『就実経営研究』第９号　抜刷
就実大学経営学部　2024年３月　発行

拡張型星座グラフによる次元削減とクラスター分類の可視化

Visualization of Dimensionality Reduction and Cluster Classification Using Expanded Constellation Graphs

藤　原　美　佳
梶　西　将　司
福　森　　護

拡張型星座グラフによる次元削減と クラスター分類の可視化

藤原美佳（岡山理科大学経営学部）
梶西将司（中国学園大学国際教養学部）
福森　護（就実大学経営学部）

Visualization of Dimensionality Reduction and Cluster Classification Using Expanded Constellation Graphs

Mika Fujiwara
Shoji Kajinishi
Mamoru Fukumori

要旨：多変量データを可視化することにより、複雑なデータの特徴を直観的に把握することが可能になる。これまで多変量データの可視化の手法は数多く開発されており、マーケティングや心理学をはじめ、社会科学や人文科学のさまざまな領域で活用されている。その中で、星座グラフを応用した手法として、拡張型星座グラフがある。本論文では、次元削減とクラスター分類の可視化を目的として拡張型星座グラフを適用し、主成分分析、多次元尺度法、t-SNE、UMAPと比較することにより、結果の検証を行う。

Abstract: Visualization of multivariate data makes it possible to intuitively grasp the characteristics of complex data. Many methods have been developed and utilized in various fields of social sciences and humanities, including marketing and psychology. Among these methods is the expanded constellation graphs, which is an application of the constellation graphs. In this paper, extended constellation graphs are applied for dimensionality reduction and cluster classification, and the results are validated by comparing them with principal component analysis, multidimensional scaling methods, t-SNE and UMAP.

1．はじめに

複雑な構造を持つ多変量データを低次元に落とし込んで可視化することにより、データの特性を直観的かつ容易に把握することで、数値解析だけでは得られない知見を得ることが可能になる。例えば、Chernoff（1973）は、最大18変数の多変量のデータを顔のパーツに割り当てて表現する方法を提案した。この方法では、類似のパターンを持つサンプルを顔の類似性により直観的に判断する

ことが可能になる。また、ベクトルを連結するというアイデアに基づく手法として、連結ベクトルグラフ（脇本・田栗, 1974）や星座グラフ（Wakimoto & Taguri, 1978）などがある。連結ベクトルグラフは、昇順にソートされた１つの目的変数と対応関係を保持して並び替えられた複数の説明変数に対して、変数ごとにサンプルの数だけベクトルを連結する方法であり、目的変数と説明変数との関連や説明変数間の関連をベクトル線の類似により直観的に判断できる。また、星座グラフは、変数の値を角度に変換し、サンプルごとに変数のベクトルを連結する方法である。連結されたベクトルの最終点の位置により、サンプルごとに変数の平均や分散を表現することができ、さらにベクトル線の形状（パス）により、変数のパターンの分類が容易にできるため、教育データをはじめ、幅広い分野での活用が可能になる。この星座グラフを応用して、福森・藤原（2020）は、変数の位置を角度に変換するアイデアを提案し、さらにFujiwara *et al.*（2021）はそのアイデアをもとに拡張型星座グラフを開発した。また、藤原ほか（2022）は、R Shinyを用いて、星座グラフ及び拡張型星座グラフのソフトウェアの開発を行った。

拡張型星座グラフは、０度から180度までの半円の円周上に変数を配置し、変数の方向（角度）に向かって０度方向に配置された変数から順に、データの持つ値によって決定される長さのベクトル線を描いて連結する。このとき、変数を配置する順番は任意であるが、例えば、Fujiwara *et al.*（2021）は、因子分析の負荷量の大きさを指標にする方法を用いている。拡張型星座グラフでは、変数の配置によって結果が変わるため、最適な配置については慎重に検討する必要がある（Fukumori *et al.*, 2021）。

ここで、円周上の変数の間隔を等間隔に配置する場合、p個の変数からなるサンプル数nの多変量データx_{ij}の各変数に対するベクトルの角度（φ_j）は、

$$\varphi_j = \frac{j-1}{p-1}\pi, \quad j = 1,2,\dots,p$$

となる。

求められたφ_jを使用して、i番目のサンプルにおける拡張型星座グラフの最終点の座標は以下のように表される。

$$(\alpha_i, \beta_i) = \frac{1}{p}\left(\sum_{j=1}^{p}\frac{x_{ij}}{U_j}\cos\varphi_j, \sum_{j=1}^{p}\frac{x_{ij}}{U_j}\sin\varphi_j\right)$$

また、変数の間隔を決定するための方法の一つとして、Fujiwara *et al.*（2021）は、以下のように因子負荷量の値によりφ_jを求める方法を提案している。

$$\varphi_j = \frac{a_u - a_j}{a_u - a_l}\pi$$
$$a_u = \max_{1\le j\le p} a_j, \quad a_l = \min_{1\le j\le p} a_j$$

ここで、a_jは因子分析によって得られた因子負荷量の値からなるベクトルaのj番目の値である。

拡張型星座グラフでは、連結されたベクトルの最終点が０度方向（右側方向）に布置するか、

180度方向（左側方向）に布置するかによってどちらの方向に配置された変数の特徴を有するかがわかる。また、最終点の高さによって平均の大きさを、パスの形状や長さによって各変数の値の大きさを把握できる。

本論文では、この拡張型星座グラフを次元削減及びクラスター分類の可視化を目的として適用し、主成分分析（以下、PCA）、多次元尺度法（以下、MDS）、t-SNE、UMAPと比較することにより、結果の検証を行う。

２．方法

２－１．使用したデータ

本論文では、スコッチウィスキー86銘柄のフレーバーについて収集されたオープンデータを使用した。このデータは、スコットランドにあるストラスクライド大学が提供するもので、12項目のフレーバー（Medicinal, Smoky, Body, Spicy, Winey, Nutty, Malty, Honey, Fruity, Sweetness, Floral, Tobacco）で構成されている。それぞれの要素は０点～４点までの５段階評価となっている。

２－２．手法の適用

スコッチウィスキーのフレーバーデータに対して、k-means法を適用するために、まずエルボー法により、クラスター内の距離の平方和のプロットを参考にクラスター数を決定した。図１は、エルボー法によるプロットである。図１を参考に、クラスター数を４と決定し、k-means法によりスコッチウィスキー86銘柄を４つのクラスターに分類した。その結果、クラスター１には８銘柄、クラスター２には7銘柄、クラスター３には27銘柄、クラスター４には44銘柄が分類された。

次に、拡張型星座グラフ、PCA、MDS、t-SNE、UMAPの５つの手法により２次元座標上に86銘柄をプロットし、k-means法によって分類されたクラスターを可視化した。

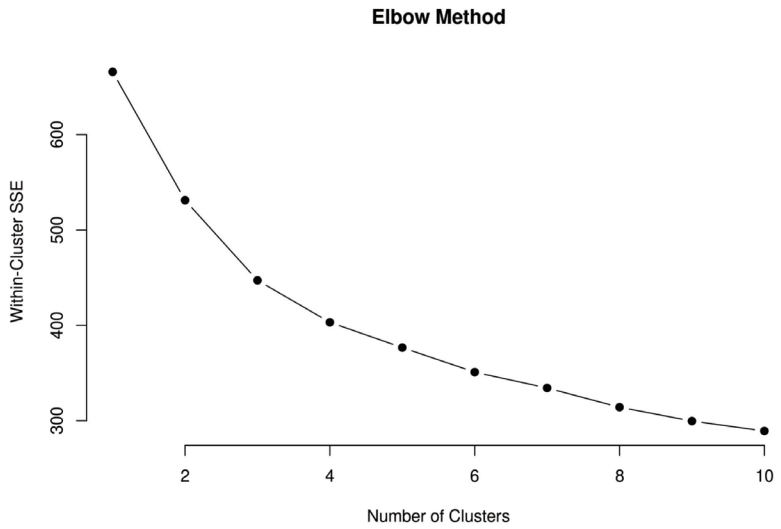

図１．エルボー法によるプロット

ここで、拡張型星座グラフと比較した各手法の概要は以下の通りである。

1）PCA

PCAは、n個のサンプルについて、p個の変量x_j, j=1, 2,⋯, p (p≥2)があるとき、データx_{pi} (i=1, 2,⋯, n)の総合的な指標として、p個の変量x^{T}=(x_1, x_2, ⋯, x_p)に適当な重みの係数a^{T}=(a_1, a_2, ⋯, a_p)を与えた合成変量z

$$z = a_1x_1 + a_2x_2 + \cdots + a_px_p = \boldsymbol{a}^T\boldsymbol{x}$$

を考え、以下の条件のもとで、合成変量zの分散を最大化する重みの係数aを求める方法である。この合成変数zを主成分と呼ぶ。

$$\sum_{i=1}^{\cdot} a_i^2 = \boldsymbol{a}^T\boldsymbol{a} = 1$$

PCAは、多変量データをより少ない次元（主成分）に合成する手法であるため、次元削減の手法として、マーケティング、心理学などをはじめとする、社会科学、人文科学などの領域を中心に幅広く用いられている。

なお、PCAは、線形構造のデータを使用することが前提となっているため、非線形データへの使用は適していない。非線形データに対しては、以下に示すMDS、t-SNE、UMAPなどの手法が用いられる。

2）MDS

MDSは、類似したデータを低次元空間に配置して、視覚的に類似関係を把握するための手法で、PCAと同様の目的で使用される。PCAとの大きな違いとしては、PCAが分散共分散行列から固有値と固有ベクトルを求めるのに対して、MDSでは距離行列（非類似度行列）を利用して生成された内積行列を用いる点である。

MDSの考え方は以下の通りである（長畑，2018）。

まずサンプル間の距離について、サンプル$i(x_i = (x_{i1}, \cdots, x_{ip})^T)$とサンプル$j(x_j = (x_{j1}, \cdots, x_{jp})^T)$との距離を$d_{ij}$で表すと、以下の式のようになる。

$$d_{ij}^2 = \left\|x_i - x_j\right\|^2 = \|x_i\|^2 + \left\|x_j\right\|^2 - 2x_i^Tx_j$$
$$B = (x_i^Tx_j)_{n\times n} = XX^T$$

この行列B=XX^Tの固有値、固有ベクトルをλ、uとすると、

$$XX^Tu = \lambda u$$

が成立する。

MDSは計量MDSと非計量MDSに大別される。計量MDSは距離データに基づいて低次元に個々のサンプルを配置する方法であり、非計量MDSは必ずしも距離データとはいえない類似度（相関係数など）に基づいて距離を測定し、サンプルを低次元に配置する方法である。

計量MDSにおいて、上記のd_{ij}^2からBの(i, j)成分b_{ij}は距離のデータd_{ij}を用いて以下の式のように求められる。

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

流れとして次のように考える。

サンプル間の距離(d_{ij})を与える　⇒　対応したもとのデータの座標(b_{ij})の計算

⇒　Bの固有値と固有ベクトル(λ, u)の計算　⇒　各サンプルの座標を計算し、配置する

次に、非計量MDSでは、クラスカルにより提案された次のストレス

$$S_{stress} = \sqrt{\frac{\sum_{i \neq j} (\delta_{ij} - d_{ij})^2}{\sum_{i \neq j} d_{ij}^2}}$$

を与えられたd_{ij}に対し、δ_{ij}について最小化し、その時の座標を用いてデータを配置する。

ここで、δ_{ij}は、非類似度データに対応するとして与えられた距離である。

3）t-SNE

t-SNEは、非線形な次元削減の手法の一つである。PCAとは異なり、低次元空間に写像する行列を求めることなく、高次元空間で近いデータは低次元空間でも近くに配置し、高次元空間で遠いデータは低次元空間でも遠くに配置する方法である。

t-SNEの手順は以下の通りである。

①　正規分布を仮定して、次のようにデータx_i, x_jから確率p_{ij}を算出する。

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

ここで、Nは全データ数を表す。

②　N個のデータ点を低次元空間にランダムに配置する。

③　t分布（自由度１）に基づき、高次元空間におけるデータ点x_i, x_jに対応する低次元空間のデータ点z_i, z_jから、次のように確率q_{ij}を算出する。

$$q_{ij} = \frac{(1 + \left\|z_i - z_j\right\|^2)^{-1}}{\sum_{k \neq i} (1 + \|z_k - z_i\|^2)^{-1}}$$

④　確率p_{ij}、q_{ij}の２つの分布が近くなるように、再配置を行う。

⑤　ステップ③、④を、結果が収束するまで繰り返す。

4）UMAP

UMAPは、高次元データを局所構造と大域構造を保持したまま低次元データに変換する方法で、次元削減の手法の一つである。

UMAPの流れは、以下の２つのステップである。

① 重み付きk近傍グラフをつくる。

② グラフを低次元に配置する。

なお、通常のk近傍法では計算量が多くなるため、UMAPを実行する際には近似法を用いることが多い。

３．結果

拡張型星座グラフを描くにあたり、最初に変数の配置を決めるために、フレーバーの12項目に対する探索的因子分析（最尤法・プロマックス回転）を行った。第１因子の因子負荷量は表１の通りである。なお、第１因子の固有値は3.2371、寄与率は26.98%であった。

負荷量から、Medicinal、Smoky、Tobacco、Body、Spicyが負の値、Floral、Sweetness、Honey、Fruity、Winey、Nutty、Malty量が正の値になっている。特に、Medicinal、Smokyが負の方向に大きい値となっており、Floral、Sweetnessが正の方向に大きい値となっているため、「スモーキー系フレーバー vs フローラル系フレーバー」の因子と解釈される。

そこで、この負荷量の値に基づいて、負荷量の高い順に180度方向から配置して、拡張型星座グラフを描いた。図２は、クラスター別に形状を変えた最終点の星を描いたものである。この図において、０度方向にスモーキー系のフレーバーに関する項目が配置され、180度方向にフローラル系のフレーバーに関する項目が配置されている。図において、最終点のプロットが「○」で描かれた銘柄は、他の銘柄と比べて０度方向にプロットされていることから、スモーキー系のフレーバーが特徴のウィスキーであることがわかる。また、「●」で描かれた銘柄は、90度付近で、他と比べて中心からの距離が比較的遠い位置にプロットされていることから、フレーバーの強い傾向は持たないが、全体的にリッチなフレーバーの銘柄であることがわかる。また、「△」で描かれた銘柄は、90度方向よりやや180度方向にプロットされていることから、ややフローラル系のフレーバーであり、リッチな銘柄からそうでないものも幅広く含まれている銘柄であると言える。最後に、「×」で描かれた銘柄は、比較的中心からの距離が近い付近にプロットされていることから、フローラル系のフレーバーが特徴的であるが、全体的にフレーバーが強くない銘柄であることがわかる。そこで、フレーバーの特徴を詳細に分析するために、図３にパスを表示した拡張型星座グラフを示す。

表１．第１因子の負荷量

変　数	因子1
Body	-0.4815
Sweetness	0.4182
Smoky	-0.8047
Medicinal	-0.8816
Tobacco	-0.4933
Honey	0.3834
Spicy	-0.0881
Winey	0.0786
Nutty	0.0962
Malty	0.2552
Fruity	0.3689
Floral	0.5764

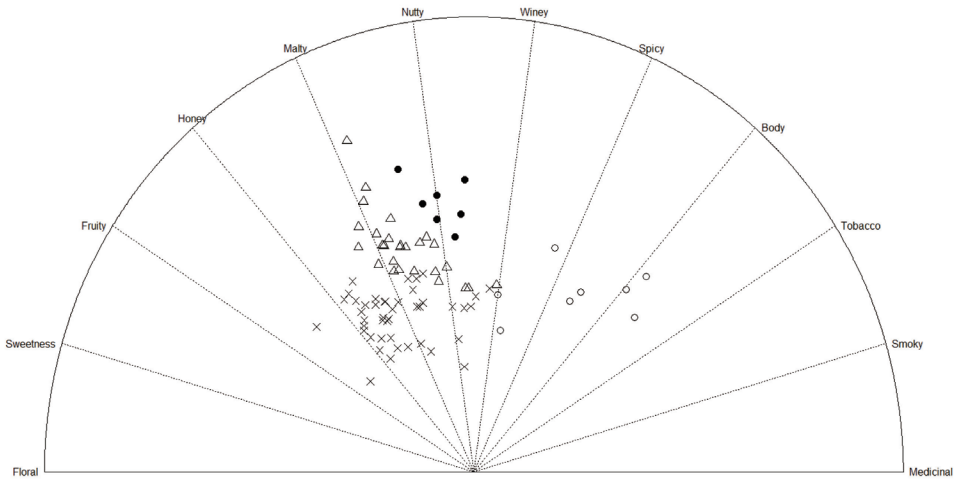

図２．拡張型星座グラフによる可視化（最終点）

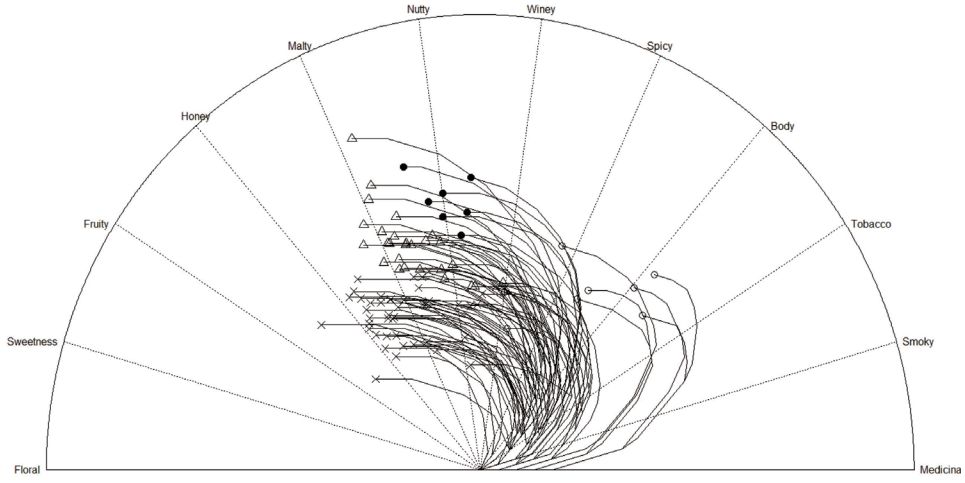

図３．拡張型星座グラフによる可視化（パスあり）

パスを描くことにより、各クラスターの傾向はある程度把握が可能である。「○」のクラスターに関しては、SmokyやBodyのベクトルが長い銘柄が多いことが読み取れ、「●」のクラスターに関しては、Medicinal、Tobaccoなどを除き、均一的にやや長いベクトルであることが読み取れる。また、「△」のクラスターに関しては、フローラル系のフレーバーのベクトルが長めであること、「×」のフレーバーは、SweetnessやFloralのベクトルが長く他のフレーバーのベクトルがかなり短いことが読み取れる。

クラスターの特徴の違いを把握するために、図４に、各クラスターの平均により描いた拡張型星座グラフを示す。この図から、「○」のクラスターはSmokyとBodyが特徴のスモーキー系ウィスキーのクラスター、「●」のクラスターは全体的にリッチでWiney、Bodyが特徴のウィスキーのクラスター、「△」のクラスターはFruity、Sweetnessが特徴的ではあるが、スモーキー系以外のフレーバーが全体的にリッチなクラスター、「×」のクラスターはSweetness、Floralが強いが他のフレーバーが極端に弱いことが読み取れる。

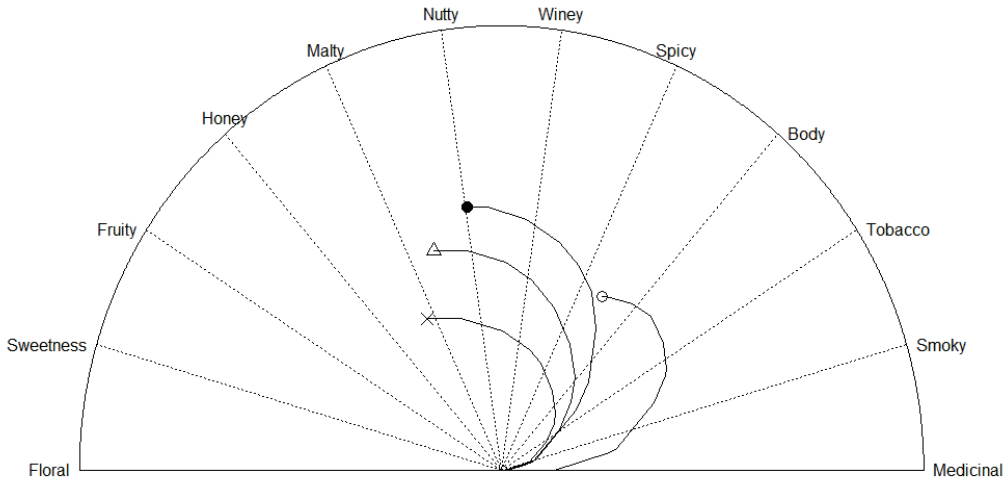

図４．拡張型星座グラフによる可視化（クラスターの平均）

このように、拡張型星座グラフから、「スモーキー系 vs フローラル系フレーバー」の軸で見た時の全体的な傾向が読み取れるだけではなく、パスの長さから、個々のフレーバーの特徴も読み取れ、またクラスターの特徴を直観的に把握できるという利点を持つ。

次に、PCA、MDS、t-SNE、UMAPによる結果と比較することで、拡張型星座グラフの結果の検証を行う。

まず、PCAを用いて、86銘柄を２次元座標上にプロットしたものを図５に示す。また図６にバイプロットの結果を示す。図５から、明確にクラスターの分類が表現されていることがわかる。また、主成分軸を解釈することにより個々の銘柄やクラスターの特徴を読み取ることが可能であり、さらに、図６のようにバイプロットを描くことにより、各フレーバーと銘柄やクラスターとの関係が把握できる。

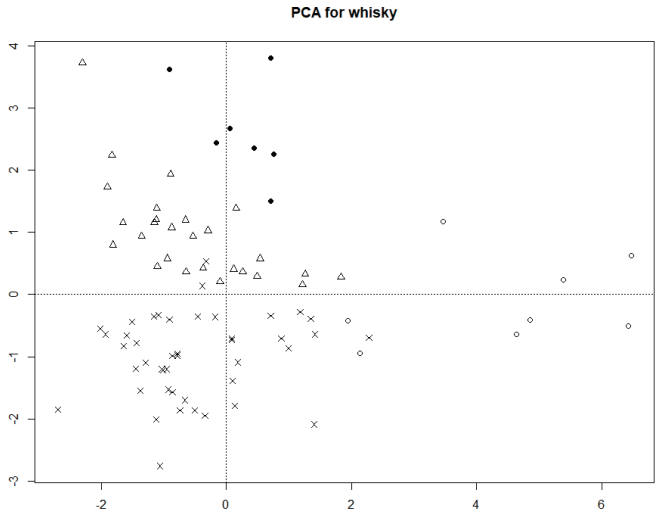

図５．PCAによる86銘柄のプロット

例えば、58、59、4などの銘柄はスモーキー系のフレーバーが強いことや７、47などの銘柄はフローラル系のフレーバーが強いことがわかる。このことから、PCAの結果と拡張型星座グラフの結果の整合性は取れていると言える。

次に、MDSを用いて86銘柄を２次元座標上にプロットしたものを図７に示す。この図から、PCA（図５）と同様に、明確にクラスターの分類が表現されていることがわかる。各クラスターや銘柄のフレーバーの特徴をグラフから読み取ることは難しいが、次元削減やクラスター分類という目的においての有効性は高いと言える。

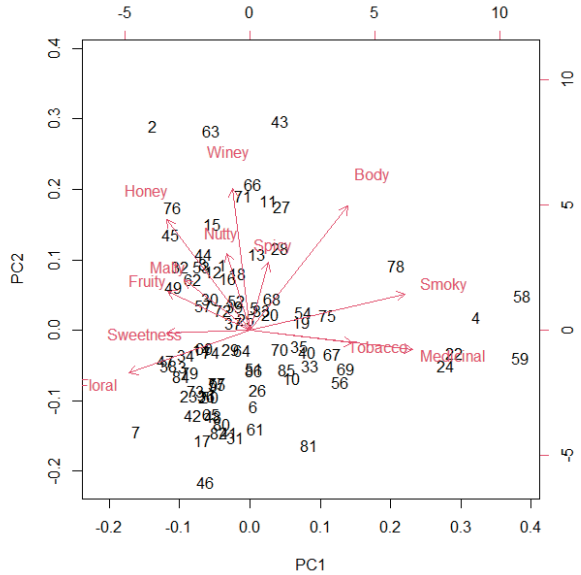

図６．主成分分析のバイプロット

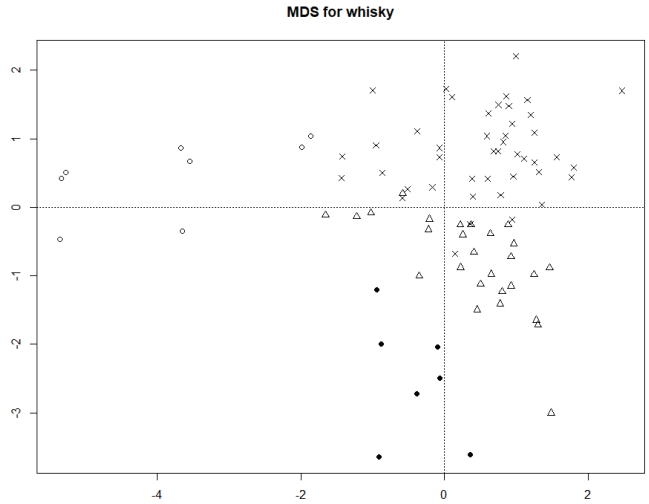

図７．MDSによる86銘柄のプロット

次に、t-SNEを用いて、86銘柄を２次元座標上にプロットしたものを、図８に示す。この図より、本データにおいては、クラスターは明確には分類されておらず、また、クラスターや銘柄の特徴を読み取ることも困難であることがわかる。

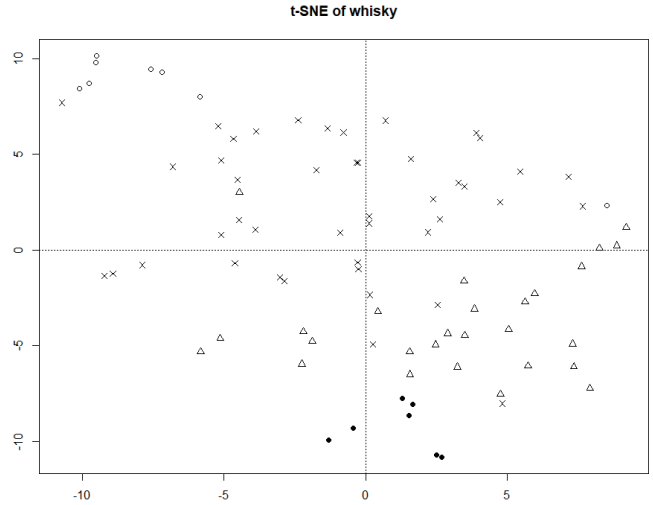

図８．t-SNEによる86銘柄のプロット

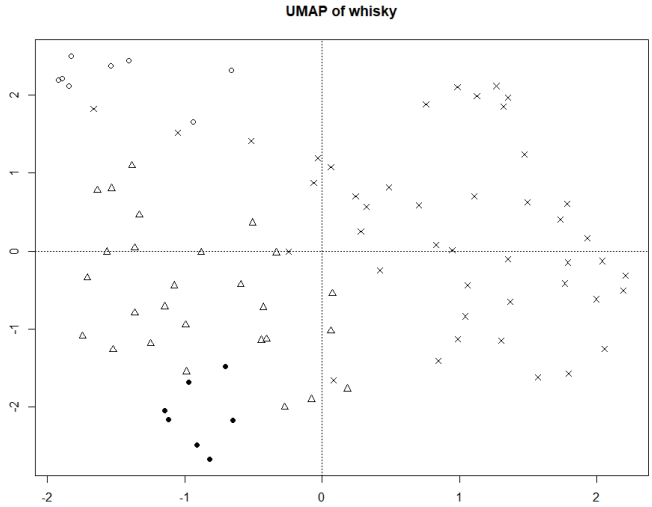

図９．UMAPによる86銘柄のプロット

次に、UMAPを用いて、86銘柄を２次元座標上にプロットしたものを、図９に示す。この図から、t-SNEよりはクラスターが視覚的に表現されているが、PCAやMDSと比べると明確に表現されているとは言えない。また、クラスターや銘柄のフレーバーの詳細な特徴を読み取ることはt-SNE同様に困難であることがわかる。

４．まとめ

本論文では、次元削減とクラスター分類の可視化を目的として拡張型星座グラフを適用し、PCA、MDS、t-SNE、UMAPと比較することにより、結果の検証を行った。12変数からなるスコッチウィスキーのフレーバーデータに対して拡張型星座グラフを適用した結果、クラスターを視覚的に表現することができ、さらにクラスターや銘柄のフレーバーの特徴の把握ができた。結果を検証するために、PCA、MDS、t-SNE、UMAPとの比較を行った結果、PCAによるバイプロットとほぼ同様の分析が可能であり、MDSと同様にクラスターの分類を表現することができた。なお、今回のデータにおいては、t-SNE及びUMAPでは、クラスターの分類は明確には表現されなかった。また、拡張型星座グラフは、フレーバーの強さをベクトルの長さで表現するため、他の手法では表現しにくいフレーバー個々の特徴を把握できるというメリットも確認できた。

グラフ手法は、本来、柔軟性が高いため、使いながら成長する側面を持っている。拡張型星座グラフは、分野によって多様な使い方が可能である。今後、使用後の問題点や要望、有用性についての意見を得て、できるだけ多くの分野で使用できる環境を整えたいと考えている。拡張型星座グラフが今後さらに普及し、データ解析において有効に利用されることを期待したい。

引用文献

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, J*ournal of the American Statistical Association*. 68, 361–368.

Fujiwara, M., Kajinishi, S. & Kurihara, K. (2021). Visualization of multivariate data using expanded constellation and expanded kanji graphs and their application to clus-tering, *Journal of Environmental Science for Sustainable Society*, 10(1), 1–8.

藤原美佳・梶西将司・栗原考次（2022）.「星座グラフを用いた多変量データ可視化のためのソフトウェア」『計算機統計学』, 34(2), 99–112.

Fukumori, M., Fujiwara, M., & Kajinishi, S. (2021). A study on the placement of variables in a modified constellation graph, *CHUGOKUGAKUEN journal*, 20, 1–6.

福森護・藤原美佳（2020）.「星座グラフの変形による多変量データの可視化とクラスター分析への応用」『中国学園』, 19, 45–52.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing Data using t-SNE, *Journal of Machine Learning Research*, 9, 2579–2605.

長畑秀和 (2018).『R で学ぶデータサイエンス』, 朝倉書店.

Wakimoto, K. & Taguri, M. (1978). Constellation graphical method for representing multi-dimensional data, *Annals of the Institute of Statistical Mathematics,* 30, Part A, 77–84.

脇本和昌・後藤昌司・松原義弘 (1979).『多変量グラフ解析法』, 朝倉書店.

脇本和昌・田栗正章 (1974).「連結ベクトルパターンによる重相関度の表現」『日本統計学会誌』, 5(1), 9–24.