多次元尺度構成法についての研究
～主成分分析法との比較～

2004MM021　堀井　里佳子
指導教員: 松田　眞一

1　はじめに

データを出来るだけ損失を少なくし、そのデータの持つ意味を解析する方法をたくさん学んできた。しかし、同じデータを使っても、解析法が異なれば結果も異なる場合がある。どこが違うから同じデータを使っても結果が異なるのかを追求してみたくなった。そこで本研究では多次元尺度構成法の理論を学ぶとともに今まで何度も使用した主成分分析法と多次元尺度構成法の比較を行っていきたいと思う。

2　多次元尺度構成法の理論

2.1　計量的多次元尺度構成法

ここでは、Torgerson法について述べる。このTorgerson法では、Young-Householderの定理の概念を用いて空間への配置を行う。Young-Householderの定理とは、$(m-1)$次の対称行列$C=(c_{jk})$を、

$$c_{jk}=\tfrac{1}{2}(s_{jm}^2+s_{km}^2-s_{jk}^2)$$

と、定義する。この、$(m-1)$次の対称行列$C=(c_{jk})$が非負定符号かつrank$C=r$ならば、その行列s_{jk}はr次元ユークリッド空間における2点jとkの距離とみなすことができるという定理である。類似度データs_{jk}がこの定理を満たすなら、s_{jk}はr次元ユークリッド空間で表された距離行列となる。ここで、r次元ユークリッド空間におけるm個の点の配置を$m\times r$の座標行列$X=(x_{jt})$、点jの座標ベクトルを$\mathbf{x}_{(j)}=(x_{j1},x_{j2},\ldots,x_{jr})^{'}$、重心ベクトルを$\bar{x}$とすると、$\mathbf{X}=(x_{(1)},x_{(2)},\ldots,x_{(m)})^{'}$、$\bar{x}=\frac{1}{m}\sum_{j=1}^{m}x_{(j)}$となり、以下の式が成立する。

$s_{jk}^2=\sum_{t=1}^{r}(x_{jt}-x_{kt})^2=s_{oj}^2+s_{ok}^2-2x_{(j)}^{'}x_{(k)}$
(s_{oj}、s_{ok}はそれぞれ点j、kまでの距離)

ここで、$\mathbf{1}=(1,1,\ldots,1)^{'}$を使うと、

$S^{(2)}=\mathbf{s1}^{'}+\mathbf{1s}^{'}-2XX^{'}$
($S^{(2)}=s_{jk}^2$, $\mathbf{1}=(1,1,\ldots,1)^{'}$、$\mathbf{s}=(s_{o1},\ldots,s_{om})^{'}$)

と表すことができる。ここで、クロネッカー記号を用いて二重中心化をすると

$$-\tfrac{1}{2}HS^{(2)}H=HXX^{'}H^{'}\qquad(H=(\delta_{jk}-\tfrac{1}{m}))$$

となる。ここで、$(B=b_{jk})$を

$$B=-\tfrac{1}{2}HS^{(2)}H$$

とし、s_{jk}が距離行列であることを利用すると、

$$s_{jk}^2-s_{j\cdot}^2-s_{\cdot k}^2+s_{\cdot\cdot}^2$$

となる。また、s_{jk}はr次元ユークリッド空間における2点jとkの距離なのでb_{jk}行列は以下のように表記できる。

$$b_{jk}=\frac{1}{2}(\frac{1}{m}\sum_{j=1}^{m}s_{jk}^2+\frac{1}{m}\sum_{k=1}^{m}s_{jk}^2-\frac{1}{m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}s_{jk}^2-s_{jk}^2)$$

以上から求められたB行列が非負定符号であり、なおかつrank$=r$であるならば、

$$B_{m\times m}=A_{m\times r}A_{r\times m}^{'}$$

と分解でき、r次元ユークリッド空間における座標行列Aを定めることができる。

次に、$Bx_t=\lambda_t x_t$となるように固有値λを定める。このとき、

$$a_t=\sqrt{\lambda_t}x_t\qquad(t=1,2,\ldots,r)$$
$$A_{m\times r}=(a_1,a_2,\ldots,a_r)$$

とおけば、空間配置が決まる。

2.2　非計量的多次元尺度構成法

ここでは、Kruscal法について述べる。非計量的多次元尺度構成法のモデルは、測定されたデータs_{jk}が高々、順序データで与えられることから

$$s_{jk}\sim\tilde{d}_{jk}\simeq d_{jk}=\left[\sum_{t=1}^{T}|x_{jt}-x_{kt}|^r\right]^{\frac{1}{r}}$$

とし、これが定義される空間に求める座標x_{jk}を定める。このモデルの$\sim$は、単調性と呼ばれ、測定された順序データs_{jk}を距離データ$\tilde{d}_{jk}$に変換することを表す。そして、$\simeq$は、最小二乗的な近似をすることを表している。求めるべき座標Xから計算されるd_{jk}と$\tilde{d}_{jk}$の差の二乗和が最小になるように座標x_{jk}を定めるという意味である。最右辺はミンコフスキー距離である。

非類似性データs_{jk}から求められた座標x_{jk}のモデルへの適合度を評価する必要がある。測定されたデータs_{jk}から生成された距離データd_{jk}は近似的なのでどうしてもずれが生じてしまう。そのずれの度合い、つまり非適合度（ストレスと呼ぶ）を以下のように定義して評価する。

$$\eta_1=\sqrt{\frac{\sum_{j,k}(d_{jk}-\tilde{d}_{jk})^2}{\sum_{j,k}d_{jk}^2}}$$

$$\eta_2 = \sqrt{\frac{\sum_{j,k}(d_{jk}-\bar{d}_{jk})^2}{\sum_{j,k}(d_{jk}-d)^2}}$$

2.3 準計量的多次元尺度構成法

ここでは、数量化　類を扱う。準計量的多次元尺度構成法は、測定された類似度データのユークリッド距離d_{jk}、つまり$d_{jk} = |x_j - x_k|$ の類似度が大きいもの同士は値を小さく、類似度が小さいもの同士は値を大きくするという概念に基づき、式を立てその式の最適解の固有値、固有ベクトルを求めることにより、空間配置を求める。

3 主成分分析法との比較

3.1 理論的比較

一番の違いはデータから行列を作る際に主成分分析法では分散共分散行列を作り、固有値と固有ベクトルを求めるのに対し、計量的多次元尺度構成法では非類似度を表すデータ（もし、間隔尺度で表されたのならば比例尺度に変換する）からYoung-Householderの定理を用いて、非類似度データが距離データであることを示す。そして、距離データであることを利用して内積B行列を生成し、この内積B行列が非負定符号であり、ランクrであれば、ｒ次元ユークリッド空間における座標行列Aを定めている。つまり、行列の作り方の違いが結果の違いを生むのではないかと思われる。

3.2 Rを使ったデータ解析による比較

一般的なデータを用いて解析したところ第一、第二主成分と計量的多次元尺度構成法の第一、第二固有ベクトルの値に差が見られなかった。プロット図も同じになった。そこで、主成分分析法と計量的多次元尺度構成法はRで解析すると差が見られないのではないかという仮説を立てた。その仮説を検討するために元々の形がわかる幾何学的なデータを用いて解析をすることにした。出来るだけ低次元で考えたいので立方体(三次元)の頂点の座標をデータとした。結果は図1、図2に示す。

図を見ると、一見違いがあるように見える。それぞれの主成分の分散の値は第一主成分も第二主成分も1.142857で一致し、結果的にはどちらを第一主成分に採用しても理論的には問題がないがプログラムの都合上、このような結果になっていると思われる。一方、計量的多次元尺度構成法は、軸という概念を持たないため、固有ベクトルの値をそのまま座標に投影している。計量的多次元尺度構成法の固有値は第一固有値も第二固有値も8で二つに差がないので描く平面は一意には決まらない。つまり、平面を適当なところに決めてプロットしていることになる。こちらもプログラムの都合上、偶然的にこのような結果になったのだと思われる。一見、主成分分析法と計量的多次元尺度構成法には差があるように見えるが、計量的多次元尺度構成法の点3、4、7、8が作る平面に基準を合わせたとしたら、主成分分析法の結果と同じになる。つまり、差があるように見えているが実はないということがわかる。ユークリッド距離を用いて元のデータを変換させて距離行列を作り解析すると、主成分分析法と計量的多次元尺度構成法の差はないということがわかった。田中・垂水[5]の主座標分析の理論からも言える。

また、ここで使用したデータに少しだけ長さを持たせて(平均1、分散0.001の乱数を加えた。)同じように解析してみたところ、二つの結果に差がないということがわかった。

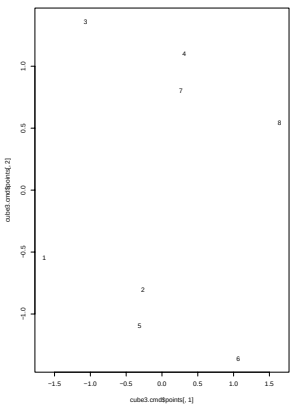

図 1: 多次元尺度構成法

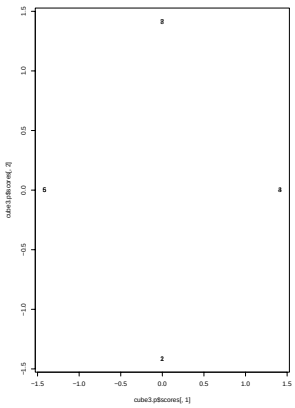

図 2: 主成分分析法

次に主成分分析法と非計量的多次元尺度構成法を一般データを用いて比較した(非多次元尺度構成法は与えられたデータをユークリッド距離に変換しその距離行列をランク付けして順序尺度にした)ところ、プロット図には明らかに差が出た。しかし、大きく外れる値はほぼ同じであった。二次元のデータを解析すると、プロット図はほぼ同じ結果が出た。

4 おわりに

主成分分析法と計量的多次元尺度構成法はデータをユークリッド距離に変換し、Rでデータを解析をすると、主成分分析法の第一主成分と第二主成分の分散が等しく、計量的多次元尺度構成法の第一固有ベクトルと第二固有ベクトルの固有値が等しいときを除いては、解析結果に違いは見られなかった。今回はRでの解析しかやれず、ユークリッド距離以外での距離を与えての解析ができなかったので、今後は距離データを様々な方法で変換して、比較していきたいと思う。また、主成分分析法と非計量的多次元尺度構成法を比較すると明らかにプロット図には差が出たが、大きく外れる値はほぼ同じものとなった。

参考文献

[1] 斉藤堯幸：多次元尺度構成法，1980.

[2] 斉藤堯幸、宿久洋：関連性データの解析法，2006.

[3] 圓川　隆夫：多変量のデータ解析，1988.

[4] プロジェクトケース：数量化理論(4),
http://case.f7.ems.okayama-u.ac.jp

[5] 田中豊、垂水共之：統計解析ハンドブック　多変量解析, 1995.