

# INDSCALのデータサイエンスへの適用可能性

奥 喜正

## 1. はじめに

データ解析者の意志如何を問わず、IT時代においては大量のデータを容易に蓄積することが可能になってきた。そこで、蓄積された大量のデータを効率的に圧縮し、視覚的に表現する新たなデータサイエンス技術の開発が要請されている（柴田, 2001）。従来、これらの要請に対して探索的データ解析（Exploratory Data Analysis: EDA）がある程度はユーザに答えてきたが、昨今のデータベースの大型化によって EDA と併にデータサイエンス<sup>(注)</sup>、ことにデータマイニング手法を活用する必要性が増してきた（福田・森本・徳山, 2001）。それゆえ、統計的データ解析もデータサイエンスへの新たな貢献が求められる状況になってきている。現時点では統計的データ解析手法の中で、データサイエンスやデータマイニングへの貢献度はクラスター分析が圧倒的に大きい。

本稿では統計的データ解析や EDA で頻繁に使用される多変量データ解析手法の一つである多次元尺度法（Multi Dimensional Scaling: MDS）のなかで（Kruskal, 1964），特に 2 相 3 元データの解析手法である INDSCAL（INdividual Difference SCALing: 個人差多次元尺度法）を数年間の日本のプロ野球成績データに適用して、3 元多次元尺度法（3 元 MDS）のデータサイエンスへの有効活用を考察する（Carroll & Chang, 1970）。そして、3 元データ解析においては、2 元 MDS よりも 3 元 MDS の INDSCAL を利用したほうが、データ圧縮過程における情報損失が小さくて済むこと、データのもつ経時的変化を適確に捉えることが多いことを指摘する。特に、3 元データ解析でのデータ圧縮過程で、データサイエンスの重要課題であるデータ濃縮（Data Compression）を INDSCAL がうまく遂行する様子を示す。なお、INDSCAL の応用例としてはソリとアラビ（1979）のそれが教義的である。

さて、データマイニングと探索的データ解析は、それらの目的において似た性質を持つ。特に、データマイニングでは以下の点で特徴づけられよう。第一に、非常に大きな

データを対象にしていること、第二にはデータ収集においてコントロールがきかないこと、第三には新しい種類のデータパターンに注目することなどである（福田 他, 2001）。これらデータマイニングの特徴を鑑みると、従来の探索的データ解析手法や統計的データ解析手法の利用形態を幾つかの点で改めてからデータサイエンスでそれらを活用する必要性が生じる。そのような観点から、本稿では実際の統計的データ解析を通じてそれらの利用時の改良点を検討しつつ、データサイエンスへの3元MDSの新たなる利用形態について INDSCAL を使用して模索する。

## 2.1 多次元尺度法について

最初に、多次元尺度法をクルスカルの方法（Kruskal, 1964）を通じて説明する。多次元尺度法とは、対象間の心理的距離など非類似性データから、潜在する少数の次元や背後にある構造を明らかにしたいときに利用される多変量データ解析手法の一つである。対象  $i, j, k, l$  における非類似性データ  $\delta_{ij}, \delta_{kl}$  が得られたとすると、それらをユーザが視覚的に認識できる距離量  $d_{ij}, d_{kl}$  に以下のようにできるだけその大小関係（順序関係）のみを保存するように単調変換する。距離量とは三角不等式

$$d_{ik} \leq d_{ij} + d_{jk} \quad \text{for } \forall i, j, k,$$

を満足する非負の量であることに留意する。そして

$$\delta_{ij} < \delta_{kl} \Rightarrow d_{ij} \leq d_{kl} \quad \text{for } \forall i, j, k, l \quad (1)$$

のようにクルスカルの単調回帰を利用して、非類似性データ量を距離量に変換する（Kruskal, 1965）。各々の対象を点として表現する空間を布置と呼び、布置の次元数を T とすると対象  $i, j$  の次元  $t$  における座標が  $X_{it}, X_{jt}$  のように求められれば

$$\hat{\delta}_{ij} \stackrel{m}{=} \hat{d}_{ij} \approx d_{ij} = \sqrt{\sum_{t=1}^T (X_{it} - X_{jt})^2}$$

が成立することになる。ここで、記号  $\stackrel{m}{=}$  は単調増加関係を示し、 $\hat{d}_{ij}$  は条件式(1)を満足する擬似距離量で、得られた布置の  $d_{ij}$  から計算される量でディスパリティとも呼ばれる。ところで、ある次元 T（通常、T = 2, 3）で完全に式(1)を満足するような布置を得ることは一般には困難である。そこで、与えられた非類似性データ  $\delta_{ij}$  の順序関係と、求められた布置から計算される  $d_{ij}$  の順序関係がどの程度、一致しているかを評価する指標にストレス S と呼ばれる次の評価規準を導入する。

$$S = \sqrt{\sum_{i < j} \sum (d_{ij} - \hat{d}_{ij})^2} / \sqrt{\sum_{i < j} \sum d_{ij}^2}$$

順序関係が完全に保存されていれば、すなわち式(1)が成立すれば、 $S=0$ となり、その適合度が悪くなるにつれて $S$ の値は大きくなり、ストレス $S$ の値は0.2未満であることが最悪でも要求される。そうでないときは、該当する次元ではデータ行列 $[\delta_{ij}]$ に十分に適合する布置は得られないとみなして現在の次元を一次元上げて、改めて布置を求めるべきである。

## 2.2 INDSCAL とは

INDSCAL (INdividual Difference SCALing: 個人差多次元尺度法) は2相3元データを解析して、全体の被験者に共通するように対象を布置する「共通対象空間」と、被験者間の違い、すなわち、個人差を表示する「被験者空間」を同時に output するモデルである。被験者間の個人差を次元に与える重みづけを変化させて表現するモデルである。被験者 $k$ の対象 $i$ と $j$ の距離 $d_{ij,k}$ は

$$d_{ij,k} = \left[ \sum_{t=1}^T w_{kt} (X_{it} - X_{jt})^2 \right]^{1/2} \quad (2)$$

となる。 $w_{kt}$  は被験者 $k$  ( $k = 1, \dots, N$ ) が次元 $t$  ( $t = 1, \dots, T$ ) を重要視する程度を示す重みである。 $X_{it}$  と  $X_{jt}$  は共通対象空間の第 $t$ 次元における対象 $i$ と対象 $j$ の座標である。すなわち、 $X_{it}$  とは被験者全体に共通する共通対象空間での対象 $i$ の $t$ 座標であるのに対して、特定の個人 $k$ から対象空間を眺めた場合には対象 $i$ の布置の当該座標は  $\sqrt{w_{kt}} X_{it}$  となって、被験者 $k$ では座標軸 $t$ が  $\sqrt{w_{kt}}$  だけ伸縮されたことになる。INDSCAL は、被験者ごとに次元 $t$ に対して、 $\sqrt{w_{kt}}$  だけ $t$ 座標軸を引き伸ばしたり縮めたりすることで被験者間の認知個人差を表現する。

ところで、式(2)の形では $d_{ij,k}$ から、 $w_{kt}$ 、 $X_{it}$ 、 $X_{jt}$ を得ることはできない。そこで、距離 $d_{ij,k}$ を、対象 $i$ と対象 $j$ を表すベクトルの内積 $b_{ij,k}$ にダブルセンタリング (Double Centering) で変換すると式(2)は

$$b_{ij,k} = \sum_{t=1}^T w_{kt} X_{it} X_{jt} + e_{ij,k}$$

と表現できる。ここで、 $e_{ij,k}$ は誤差項である。正準分解 (Canonical Decomposition Analysis) を利用すれば、 $b_{ij,k}$ が与えられたときに $T$ の値を固定すると、 $b_{ij,k}$ に最小2乗的にあてはまりのよい $w_{kt}$ 、 $X_{it}$ 、 $X_{jt}$ が得られる。つまり

$$\sum_{i,j,k} \left( b_{ij,k} - \sum_{t=1}^T w_{kt} X_{it} X_{jt} \right)^2 \rightarrow \min$$

が成立するように NILS (Nonlinear Iterative Least Squares) を使用して計算できるこ

とが知られている。このように INDSCAL はモデル構築過程で内積を利用するのである。なお、INDSCAL の場合には適合度の規準、ストレス  $S$  は被験者数を  $m$  とすると

$$S = \left[ (1/m) \sum_k \left\{ \sum_{i,j} \left[ (d_{ij,k})^2 - (\hat{d}_{ij,k})^2 \right]^2 / \sum_{i,j} (d_{ij,k})^4 \right\} \right]^{1/2}$$

のように定義される。INDSCAL 以外の 3 元多次元尺度構成法には、ハーシュマンの PARAFAC-1 (PARAell FACTors) プログラムなどがある (Harshman, 1970)。

### 3. 分析方法

1998年度から2001年度までの4年間にわたるプロ野球チーム12球団の年間成績データ、すなわち、6個の対象である成績、勝率、打率、得点、失点、本塁打数、防御率に関するデータを MDS のクルスカルの方法と INDSCAL で解析して 6 個の対象の布置を得る。

一年間の成績データは、 $12 \times 6$  型の 2 相 2 元データ行列  $Z_i (i = 1, \dots, 4)$  を形成する。これをユークリッド距離を採用して、6 個の対象間の  ${}_6C_2$  個の非類似性データを計算し単相 2 元データの  $6 \times 6$  型非類似性行列  $\Delta$  を形成する。4 年間にわたりデータ集計したので 4 個の非類似性行列  $\Delta_1, \Delta_2, \Delta_3, \Delta_4$  が作成される。さらに、これら 4 個の非類似性行列  $\Delta_k (k = 1, \dots, 4)$  を縦に並べると、2 相 3 元データの  $24 \times 6$  型の非類似性行列  $D = [d_{i,j,k}]$  が形成される。

このデータ行列  $D$  を INDSCAL で解析して、2 次元の共通対象空間  $X$  と被験者空間、 $W$  そして適合度の指標であるストレス  $S$  を得る。また、比較検討のために  $24 \times 6$  型の非類似性行列  $D$  をクルスカルの方法でも解析して 2 次元布置  $X'$  を得ることにした。他方、4 個の非類似性行列  $\Delta_k$  をそれぞれクルスカルの方法で解析して 2 次元の 4 個の布置  $X_1, X_2, X_3, X_4$  を得る。今回のデータ解析では、布置の次元はどの解析手法の場合でも一様に 2 に固定する。

このように本稿では同じデータに、3 種類の多次元尺度法によるデータ解析手法を適用してその布置、結果の相違点を比較検討し、データ圧縮過程における手法ごとの情報損失の違いをデータ濃縮という観点に留意して考察する。

### 4. 結果

各年度の成績データからクルスカルの方法によって 6 個の対象、すなわち、打率、勝率、得点、失点、本塁打数、防御率についての布置が 4 年間分の 4 個得られた (図 1, 2, 3, 4)。それぞれのストレス値は、0.107, 0.0605, 0.0843, 0.0154 であったので 2 次元布置の適合度に問題はなかった。図 1 から図 4 までを眺めると 6 個の対象の位置は

### INDSCAL のデータサイエンスへの適用可能性

年毎にかなり変位している様子が観察され、一年間のデータだけでは安定した布置は得られないことがわかった。それゆえ、4年間の成績データから計算される非類似性行列  $D$ に基づいて布置を求めることが必要になったので、INDSCALでそれをデータ解析した。INDSCALによる共通対象布置、被験者空間およびクルスカルの方法によって得られた布置を図5、図6および図7に示す。ストレス値はそれぞれ0.146および0.157で、適合度は良好とは言えなかった。INDSCALでの本稿の被験者空間  $W$ という出力は、プロ野球成績データの年時変動、すなわち、時系列的変化を捉えたものと理解できた(図6)。また、出力  $W$ から1998年度、1999年度、2001年度では次元2よりも次元1を重視することが判明した。このように3元データ圧縮過程でデータ濃縮という課題をINDSCALは首尾よく遂行する様子が窺えた。

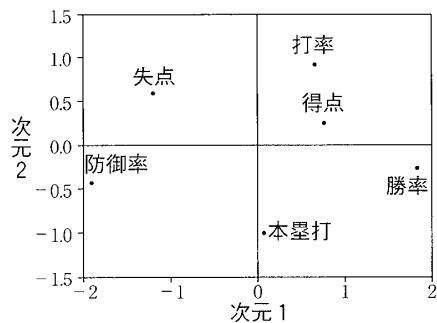


図1. クルスカルの方法による1998年度のプロ野球成績の布置

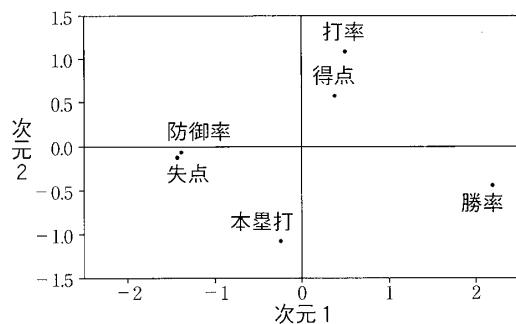


図2. クルスカルの方法による1999年度の布置

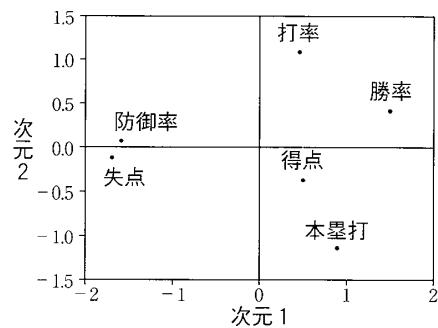


図3. クルスカルの方法による2000年度の布置

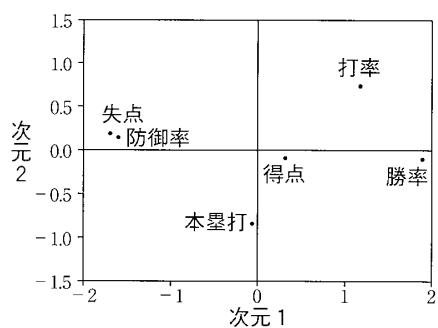


図4. クルスカルの方法による2001年度の布置

表1. INDSCAL の場合の各年度の布置のストレス値と平均ストレス値

行 列	ストレス
1998	0.190
1999	0.180
2000	0.105
2001	0.078

Averaged over matrices Stress=0.14630

表2. INDSCALによる被験者行列Wの出力

年度	次元1	次元2
1998	0.769	0.410
1999	0.855	0.273
2000	0.645	0.715
2001	0.941	0.278

Overall importance of each dimension: 0.656 0.278

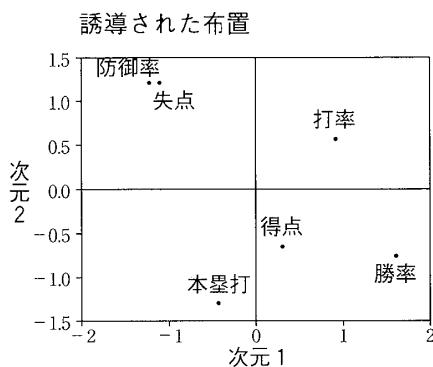


図5. INDSCALによる共通対象空間Xの2次元布置

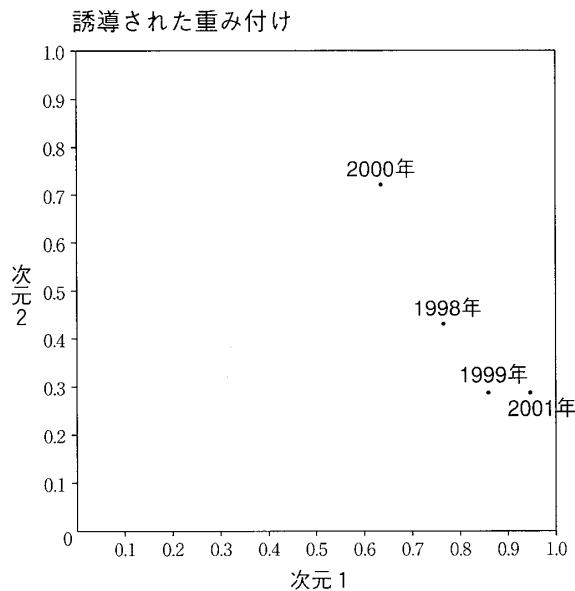


図6. INDSCALによる被験者空間W

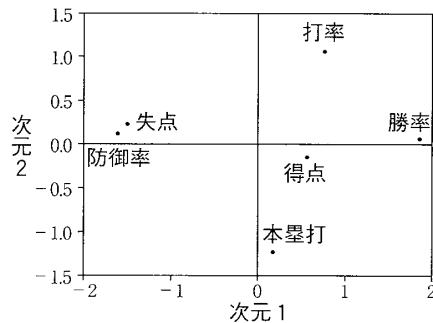


図7. クルスカルの方法による4年間での対象布置

## 5. 考 察

図5, 6, 7によれば、2相3元データをクルスカルの方法で解析した場合に比べて INDSCAL を利用した場合のほうが、被験者空間 W が出力されるだけ圧縮過程における情報損失を減らせることが示された。すなわち、2相3元データのもつ経時的変動を INDSCAL は被験者空間 W で捉えたわけである。それとともに、INDSCAL が時系列的な3元データを解析するのに適切な方法であることも窺えた。データ圧縮過程でデータサイエンスが目指すデータ濃縮、データ集約という作業に（柴田, 2001）、INDSCAL が貢献して有用な情報を形成している様子が理解できた。

ところで、一年ごとのデータにクルスカルの方法をそれぞれ適用して得た、6個の対象に関する4個の布置は（図1, 2, 3, 4），年毎に対象の布置がかなり変動していることが読み取れた。しかるに、6個の対象の相互位置関係を表現する「安定した布置」を得ることが多次元尺度法の適用目的であるとすれば、安定した布置を得るためにには一年間のデータだけでは情報不足であることが示唆された。これに対して、4年間のデータから成る2相3元データを INDSCAL によって解析した結果の共通対象空間 X は、6個の対象の、より安定した布置となって、INDSCAL は対象の安定した布置を供給したといえよう。同時に、INDSCAL の被験者空間出力 W は年毎の布置変動を示すものであるが、その時系列的な布置変動を偶然的変動とみなすべきか、それとも年毎のプロ野球成績の特徴を表すものであるかを判断することは、モデル作成者の立場に依存する。

INDSCAL のモデリングは年毎の布置変動を軸の重みの変化として捉えて、プロ野球成績の年度毎の特徴とみなして個人差、この場合には年度差を被験者空間 W で考慮するモデル構築の立場を探っている。これに対して成績の年度差を偶然的変動とみなす場合には、4年間のデータを全体として総括してから、クルスカルの方法を一回利用した

## INDSCALのデータサイエンスへの適用可能性

データ解析を実行すればよいことになる。この場合には成績の年度差は考慮されないから、データの経時的変動を無視することを覚悟しなければならない。しかしながら、集計されたデータが経時的性質を伴う場合には、時系列的な変化を無視しないで時系列的変化を軸に対する重みの変化で捉える INDSCAL のようなモデルを使用してデータ解析を行うほうがデータ圧縮時の情報損失を抑えることができるところが本稿の分析例で窺えた。このような被験者空間  $W$  で捉えるというデータの時系列変化の捉え方は、データサイエンスが強調するデータ濃縮という事項を INDSCAL が少なからず試みているとみなせる。

ところで、データサイエンスやデータマイニングの特徴を鑑みると、従来の統計的データ解析手法をデータマイニングの枠組みで使用するときには、その使用形態を工夫する必要がある。例えば、データマイニングでクラスター分析を実行するときは、完全に  $k$  個のクラスターに分別する必要はなく、おおよその数のクラスターに分別すれば十分であって、これは例外的なデータを排除することによってその目的を達成することができる。また、不要な属性の排除も重要である。このようにデータマイニングにおいてクラスター分析を使用するときは、例外的なデータの排除や不要属性の排除要求などを福田らは（2001）、利用する場合の改良点として要請した。同様にデータマイニングにおいて 3 元多次元尺度法を利用する際も、その工夫改良点が今後、検討されるべきである。3 元 MDS は、データ濃縮やデータ集約ということを遂行できる可能性があるので、従来以上にデータサイエンスという枠組みで使用されるべきであると筆者は考えている。

本稿は、従来はそれほど利用されなかった 3 元多次元尺度法のデータサイエンスへの適用が、クラスター分析と同様に有効であることを具体的なデータに INDSCAL を利用して指摘した。特に当該モデルが 3 元データに対して安定した布置を供給するとともに、データ濃縮という事項と関連するデータの時系列変動を適切に捉えたことを示して、その有用性を強調した。

### 謝 辞

稿を終えるにあたり、流通経済大学の内藤松幸君には本研究に対して甚大なる御協力を頂きましたので、ここに深謝いたします。

## 注

データサイエンスとは、データ取得段階のサンプリングから、データ解析やモデリング、検証、モデルによる問題解決という最終的な結論を提出するに至る過程まで、すなわち、データの上流から下流までをトータルに科学する学問分野をいう（柴田、2001）。情報科学とは密接に関連するがデータサイエンスは、ある特定の目的に沿って体系化された、具体的な情報であるデータを対象にする。データ圧縮時に、どのように濃縮したらその解析目的に有用な情報になり得るかというデータ濃縮、データ集約に興味が注がれる。

## 参考文献

- Arabie, P., Carroll, J. D. & Desarbo,W.S. (1987). Three-way Scaling and Clustering. Newberry Park, CA: Sage.
- 岡太・今泉（共訳）(1990). 3元データの分析，共立出版
- Borg, I and Lingoes, J.C. (1978). What Weights should Weight have in Individual Differences Scaling? *Quality and Quantity*, **12**, 233 – 237.
- Carroll, J.D. and Chang, J.J. (1970). Analysis of Indivisual Differences in Multi Dimensional Scaling via an N-way Generalization of Eckart-Young Decomposison. *Psychometrika*, **35**, 283 – 319.
- Everitt, B.S. & Rabe-Hesketh, S. (1997). The Analysis of Proximity Data. Kendall's Library of Statistics 4. London: Arnold.
- Fayyad, U.M., Piatesky-Shapiro G., Smyth P. & Uthurusamy E. (1996). Advances in Knowledge Discovery and Data Mining. The MIT Press.
- Harshman, R.A. (1970). Foundations of the PARAFAC Procedure: Models and Conditions for an Explanatory Multi-model Factor Analysis. UCLA Working Papers in Phonetics, **16**, 1 – 84.
- Kruskal, J.B. (1964). Nonmetric Multidimensional Scaling: Numerical Method. *Psychometrika*, **29**, 28 – 42.
- Kruskal, J.B. (1965). Analysis of Factorial Experiments by Estimating Transformations of the Data. *Journal of the Royal Statistical Society Series B*, **27**, 251 – 263.
- Soli, S.D. & Arabie, P. (1979). Auditory versus Phenetic Accounts of Observed Confusions between Consonant Phonemes. *Journal of the Acoustical Society of America*, **66**, 46 – 59.
- 柴田 里程 (2001). データリテラシー. 共立出版
- 福田 剛志, 森本 康彦, 徳山 豪 (2001). データマイニング. 共立出版