らない. このようなトレードオフの状況において最適なパラメータ数, すなわちモデルを選択する基準として情報量基準 AIC がある.[2]

情報量規準 AIC はモデルの最大対数尤度を $MML$, モデルに含まれる自明パラメータの数を $k$ として,

$$AIC = -2MML + 2k$$

で与えられ, この AIC を最小にするモデルが良いモデルであると判断するモデル選択の指標である.

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.

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

ここで $S^2$ は, 先に述べたように, $\widehat{d_{jk}}$ を布置された空間における対象間の Euclid 距離としたストレス $S$ の 2 乗である. また, モデルの自由パラメータ数としては, 可視化次元を $\ell$ とした場合, $n$ 次元の固有ベクトルと固有値の組が $\ell$ 個と誤差分散に相当する分がひとつあるので, 合計

$$k = \ell(n-1) + \ell + 1 = \ell n + 1$$

である. 最適な布置される空間の次元として, この $C$ が小さいものを選択するような手法である.

## 4 有効性の検証実験

### 4.1 シミュレーションデータの場合

#### 4.1.1 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.

#### 4.1.2 実験結果

結果は, ストレス値では 2 次元の場合, $S = 0.7455$ となり, 3 次元になると $S = 0.70711$ となり, 3 次元を示しているが, 本研究で提案する手法で $C$ 値を比較すると, 2 次元では $C = 55.174$, 3 次元では $C = 81.386$ となり, 2 次元を選ぶべきであると判断された.

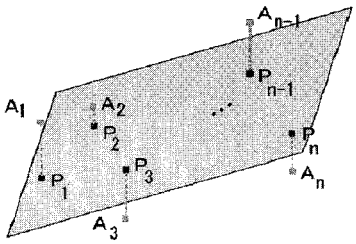


図 1:

## 4.2 実データの場合

### 4.2.1 近鉄路線

MDS 法の実データへの応用として, 鉄道運賃を 2 地点間の距離として, 運賃表から地点の相対的位置関係の復元を試みた. 一般に鉄道運賃は距離に応じて加算されていくが, 当然, 初乗り料や距離と共に正確に線形関係にある訳ではないためそれらが雑音として混入した距離データと考えられる.

幾つかの会社路線で実験を行ったが, ここでは近畿日本鉄道の例を挙げることにする. 近鉄の駅から 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 2 のような結果を得た. いくつかの駅は相対位置がずれてしまうものが見られるが, これ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 3 である.

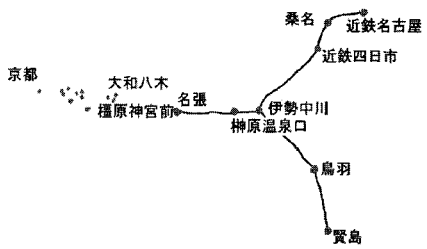


図 2:

### 4.2.2 復元結果

このとき 2 次元のストレス値は $S = 0.1350$, 3 次元のストレス値は $S = 0.1082$ になった. もちろん, これは高低差の小さい平面の鉄道路線なので, 2 次元であるべきである. 本研究で提案した $C$ 値を用いると 2 次元のときは $C = 58.0098$, 3 次元のときは $C = 82.8941$ になり, 2 次元であると判断された.

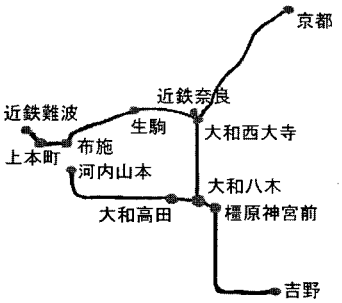


図 3:

## 5 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.

## 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).