

# Rと多次元尺度法

## 1. 多次元尺度法とは

多次元尺度法 (MDS: multi-dimensional scaling) は、個体間の親近性データを、2次元あるいは3次元空間に類似したものを近く、そうでないものを遠くに配置する方法で、データの構造を考察する方法である。

多次元尺度法は計量多次元尺度法と非計量多次元尺度法に大別される。計量多次元尺度法とは距離データを低次元に配置する方法で、非計量多次元尺度法は、順序尺度のデータの類似度あるいは距離に変換可能な親近性データを低次元に配置する方法である。

MDS にも多くのアルゴリズムが提案されているが、古典的多次元尺度法としては 1950 年代 Torgerson の貢献が大きい[4]。

多次元尺度法をイメージ的に説明するため、近畿地方の地図を図 1 に示す。図 1 では兵庫から和歌山、大阪、奈良、滋賀、京都の距離を点線で示している。このような任意の 2 点間の距離を表 1 に示す。

計量多次元の尺度法では、表 1 のような距離データから各点の座標を求め、元のデータ構造を再現することを課題としている。

例えば、表 1 から何らかの方法で求めた 2 次元の(横と縦)座標値が表 2 に示すとおりであるとする。表 2 のデータの散布図を図 2 に示す。図 2 は図 1 の地域間の相対位置関係を再現しているものである。

多次元尺度法は、データから距離(あるいは類似度)を求め、そのデータに基づいて 2、3 次元空間上の各点(個体)の座標値を求め、視覚的に

その相対関係を考察する。

図 1 近畿地方の地図

(図 1 および表 1 の作成はフリーソフト kenMap ver 8.0 を用いた。)



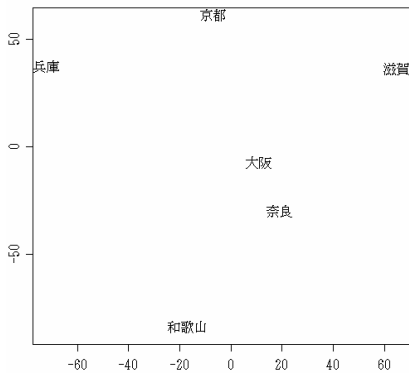
表 1 図 1 の点の間の距離(単位は km)

	兵庫	和歌山	大阪	奈良	滋賀	京都
兵庫	0	132	85	116	136	60
和歌山	132	0	68	68	144	142
大阪	85	68	0	75	32	83
奈良	116	68	75	0	79	95
滋賀	136	144	32	79	0	61
京都	60	142	83	95	61	0

表 2 表 1 から求めた点の 2 次元座標値

	横軸	縦軸
兵庫	-71.9	35.1
和歌山	-16.8	-85.9
滋賀	65.0	33.8
京都	-6.7	58.6
奈良	19.3	-32.1
大阪	11.0	-9.5

図2 表2の散布図



$$ed_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$$

$$cd_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|$$

$$md_{ij} = \sqrt[p]{\sum_{k=1}^n |x_{ik} - x_{jk}|^p}$$

距離は自由に定義できるが次の距離の公理を満たさなければならない。

$$d_{ij} \geq 0, \quad d_{ij} = d_{ji}, \quad d_{ij} + d_{jk} \geq d_{ik}$$

## 2. 距離と類似度

表3のように、 $m$ 個の研究対象(個体)に対し、 $n$ 個の項目に分けたデータがあるとする。

表3 データ  $X_{m \times n}$

	$x_1$	$x_2$	...	$x_k$	...	$x_n$
個体1	$x_{11}$	$x_{12}$	...	$x_{1k}$	...	$x_{1n}$
個体2	$x_{21}$	$x_{22}$	...	$x_{2k}$	...	$x_{2n}$
...	...	...	...	...	...	...
個体 <i>i</i>	$x_{i1}$	$x_{i2}$	...	$x_{ik}$	...	$x_{in}$
...	...	...	...	...	...	...
個体 <i>j</i>	$x_{j1}$	$x_{j2}$	...	$x_{jk}$	...	$x_{jn}$
...	...	...	...	...	...	...
個体 <i>m</i>	$x_{m1}$	$x_{m2}$	...	$x_{mk}$	...	$x_{mn}$

表3のデータが量的データの場合は、何らかの距離の定義で表4のような距離マトリクスを求めることができる。

個体*i*と個体*j*との間の距離を  $d_{ij}$  で表すことにする。

$$d_{ij} = \|\text{個体}i - \text{個体}j\|$$

距離の中で最も広く知られているのは、ユークリッド距離(ed: Euclidean distance)、市街距離(cd: city-block distance)である。市街距離はマンハッタン(manhattan)距離とも呼ぶ。これらの距離は、ミンコフスキー距離(md: Minkovski distance)で一般化できる。

表4 データ行列  $D_{m \times m}$

	個体1	個体2	...	個体 <i>j</i>	...	個体 <i>m</i>
個体1	0	$d_{12}$	...	$d_{1j}$	...	$d_{1m}$
個体2	$d_{21}$	0	...	$d_{2j}$	...	$d_{2m}$
...	...	...	...	...	...	...
個体 <i>i</i>	$d_{i1}$	$d_{i2}$	...	$d_{ij}$	...	$d_{im}$
...	...	...	...	...	...	...
個体 <i>m</i>	$d_{m1}$	$d_{m2}$	...	$d_{mj}$	...	0

距離の公理から分かるように、距離は対称性を持っているので、表4に示したように距離マトリクスは対称である。よって、距離のマトリクスは対角線の下(あるいは上)の半分のみを用いればよい。

距離の測度と逆の概念としては類似度がある。距離は値が小さいほど個体間の関連性が強いと判断するが、類似度は値が大きいほど個体間の関連性が強いと判断する。質的データの場合は、距離より類似度が多く用いられている。

類似度の測度として最も知られているのはピアソン相関係数(r)である。また、パターン類似率(ps: pattern similarity)も多用されている。

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^n (x_{jk} - \bar{x}_j)^2}}, \quad -1 \leq r_{ij} \leq 1$$

$$ps_{ij} = \frac{\sum_{k=1}^n x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^n x_{ik} \sum_{k=1}^n x_{jk}}}, \quad 0 \leq ps_{ij} \leq 1$$

これらの類似度は、次のような変換で、表 4 に示すような距離構造に変換することが可能である。

$$rd_{ij} = 1 - r_{ij}, \quad pd_{ij} = 1 - ps_{ij}$$

変数の値が 2 値(0, 1)である場合の距離をバイナリ距離と呼ぶ。

### 3. 距離関数と MDS 関数

多次元尺度分析は、距離(あるいは類似度)を求めることから始まる。

#### (1) 距離関数

R の幾つかのパッケージに距離を求める関数実装されている。距離はしばしば非類似度(dissimilarity)とも呼ばれている。R をインストールする際、自動的にインストールされ、読み込みの手続きを必要としないパッケージ stats に距離を求める関数 **dist** がある。関数 **dist** では、"euclidean"、"maximum"、"manhattan"、"canberra"、"binary"、"minkovski"距離を求めることができる。

パッケージ **mvpart** には関数 **gdist**、パッケージ **vegan** には関数 **vegdist**、パッケージ **ade4** には **dist.binary**、**dist.dudi**、パッケージ **cluster** には関数 **daisy** などの距離を求める関数がある。

#### (2) MDS 関数

##### ① 計量多次元尺度法の関数

R のパッケージ stats には、計量多次元尺度法の関数 **cmdscale** (Classical (Metric) Multidimensional Scaling) がある。この古典的多次元尺度法は、主座標分析(principal

coordinate analysis)とも呼ばれている[2]。古典的多次元尺度法に用いる距離がユークリッド距離である場合は、相関行列を用いた主成分分析と等価である。

古典的多次元尺度法は、求めた距離マトリクス  $D_{m \times m}$  を、次のような変換を施したマトリクス  $Z_{m \times m}$  の固有ベクトルを点の座標値とする。

$$z_{ij} = \frac{1}{2} \left( \sum_{i=1}^m \frac{d_{ij}^2}{m} + \sum_{j=1}^m \frac{d_{ij}^2}{m} - \sum_{i=1}^m \sum_{j=1}^m \frac{d_{ij}^2}{m^2} - d_{ij}^2 \right)$$

関数 **cmdscale** の書き式を次に示す。

```
cmdscale(d, k = 2, eig=FALES...)
```

引数 **d** は距離構造のデータで、関数 **dist** が返す結果の構造である。表 4 のような対称行列の場合は関数 **as.dist** を用いて変換することができる。

引数 **k** は次元数で、デフォルトでは 2 に設定されている。引数 **eig** は、固有値を返すか否かを指定する。デフォルトでは **FALES** になっているので固有値を返さない。引数 **eig=TRUE** にすると座標値は **\$points** に記録される。

データを用いてその使用方法を示すことにする。R のパッケージ **datasets** にはヨーロッパ主な 21 都市間の距離データ **eurodist** がある。データ **eurodist** を用いてまず各都市の座標値を求め、その座標値に基づいた 2 次元の配置図を作成するコマンドとその結果を次に示す。

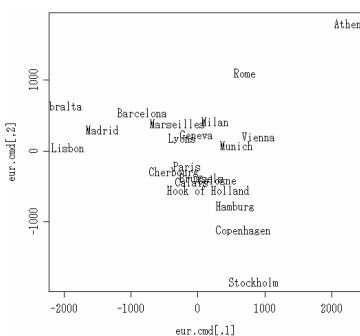
```
>(eur.cmd<-cmdscale(eurodist))
<返された結果は省略>
>plot(eur.cmd,type="n")
>text(eur.cmd,rownames(eur.cmd))
```

同じ方法で、表 1 の近畿地方の距離データを用いて確かめてみるのもよい。ただし、多次元尺度法で求めた座標値は相対的なものであるの

で、地図と見やすく対応付けるためには、軸の回転などを必要とする場合がある。

多次元尺度法の配置図における点の間の距離  $\hat{d}_{ij}$  は、用いた距離  $d_{ij}$  の推測値である。よって、その間には誤差がある。この点では、多次元尺度法を回帰分析のように「当てはめる」視点で扱うことができる。当てはまりの良さは用いた両距離のマトリクスと座標値の距離との相関係数を用いて考察することができる。 $\hat{d}_{ij}$  と  $d_{ij}$  の相関係数 2 乗値は約 0.97 であることから、多次元尺度法による 2 次元の距離の再現は大きい歪みがないと言えるであろう。

図 3 関数 cmdscale によるヨーロッパの 21 都市の 2 次元配置図

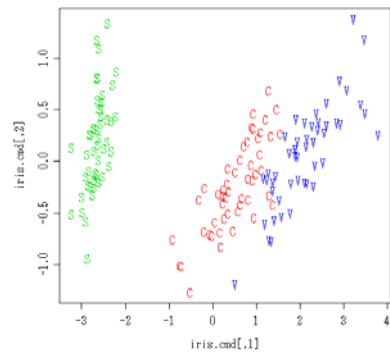


上記の例では、既存の距離データを用いているが、何らかのデータから距離を求めて関数 cmdscale を用いることも可能である。

次に iris のデータから、関数 dist を用いてユークリッド距離を求め、関数 cmdscale による多次元尺度法のコマンドおよびその結果を示す。

```
> iris.dist <- dist(iris[, -5]) #ラベルの部分を除く！
> iris.cmd <- cmdscale(iris.dist)
> plot(iris.cmd, type="n")
> iris.lab <- factor(c(rep("S", 50), rep("C", 50), rep("V", 50))) #iris[, -5]の長い記号列を略する
> text(iris.cmd, labels=iris.lab, col=unclass(iris.lab))
```

図 3 iris のユークリッド距離の多次元尺度図



## ② 非計量多次元尺度法

計量多次元尺度法では、比率尺度のデータにおける個体間の親近性を距離として計測し、距離データをユークリッド空間上で個体を配置することを前提としているが、人文社会学などで得られたデータには間隔尺度が多く、その親近性のデータを距離構造のデータに変換したとしても計量多次元尺度で用いた距離とは意味が異なる。質的データの解析をも視野に入れ、計量多次元尺度法を発展させたものが非計量多次元尺度法である。

非計量多元尺度法には、幾つかのアルゴリズムが提案されている。非計量多次元尺度で扱うデータ構造は、基本的には計量多次元尺度法と同じく距離構造のデータである。

非計量多次元尺度法では、基本的には個体間の距離  $d_{ij}$  を配置すべく  $k$  次元における距離  $\hat{d}_{ij}$  との差の 2 乗和が最小になるような座標値を求める。

カルスカル (Kruskal) は、次に示すストレス (stress) と呼ばれる統計量を最小にするように座標を定めることを提案している。

$$STRESS1 = \sqrt{\frac{\sum \sum (d_{ij} - \hat{d}_{ij})^2}{\sum \sum d_{ij}^2}}$$

この値が小さければ小さいほど、当てはまりが良いと判断する。この値を用いたカルスカルの評価目安を次の表 5 にします[3]。

表 5 ストレスによる当てはまりの評価

ストレス値	評価
0.2	悪い (poor)
0.1	まずまず (fair)
0.05	よい (good)
0.025	すばらしい (excellent)
0.00	完璧 (perfect)

R には幾つかのパッケージに非計量多次元尺度法の関数を実装されている。その中で最も基本となるのは、パッケージ MASS の中の関数 **isoMDS** と **sammon** である。

関数 isoMDS は、次に示すストレスを最小になるように座標を決める。

$$STRESS2 = \frac{\sum \sum [\theta(d_{ij}) - \hat{d}_{ij}]^2}{\sum \sum \hat{d}_{ij}^2}$$

関数 isoMDS の書き式を次に示す。

isoMDS(d, k = 2, ...)

引数 d は、距離構造を持つデータマトリクスで、k は次元の数である。これ以外にも幾つかの引数がある。関数 isoMDS は、座標値 (\$points) と最終のストレス (\$stress) を返す。ストレスは百分率になっている。

関数 sammon は、与えられた k 次元上で、次に示す重みつきストレスを最小化することで座標値を求めるアルゴリズムである。

$$STRESS3 = \frac{1}{\sum \sum_{i \neq j} d_{ij}} \sum \sum_{i \neq j} \frac{(d_{ij} - \hat{d}_{ij})^2}{d_{ij}}$$

関数 sammon の書き式は基本的には、関数

isoMDS と同じである。返される結果も isoMDS と同じく 2 項目である。ただし、ストレス (\$stress) は百分率ではない。

関数 isoMDS、sammon は初期に与えた座標値を初期値とし、用いたストレス統計量が最小になるように計算を繰り返す。デフォルトには、関数 cmdscale が指定されている。計算の繰り返しの回数は、引数で調整できる。繰り返しの回数のデフォルト値は、関数 sammon では niter = 100、関数 isoMDS では maxit = 50 になっている。

また、パッケージ vegan には非計量多次元尺度法の関数 **metaMDS** がある[1]。田口らが提案した多次元尺度構成法の R プログラム (NMDS) も公開されている[5]。

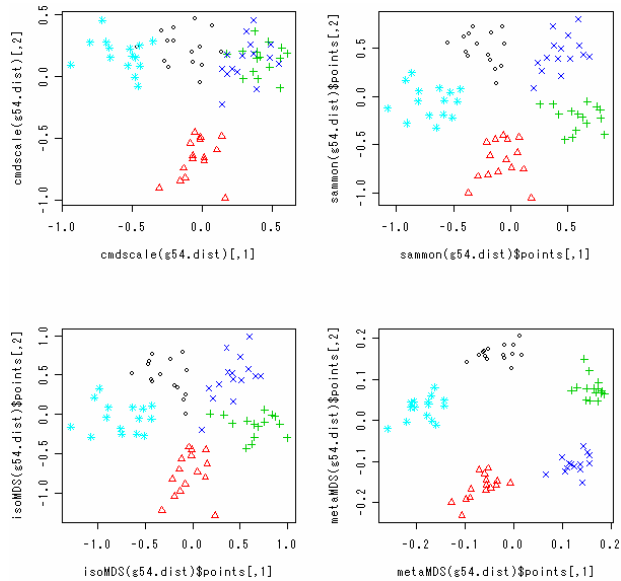
主成分分析、計量多次元尺度法 (cmdscale)、非計量多次元尺度法 (sammon、isoMDS、metaMDS) を同じデータセットと各関数のデフォルト値に基づいたコマンドの使用例およびその 2 次元の配置図を次に示す。

用いたデータは、Edwards and Oman が作成した 75 個体 5 変数の人工データである。ただし、5 変数の中の第 5 列は個体が所属するグループのラベルである。合計 5 グループで、それぞれのグループは 15 個体によって構成されている。データを作成するプログラムは、参考文献から取得できる[1]。

```
> par(mfrow=c(2,2))
> P<-c(1,2,3,4,8)[unclass(g54[,5])]
> C<-unclass(g54[,5])
> g54.dist<-dist(g54[, -5])
> library(MASS) #sammon と isoMDS を用いるため
> plot(cmdscale(g54.dist),pch=P,col=C)
> plot(sammon(g54.dist)$points,pch=P,col=C)
> plot(isoMDS(g54.dist)$points,pch=P,col=C)
> library(vegan) #metaMDS を用いるため
```

```
> plot(metaMDS(g54.dist)$points,pch=P,col=C)
```

図4 6種類の2次元の配置図



- [1] Edwards, J. and Oman, P (2003): Dimensional Reduction for Data Mapping—A practical guide using R, R News, Vol. 3/3, 2-7. [http://cran.r-project.org/doc/Rnews/Rnews\\_2003-3.pdf](http://cran.r-project.org/doc/Rnews/Rnews_2003-3.pdf)
- [2] Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, Vol. 53, 325-328.
- [3] Kruskal, J. B. (1964): Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, Vol. 29, 1-27.
- [4] Torgerson, W. S. (1958): Theory and Methods of Scaling. New York, Wiley.
- [5] 田口善弘・大野克嗣・横山和成(2001):非計量多次元尺度構成法への期待と新しい視点、統計数理、第49巻第1号 133-153(R用のプログラム nMDS は現時点では次のサイトから入手可能。[http://tag.cocolog-nifty.com/tags\\_diary/](http://tag.cocolog-nifty.com/tags_diary/))