ランダム行列の非線形手法によるパターン認識への応用

理学専攻・情報科学コース 2040653 廣田 梨那

1 はじめに

近年, コンピュータやインターネットが発達し, 大量のデータを集めることが容易になった. ランダムに収集された大量のデータは, 本来何らかの目的をもって集められる一方で, データの大きさや乱雑さが邪魔をして, 重要な特徴を抽出するのが難しい. クラスタリングは, 機械学習における教師なし学習の一種で, データ間の類似度に基づいて, データをグループ分けする手法である. これにより, 大規模かつ複雑なデータの特徴を把握することができる. 本研究では, カーネル行列にランダム行列の理論を応用することで, クラスタリングにおける最適なクラスタ数を推定する手法を提案する. 具体的には, ランダムなガウスカーネル行列を Wishart 行列とみなし, Marchenko-Pastur 則を適用する. これにより, 識別の難しい非線形データに対しても, データの特徴を把握し, 最適なグループ分けが可能になると考えられる.

2 ガウスカーネルとカーネル行列

2.1 ガウスカーネル

ガウスカーネルとは, 以下の式で表される 2 変数関数である. ただし, β は正の定数であらかじめ適当に定められるパラメータである.

$$k(x,y) = \exp\big(-\boldsymbol{\beta}||x-y||^2\big),$$

ガウスカーネルを用いて特徴空間に写像した行列は相関行列と同じような振る舞いをすることが知られている.

2.2 カーネル行列

$$K = \begin{pmatrix} k(x_1,x_1) & k(x_2,x_1) & \cdots & k(x_n,x_1) \\ k(x_1,x_2) & k(x_2,x_2) & \cdots & k(x_n,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_1,x_n) & k(x_2,x_n) & \cdots & k(x_n,x_n) \end{pmatrix},$$

ガウスカーネルの二次形式は以下のように常に非負, つまり正定値性を持っている.

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K_{ij} \geq 0, \quad \forall n \in \mathbb{N}, \forall \alpha_i \in \mathbb{R}$$

3 Wishart 行列と Marchenko-Pastur 則

3.1 Wishart 行列

一般にランダム行列とは確率変数を要素にもつ行列であり, その代表例として Wishart 行列が挙げられる. $G_{N,M}$ を $N \times M$ の Ginibre 行列, つまり平均 0 分散 1 の標準ガウス分布に従う変数を要素に持つランダム行列とし, その $N \times N$ の共分散行列を,

$$W_N = \frac{1}{N} {G_{N,M}}^T G_{N,M}$$

とする. これは Wishart 行列と呼ばれる.

3.2 Marchenko-Pastur 則

Wishart 行列の漸近固有値分布は Marchenko-Pastur 分布であり, $N \times M$ の Ginibre 行列の漸近的縦横比, つまり $\lambda = \frac{N}{M}$ のとき, その漸近固有値分布は以下の式で与えられる.

$$d\pi_\lambda(x) = \frac{\sqrt{-(x-\lambda_-)(x-\lambda_+)}}{2\pi\lambda x} 1_{[\lambda_-,\lambda_+]}(x)dx + \max\{0, 1-\lambda\}\delta_0(x),$$

ただし, $\lambda_\pm = (1 \pm \sqrt{\lambda})^2$ とする.
ランダム行列の行列サイズが十分に大きい時に, 固有値経験分布が一定の分布に近づく. これを漸近固有値分布という.

3.3 Catalan 数

Marchenko-Pastur 分布において, $\lambda = 1$ のとき, その n 次モーメントは Catalan 数 C_n と呼ばれる自然数列となる.

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!}, \quad n \geq 1 = 1, 2, 5, 14, 42, 132, 429, \cdots.$$

4 最適なクラスタ数推定の手法について

4.1 実験概要

ガウスカーネルのパラメータ β の変化によるカーネル行列の鎖が TDA のフィルトレーションの候補になり得るのではないかと考えている. TDA は位相的データ解析のことで, 与えられたノイズを含む幾何学的データから単体複体のフィルトレーションを考え, 位相的性質に重要度を割り当てることにより, データの本質部とノイズ部を識別するものである. これは, ランダム行列によりノイズ部を除去し, データモデルを明らかにする流れと共通していると考えられる. つまり, カーネルのパラメータ β とデータ半径 r の関係性を考えることが, 最適なクラスタ数を推定することに繋がるのではと考えている.

4.2 手順 1: クラスタ数と優固有値の一致

先行研究 [1] より, ランダムカーネル行列に内在するノイズは適当なスケーリング変換により Marchenko-Pastur 分布を用いて推定可能であり, 固有値分布と Marchenko-Pastur 分布とを比較することによりノイズ部を推定できる. すなわち優固有値を取り除き, 残りの固有値のモーメントが Catalan 数に近づけるように β を変化させることで, データの本質的次元部とノイズ部を切り分けることが可能であると考えられる. 具体的には, 取り除く優固有値の個数とクラスタ数が一致することを利用する.

4.3 手順 2: パラメータ β とデータ半径 r の関係

シミュレーションデータを用いた以下の数値計算により, パラメータ β とデータ半径 r には $-\frac{1}{2}$ の冪乗則が成り立つことが明らかになった. これにより, β から r を算出して最適なクラスタ数を推定できると考えられる.

4.4 実験例

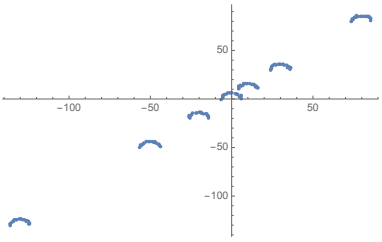

図1　パラメータ $\boldsymbol{\beta}$: 0.00001∼0.002

表1　β と r の関係

C	6→ 5	5→ 4	4→ 3	3→ 2
$\boldsymbol{\beta}$	0.00035	0.00013	0.000045	0.0000125
r	8.3	8.9	15.9	29.6

$$\log r = -0.5032 \log \boldsymbol{\beta} + \text{Const.}$$

$$r \propto \boldsymbol{\beta}^{-\frac{1}{2}}$$

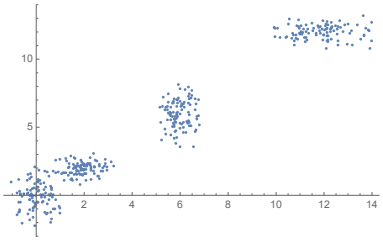

図2　パラメータ $\boldsymbol{\beta}$: 0.00001∼0.03

表2　β と r の関係

C	4→ 3	3→ 2	2→ 1
$\boldsymbol{\beta}$	0.0092	0.00195	0.000066
r	0.2	1.31	2.56

$$\log r = -0.4646 \log \boldsymbol{\beta} + \text{Const.}$$

$$r \propto \boldsymbol{\beta}^{-\frac{1}{2}}$$

図 1, 図 2のデータに対してクラスタリングを行った. データの点の周りに小さい円を描き, その半径を r とする. r を大きくすればするほど, 隣の円と重なり, クラスタ数が少なくなり最後は一つの塊になる. 何個の塊の状態が続くかを捉えることで自動的にクラスタ数を決定することができる. β の見つけ方は, 優固有値を取り除いた際に残りの固有値のモーメントが Catalan 数に近づくように β を動かしている. 表 1, 表 2より, 回帰分析を用いると r は $\boldsymbol{\beta}^{-\frac{1}{2}}$ に近似されることが分かる. 表 3は, 図 2のデータに対して, r を抽出せず β の値のみから最適なクラスタ数を推定したものである. この場合, 最適なクラスタ数は 2, 次に 3 と読み取れる.

表3　最適なクラスタ数推定

C	6→ 5	5→ 4	4→ 3	3→ 2	2→ 1
$\boldsymbol{\beta}$	0.022	0.014	0.0092	0.00195	0.000066
r	6.74200	8.45154	10.4257	22.6455	123.091

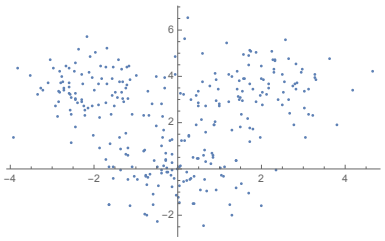

図3　パラメータ $\boldsymbol{\beta}$: 0.001∼0.03

表4　最適なクラスタ数推定

C	6→ 5	5→ 4	4→ 3	3→ 2	2→ 1
$\boldsymbol{\beta}$	0.0284	0.0251	0.0235	0.00126	0.00113
r	5.93391	6.31194	6.52328	28.1718	29.7482

表 4は, 図 3のデータに対して, r を抽出せず β の値のみから最適なクラスタ数を推定したものである. この場合, 最適なクラスタ数は 3 と読み取れる. r の値が大きく変化してもクラスタ数が変化しないところが最適なクラスタ数と推定できる. このデータは実際に正三角形の頂点に対してノイズを入れて作成したデータであるため, この実験結果は事前の予想とも一致している.

5　おわりに

先行研究 [1] で得られた結果, 今回の数値計算で得られた結果を合わせて, 実際の r の値を調べることなく, カーネル行列のパラメータ β の動きを見ることで, 最適なクラスタ数を推定できることが分かった.
数値計算により, パラメータ β とデータ半径 r には $-\frac{1}{2}$ の冪乗則が成り立つことを明らかにしたことは, 本研究の独創的な部分である. TDA では, 幾何データは Persistent Homology を経由して, Persistent 図に展開される. 今後の課題としては, Persitent 図のプロットをランダム行列の言葉で記述することを目指す.

参考文献

[1] 根本優花, ” パラメータの変動に伴うガウス型カーネル行列の固有値の挙動”, お茶の水女子大学理学専攻修士論文, 2020.