修正非線形マッピングと そのクラスター分析への応用

Modified Non-Linear Mapping and its Application to Cluster Analysis

(1996年3月26日受理)

福森　護　　奥村　英則
Mamoru Fukumori　Hidenori Okumura

Key words: 修正非線形マッピング，バイプロット，クラスター分析

1. はじめに

p次元空間の点を2次元空間にグラフ表現することを目的とした手法は数多い。Everitt(1978)はこれらの手法をオーディネーション手法と呼び，主成分プロット，主座標プロット(1966)，バイプロット(1971)，ノンメトリック多次元尺度構成法(1964)，非線形マッピング(1969)をその代表的なものとして取り上げて解説している。これらの手法のうち，主成分プロット，主座標プロット，バイプロットの3つの手法はすべて固有値と固有ベクトルに基づく方式であり，他の2つの手法はある特定の基準を反復アルゴリズムを用いて最小にする方式である。福森(1994)は，従来の非線形マッピングが2点間の距離が保持されるように次元を縮小することに注目し，(2点間の距離だけではなく) 2点間の角度が保たれるように次元を縮小する修正非線形マッピングを提案した。そして，open/closed book data(1979)を修正非線形マッピング，主成分プロット，従来の非線形マッピングに適用することにより，低次元に落としたときに元の次元の構造がどの程度保持されているかを比較検討し，修正非線形マッピングが最も優れていることを示した。

本研究では，修正非線形マッピングのクラスター分析における有効性について，元のデータのクラスター構造が低次元に落とした場合にどの程度再現できるかを，従来の非線形マッピングおよびバイプロットとの比較によって検討する。

2　各手法の概要

2．1　バイプロット

バイプロットは，2点間の距離の測度で示される個体間の関連性，および変数の共分散や相関で示される変数間の関連性を同時に座標上にグラフ表現する方法である。本手法を要約すると以下のようになる。

階級 r のデータ行列 X は特異値分解により

$$X=\sum_{i=1}^{r}\sqrt{\lambda_i}\,\mathrm{p}_i\mathrm{q}',\quad \lambda_1\geq\lambda_2\geq\cdots\lambda_r>0$$

と表される。ここで，λ_i は $X'X$ または XX' の第 i 固有値であり，p_i は λ_i に対応する第 i 固有ベクトル，また q_i は $X'X$ の固有ベクトルである。これを利用して X をランク2の行列

$$X_{(2)}=\sqrt{\lambda_1}\mathrm{p}'_1\mathrm{q}_1+\sqrt{\lambda_2}\mathrm{p}'_2\mathrm{q}_2$$

での近似を考える。このとき $n\times2$ 行列 H と $p\times2$ 行列 G を

$$H=(\mathrm{q}_1,\ \mathrm{q}_2),\quad G=(\sqrt{\lambda_1}\mathrm{p}_1,\ \sqrt{\lambda_2}\mathrm{p}_2)$$

または

$$H=\frac{1}{\sqrt{n}}(\sqrt{\lambda_1}\mathrm{q}_1,\ \sqrt{\lambda_2}\mathrm{q}_2),\quad G=\sqrt{n}\,(\mathrm{p}_1,\ \mathrm{p}_2)$$

とおく。前者の場合には G の n 個の要素，後者の場合には H の p 個の要素を2次元座標にプロットする。これらのプロットによって G の場合は X の行についてのクラスターを，H の場合はその列についてのクラスターを視覚的に表現する。

2．2　非線形マッピング

非線形マッピングは，2点間の距離ができるだけ保持されるように次元の縮小を行い，より低次元の座標上でグラフ表現を行う手法である。本手法を要約すると以下のようになる。

n 個の p 変量データ

$$\mathrm{x}_1=(x_{11},\ x_{12},\cdots,\ x_{1p})$$
$$\mathrm{x}_2=(x_{21},\ x_{22},\cdots,\ x_{2p})$$
$$\cdots\cdots$$
$$\mathrm{x}_n=(x_{n1},\ x_{n2},\cdots,\ x_{np})$$

が与えられたとき，p 次元での距離 d_{ij}^{*} を $d_{ij}^{*}=\{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2\}^{\frac{1}{2}}$ で与える。一方，2次元での n 個の任意の点

$$\mathrm{y}_1=(y_{11},\ y_{12})$$
$$\mathrm{y}_2=(y_{21},\ y_{22})$$

$$\cdots\cdots$$

$$\mathrm{y}_n = (y_{n1}, y_{n2})$$

とするとき，2次元での距離 d_{ij} を $d_{ij}=\{\sum_{k=1}^{2}(y_{ik}-y_{jk})^2\}^{\frac{1}{2}}$ で与える。このとき，各 $i(1\leq i\leq n)$ に対して

$$\frac{1}{\sum_{i\neq j}^{n} d_{ij}^{*}} \sum_{i\neq j}^{n} \frac{(d_{ij}^{*}-d_{ij})^2}{d_{ij}^{*}}$$

を最小にする $(\tilde{y}_{i1},\tilde{y}_{i2})$ を求める。そしてこれら n 個の点 $(\tilde{y}_{i1},\tilde{y}_{i2})(1\leq i\leq n)$ を2平面上にプロットする。

このように，非線形マッピングにおける基準は，重み付き平方和であり，より近くにある点に対して大きな重みを与えるという意味において，局所的な距離を保持するように次元を縮小する方法であるといえる。

2．3　修正非線形マッピング

非線形マッピングでは，局所的な2点間の距離をできるだけ保存するように次元を縮小する方法であったが，ここでは，類似の方向を持つ2点にできるだけ大きなウェイトを与えるような基準を考える。

データ行列 $X=(\mathrm{x}_1,\cdots,\mathrm{x}_p)'$ に対して，$s_{ij}=\mathrm{x}_j'\cdot\mathrm{x}_j=\|\mathrm{x}_i\|\|\mathrm{x}_j\|\cos\alpha_{ij}$ とする。また n 個の任意の実数の点 $\mathrm{y}_i=(y_{i1},y_{i2})(i=1,\cdots,n)$ に対して，$s_{ij}^{*}=\|\mathrm{y}_i\|\|\mathrm{y}_j\|\cos\alpha_{ij}^{*}$ とする。ここで $\|\cdot\|$ はユークリッドノルムであり，α_{ij}, α_{ij}^{*} は，それぞれ x_i, x_j と y_i, y_j のなす角である。

このとき，

$$\sum_i \sum_j w_{ij}(s_{ij}-s_{ij}^{*})^2,$$

という基準を定義し，これを最小にするように次元を縮小することを考える。ここで，w_{ij} は，類似の方向を持つベクトルにできるだけ大きなウェイトを与えるように，

$$w_{ij}=\{(1+\cos\alpha_{ij})/2\}^{k}$$

を考える。

3　数　値　例

上記の3つの手法の有効性について調べるために，乱数データに対して各手法を適用した。乱数データは，以下のような手順で作成された。

まず，3つの群の3変数のデータを次のように想定した。3個の3次元確率ベクトル X,Y,Z は

それぞれ独立で3変量正規分布 $N((0,0,0),I),N((3,3,0),I),N((3,3,6),I)$ に従うものとした。ここで I は単位行列である。X,Y,Z は各群に対応するものとし，データはこれらの確率ベクトルの実現値であると考える。実際のデータの作成は，共分散分散行列が単位行列であることより確率ベクトルのそれぞれの変量は平均は異なるが独立で分散1の正規分布に従うので，正規乱数を用いた。以上の手順により，3つの群に対してそれぞれの分布に従う10個の乱数データを作成した。

作成された乱数データに対して，バイプロット，非線形マッピング，修正非線形マッピングを適用した。ただしバイプロットでは，$(\sqrt{\lambda_1}p_1,\sqrt{\lambda_2}p_2)$ の値をプロットし，修正非線形マッピングでは，ウェイト w_{ij} のパラメータ k の値は1を使用した。

表1～表3は，それぞれの手法によって2次元空間に落とされたときの座標を示している。

表1．バイプロットによる座標

	1 VAR 1	2 VAR 2
1	-3.799	-1.584
2	-3.351	-1.997
3	-4.039	-2.059
4	-3.219	-3.624
5	-3.927	-1.383
6	-3.144	-4.131
7	-3.411	.761
8	-2.301	.546
9	-2.918	-.314
10	-3.271	-1.202
11	-.898	1.213
12	-1.010	2.015
13	-1.271	3.103
14	.590	1.385
15	.144	3.056
16	-.878	1.760
17	1.174	3.624
18	-1.975	2.402
19	-3.067	2.123
20	-.749	2.894
21	3.481	-2.674
22	5.151	-.447
23	4.651	-1.347
24	4.385	-.138
25	4.386	-2.549
26	3.828	-.681
27	4.763	-.914
28	4.933	.248
29	2.358	-.104
30	3.384	.017

表2．非線形マッピングによる座標

	1 VAR 1	2 VAR 2
1	-.561	-.597
2	-.093	-1.541
3	-1.016	-1.393
4	.046	-3.133
5	-1.278	-.351
6	.295	-3.610
7	-.152	1.728
8	1.015	1.402
9	.325	.708
10	.045	-.233
11	2.377	2.040
12	2.254	3.083
13	1.953	4.288
14	3.877	2.238
15	3.665	3.934
16	2.434	2.523
17	4.595	4.567
18	1.090	3.272
19	-.213	3.362
20	2.456	4.163
21	6.653	-1.841
22	8.874	.539
23	7.812	-.695
24	7.582	.776
25	7.589	-1.736
26	6.998	.182
27	8.054	.003
28	8.124	1.245
29	5.567	.732
30	6.585	.932

表3．修正非線形マッピングによる座標

	1 VAR 1	2 VAR 2
1	-.547	-.813
2	.250	-1.439
3	-.664	-1.154
4	.692	-3.040
5	-.677	-.638
6	.367	-3.336
7	-.155	1.771
8	1.043	1.463
9	.578	.440
10	.018	-.475
11	2.406	2.076
12	2.235	2.933
13	1.958	4.057
14	3.870	2.246
15	3.508	3.967
16	2.470	2.637
17	4.501	4.521
18	1.515	3.366
19	.700	3.334
20	2.488	3.858
21	6.707	-1.755
22	8.345	.437
23	7.893	-.444
24	7.608	.749
25	7.576	-1.632
26	7.014	.206
27	7.937	-.022
28	8.126	1.127
29	5.583	.773
30	6.567	.891

また，図１～図３は，表１～表３の座標をプロットしたものである。

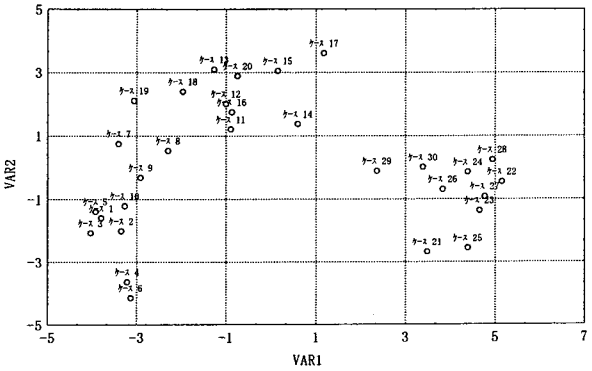

図１．バイプロットによるプロット

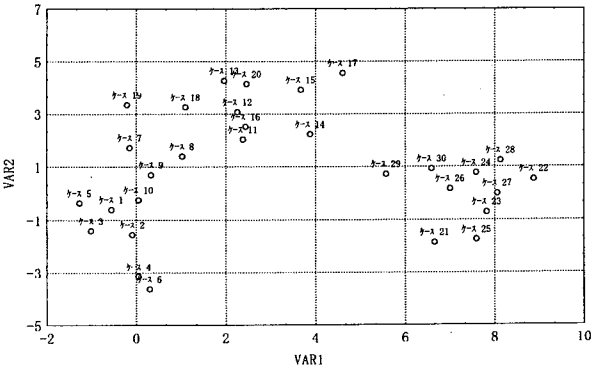

図２．非線形マッピングによるプロット

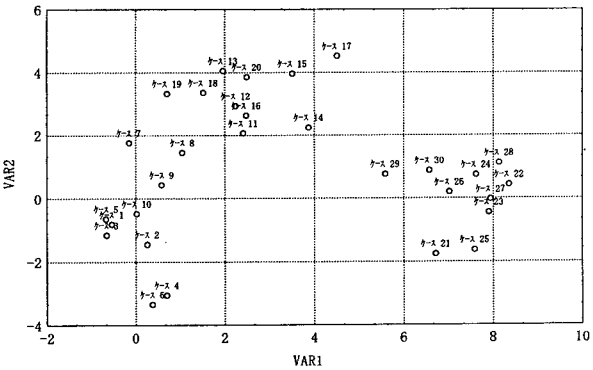

図３．修正非線形マッピングによるプロット

図１～図３から，全体のプロットの傾向は各手法で類似しているように見える。そこで，それぞれの違いをより詳細に調べるために，表１～表３のデータに対して，デンドログラムを描き，それらの比較を行った。デンドログラムは，距離係数としてユークリッド距離を用い，群平均法によって描かれた。図４はバイプロットによるデンドログラムを示しており，図５は非線形マッピングによるデンドログラムを示している。また，図６は修正非線形マッピングによるデンドログラムを示している。まず図４および図５を見ると，ケース７，ケース８，ケース９がケース11～ケース20ま

でのクラスターに連結していることがわかる。図1を見るとケース7～ケース9は2つの群の間にあるもので，どちらの群に判別されるかの判断が困難な位置に存在しているが，今回作成された3群の構造については正しい連結が得られているとは言えない。次に，図6を見ると，3つの群の構造が元の3次元のデータの構造を正しく表現できていることがわかる。これらの結果より，本論文で提案した修正非線形マッピングが今回のデータに対しては，最も正確に元の次元のクラスター構造を再現できていることが示された。

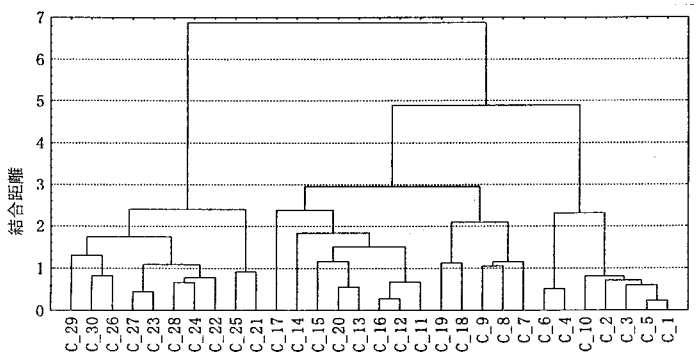

図4．バイプロットによるデンドログラム

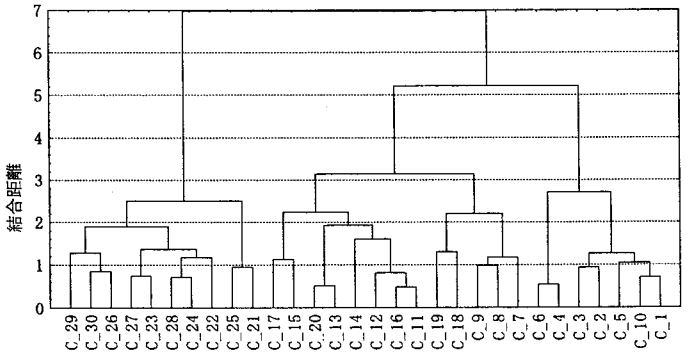

図5．非線形マッピングによるデンドログラム

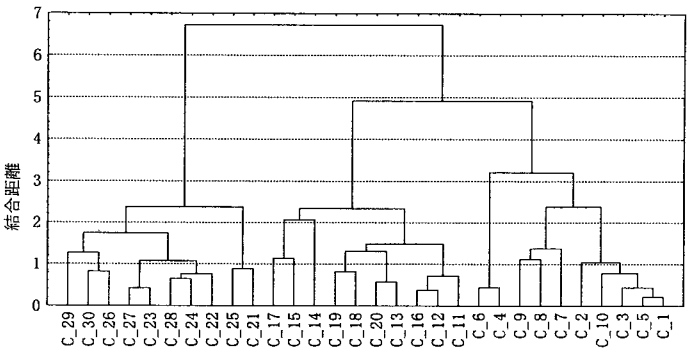

図6．修正非線形マッピングによるデンドログラム

4 ま と め

本論文では，多変量データを2次元空間にプロットする手法について検討した。これらの手法の最大の利点は多次元データの構造を2次元空間で視覚的にとらえられることである。多変量データのデータ解析においては，解釈の容易さと簡明さが重要となる。多次元データを2次元で表現することは，簡明さという利点を持っており，また，解釈も容易になる。その意味においても，オーディネーション手法の有効性について検討することは必要である。本論文では，バイプロットと非線形マッピングを取り上げ，修正非線形マッピングとそれら従来の手法との比較を行った。数値例の結果，今回用いたデータに対しては，修正非線形マッピングが，2次元空間上に最も正確に元の次元の構造を表現できていることが示された。

本論文で示した，非線形マッピングなどのオーディネーション手法で得られた2次元のデータを用いてデンドログラムを描くという方法は，データの構造を把握する上で非常に有効なものであるといえ，今後さらに詳細な検討が必要である。多次元データを2次元で表現したのでは，データの複雑な構造をとらえるためには十分とはいえない。今後，データの構造を保存するのに最適な次元数を決定する方法やその構造の可視化など，残された問題点も多い。また，記述主体のグラフ表現法を併用することにより，より詳細な解釈ができることが予想され，多変量データのクラスタリングへのグラフ手法の応用について系統的な研究を進める必要があるといえる。

参 考 文 献

〔1〕 Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically. J.Amer.Statist.Assoc., 68, 361-368.

〔2〕 Everitt, B. (1978). Graphical techniques for multivariate data. Gower Publishing Ltd.

〔3〕 Fukumori, M. and Tanaka, Y. (1994). Modified non-linear mapping and its application to the problem of detecting influencial subsets. Proceedings of the Fifth Japan-China Symposium on Statistics, 76-79.

〔4〕 Fukumori, M. (1995). Evaluation of multivariate graphs. Journal of Chugoku Junior College, 26, 53-66.

〔5〕 Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325-338.

〔6〕 Kruskal, J. B. (1964). Non-metric multidimensional scaling : a numerical method. Psychometrika, 29, 115-129.

〔7〕 Sammon, J. W. (1969). A non-linear mapping for data structure analysis. IEEE Trans. Computers, C18, 401-409.