目白大学　経営学研究
創刊号　　2003年3月　43－50

クラスタリングされた大量データの低次元表示法*

An Integrated Method for Visualizing Categorical Data via Multidimensional Scaling and Procrustes Transformation

橋口　博樹　　宿久　洋
(Hiroki Hashiguchi and Hiroshi Yadohisa)

【要　旨】

本論文では，対象に関する多変量データが与えられ，かつ，対象がある基準でいくつかのクラスタに分割可能な場合について，多次元尺度構成法 (MDS) と Procrustes 変換を併用した解析法を提案する．各クラスタについて，所属する対象を代表する値（例えば，平均値）を１つ定める．この代表値に MDS を適用することにより，代表値を表す点を低次元のユークリッド空間上に付置する．さらに，Procrustes 変換を用い，全データを代表値が付置された空間に射影し，射影されたデータの共分散行列によって，代表値を表す点の周りに確率楕円を描画する．この確率楕円でクラスタ内のデータの広がりを低次元に再現する．

キーワード：多次元尺度構成法 (MDS)，Procrustes 変換，確率楕円，固有値・固有ベクトル．

はじめに

多次元尺度構成法 (Multidimensional Scaling, 以下，MDS と略す) は，社会学，心理学，マーケティング分野，パターン認識の分野などで広く用いられている方法である．この方法は，対象間の類似性や非類似性が与えられたとき，類似関係をできるだけ再現するように低次元ユークリッド空間内に対象を付置する手法である．対象間の関係を空間の位置関係で表現することによって，個々の関係や，個と全体との関係がより見えてくる．MDS で扱うデータには，あらかじめ対象についての多変量データが与えれていて，類似性や非類似性を計算する場合や，そもそも類似性や非類似性しか与えられていない場合の２つがある．前者は，高次元から低次元へ縮約するという意味では主成分分析に似た考え方であるが，MDS では距離関係を再現しようという意図がより働く．

一方，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，一方を他方に近づける変換である．この Procrustes という用語はギリシャ神話に由来するが，統計の分野では Hurley and Cattell (1962) により初めて用いられた．しかし，この種の問題は

目白大学経営学部
受理日　2003年3月20日

他にも Green (1952), Schönemann (1966) などで扱われている.

本論文では，対象の多変量データが与えられて，何らかの基準でクラス分けされたデータを扱う．ここで基準とは，各対象に固有のラベルがついている場合や，k-means 法などのクラスタリング手法でクラス分けされていてもよい．このような状況で，クラスタの代表値（例えば平均）間の距離関係を保ちつつ，クラスタ間の関係を低次元のユークリッド空間上に表すことを考える．さらに，クラスタ内のデータの広がり（分散）を考慮して，高次元での広がりを低次元に再現することも考える．これは，特に対象数がクラスタ数よりも極めて多い場合，クラスタの広がりを含めた位置関係を見たほうが，計算コストの削減以上に，低次元化の効果（解釈のし易さ，見易さ）が明確になる．

そこで，これら 2 つの MDS と Procrustes 変換を次のように適用する方法を提案する．まず，クラスタの代表値を MDS で次元縮約する．次に Procrustes 変換を使って，代表値の低次元化の線形変換を求め，各クラスタ内対象をさらに線形変換する．線形変換された座標値でクラスタ内の分散，共分散を計算すればクラスタ内の広がりを求めることができる．この低次元化の次元を 2，3 次元とすれば，楕円，楕円体が広がりを表現し可視できる．

2 MDS と Procrustes 変換の併用

N 個の対象が k-means 法等のクラスタリング法により n 個のクラスタ $C_i\ (i = 1, \ldots, n)$ に分類されているとする．もちろん, 属性情報などにより予め分類されていてもかまわない．ただし，対象の数 N はクラスタの数 n より十分大きい（$N \gg n$）と仮定する．また，クラスタ $C_r (r = 1, \ldots, n)$ の代表点の座標 $\boldsymbol{y}_r$ が定められているものとする (通常は, 各クラスタに属するデータの平均である). クラスタ C_r と C_s の間の非類似性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．これらの代表点を付置することを考えると，その数は $n \ll N$ なので大幅に少ない数で MDS を用いることができる．

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, 各クラスタ内での対象の関係などは無視されてしまう．そこで, クラスタ内の対象の広がりを表す確率超楕円体を描画することでクラスタ内の広がりを表現する．クラスタ内の全対象は，Procrustes 変換 (Hurley and Cattell, 1962) の適用によって代表点の付置された空間へ射影され，それらの共分散行列が確率超楕円体をつくることになる．

2.1 古典的な MDS によるクラスタの代表値の付置

クラスタ $C_i\ (i = 1, \ldots, n)$ の代表点の座標 $\boldsymbol{y}_r \in \mathbf{R}^p\ (n > p)$ に対して古典的な MDS を適用する（Cox and Cox (2001) の 2 章参照）．つまり，クラスタ間の非類似性 d_{rs}^2:

$$d_{rs}^2 = (\boldsymbol{y}_r - \boldsymbol{y}_s)^T (\boldsymbol{y}_r - \boldsymbol{y}_s)$$

に対して，$-\frac{1}{2} d_{rs}^2$ を (r, s)-成分にもつ行列を $\boldsymbol{D}$ と書き，さらに $\boldsymbol{D}$ を中心化する行列

$\boldsymbol{H} = \boldsymbol{I} - 1/n\ \mathbf{1}\mathbf{1}^T$ を使って $\boldsymbol{HDH}$ のスペクトル分解 $\boldsymbol{V\Lambda V}$ を求める．ただし，T はベクトル，行列の転置を，$\boldsymbol{I}$ は単位行列，$\mathbf{1} = (1, 1, \ldots, 1)^T$ を表す．また，$\boldsymbol{HDH}$ の固有値を $\lambda_1, \lambda_2, \ldots, \lambda_n$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0)$，$\lambda_i$ に対応する長さ 1 の固有ベクトルを $\boldsymbol{v}_i$ としたとき，$\boldsymbol{\Lambda}$ と $\boldsymbol{V}$ は以下の通りである．

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \quad \boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]. \tag{1}$$

このスペクトル分解から，$\boldsymbol{y}_r$ に対応する付置座標 $\boldsymbol{x}_r \in \mathbf{R}^q$ が，$\boldsymbol{x}_r \in \mathbf{R}^q$ を並べた行列を $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ として，次の (2) 式で具体的に得られる．

$$\boldsymbol{X}^T = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q]\,\mathrm{diag}(\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}} \ldots, \lambda_q^{\frac{1}{2}}). \tag{2}$$

結局，この古典的な MDS では $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{HDH}$ となる $\boldsymbol{X}$ を見つけるために，$\boldsymbol{HDH} = \boldsymbol{V\Lambda V}$ とスペクトル分解している．ただし Cox and Cox (2001) で紹介されているように，上に述べた古典的な MDS 以外でも多くの MDS の方法が提案されていることに注意する．

2.2　Procrustes 変換によるによるクラスタ内データの低次元化

クラスタ C_r $(r = 1, \ldots, n)$ の代表点 $\boldsymbol{y}_r \in \mathbf{R}^p$ に対して，MDS の適用により求められた付置座標を $\boldsymbol{x}_r \in \mathbf{R}^q$ と表す．もちろん $p > q$ である．次元数の違いを無くすために，$\boldsymbol{x}_r$ の末尾に $p - q$ 個の 0 を加えた p 次元ベクトルを新たに $\begin{pmatrix} \boldsymbol{x}_r \\ \mathbf{0} \end{pmatrix} \in \mathbf{R}^p$ とおく．この 2.2 節では，混乱のない限り記号の煩雑さを避けるため $\begin{pmatrix} \boldsymbol{x}_r \\ \mathbf{0} \end{pmatrix} \in \mathbf{R}^p$ を新たに $\boldsymbol{x}_r \in \mathbf{R}^p$ と書くことにする．

この $\boldsymbol{x}_r \in \mathbf{R}^p$ に回転, 拡大縮小, 平行移動を施したベクトル

$$\rho\boldsymbol{A}^T\boldsymbol{y}_r + \boldsymbol{b} \tag{3}$$

と $\boldsymbol{x}_r$ との 2 乗距離の和

$$R^2 = \sum_{r=1}^{n} (\boldsymbol{x}_r - \rho\boldsymbol{A}^T\boldsymbol{y}_r - \boldsymbol{b})^T(\boldsymbol{x}_r - \rho\boldsymbol{A}^T\boldsymbol{y}_r - \boldsymbol{b}) \tag{4}$$

を最小にする $\boldsymbol{A}, \rho, \boldsymbol{b}$ を求める．ここで, $\boldsymbol{A}$ は回転を表す直交行列, ρ は拡大縮小を表す実数, $\boldsymbol{b}$ は平行移動を表すベクトルである．R^2 を最小化する $\widehat{\boldsymbol{A}}, \widehat{\rho}, \widehat{\boldsymbol{b}}$ は，それぞれ以下のように求められる (Cox and Cox (2001), Gower and Hand (1996) 参照)．

$$\widehat{\boldsymbol{A}} = \boldsymbol{VU}^T, \quad \widehat{\rho} = \frac{\mathrm{tr}(\widehat{\boldsymbol{A}}\boldsymbol{X}^T\boldsymbol{Y})}{\mathrm{tr}(\boldsymbol{Y}^T\boldsymbol{Y})}, \quad \widehat{\boldsymbol{b}} = \boldsymbol{x}_0 - \rho\widehat{\boldsymbol{A}}^T\boldsymbol{y}_0. \tag{5}$$

ここで, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]$, $\boldsymbol{U}$ と $\boldsymbol{V}$ は，$\boldsymbol{X}^T\boldsymbol{Y}$ の特異値分解

$$\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{6}$$

の直交行列, $\boldsymbol{\Lambda}$ は対角行列を表す. さらに $\boldsymbol{x}_0 = \sum_{r=1}^{n} \boldsymbol{x}_r/n$, $\boldsymbol{y}_0 = \sum_{r=1}^{n} \boldsymbol{y}_r/n$, である.

もし, $\boldsymbol{X}^T\boldsymbol{Y}$ が非特異であれば $\widehat{\boldsymbol{A}}, \widehat{\rho}$ は更に以下のように書くことができる.

$$\widehat{\boldsymbol{A}} = (\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y})^{1/2}(\boldsymbol{X}^T\boldsymbol{Y})^{-1}, \quad \widehat{\rho} = \frac{\mathrm{tr}(\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y})^{1/2}}{\mathrm{tr}(\boldsymbol{Y}^T\boldsymbol{Y})}. \tag{7}$$

2.3 Procrustes 変換後の共分散構造

Procrustes 変換後のクラスタ内の共分散構造は, 変換前の構造を $\widehat{\rho}^2$ だけ拡大・縮小し, $\widehat{\boldsymbol{A}}$ により回転したものである. つまり次の性質がある.

性質 2.1 各クラスタ C_r $(r = 1, \ldots, n)$ のクラスタ内共分散行列を Cov_r とする. また, (5) 式の $\widehat{\boldsymbol{A}}, \widehat{\rho}, \widehat{\boldsymbol{b}}$ によって, C_r 内の全対象データを Procrustes 変換した後の共分散行列を Cov'_r とする. なお, Cov_r と Cov'_r のサイズはともに $p \times p$ であることに注意する. このとき次の 2 つが成立する.

(1) $\mathrm{Cov}'_r = \widehat{\rho}^2 \widehat{\boldsymbol{A}}\, \mathrm{Cov}_r\, \widehat{\boldsymbol{A}}^T$ が成立する.

(2) 上の (1) から, Cov'_r の大きい順での第 i 番目の固有値 λ'_i と, 同様に Cov_r の第 i 番目の固有値 λ_i には $\lambda'_i = \widehat{\rho}^2\lambda_i$ が成立する. このことから Cov'_r と Cov_r の寄与率は等しいので, 累積寄与率も等しい.

性質 2.1 から, MDS・Procrustes 変換前後の違いは, 座標軸の回転, $\widehat{\rho}$ による縮小・拡大のみだけなので, これらの変換前後のクラスタ内の共分散構造は変化がないと言ってもよい. そこで, この共分散構造を $\mathbf{R}^q$ で反映させるために, 確率楕円を Cov'_r から次のように構成し描画する. まず, Cov'_r を

$$\mathrm{Cov}'_r = \left(\begin{array}{c|c} \mathrm{Cov}'_r(q) & * \\ \hline * & * \end{array}\right) \tag{8}$$

と分割する. ここで, $\mathrm{Cov}'_r(q)$ は, 逆行列の存在が仮定された $q \times q$ の行列である. この $\mathrm{Cov}_r(q)$ によって, $\mathbf{R}^q$ に描画される超楕円体の領域 W_c(c は適当な正の定数) は, 以下のように表すことができる.

$$W_c = \{\boldsymbol{z} \in \mathbf{R}^q \mid (\boldsymbol{z} - \boldsymbol{x}_r)^T\, \mathrm{Cov}_r(q)^{-1}\, (\boldsymbol{z} - \boldsymbol{x}_r) \leq c^2\} \tag{9}$$

ただし, $\boldsymbol{x}_r \in \mathbf{R}^q$ は, クラスタ C_r の代表値 $\boldsymbol{y}_r$ を MDS により q 次元空間に付置した座標を表す.

3　数値例

先の 2 節で述べた提案法を数式処理システム Mathematica 4.0.2 上に実装した．Mathematica 上のプログラムでは MDS，Procustes 変換に加え，2，3 次元での描画も行っている．この節では，Mathematica 上で正規乱数を使っての基本的な実験を行う．

3.1　クラスタリングされた 3 次元データの 2 次元付置

図 1 には，正規母集団 $N(\boldsymbol{\mu}_i, \boldsymbol{I})$ $(i = 1, \ldots, 4)$ からの乱数ベクトルをプロットしている．平均ベクトルはそれぞれ，$\boldsymbol{\mu}_1 = (0,0,0)^T$, $\boldsymbol{\mu}_2 = (0,0,10)^T$, $\boldsymbol{\mu}_3 = (10,0,0)^T$, $\boldsymbol{\mu}_4 = (0,10,0)^T$ であり，$\boldsymbol{I}$ は 3 次の単位行列である．また，各クラスタの標本数は 100 である．得られた乱数ベクトルから各クラスタの標本平均を求め，これらを各クラスタの代表値 $(\boldsymbol{y}_r)$ とした．代表値 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_4$ に MDS を適用し，2 次元での対応する座標値 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4$ を得，2.2 節で述べた方法により線形変換をもとめた．この線形変換を 400 個のデータすべてに行い，2 次元に付置した図が図 2 である．これら 2 つの図から，MDS・Procustes 変換がうまく機能し，3 次元でのデータが 2 次元でも重なり合うことなく描画されている．

図 3，4 では，各クラスタ内の標本共分散行列に基づく W_1 と W_3 の確率楕円を図示し，図 2 中のクラスタ内のデータの広がりを示している．

3.2　クラスタリングされた 6 次元データの低次元付置

平均ベクトルを $\boldsymbol{\mu}_1 = (0,0,0,0,0,0)^T$, $\boldsymbol{\mu}_2 = (0,0,10,10,10,10)^T$, $\boldsymbol{\mu}_3 = (0,0,0,10,0,0)^T$, $\boldsymbol{\mu}_4 = (0,10,0,10,0,10)^T$, 共分散行列を 5 次の単位行列として，各クラスタ 100 個の正規乱数データを MDS・Procrustes 変換した．図 5，6 において，2 次元付置の確率楕円 W_1，W_3 を示す．クラスタ 1 とクラスタ 3 が次元縮約により重なっている．一方で 3 次元での確率楕円 W_1，W_3 を図 7，8 に示すが，この場合は，クラスタ 1 とクラスタ 3 はうまく分離されている．そこで MDS の固有値の寄与率を調べたところ，(第 1 固有値の寄与率, 第 2 固有値の寄与率, 第 3 固有値の寄与率, 第 4 固有値の寄与率) = (42%, 31%, 26%, 0%) であった．このことは，2 次元よりも 3 次元で元データがうまく再現されることを意味する．この数値例は，寄与率がデータの再現性を示す尺度になること，および，この点を考慮した解析の必要性を示唆している．

3.3　まとめと課題

本論文ではクラスタリングされたデータの低次元表示方法について，MDS と Procrustes 変換を併用する方法を提案し，いくつかの性質を調べシミュレーション実験を行った．今回の提案法は，別の見方をすれば，クラスタ内で主成分分析を行い，その後，各々の結果を同一空間にどう表示すればよいかという問題を扱ったとも見て取れる．したがって，今回は共分散行列

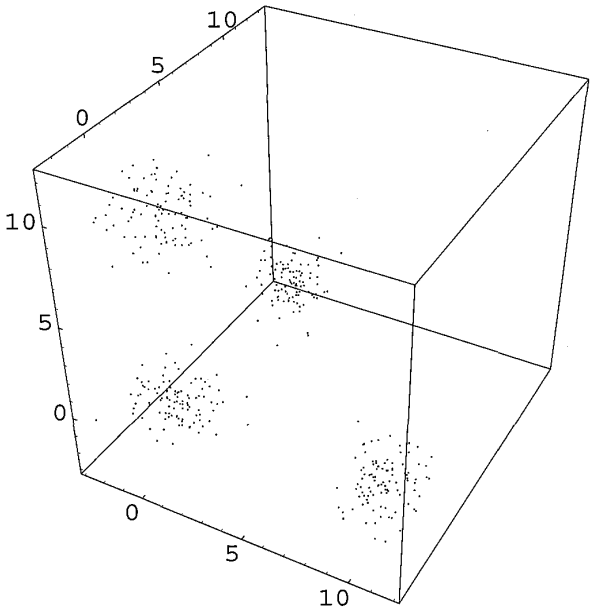

図 1　4 クラスタ 3 次元の正規乱数データ

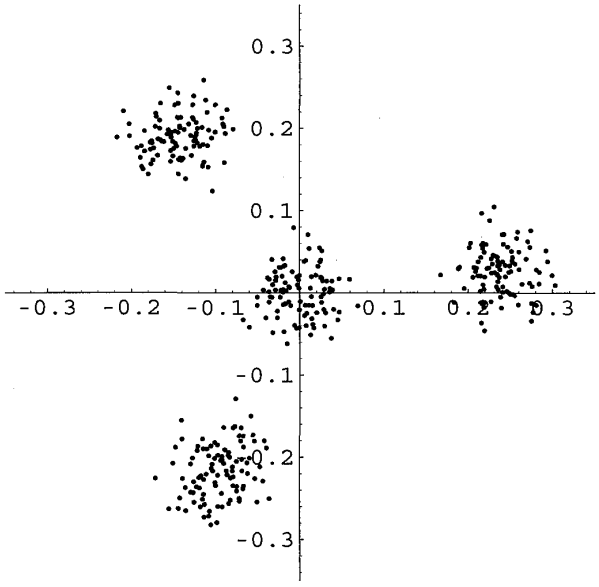

図 2　図 1 の MDS・Procrustes 変換による 2 次元付置

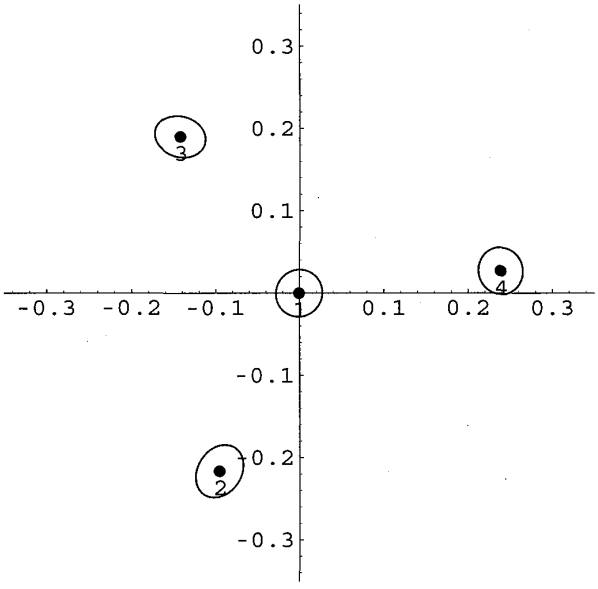

図 3　図 2 の確率楕円 W_1

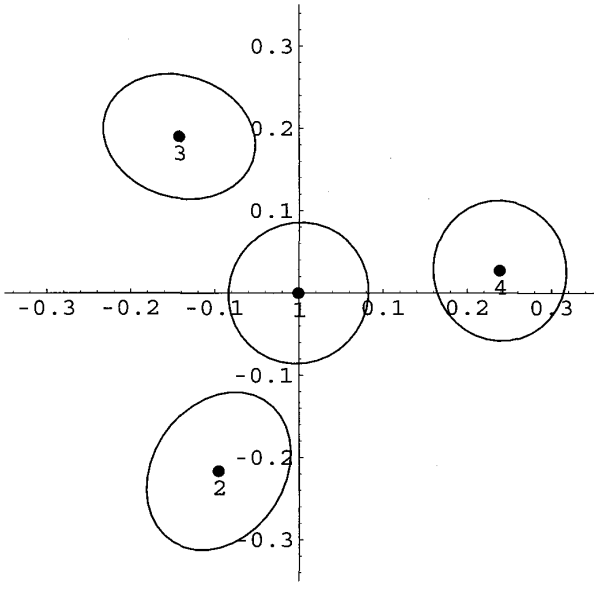

図 4　図 2 の確率楕円 W_3

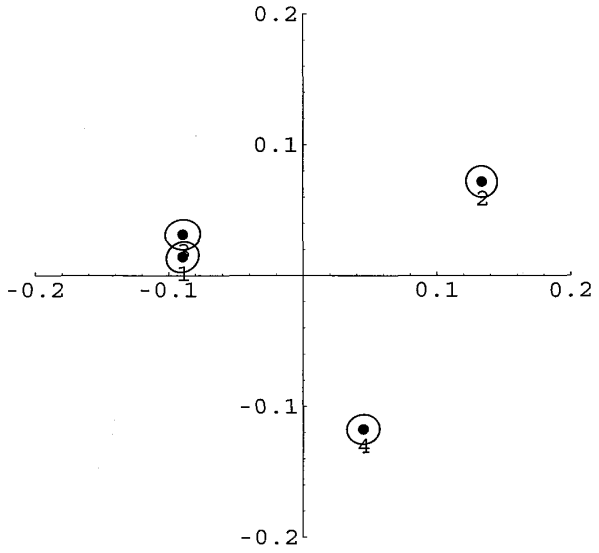

図 5　6 次元 4 クラスタ例の 2 次元確率楕円 W_1

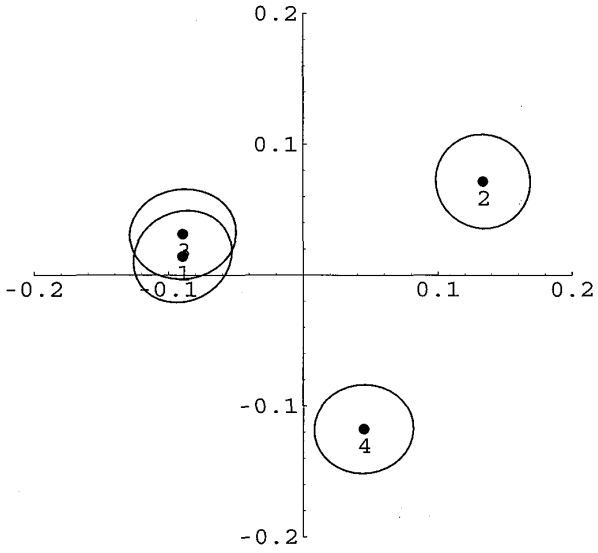

図 6　6 次元 4 クラスタ例の 2 次元確率楕円 W_3

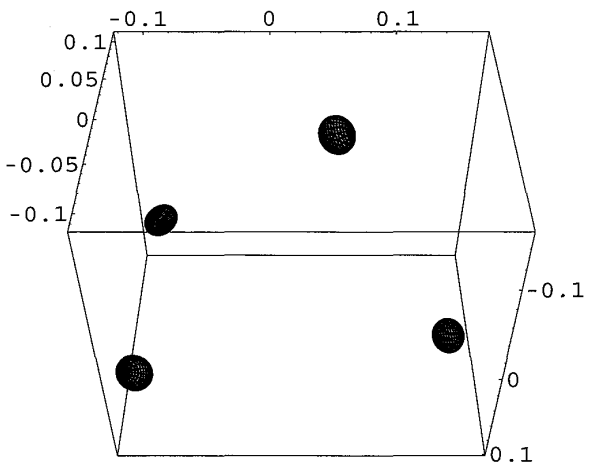

図 7　6 次元 4 クラスタ例の 3 次元確率楕円 W_1

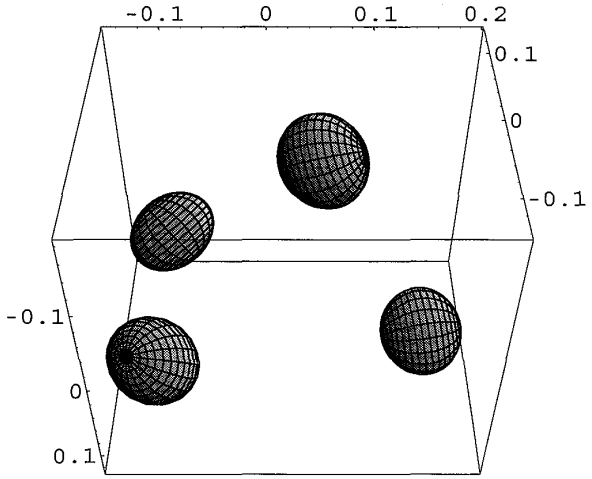

図 8　6 次元 4 クラスタ例の 3 次元確率楕円 W_3

を扱ったが，相関行列でも全く同様の議論が展開できる．さらに主成分分析では，寄与率によって次元縮約による情報の損失を定量化しているが，提案法でもその種の議論をしなければならない．このことは，3.2 節の実験からもその必要性が読み取れる．この情報損失を定量化し理論性質を調べた上で，実データの解析を行いたいと考えている．

また，実データを解析する際には，予備解析としてクラスタ内で主成分分析をしておくことは重要で，提案法で得られた解析結果のより深い解釈に繋がることが期待できる．それゆえ，主成分分析と提案法を融合し，実データの解析を行うことは大変興味深い課題である．

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 t test a hypothesized factor sructure. *Behavioral Science*, **7**, 258–262.

[7] Schönemann, P. H. (1966). A generalized solution of the orthognal Procrustes problem, *Psychometrika*, **31**, 1–10.

[8] 宿久洋, 橋口博樹 (2003), 類似度に基づく大量データの表示, 2002 年度統計数理研究所プロジェクト研究による研究会「統計科学情報の高度利用」(報告集掲載予定)