Rによる多次元尺度分析 2014年１月

（金明哲　Rによるデータサイエンス、第４章、pp97～106）

多次元尺度法(Multi-Dimensional Scaling, MDS)：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。

（データの個体間の類似度、あるいは距離を求め、そのデータを２～３次元平面上にプロットする方法でデータの構造やパターンの形成などを察する方法。

距離と類似度

ｎ個の研究対象（個体）、ｐ個の項目データXn\*pがあるとする。データが量的データの場合は、何らかの距離の定義で距離行列を求めることができる。（ユークリッド距離、市街距離（マンハッタン距離）、ミンコフスキー距離）

個体iと個体ｊの距離をｄijとすると、

 dij = ||個体i －個体j｜｜

ユークリッド距離

$$ed\_{ij}=\sqrt{\sum\_{k=1}^{p}(x\_{ij}-x\_{jk}})2$$

類似度の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も類似度が多く用いられる。）

ピアソン相関係数ｒ

パターン類似度ps $ps\_{ij}=\frac{\sum\_{k=1}^{p}(x\_{jk}-\overbar{x}\_{j})(x\_{jk}-\overbar{x}\_{j})}{\sqrt{\sum\_{k=1}^{p}x^{2}\_{jk}\sum\_{k=1}^{p}x^{2}\_{jk}}}$, 0≦psij≦1

類似度と距離の変換

　　　　rdij = 1 – rij, pdij=1-psij

Rで利用可能な距離関数　stat パッケージのdist()関数など

計量的MDS

分析プロセス　１．距離を求める

　　　　　　　２．座標軸を求める

　　　　　　　３．２～３次元上で個体を配置する（散布図を作成する）

　　　　　　　４．信頼性について考察する

古典的MDS　計量MDSの関数 cmdscale(d, k=2, eig=F )

（例）iris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MDS分析

> iris.dist<-dist(iris[,-5])

> head(iris.dist)

[1] 0.5385165 0.5099020 0.6480741 0.1414214 0.6164414 0.5196152

> iris.cmd<-cmdscale(iris.dist)

> head(iris.cmd)

 [,1] [,2]

[1,] -2.684126 0.3193972

[2,] -2.714142 -0.1770012

[3,] -2.888991 -0.1449494

[4,] -2.745343 -0.3182990

[5,] -2.728717 0.3267545

[6,] -2.280860 0.7413304

> plot(iris.cmd, type="n")

> iris.lab<-factor(c(rep("S", 50), rep("C", 50), rep("V", 50)))

> text(iris.cmd, labels=iris.lab, col=unclass(iris.lab))



※ユークリッド距離を用いたcmdscaleのこの結果は、主成分分析の主成分得点の散布図と等価である。(Gower, J.C. 1966, Biometrika, 53, 325-328)

> iris.pca<-princomp(iris[,-5])

> plot(iris.pca$scores, type="n")

> text(iris.pca$scores, label=iris.lab, col=unclass(iris.lab))



文体特性項目データの主成分分析

> style.pca <- princomp( Zscore.df[, 3:12])

> plot(style.pca$score, type="n")

> author.lab<-strtrim(Zscore.df$author, 2)

> text(style.pca$score, label=author.lab)



文体特性項目データのMDS多次元尺度法

> style.dist<-dist( Zscore.df[, 3:12] ) # デフォルトではmethod=”euclidian”, p=2

> style.cmd <- cmdscale( style.dist )

> plot( style.cmd, type="n" )

> author.lab<-strtrim(Zscore.df$author, 2)

> text(style.cmd, label=author.lab)



特徴的な作家のみ表示

> style.dist<-dist( Zscore.df[, 3:12] )

> style.cmd <- cmdscale( style.dist )

> plot( style.cmd, type="n", xlim=c(-60, 60), ylim=c(-40, 40) )

> author.lab<-strtrim(Zscore.df$author, 1)

> text(style.cmd[c(8:19, 30:41, 54:63, 68:77, 90:102)], label=author.lab[c(8:19, 30:41, 54:63, 68:77, 90:102)])

コンラッド　8:19

ドイル　　　30：41

ヘンリ・ジェイムズ　54：63

メルヴィル　68：77

ウッドハウス　90：102

# コンラッド作品のみ赤で表示する

> # コンラッドデータ(8:19)を赤で表示する

> col.conrad <- rep( c( 1,2,1 ), c( 7, 12, 83 ) )# 表示色を表すベクトルを作成

> text(style.cmd, label=author.lab, col=col.conrad )



