レーダーチャートの可視化について

2017SS005 藤村真菜
指導教員：塩濱敬之

1　はじめに

データの可視化は, 情報をより直感的に理解するという面において必要不可欠であり, データの特徴を理解し分析するための有効な手段である. 一般に 4 次元以上の多変量データになると, 紙面上に可視化することが困難になるため, これまで多くの多変量データの可視化手法が提案されてきた ([1],[2]). これらの手法には, 散布図を行列に配置してプロットする散布図行列から始まり, チャーノフの顔や, スティック図等のアイコンを用いる方法, 平行座標プロットなどが含まれる.

レーダーチャートとは, 多変量データの可視化の手法の 1 つであり, スパイダープロット (蜘蛛の巣チャート) など, 様々な呼び名で使われている. 複数変数の取る値を変数の次元で構成される多角形の内部にプロットすることで, 異なる対象との比較や, 特徴的な変数の抽出など多変量データの視覚的な解釈を与えることが可能となる.

多角形の内部の面積は, そのデータの個体や対象のもつ総合指標を構成すると考えることができる. そのため, 本研究では, 多変量データの統合指標をレーダーチャートの面積として提案する. この多角形の面積を以降 (AoP; Area of Polygon) と呼ぶ. AoP は, 多角形の各頂点にどの変数を配置するかによって変動するため, AoP によるインデックス化を実行するためには, AoP を最大にするような変数の並び方を考慮しないとならない. そこで, 本研究目的を AoP を最大にする変数の並び方の組み合わせを求めることと, AoP の統計的性質を調べることに設定する.

2　提案手法

d 次元のデータ $\boldsymbol{X} \in \mathbb{R}^d$ を考える. 変数のインデックス集合を $[1,d] = \{1,2,\ldots,d\}$ と表し, その任意の並べ替えの写像を $\pi : [1,d] \to [1,d]$ と表す. このとき d 変量の並べ替えの集合は対称群 S_d として知られていて, その位数は $d!$ である. 今, レーダーチャート上の変数の並び替えを想定しているため, 巡回群と反転して同等になる並び替えは考慮しなくてもよい, そこで, 対称群から巡回群と反転を除いた, 円周上のすべての変数の並べ替えからなる集合を $\tilde{S}_d$ と表す. その位数は $\frac{1}{2}(d-1)!$ である.

一般にレーダーチャートを用いる場合は変数を平均が ℓ, 標準偏差が s になるように標準化を行う. このとき, レーダーチャートの始点を終点をそれぞれ 0, $\ell_{\max}$ とし, ℓ は

チャートの各変数の中心を表すとする. 以降, 一般性を失うことなく, $s = 1$ とし, 確率変数 X は平均 0 ベクトル, ρ_{ij} を要素とする相関係数行列をもつとする.

並べ替え $\pi \in \tilde{S}_d$ に対して, 第 $k, k = 1,\ldots,n$ 番目のデータの AoP は次式で与えられる.

$$A_k(\pi) = \frac{1}{2}\sin\left(\frac{2\pi}{d}\right)\sum_{i=1}^{d}(\ell + X_{k,\pi(i)})(\ell + X_{k,\pi(i+1)}),$$

n 個の標本データが与えられたとき, 平均 AoP は次のように定義する.

$$\bar{A}_n(\pi) = \frac{1}{2n}\sin\left(\frac{2\pi}{d}\right)\sum_{i=1}^{d}\sum_{k=1}^{n}(\ell + X_{k,\pi(i)})(\ell + X_{k,\pi(i+1)}). \tag{1}$$

このとき, AoP とその標本平均 (1) は, 変数の並べ替え π に応じて定まる.

平均 AoP の最大化問題を次のように定式化する.

$$\text{maximize} \quad \bar{A}_n(\pi) \quad \text{subject to: } \pi \in \tilde{S}_d.$$

ここで $\boldsymbol{X}$ は d 変量標準正規分布に従う確率変数であるから, 上記の最大化問題は以下と同等である.

$$\hat{\pi}_n^* = \underset{\pi \in \tilde{S}_d}{\operatorname{argmax}} \sum_{i=1}^{d} \hat{r}_{\pi(i),\pi(i+1)}.$$

ここで, $\hat{r}_{X_i,X_j}$ は変数 X_i と X_k の標本相関係数である. 標本 AoP を最大化する並べ替え $\hat{\pi}_n^*$ に対して, k 番目の標本 AoP と n 組の平均 AoP を以下のように表す.

$$A_k(\hat{\pi}_n^*), \quad \text{and} \quad \bar{A}_n(\hat{\pi}_n^*),$$

一般に $\tilde{S}_d$ の位数は, d が増えると発散するため, $\tilde{S}_d$ のすべての組み合わせに対して最適解を発見することは, 多くても $d = 10$ 程度までであり, それ以上の次元になると, 最適化ソルバーを用いて最適解を得ることになる.

3　AoP の統計的性質

(X, Y) がそれぞれ平均 0, 分散 1 の標準正規分布に従い, その相関係数が ρ であるとき, 積の確率変数 $Z = XY$ の密度関数は以下のように与えられる [3].

$$f_Z(z) = \frac{1}{\pi\sqrt{1-\rho^2}}\exp\left[\frac{\rho z}{1-\rho^2}\right]K_0\left(\frac{|z|}{1-\rho^2}\right), \tag{2}$$

ここで, $K_0(\cdot)$ は 0 次の第 2 種修正ベッセル関数である.

$$K_0(x) = \int_0^\infty \cos(x\sinh t)dt = \int_0^\infty \frac{\cos(xt)dt}{\sqrt{t^2+1}}.$$

確率変数 Z の平均は ρ であり, 分散は $1+\rho^2$ である. また, $A_1(\pi)$ の平均と分散は次のように与えられる.

$$\begin{aligned} E[A_1(\pi)] =& \frac{d\ell^2}{2}\sin\left(\frac{2\pi}{d}\right) + \frac{1}{2}\sin\left(\frac{2\pi}{d}\right)\sum_{i=1}^{d}\rho_{\pi(i),\pi(i+1)} \\ =:& \bar{A} + \mu_A(\pi) \end{aligned} \tag{3}$$

$X_{\pi(q)}$ について, $q > d$ のとき $X_{\pi(q-d)}$ と表し, $q < 1$ のとき $X_{\pi(q+d)}$ と表すとすると, $A_1(\pi)$ の分散は次のように与えられる.

$$\sigma^2_{A(\pi)} := \frac{1}{4}\sin^2\left(\frac{2\pi}{d}\right)\Bigg[4\ell^2\sum_{i,j=1}^{d}\rho_{\pi(i),\pi(j)} + \sum_{i=1}^{d}(1+\rho^2_{\pi(i),\pi(i+1)}) + 2\sum_{i=1}^{d}\rho_{\pi(i),\pi(i-1)}\rho_{\pi(i),\pi(i+1)} + \rho_{\pi(i-1),\pi(i+1)} \quad (4)$$
$$+ \sum_{\substack{i,j=1 \\ (|i-j|\neq\{0,1\})}}^{d}\rho_{\pi(i),\pi(j)}\rho_{\pi(i+1),\pi(j+1)} + \rho_{\pi(i),\pi(j+1)}\rho_{\pi(i+1),\pi(j)}\Bigg],$$

ここで, $\mathrm{Cov}(Z_1, Z_2) = \rho_{12}\rho_{23} + \rho_{13}$, また $\mathrm{Cov}(Z_1, Z_3) = \rho_{13}\rho_{24} + \rho_{14}\rho_{23}$ を用いた. このとき, (1) 式で定義された AoP に対して, 次の中心極限定理が成り立つ.

定理 1 $\boldsymbol{X}$ は d 変量標準正規分布に従い, 相関係数行列 $\boldsymbol{\Sigma} = [\rho_{ij}]_{i,j=1,\ldots,d}$ をもつとする. このとき, $n \to \infty$ とすると,

$$\sqrt{n}\left(\bar{A}(\hat{\pi}^*_n) - \bar{A}\right) \to_d N\left(\mu_A(\pi^*), \sigma^2_{A(\pi^*)}\right),$$

ただし, $\bar{A}$ および $\mu_A(\pi^*)$ は (3) 式で, $\sigma^2_{A(\pi^*)}$ は (4) で定義した.

次に, 2 つの並べ替え, π と ν を考える. この並べ替えに対して次の中心極限定理が成り立つ.

定理 2 定理 1 と同じ条件のもとで, $n \to \infty$ とすると次の中心極限定理が成り立つ.

$$\sqrt{n}\begin{pmatrix}\bar{A}_n(\pi) - \bar{A} \\ \bar{A}_n(\nu) - \bar{A}\end{pmatrix} \to_d N\left(\begin{pmatrix}\mu_A(\pi) \\ \mu_A(\nu)\end{pmatrix}, \begin{pmatrix}\sigma^2_A(\pi) & \sigma_{\pi,\nu} \\ \sigma_{\pi,\nu} & \sigma^2_A(\nu)\end{pmatrix}\right),$$

ここで, $\sigma_{\pi,\nu}$ は以下で定義される.

$$\sigma_{\pi,\nu} = \ell^2\sin^2\left(\frac{2\pi}{d}\right)\sum_{i,j=1}^{d}\rho_{\pi(i),\nu(j)} + \frac{1}{4}\sin^2\left(\frac{2\pi}{d}\right)\sum_{i,j=1}^{d}\left[\rho_{\pi(i),\nu(j)}\rho_{\pi(i+1),\nu(j+1)} + \rho_{\pi(i),\nu(j+1)}\rho_{\pi(i+1),\nu(j)}\right].$$

定理 2 を用いることで, 2 つの並べ替え π と ν による平均 AoP 差の検定が可能になる.

4 データ解析

機械学習の識別問題で良く用いられるワインデータ ($n = 178$, $d = 13$) を用いたデータ解析を行う. 与えられたデータ順を ν とし, 最適な並び替えを $\hat{\pi}^*_n$ とすると, それぞれの並び方は以下の通りである.

$$\nu = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13],$$
$$\hat{\pi}^*_n = [1, 10, 2, 8, 4, 3, 5, 9, 6, 7, 12, 11, 13].$$

AoP はそれぞれ, $\bar{A}_n(\nu) = 75.926$ と $\bar{A}_n(\hat{\pi}^*_n) = 76.933$ が得られる. 並び替えによらない AoP が $\bar{A} = 75.518$ であることから, 初期配置がある程度レーダーチャートの良い並び順になっていることが分かる. 平均差の検定統計量は, $Z = 0.766$ となり, 10% 水準で統計的に有意ではないが, この検定の検出力はさほど高くないことと, 標本数や次元の大きさを考慮してもその差は無視できない.

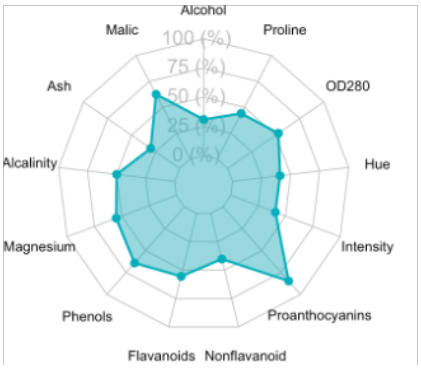

図 1 ID: 111, AoP = 73.87, Rank = 86

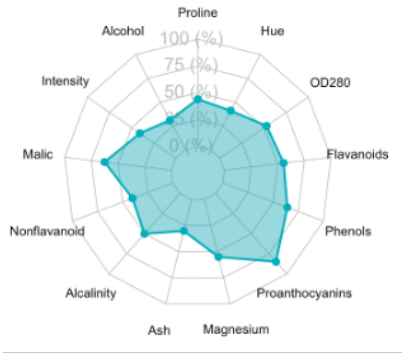

図 2 ID: 111, AoP = 77.67, Rank = 111

最適な並び替えを用いることで, 初期状態の変数配置と AoP が大きく変わる例を図 1, 2 に示した.ID が 111 のデータは初期状態の変数配置では, AoP が 73.87 であったのが, 最適な並べ替えの下では, AoP が 77.67 へと変化し, そのためランクが 86 から 111 に上昇した.

5 まとめ

レーダーチャートの多角形の平均面積を最大にする, 変数の並べ方の問題について考え, その統計量の標本分布を導出した.

参考文献

[1] Chen, C. H., Härdle, W. K., & Unwin, A. (Eds.). (2007). Handbook of data visualization. Springer Science & Business Media.

[2] Jacoby, W. G. (1988). Statistical graphics for visualizing multivariate data. Vol. 120. Sage.

[3] Nadarajah, S. and Pogány T. K. (2016). On the distribution of the product of correlated normal random variables. Comptes Rendus Mathematique, 354, 201–204.