クラスタリングと特徴抽出の融合[†]

宇野 光平*

1. はじめに

クラスタリングは、観測されたデータをいくつかのグループに分割する手法である。一見すると教師なしの分類問題とも考えられるが、データ解析の道具としてクラスタリングを捉えれば、データ解析時にクラスタリングが果たす役割は教師あり分類問題とは大きく異なるのではないだろうか？ クラスタリングを用いる目的の一つが、データ内に隠された構造を探索的に発見することである [1]。これはすなわち、仮説の創出である。医学分野において疾患のサブタイプを発見することなどが具体例として挙げられる [2]。こうした探索的な利用はクラスタリング単体で運用するのではなく、疾患サブタイプごとのメカニズム解明などの前工程として用いられることが多い。そのため、探索的クラスタリングによって推定された各クラスターがどのような特徴を持つグループであるかを理解することは、その後の解析を進める上で重要な意味を持つ。このとき観測データに共変量が存在すればクラスターの特徴を推察することは可能だが、探索的なデータ解析において共変量が存在しない場合も少なくない。そのため、変数（特徴量）からクラスターの特徴を把握することが必要となる。

データの変数（特徴量）の数が3以下の場合は、散布図によるデータの可視化がクラスター特徴の把握に有用である。各クラスターと変数の関係だけでなく、クラスター同士の位置関係も可視化されることでクラスターの特徴が表れるからである。しかし、変数の数が4以上など一般化された場合、散布図による可視化は困難である。そのため、散布図からクラスター特徴の把握するためには、特徴抽出などでデータを視覚化可能にすることが求められる。

本稿では、特徴抽出による視覚化と探索的クラスタリングの親和性および二つの手法を組み合わせた手法を中心に述べる。2節では特徴抽出の代表例である主成分分析とクラスタリングの代表例である k-means クラスタリング、そしてそれらを組み合わせた Reduced k-means クラスタリングとその派生を解説する。3節では特徴抽出ではなく、スパース推定による特徴選択とクラスタリングについて述べる。4節では近年発展著しい機械学習における次元縮約とクラスタリングとの関係について言及するとともに、本稿で扱う手法の特徴を総括する。

2. クラスタリングと特徴抽出

代表的な特徴抽出法の一つが、主成分分析である。サンプルサイズが N 、変数の数が P の標準化されたデータ行列 $\mathbf{X} \in \mathbb{R}^{N \times P}$ があるとする。本稿ではデータ行列は全て標準化されたものとする。いま主成分数を M としたとき、直交射影行列を用いた射影モデルとしての主成分分析の目的関数は、

$$PCA_{\text{projection}}(\mathbf{A}|\mathbf{X}) = \|\mathbf{X} - \mathbf{X}\mathbf{\Pi}_\mathbf{A}\|_F^2 = \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^\top\|_F^2 \quad \text{s.t. } \mathbf{A}^\top\mathbf{A} = \mathbf{I}_M \quad (1)$$

と書くことができる。ただし $\|\mathbf{X}\|_F$ は行列 $\mathbf{X}$ のフロベニウスノルムであり、 $\mathbf{\Pi}_\mathbf{A} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top$ は行列 $\mathbf{A} \in \mathbb{R}^{P \times M}$ の射影行列、そして $\mathbf{A}$ は主成分の重み（主成分負荷量とも呼ばれる）を表す。 $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_M$ から $\mathbf{A}$ は列直交であるといえる。また低ランク近似モデルとしての主成分分析の目的関数は、

$$PCA_{\text{low-rank}}(\mathbf{F}, \mathbf{A}|\mathbf{X}) = \|\mathbf{X} - \mathbf{F}\mathbf{A}^\top\|_F^2 \quad \text{s.t. } \mathbf{A}^\top\mathbf{A} = \mathbf{I}_M \quad (2)$$

と書くことができる。重み $\mathbf{A}$ が変数と主成分との関係を表す行列であるのに対し、主成分得点 $\mathbf{F} \in \mathbb{R}^{N \times M}$ は個体と主成分との関係を表し、 $\mathbf{F}$ は低次元空間上に射影されたデータ $\mathbf{X}$ である。また $\mathbf{F} = \mathbf{X}\mathbf{A}$ とすることで、低ランク近似モデルと射影モデルの目的関数が同一であることがわかる。主成分分析によって得られた $\mathbf{F}$ にクラスタリングを適用することで、低次元空間上でのクラスタリング、そしてクラスターの可視化が可能となる。

代表的なクラスタリング手法の一つが k-means クラスタリングであり、目的関数は

$$KM(\mathbf{U}, \mathbf{C}|\mathbf{X}) = \sum_{n=1}^N \sum_{k=1}^K u_{nk} \|\mathbf{x}_n - \mathbf{c}_k\|_2^2 = \|\mathbf{X} - \mathbf{U}\mathbf{C}\|_F^2 \quad (3)$$

と書ける。ここで u_{nk} は個体 n がクラスター k に所属するかどうかを $\{0, 1\}$ で表し、 $\mathbf{U} \in \mathbb{R}^{N \times K}$ は u_{nk} を n 行 k 列の要素とする行列である。また $\mathbf{C} \in \mathbb{R}^{K \times P}$ はクラスター毎の中心ベクトルを表すセントロイド行列である。 K はクラスター数を表す正の整数で、分析時に決定する必要がある。

$P \leq 3$ のような低次元のデータに対するクラスタリングはクラスターの可視化が可能である一方、変数の数である P が大きい場合は散布図によるクラスターの可視化が難しい。ただし、以下のように k-means クラスタリングの前処理として主成分分析を適用することで、4次元以上のデータに対しても可視化が容易となる。

1. $\mathbf{X}$ に主成分分析を適用し、主成分得点 $\mathbf{F}$ を得る。

[†] Clustering Meets Feature Extraction
Kohei UNO

* 名古屋大学大学院医学系研究科
Nagoya University Graduate School of Medicine

2. **F** に k-means クラスタリングを適用する.

しかし, こうした分析プロセスには一つ問題がある. 主成分分析による特徴抽出は, クラスターの情報を考慮しないため特徴抽出をした結果がクラスターを捉えるために適しているかどうかはわからない.

このような要請に基づき, 特徴抽出とクラスタリングを同時推定する手法が開発された. これらの手法は部分空間クラスタリング (subspace clustering) と呼ばれ多くの手法が提案されている [3]. 特に本稿では主成分分析と k-means クラスタリングを組み合わせることに伴う同時推定手法に着目する.

2.1 Reduced k-means クラスタリング

2.1.1 モデルと目的関数

Reduced k-means クラスタリング (RKM) は, 主成分分析による特徴抽出と k-means クラスタリングによるクラスタリングを同時に行う手法である [4]. RKM の目的関数は

$$RKM(\mathbf{U}, \mathbf{G}, \mathbf{A}|\mathbf{X}) = \|\mathbf{X} - \mathbf{UGA}^\top\|_F^2 \quad \text{s.t.} \quad \mathbf{A}^\top \mathbf{A} = \mathbf{I}_M \quad (4)$$

である. ここで $\mathbf{U} \in \mathbb{R}^{N \times K}$ は個体の所属を表す行列であり, $\mathbf{G} \in \mathbb{R}^{K \times M}$ はクラスターと主成分との関係を表す得点行列, $\mathbf{A} \in \mathbb{R}^{P \times M}$ は重み行列である. 低ランク近似としての目的関数 (2) 式における **F** が, RKM の目的関数である (4) 式では **UG** に置き換わっていると考えることができる. また (4) 式はフロベニウスノルムとトレースの関係 $\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{X}^\top \mathbf{X})$ を用いることで,

$$\begin{aligned} PCA_{\text{projection}}(\mathbf{A}|\mathbf{X}) + KM(\mathbf{U}, \mathbf{G}|\mathbf{XA}) \\ &= \|\mathbf{X} - \mathbf{XAA}^\top\|_F^2 + \|\mathbf{XA} - \mathbf{UG}\|_F^2 \\ &= \text{tr}(\mathbf{X}^\top \mathbf{X}) - 2\text{tr}(\mathbf{A}^\top \mathbf{X}^\top \mathbf{UG}) + \text{tr}(\mathbf{G}^\top \mathbf{U}^\top \mathbf{UG}) \\ &= \|\mathbf{X} - \mathbf{UGA}^\top\|_F^2 \\ &= RKM(\mathbf{U}, \mathbf{G}, \mathbf{A}|\mathbf{X}) \end{aligned} \quad (5)$$

と書き換えることができるため, **X** の低次元空間による射影とデータの低次元空間上での得点 **XA** へのクラスタリングを, 同時に行っていることがわかる.

2.1.2 アルゴリズム

RKM のパラメータ推定は目的関数 (4) 式の最小化によって達成される. ただし推定すべきパラメータが **U, G, A** の3つ存在するため, 一度に推定することは難しい. そこで, 1つのパラメータに着目した最適化を行いその他のパラメータは固定することで全てのパラメータを推定する交互最小二乗法によって3つのパラメータを推定する [5].

まず **U** 以外のパラメータは固定し, **U** の更新を検討する. **U** の更新は, 各行についての目的関数

$$RKM(u_n|\mathbf{G}, \mathbf{A}, \mathbf{x}_n) = \|\mathbf{x}_n - \mathbf{AG}^\top \mathbf{u}_n\|_2^2$$

を最小にするような $\mathbf{u}_n$ を探索することで達成される. $\mathbf{u}_n \in \mathbb{R}^K$ には要素の一つのみが 1, それ以外の要素は 0 という制約があるため, $\mathbf{u}_n$ は K 通りのみ考えればよいことがわかる. そこで $\mathbf{u}_n$ の考えうる全ての通りを要素とする集合を $\mathcal{U}$

とすると,

$$\hat{\mathbf{u}}_n = \arg \min_{\mathbf{u} \in \mathcal{U}} \|\mathbf{x}_n - \mathbf{AG}^\top \mathbf{u}\|_2^2 \quad (6)$$

を考えることで, **U** の各行が更新される.

つづいて **G** の更新を考える. **G** には制約がないため, トレース

$$\begin{aligned} RKM(\mathbf{G}|\mathbf{U}, \mathbf{A}, \mathbf{X}) \\ &= \text{tr}(\mathbf{X}^\top \mathbf{X}) - 2\text{tr}(\mathbf{X}^\top \mathbf{UGA}^\top) + \text{tr}(\mathbf{G}^\top \mathbf{U}^\top \mathbf{UG}) \end{aligned}$$

に対する偏微分によって導出される方程式から **G** を更新すればよい [6]. すなわち

$$\frac{\partial RKM(\mathbf{G}|\mathbf{U}, \mathbf{A}, \mathbf{X})}{\partial \mathbf{G}} = -2\mathbf{U}^\top \mathbf{XA} + 2\mathbf{U}^\top \mathbf{UG} = \mathbf{O}_{K \times M}$$

から, **G** の更新式

$$\hat{\mathbf{G}} = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{XA} \quad (7)$$

を得る.

最後に **A** の更新について考える. **A** には列正規直交制約 $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_M$ が存在するため, 制約を満たした上での最適化が必要となる. const_A を **A** と関係のない項とすると, 目的関数は

$$RKM(\mathbf{A}|\mathbf{U}, \mathbf{G}, \mathbf{X}) = \text{const}_A - 2\text{tr}(\mathbf{X}^\top \mathbf{UGA}^\top)$$

となることから, **A** の更新はトレース

$$\text{tr}(\mathbf{X}^\top \mathbf{UGA}^\top)$$

の最大化によって達成される. これは直交プロクラステス問題 [7], あるいはその一般化として考えることができる ten Berge の定理 [8] によって解くことができ, 特異値分解

$$\mathbf{X}^\top \mathbf{UG} = \mathbf{KAL}^\top \quad (8)$$

の左, 右特異行列を用いることで, **A** の更新式

$$\hat{\mathbf{A}} = \mathbf{KL}^\top \quad (9)$$

がわかる.

RKM のアルゴリズムをまとめると, 以下のアルゴリズムになる.

1. **U, G, A** を制約を満たしたうえで初期化する.
2. (6) 式によって **U** の各行を更新する.
3. (7) 式によって **G** を更新する.
4. (9) 式によって **A** を更新する.
5. 収束基準を満たしているか確認し, 満たしている場合計算を終了する. 満たしていない場合, 2. に戻る.
5. にて記述している通り, 収束するまで各パラメータの更新を続ける交互最小二乗法では, 収束したとみなす収束基準を設定する必要がある. そこで, 2. から 4. までを一反復と考え, t 回目反復時のパラメータを $\Theta^{(t)} = \{\mathbf{U}^{(t)}, \mathbf{G}^{(t)}, \mathbf{A}^{(t)}\}$ とした場合の条件

$$RKM(\Theta^{(t)}|\mathbf{X}) - RKM(\Theta^{(t+1)}|\mathbf{X}) < \epsilon$$

を満たしたとき収束したとみなす. 具体的には $\epsilon = 10^{-5}$ など非常に小さな値に設定する. RKM は k-means クラスタリング

図1 ユーロッパ食材データの低次元散布図

と同様に目的関数が非凸であることから、推定したパラメータが大域解である保証がない。そのため、1. から 5. までの手順を複数回繰り返し、最も目的関数が小さい場合を大域解とみなす。

2.1.3 解析例

RKM の解析例を Hartigan によるヨーロッパの食材データを用いて紹介する [9]。このデータは、16 カ国におけるコーヒーやバター、ヨーグルトなどの食材 20 種を家庭で持っている世帯の割合を示している。RKM および k-means クラスタリングによって 20 種の食材をいくつかのグループにわけ、図 1 における左図は主成分 2 として主成分分析を適用した後、クラスター 3 として k-means クラスタリングを適用した場合の散布図であり、右図は主成分 2、クラスター 3 として RKM を適用した場合の散布図である。RKM の場合、よりクラスター内平方和が小さくなるように各食材を表す点が比較的密集していることがわかる。

左図の第 1 主成分は主にイタリア、ルクセンブルク、イギリス、スイス、スペイン、アイルランドが正に影響があり、右図の第 1 主成分は主にイタリア、ルクセンブルク、ポルトガル、スペインが正に強い影響があった。また、第 2 主成分は共にスウェーデン、デンマーク、ノルウェー、フィンランドが正に強い影響があった。散布図から主成分が共に高い（散布図の右上に位置する）クラスターにはティーバッグや果物、バターなど所属し、これらがヨーロッパで広く普及していることが示唆された。対照的に、主成分が共に低いクラスターには冷凍魚やヨーグルトなどが所属し、これらは一部の家庭でのみ親しまれていることがわかった。

2.2 RKM のファジィクラスタリングへの拡張

k-means クラスタリングは、個体は一つのクラスターにしか属さず二つ以上のクラスターに同時に所属することはないという強い仮定を想定している。この仮定を緩和し、k-means クラスタリングをより現実的なモデルへと拡張し所属の曖昧さを許容した手法が fuzzy c-means クラスタリング (FCM) である [10, 11]。FCM の目的関数は以下で表すことが出来る。

$$FCM(\mathbf{U}, \mathbf{C}|\mathbf{X}) = \sum_{n=1}^N \sum_{k=1}^K u_{nk}^\alpha \| \mathbf{x}_n - \mathbf{c}_k \|^2_2$$

$$\text{s.t. } 0 \leq u_{nk} \leq 1, \mathbf{U} \mathbf{1}_K = \mathbf{1}_N \quad (10)$$

(10) 式における $\mathbf{c}_k$ はクラスター k の中心ベクトルを表し、 $\mathbf{1}_N$ は要素がすべて 1 の N 次元ベクトルを表す。 u_{nk} は個体 n の

クラスター k への所属度を表すため、k-means クラスタリングと同様と考えることができる。しかし、 u_{nk} が 0 もしくは 1 ではなく 0 以上 1 以下、そして各 n について要素の総和が 1 となっている点で、 u_{nk} に関する制約が緩和されている。 $\alpha \geq 1$ は所属の曖昧さを制御するハイパーパラメータであり、 $\alpha = 1$ のとき k-means クラスタリングと一致する。

また、異なる定式化としてエントロピー正則化による FCM が提案されている [12]。この定式化における目的関数は

$$FCM_{\text{entropy}}(\mathbf{U}, \mathbf{C}|\mathbf{X}) = \sum_{n=1}^N \sum_{k=1}^K u_{nk} \left\{ \| \mathbf{x}_n - \mathbf{c}_k \|^2_2 + \lambda^{-1} \log u_{nk} \right\}$$

$$\text{s.t. } 0 \leq u_{nk} \leq 1, \mathbf{U} \mathbf{1}_K = \mathbf{1}_N \quad (11)$$

である。 $\lambda > 0$ によって所属の曖昧さが制御され、 λ が大きくなるほど k-means クラスタリングに近づく。(11) 式による FCM の定式化は、通常の FCM である (10) 式とは異なり中心行列 $\mathbf{C}$ の更新が k-means クラスタリングと同一であるため、k-means クラスタリングから FCM への拡張が容易なことが特徴の一つである。

西田は、エントロピー正則化による FCM と PCA を組み合わせることで RKM を拡張した [13]。この Reduced fuzzy c-means clustering (RFCM) の目的関数は

$$RFCM(\mathbf{U}, \mathbf{G}, \mathbf{A}|\mathbf{X}) = \sum_{n=1}^N \sum_{k=1}^K u_{nk} \left\{ \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_k \|^2_2 + \lambda^{-1} \log u_{nk} \right\}$$

$$\text{s.t. } 0 \leq u_{nk} \leq 1, \mathbf{U} \mathbf{1}_K = \mathbf{1}_N \quad (12)$$

であり、アルゴリズムは $\mathbf{U}$ の更新以外 RKM と同一である。 $\mathbf{g}_k \in \mathbb{R}^M$ は $\mathbf{G} \in \mathbb{R}^{K \times M}$ の k 行ベクトルであり、クラスター k と各主成分の関係を表すベクトルである。 $\mathbf{U}$ の更新は、制約条件 $0 \leq u_{nk} \leq 1, \mathbf{U} \mathbf{1}_K = \mathbf{1}_N$ を満たすためにラグランジュの未定乗数法によって最適化を行う。すなわち

$$L = \sum_{n=1}^N \sum_{k=1}^K u_{nk} \left\{ \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_k \|^2_2 + \lambda^{-1} \log u_{nk} \right\}$$

$$+ \sum_{n=1}^N \eta_n \left(\sum_{l=1}^K u_{nl} - 1 \right) \quad (13)$$

を (12) 式の代わりに最小化するような $\mathbf{U}$ を見つける。ここで η_n はラグランジュ乗数である。

$$\frac{\partial L}{\partial u_{nk}} = \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_k \|^2_2 + \lambda^{-1} (\log u_{nk} + 1) + \eta_n = 0$$

$$\Rightarrow \log u_{nk} + 1 = -\lambda \left(\| \mathbf{x}_n - \mathbf{A} \mathbf{g}_k \|^2_2 + \eta_n \right)$$

$$\Rightarrow u_{nk} = \exp(-1 - \lambda \eta_n) \exp \left(-\lambda \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_k \|^2_2 \right) \quad (14)$$

と制約条件 $\mathbf{U} \mathbf{1}_K = \mathbf{1}_N$ 、つまり $\sum_{l=1}^K u_{nl} = 1$ を考慮すると

$$\exp(-1 - \lambda \eta_n) = \sum_{l=1}^K \exp \left(-\lambda \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_l \|^2_2 \right)^{-1}$$

となることがわかる。この式を (14) 式に代入し、更新式

$$\hat{u}_{nk} = \frac{\exp \left(-\lambda \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_k \|^2_2 \right)}{\sum_{l=1}^K \exp \left(-\lambda \| \mathbf{x}_n - \mathbf{A} \mathbf{g}_l \|^2_2 \right)} \quad (15)$$

を得る。これらの更新をまとめると、RFCM のアルゴリズムは以下となる。

1. $\mathbf{U}, \mathbf{G}, \mathbf{A}$ を制約を満たしたうえで初期化する。
2. (15) 式によって $\mathbf{U}$ の各行を更新する。
3. (7) 式によって $\mathbf{G}$ を更新する。
4. (8) 式によって $\mathbf{A}$ を更新する。
5. 収束基準を満たしているか確認し、満たしている場合計算を終了する。満たしていない場合、2. に戻る。

2.3 その他の同時推定アルゴリズム

本節ではこれまで、k-means クラスタリングと主成分分析を同時に推定するアルゴリズムである RKM, そしてそのファジィ化である RFCM を解説した。k-means クラスタリングとその他の手法を同時に推定した手法は、RKM だけでなく数多く提案されている。例えば主成分分析と同様に次元縮約を目的として用いられる因子分析とクラスタリングを同時に推定する手法が提案されている [14, 15]。その他にもバイプロット [16, 17] や多次元尺度構成法 [18]、対応分析 [19]、関数主成分分析 [20] など様々な次元縮約法とクラスタリングを同時に推定することで、解釈性の向上を目指した手法が提案されている。これらクラスタリングと次元縮約の同時推定を網羅的に解説した論文としては、[21] が詳しい。

3. クラスタリングと特徴選択

ここまで、主成分分析によって前処理した後にクラスタリングを適用することの利点を説明した。また、次元縮約法とクラスタリングを同時に推定する手法を、RKM を中心に紹介した。しかし、これらの方法では主成分分析などで抽出された特徴量がどのようなものであるかという解釈を行う必要があり、結果に曖昧さが残る。さらに、データが高次元の場合には主成分の解釈自体が難しい。こうした問題を解決できる方法の一つが、特徴選択である。既に与えられた特徴量の中からクラスタリングに必要と思われる特徴量を選ぶことで解釈を行う余裕が生まれ、さらにはクラスタリングに関わらない特徴を除くことで分類精度の向上も見込める。

3.1 スパース推定

スパース推定はパラメータの一部を 0 と推定することで特徴選択ができる手法であるため、高次元データに対しても有効と考えられている。スパース推定流行の起点となったスパース回帰である Lasso [22] の目的関数は

$$\begin{aligned} \text{Lasso}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) &= \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \\ &= \frac{1}{2} \left\| \mathbf{y} - \sum_{p=1}^P \mathbf{x}_p \beta_p \right\|_2^2 + \lambda \sum_{p=1}^P |\beta_p| \end{aligned} \quad (16)$$

である。ただし $\mathbf{y} \in \mathbb{R}^N$ は目的変数、 $\mathbf{X} \in \mathbb{R}^{N \times P}$ は説明変数、 $\boldsymbol{\beta} \in \mathbb{R}^P$ は回帰係数である。 $\lambda \geq 0$ はスパース性を調整するハイパーパラメータである。

通常の回帰分析であれば微分によって勾配を計算することで、パラメータである $\boldsymbol{\beta}$ を推定することができる。しかし $|\beta_p|$

は微分不可能であることから、微分ではなく劣微分を考える [23]。劣微分の値は集合となることが特徴で、微分可能な点において劣微分の値は要素が 1 つの集合、すなわち微分の値と一致する。例えば $|\beta_p|$ の β_p における劣微分は以下である。

$$\frac{\partial |\beta_p|}{\partial \beta_p} = \begin{cases} 1 & \beta_p > 0 \\ -1 & \beta_p < 0 \\ [-1, 1] & \beta_p = 0 \end{cases} \quad (17)$$

微分可能な点である $\beta_p \neq 0$ の場合には微分と同様の値であり、微分不可能な点である $\beta_p = 0$ の場合には集合 $[-1, 1]$ になっていることがわかる。

ここで、 $\mathbf{X}^\top \mathbf{X}$ が対角行列、すなわち $\mathbf{X}$ の各列が互いに無相関であるとする、Lasso の $\boldsymbol{\beta}$ を解析的に解くことができる。劣微分から方程式

$$\begin{aligned} \frac{\partial \text{Lasso}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})}{\partial \beta_p} &= -\mathbf{x}_p^\top \mathbf{y} + \mathbf{x}_p^\top \mathbf{x}_p \beta_p + \lambda \frac{\partial |\beta_p|}{\partial \beta_p} = 0 \\ \Rightarrow \beta_p &= \frac{\left(\mathbf{x}_p^\top \mathbf{y} - \lambda \frac{\partial |\beta_p|}{\partial \beta_p} \right)}{\mathbf{x}_p^\top \mathbf{x}_p} \end{aligned}$$

が得られるため、Lasso の推定式は

$$\hat{\beta}_p = \begin{cases} \frac{\mathbf{x}_p^\top \mathbf{y} - \lambda}{\mathbf{x}_p^\top \mathbf{x}_p} & \beta_p > 0 \\ \frac{\mathbf{x}_p^\top \mathbf{y} + \lambda}{\mathbf{x}_p^\top \mathbf{x}_p} & \beta_p < 0 \\ 0 & |\mathbf{x}_p^\top \mathbf{y}| \leq \lambda \end{cases} \quad (18)$$

となる。

Lasso は各変数に対する回帰係数を 0 にするが、Group Lasso はいくつかのグループにまとめられた変数すべてを 0 にする [24]。1, ..., P の変数を G 個のグループに事前に振り分けたとすると、Group Lasso の目的関数は以下で書くことができる。

$$\begin{aligned} \text{GLasso}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) &= \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^G \|\boldsymbol{\beta}_g\|_2 \\ &= \frac{1}{2} \left\| \mathbf{y} - \sum_{g=1}^G \mathbf{X}_g \boldsymbol{\beta}_g \right\|_2^2 + \lambda \sum_{g=1}^G \|\boldsymbol{\beta}_g\|_2 \end{aligned} \quad (19)$$

$\mathbf{X}_g$ と $\boldsymbol{\beta}_g$ は、それぞれ変数グループ g のデザイン行列と回帰係数を表す。

Group Lasso のペナルティ項である $\|\boldsymbol{\beta}_g\|_2$ も微分不可能であるため、劣微分によって更新式を導出する必要がある。 $\|\boldsymbol{\beta}_g\|_2$ の $\boldsymbol{\beta}_g$ についての劣微分は、 $\boldsymbol{\beta}_g \neq \mathbf{0}$ のとき

$$\frac{\partial \|\boldsymbol{\beta}_g\|_2}{\partial \boldsymbol{\beta}_g} = \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|_2}$$

である。また、 $\boldsymbol{\beta}_g = \mathbf{0}$ のときは

$$\left\| \frac{\partial \|\boldsymbol{\beta}_g\|_2}{\partial \boldsymbol{\beta}_g} \right\|_2 \leq 1$$

となる。いま各 g について $\mathbf{X}_g^\top \mathbf{X}_g = \mathbf{I}$ を仮定することで、 $\boldsymbol{\beta}_g$ の更新式を陽に書くことができる。 $\boldsymbol{\beta}_g \neq \mathbf{0}$ とすると、方程式

$$\begin{aligned}
 & \frac{\partial GLasso(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})}{\partial \boldsymbol{\beta}_g} \\
 &= \frac{\partial}{\partial \boldsymbol{\beta}_g} \left(\frac{1}{2} \left\| \mathbf{y} - \sum_{\ell \neq g}^G \mathbf{X}_\ell \boldsymbol{\beta}_\ell - \mathbf{X}_g \boldsymbol{\beta}_g \right\|_2^2 + \lambda \|\boldsymbol{\beta}_g\|_2 \right) \\
 &= -\mathbf{X}_g \mathbf{r}_{-g} + \mathbf{X}_g^\top \mathbf{X}_g \boldsymbol{\beta}_g + \lambda \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|_2} = \mathbf{0} \\
 &\Rightarrow \left(1 + \frac{\lambda}{\|\boldsymbol{\beta}_g\|_2} \right) \boldsymbol{\beta}_g = \mathbf{X}_g \mathbf{r}_{-g} \\
 &\Rightarrow \boldsymbol{\beta}_g = \left(\frac{\|\boldsymbol{\beta}_g\|_2}{\|\boldsymbol{\beta}_g\|_2 + \lambda} \right) \mathbf{X}_g \mathbf{r}_{-g} \quad (20)
 \end{aligned}$$

が得られる。ただし、

$$\mathbf{r}_{-g} = \mathbf{y} - \sum_{\ell \neq g}^G \mathbf{X}_\ell \boldsymbol{\beta}_\ell$$

である。 $\|\boldsymbol{\beta}_g\|_2$ の値が不明であることから、両辺の L2 ノルムをとることで以下を得る。

$$\begin{aligned}
 \|\boldsymbol{\beta}_g\|_2 &= \left(\frac{\|\boldsymbol{\beta}_g\|_2}{\|\boldsymbol{\beta}_g\|_2 + \lambda} \right) \|\mathbf{X}_g \mathbf{r}_{-g}\|_2 \\
 &= \|\mathbf{X}_g \mathbf{r}_{-g}\|_2 - \lambda
 \end{aligned}$$

これを (20) 式に代入することで、

$$\boldsymbol{\beta}_g = \left(1 - \frac{\lambda}{\|\mathbf{X}_g \mathbf{r}_{-g}\|_2} \right) \mathbf{X}_g \mathbf{r}_{-g} \quad (21)$$

を得る。

また $\boldsymbol{\beta}_g = \mathbf{0}$ の場合、 $\|(\partial \|\boldsymbol{\beta}_g\|_2) / \partial \boldsymbol{\beta}_g\|_2 < 1$ であるから

$$\|\mathbf{X}_g \mathbf{r}_{-g}\|_2 \leq \lambda \quad (22)$$

となる。(21) 式と (22) 式を組み合わせることで $\boldsymbol{\beta}_g$ の更新式

$$\hat{\boldsymbol{\beta}}_g = \left(1 - \frac{\lambda}{\|\mathbf{X}_g \mathbf{r}_{-g}\|_2} \right)_+ \mathbf{X}_g \mathbf{r}_{-g} \quad (23)$$

がわかる。

3.2 k-means クラスタリングのスパース推定

Group Lasso を用いて k-means クラスタリングの変数選択を行う手法が提案され、この手法は正則化 k-means クラスタリング (RegKM) と呼ばれる [25]。RegKM の目的関数は以下である。

$$\begin{aligned}
 RegKM(\mathbf{U}, \mathbf{C}|\mathbf{X}) &= KM(\mathbf{U}, \mathbf{c}_p|\mathbf{x}_p) + \lambda \sum_{p=1}^P \|\mathbf{c}_p\|_2^2 \\
 &= \frac{1}{2} \sum_{p=1}^P \|\mathbf{x}_p - \mathbf{U} \mathbf{c}_p\|_2^2 + \lambda \sum_{p=1}^P \|\mathbf{c}_p\|_2^2 \quad (24)
 \end{aligned}$$

Group Lasso の目的関数と比較すると、目的変数にあたる $\mathbf{y}$ が $\mathbf{x}_p$ に、デザイン行列と回帰係数にあたる $\sum_{g=1}^G \mathbf{X}_g \boldsymbol{\beta}_g$ が $\mathbf{U} \mathbf{c}_p$ になっていることがわかる。ところが、一般に $\mathbf{U}^\top \mathbf{U} \neq \mathbf{I}$ であるため、 $\mathbf{c}_p$ の更新式を陽に表すことができない。そこで Majorization を用いた更新を行う [26]。すなわち常に

$$RegKM(\mathbf{c}_p|\mathbf{U}, \mathbf{x}_p) \leq Maj(\mathbf{c}_p|\mathbf{U}, \mathbf{x}_p)$$

となる関数 $Maj(\mathbf{c}_p)$ を見つけ、 $RegKM(\mathbf{c}_p|\mathbf{U}, \mathbf{x}_p)$ 最小化の代

わりに $Maj(\mathbf{c}_p|\mathbf{U}, \mathbf{x}_p)$ を最小化することで $\mathbf{c}_p$ を更新する。具体的にはテイラー展開による関数の二次近似

$$\begin{aligned}
 Maj(\mathbf{c}_p|\mathbf{U}, \mathbf{x}_p) &= KM(\mathbf{c}_p^{\text{old}}|\mathbf{U}, \mathbf{x}_p) - (\mathbf{c}_p - \mathbf{c}_p^{\text{old}})^\top \mathbf{a} \\
 &\quad + \frac{1}{2} \gamma (\mathbf{c}_p - \mathbf{c}_p^{\text{old}})^\top (\mathbf{c}_p - \mathbf{c}_p^{\text{old}}) + \lambda \|\mathbf{c}_p\|_2 \quad (25)
 \end{aligned}$$

を代わりに最小化する。ただし

$$\begin{aligned}
 \mathbf{a} &= -\nabla KM(\mathbf{c}_p^{\text{old}}|\mathbf{U}, \mathbf{x}_p), \\
 \gamma &= \frac{1}{2} (1 + 10^{-6}) \times \max \left(\sum_{n=1}^N u_{n1}, \dots, \sum_{n=1}^N u_{nK} \right)
 \end{aligned}$$

とする。(25) 式の $\mathbf{c}_p$ についての劣微分を考えることで、 $\mathbf{c}_p$ の更新式

$$\hat{\mathbf{c}}_p = \frac{1}{\gamma} (\gamma \mathbf{c}_p^{\text{old}} + \mathbf{a}) \left(1 - \frac{\lambda}{\|\gamma \mathbf{c}_p^{\text{old}} + \mathbf{a}\|_2} \right)_+ \quad (26)$$

を得る。 $\mathbf{c}_p$ の推定は以下の 4 段階を繰り返す反復計算によって達成される。

1. $\mathbf{c}_p^{\text{old}}$ をランダムに作成する。
2. $\mathbf{c}_p$ を更新する。
3. $\mathbf{c}_p^{\text{old}} = \mathbf{c}_p$ とする。
4. 収束判定を行う。

RegKM の分析例を示すため、白血病のマイクロアレイデータを用いる [27, 28]。白血病データは急性骨髄性白血病と急性リンパ性白血病の患者計 72 名、遺伝子 7128 個の高次元データである。クラスタリングによって患者をいくつかのグループに分類する問題を考える。 $k = 2$, $\lambda = 50$ として RegKM を適用したところ、7128 個の遺伝子から 267 個の遺伝子を選択された。選択された遺伝子と患者の関係性を図 2 のようにヒートマップで可視化すると、2 つのグループの存在が示された。このグループ構造は、急性骨髄性白血病と急性リンパ性白血病との違いを反映したものとなっている。通常の k-means クラスタリングの場合、二つの疾患を正しく分類できた精度が 68% であったのに対して、RegKM では分類精度が 85% であった。そのため RegKM は探索的なアプローチとして優れているだけでなく、分類精度の観点からも優れた特徴を持つことが示唆された。

4. まとめ

本稿ではクラスタリングと特徴抽出を組み合わせることで、探索的なグループ構造の推定が容易となることを、主成分分析と k-means クラスタリングを中心に紹介した。また、この 2 つの手法を同時に推定する RKM は、クラスター内平方和がより小さくなる傾向があることから解釈しやすい利点があることをデータへの適用によって示した。2.3 節にて述べたようにこうした次元縮約法とクラスタリングの同時推定は近年活発に研究されており、そのファジィクラスタリングへの拡張も検討されている。

クラスタリングと相性の良い分析手法は、主成分分析などによる特徴抽出法だけではない。高次元データを扱う場合の一般的なツールとなったスパース推定も、クラスタリングと併せて用いるのに適していると考えられる。3 節では、スパー

図2 選択された遺伝子でのヒートマップ

ス推定について劣微分を用いた更新式の導出を中心に解説し、Group Lasso による k-means クラスタリングの変数選択である RegKM を紹介した。マイクロアレイデータへの適用例から、変数を選択することで高次元データであっても解釈が可能となるだけでなく、クラスタリングに関係があるだろう変数のみを選択することによって、分類精度が向上する可能性があることも示された。クラスタリングのスパース推定も近年研究が進められているテーマの一つであり、今後ファジィクラスタリングへ拡張されることが期待されている。

クラスタリングと特徴抽出の親和性は、機械学習においても注目されている。探索的分析を目的とするクラスタリングを行う前工程として、データを可視化する次元縮約法の研究が盛んに行われている。特に近年広く普及した手法として、t-SNE (t-Distributed Stochastic Neighbor Embedding) が挙げられる [29]。機械学習では画像や単語などのデータを大量に扱うことがあるが、それらデータどうしの関係性を可視化しグループ構造を見出すための前処理として t-SNE は分野を問わず用いられている。近年では UMAP (Uniform Manifold Approximation and Projection) という新たな可視化手法も t-SNE に代わり利用され始めており [30]、ますます研究が加速する可能性がある。

本稿で紹介したように、近年注目されている研究の方向性の一つとして、解釈性がキーワードとなっている。大規模なデータが収集可能となった一方で、そのデータの背後にある関係性などが未知の場合には、データを探索的に分析し仮説を生み出せるような結果を導く手法が必要とされるからである。こうした研究の潮流を考えると、ファジィ性は近年の大規模データ分析や機械学習における解釈性に新たな示唆を与える可能性があり、今後の研究の発展に期待したい。

参考文献

[1] U. von Luxburg, R. C. Williamson, and I. Guyon: "Clustering: Science or Art?," *Proc. of ICML Workshop on Unsupervised and Transfer Learning*, Jun. 26-Jul. 1, Edinburgh, pp. 65-79, 2012.
 [2] S.-M. Fereshtehnejad, Y. Zeighami, A. Dagher, and R. B. Postuma: "Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression," *Brain*, Vol.140, pp. 1959-1976,

2017.
 [3] H. P. Kriegl, P. Kröger, and A. Zimek: "Subspace clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol.2, pp. 351-364, 2012.
 [4] G. De Soete and J. D. Carroll: "K-means clustering in a low-dimensional Euclidean space," in *New Approaches in Classification and Data Analysis*, Springer, pp. 212-219, 1994.
 [5] 森裕一, 黒田正博, 足立浩平: 最小二乗法・交互最小二乗法, 共立出版, 2017.
 [6] J. R. Magnus and H. Neudecker: *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, 2019.
 [7] P. H. Schönemann: "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, Vol.31, pp. 1-10, 1966.
 [8] J. M. Ten Berge: "A generalization of Kristof's theorem on the trace of certain matrix products," *Psychometrika*, Vol.48, pp. 519-523, 1983.
 [9] J. A. Hartigan: *Clustering Algorithms*, John Wiley, 1975.
 [10] J. C. Dunn: "A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters," *J. of Cybernetics*, Vol.3, No.3, pp. 32-57, 1973.
 [11] J. C. Dunn: "Well-separated clusters and optimal fuzzy partitions," *J. of Cybernetics*, Vol.4, No.1, pp. 95-104, 1974.
 [12] 宮本定明, 馬屋原一孝, 向殿政男: "ファジィ c-平均法とエントロピー正則化法におけるファジィ分類関数," *日本ファジィ学会誌*, Vol.10, pp. 548-557, 1998.
 [13] 西田豊: "曖昧さを表現する概念形成モデル," *知能と情報*, Vol.22, pp. 434-442, 2010.
 [14] Z. Ghahramani and G. E. Hinton: "The EM algorithm for mixtures of factor analyzers," *University of Toronto Technical Report*, CRG-TR-96-1, 1996.
 [15] K. Uno, H. Satomura, and K. Adachi: "Fixed factor analysis with clustered factor score constraint," *Computational Statistics & Data Analysis*, Vol.94, pp. 265-274, 2016.
 [16] 山下直人, 前川眞一: "K-平均クラスタリングによる個体・変数の分類を伴うパイプロット," *データ分析の理論と応用*, Vol.2, pp. 33-51, 2012.
 [17] N. Yamashita and S.-I. Mayekawa: "A new biplot procedure with joint classification of objects and variables by fuzzy c-means clustering," *Advances in Data Analysis and Classification*, Vol.9, pp. 243-266, 2015.
 [18] K. Tanioka and H. Yadohisa: "Discriminant coordinates for asymmetric dissimilarity data based on radius model," *Behaviormetrika*, Vol.43, pp. 1-17, 2016.
 [19] M. van de Velden, A. I. D'Enza, and F. Palumbo: "Cluster correspondence analysis," *Psychometrika*, Vol.82, pp. 158-185, 2017.
 [20] M. Yamamoto: "Clustering of functional data in a low-dimensional subspace," *Advances in Data Analysis and Classification*, Vol.6, pp. 219-247, 2012.
 [21] M. Mitsuhiro and H. Yadohisa: "A unified representation of simultaneous analysis methods of reduction and clustering," *Japanese J. of Statistics and Data Science*, Vol.1, pp. 393-412, 2018.
 [22] R. Tibshirani: "Regression shrinkage and selection via the lasso," *J. of the Royal Statistical Society: Series B*, Vol.58, pp. 267-288, 1996.
 [23] 川野秀一, 松井秀俊, 廣瀬慧: スパース推定法による統計モデリング, 共立出版, 2018.
 [24] M. Yuan and Y. Lin: "Model selection and estimation in regression with grouped variables," *J. of the Royal Statistical Society, Series B*, Vol.68, pp. 49-67, 2006.
 [25] W. Sun, J. Wang, and Y. Fang: "Regularized k-means clustering of high-dimensional data and its asymptotic consistency," *Electronic J. of Statistics*, Vol.6, pp. 148-167, 2012.
 [26] Y. Yang and H. Zou: "A fast unified algorithm for solving group-lasso penalized learning problems," *Statistics and Computing*, Vol.25, pp. 1129-1141, 2015.
 [27] B. Efron and T. Hastie: *Computer Age Statistical Inference*, Cam-

- bridge University Press, 2016.
- [28] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander: “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, Vol.286, pp. 531-537, 1999.
- [29] L. van der Maaten and G. Hinton: “Visualizing data using t-SNE,” *J. of Machine Learning Research*, Vol.9, pp. 2579-2605, 2008.
- [30] L. McInnes, J. Healy, and J. Melville: “Umap: Uniform manifold approximation and projection for dimension reduction,” arXiv preprint, arXiv:1802.03426, 2018.

(2021 年 4 月 14 日 受付)

[問い合わせ先]

〒461-8673 愛知県名古屋市東区大幸南一丁目 1 番 20 号

名古屋大学 大学院医学系研究科

E-mail: kohei.uno.stat@gmail.com

—— 著 者 紹 介 ——

うの こうへい
宇野 光平 [非会員]

2019 年大阪大学大学院人間科学研究科修了。
現在、名古屋大学大学院医学系研究科特任助教。日本分類学会などの会員。