多次元尺度構成法

多次元データについて、各サンプルの類似性を低次元のユークリッド空間（距離）で表現することで視覚的な表現を試みる。なお、本レジュメの目的は、教科書の文章について論理のトビがあると思われる部分を補完すること、そして、わかりづらい表現になっているところに注釈を加えることの２つである。従い、高度な計算を追うものではないことを初めに宣言しておく。そして、今回わからないことが多く、まとめきれないところが多々存在するため、改めて編集し共有させて欲しい。

1　適用例と解析ストーリー

ここは読み合わせとする。

2　非計量 MDS の解析方法

解析対象とするデータについて、そのデータ間の距離が距離の公理を満たさない場合を考える。そこで、本節においては絶対値に意味がない順序尺度のデータ[*1]を距離という定量的なデータに落とし込み、それらを比較、分析する方法を学ぶ。

今回は非計量 MDS の代表的な方法であるクラスカルの方法を取り扱う。この手法により、データ対象間の類似度（の大小）を点間の距離として視覚的に捉えられるようにしたい。順序尺度である親近性 s_{ij} を距離データ d^*_{ij} に変換する。その際

$$s_{ij} < s_{kl} \text{ならば } d^*_{ij} \geq d^*_{kl} \tag{1}$$

$$s_{ij} = s_{kl} \text{ならば } d^*_{ij} = d^*_{kl} \tag{2}$$

を満たすようにする。式 1 は d^*_{ij} が s_{ij} に関して単調減少になってほしいということを意味している。ここで距離 d^*_{ij} は

$$d^*_{ij} \simeq d_{ij} = \left(\Sigma_{m=1}^{P} |x_{im} - x_{jm}|^t\right)^{\frac{1}{t}} \tag{3}$$

というミンコフスキー距離[*2]で定義されている。P は次元数を表している。$t = 2$ の場合が我々のよく知るユークリッド距離となる。今回我々は最適化計算によって得られた座標で作られる d_{ij} と親近性に適当な変換[*3]を施すことによって作られる d^*_{ij} との差を小さくすることを考えていきたい。そのために以下の式で表現されるストレスと呼ばれる量

$$S = \sqrt{\frac{\Sigma\Sigma_{i<j}(d_{ij} - d^*_{ij})^2}{\Sigma\Sigma d^2_{ij}}} \tag{4}$$

を最小化することを考える。S の最小化は gradient 計算による最急降下法[*4]を用いて実行する。当然 Iterative に計算を回すことになるが、計算実行にあたって以下の 3 点がポイントになる。

[*1] 成績の優、良、可、不可をイメージしてもらうと良いでしょう。

[*2] 余談ですが、「ミンコフスキー距離」と「ミンコフスキー空間で定義される距離」は全くの別物です。後者は相対性理論の基本的な概念である「世界線」のことを表しています。世界線はデカルト座標に時間の要素を加えた四元空間上（ミンコフスキー空間）の事象と事象との間の距離として定義されるものとなっています。

[*3] 変換の詳細は残念ながらわかりませんでした。しかし、負号をつける、もしくは逆数をとるなどの操作が考えられるでしょう。

[*4] このレジュメではない別の資料を参照します。リンクは個別に送りますので、そちらのサイトを参照してください。もちろんよく知っている手法とは思いますが、簡単な例で手を動かして理解を深めたいと思います。

(1)S が最小化されているか？
(2)$S-P$ の関係をプロットしたときに、どの次元で S が急激に減少したかはっきり見えるか？[*5]
(3) 布置の解釈ができる次元か？[*6]

求められた座標を散布図に表したり、ストレスをプロットしたり、d_{ij}^{*} の s_{ij} に関する単調減少性を確かめたりする[*7]ことで最適化計算の終了判断や類似度の可視化を行い、さらにはデータの解釈を行う。具体的な解釈は教科書を参考にされたい[*8]。布置したものに対する解釈についてはセンスを問われるので、そこが難しいところなのだと思う。

3　計量 MDS の考え方

解析対象とするデータ間の距離が距離の公理を満たす場合は、計量 MDS を用いて解析的にデータの座標を求めることができる。

今回は計量 MDS の代表的な方法であるトガーソンの方法を取り扱う。2 点 x_i, x_j のデータ間距離（ユークリッド距離）d_{ij} は次元数を P[*9]とすると

$$d_{ij}=\sqrt{\Sigma_{m=1}^{P}(x_{im}-x_{jm})^2} \tag{5}$$

である。これら 2 点について、任意の位置 k を原点とし、その原点からの内積を

$$z_{ij}=\boldsymbol{x_i'}\cdot\boldsymbol{x_j}=\Sigma_{m=1}^{P}x_{im}x_{jm} \tag{6}$$

と定義する。ここで内積を持ち出した理由は d_{ij} から z_{ij} を求めて、そこから x_i, x_j を決定するという流れに持ち込みたいからである。我々に与えられるのは d_{ij} のみであり、そこから最適化された x_i, x_j を求めるために、z_{ij} を中継するというステップを踏む。d_{ij} と z_{ij} を結びつける計算は、あまりにも初等的であるため、本レジュメでは取り扱わない。しかしながら、我々は次の表式

$$z_{ij}=\frac{1}{2}(d_{ik}^2+d_{jk}^2-d_{ij}^2) \tag{7}$$

を得ることができる。この z_{ij} を成分とする $n\times n$ 行列 $\boldsymbol{Z}$ は、求めるべき座標を成分とする $n\times P$ 行列 $\boldsymbol{X}$ とすると、内積の関係から

$$\boldsymbol{Z}\simeq\boldsymbol{X}\boldsymbol{X'} \tag{8}$$

と表すことができる。今回我々は元の距離データ d_{ij} から計算される z_{ij} と最適化計算により求められる座標（次元削減されたもの）から作った内積との差を最小化することを考える。そこで

$$Q=\Sigma_i\Sigma_j\left(z_{ij}-\Sigma_{m=1}^{P}x_{im}x_{jm}\right)^2 \tag{9}$$

の最小化を考える。式 9 において、z_{ij} は次元削減前の距離データから計算される内積であり、$\Sigma_{m=1}^{P}x_{im}x_{jm}$ は次元削減後に最適化された座標から計算される内積である[*10]。

[*5] S が急激に折れ曲がる点のことを「肘」と呼ぶようです。教科書の図 11.1 を参照してください。
[*6] 詰まるところ $P=2,3$ が選ばれるオチです。
[*7] シェパードダイアグラムによる解釈のことを述べていますが、私はこのダイアグラムの描き方がわかりませんでした。
[*8] $p.168-170$ を確認しましょう。
[*9] 教科書で説明されている駅や都市間の距離という例であれば $P=2$ となりますが、このレジュメでは一般的な場合について議論をすることにします。
[*10] この表式における P は次元削減前の P とは異なるものです。

ここで迷子になった。式 8 と後のエッカートーヤング分解[*11]によって、行列 $\boldsymbol{X}$ がわかるなら、最適化計算など不要ではないか？[*12]

次に、内積の原点をデータの重心で定義する場合について考える。原点を重心にすることで、距離の誤差の影響を抑えることができる。[*13]この場合について、内積 z_{ij} と距離 d_{ij} の関係を計算し、式 7 と異なる結果を得ることにする。正直に言って答えを見てしまった。

まず重心を原点とすることから

$$\overline{x}_{\cdot m} = 0 \ \rightarrow \ \Sigma_{i=1}^{n} x_{im} = 0 \tag{10}$$

である。

地点 i と j を原点とした時の内積 z_{ij} は余弦定理より

$$z_{ij} = \frac{1}{2}(d_{i0}^2 + d_{j0}^2 - d_{ij}^2) \tag{11}$$

である。ここで d_{i0}, d_{j0} はそれぞれ地点 i, j における原点からの距離である。座標成分で作った内積 $z_{ij} = \Sigma_{m=1}^{P} x_{im} x_{jm}$ について次の関係が成り立つ。i について総和をとることを考える。

$$\Sigma_{i=1}^{n} z_{ij} = \Sigma_{i=1}^{n} \Sigma_{m=1}^{P} x_{im} x_{jm} = \Sigma_{m=1}^{P} \Sigma_{i=1}^{n} x_{im} x_{jm} = \Sigma_{m=1}^{P} \left(\Sigma_{i=1}^{n} x_{im}\right) x_{jm} = 0 \ (\because \text{式 } 10) \tag{12}$$

総和の入れ替えがポイントであろう。同様の計算によって j の総和についても

$$\Sigma_{j=1}^{n} z_{ij} = 0 \tag{13}$$

という結果が得られる。ここで式 11 について両辺に 3 通りの方法で Σ をかます。

(1) i について

$$\Sigma_{i=1}^{n} z_{ij} = \frac{1}{2}\left(\Sigma_{i=1}^{n} d_{i0}^2 + \Sigma_{i=1}^{n} d_{j0}^2 - \Sigma_{i=1}^{n} d_{ij}^2\right) = \frac{1}{2}\left(\Sigma_{i=1}^{n} d_{i0}^2 + n d_{j0}^2 - \Sigma_{i=1}^{n} d_{ij}^2\right) = 0 \ (\because \text{式 } 12) \tag{14}$$

(2) j について

$$\Sigma_{j=1}^{n} z_{ij} = \frac{1}{2}\left(\Sigma_{j=1}^{n} d_{i0}^2 + \Sigma_{j=1}^{n} d_{j0}^2 - \Sigma_{j=1}^{n} d_{ij}^2\right) = \frac{1}{2}\left(n d_{i0}^2 + \Sigma_{j=1}^{n} d_{j0}^2 - \Sigma_{j=1}^{n} d_{ij}^2\right) = 0 \ (\because \text{式 } 13) \tag{15}$$

(3) $i,\ j$ について

$$\Sigma\Sigma_{i,j} z_{ij} = \frac{1}{2}\left(\Sigma\Sigma_{i,j} d_{i0}^2 + \Sigma\Sigma_{i,j} d_{j0}^2 - \Sigma\Sigma_{i,j}^{n} d_{ij}^2\right) = \frac{1}{2}\left(n\Sigma_i d_{i0}^2 + n\Sigma_j d_{j0}^2 - \Sigma\Sigma_{i,j}^{n} d_{ij}^2\right) = 0 (\because \text{式 } 12 \text{ or } 13) \tag{16}$$

ここで式 14, 15 について、それぞれ左辺の d_{ij} 項部分を右辺に移項してから、両辺を足すと

$$d_{i0}^2 + d_{j0}^2 = \frac{1}{n}\Sigma_i^n d_{ij}^2 + \frac{1}{n}\Sigma_j^n d_{ij}^2 - \frac{1}{n}\left(\Sigma_i^n d_{i0}^2 + \Sigma_j^n d_{j0}^2\right) \tag{17}$$

[*11] 輪講内の議論を通じて、大事なポイントを忘備録として残しておきます。実は、正方行列 $\boldsymbol{Z}$ は、その固有値方程式を解くことによって得られる固有値や固有ベクトルを使うことで、主成分的に近似が可能であるということが一般的に知られています。$\boldsymbol{Z}$ はその固有値を並べた対角行列に、左から固有ベクトルを縦に並べた行列をかけて、右からその転置行列をかけた形で近似的に表現されます。

[*12] 遠い将来に理解を深めたところで、もう一度編集します。いつになるかわかりませんが。

[*13] ごめんなさい。ちんぷんかんぷんでした。平均を使うことで、誤差の分散が小さくなるだろう、くらいのイメージしか持ち合わせていません。

が得られる。ところで式 16 より

$$\Sigma_i^n d_{i0}^2 + \Sigma_j^n d_{j0}^2 = \frac{1}{n}\Sigma\Sigma_{i,j} d_{ij}^2 \tag{18}$$

が成り立つ。

準備が整ったので、いよいよ z_{ij} を求める。式 17 を式 11 に代入して

$$z_{ij} = \frac{1}{2}(\frac{1}{n}\Sigma_i^n d_{ij}^2 + \frac{1}{n}\Sigma_j^n d_{ij}^2 - \frac{1}{n}\left(\Sigma_i^n d_{i0}^2 + \Sigma_j^n d_{j0}^2\right) - d_{ij}^2) \tag{19}$$

最後に式 19 に式 18 を代入すると

$$z_{ij} = \frac{1}{2}(\frac{1}{n}\Sigma_i^n d_{ij}^2 + \frac{1}{n}\Sigma_j^n d_{ij}^2 - \frac{1}{n}\left(\frac{1}{n}\Sigma\Sigma_{i,j} d_{ij}^2\right) - d_{ij}^2) = \frac{1}{2}(\frac{1}{n}\Sigma_i^n d_{ij}^2 + \frac{1}{n}\Sigma_j^n d_{ij}^2 - \Sigma\Sigma_{i,j}\frac{d_{ij}^2}{n^2} - d_{ij}^2) \tag{20}$$

となる。以上より、原点を重心に据えた場合の z_{ij} の表式が得られた。