モーメント法によるノイズ推定を用いたスペクトラルクラスタリング

理学専攻 情報科学コース　茨木志織　(指導教員：吉田裕亮)

1　はじめに

データをまとまりごとに集めてグループ分けをすることは, クラスタリングと呼ばれている. クラスタリングは, 階層的手法と非階層的手法の 2 つに大きく分類され, 非階層的手法の代表例に $K-$ 平均法がある. $K-$ 平均法は非常に有効なクラスタリング手法ではあるが, 反復演算を必要とする点と, 収束解が必ずしも目的関数を最適にするものではないという欠点がある. スペクトラルクラスタリングでは, クラスタリングの問題を固有値問題として定式化することによって, これらの問題点を避けるアルゴリズムを構成することが出来る.

本研究では, データから得られるカーネル行列のスペクトル分布を, ランダム行列理論で知られているスペクトル分布と, モーメント法を用いて照らし合わせ, データの構造部を推定する. そして, 構造部のスペクトルのみを用いて, スペクトラルクラスタリングを行うことにより, その精度を向上させる手法を試みる.

2　ランダム行列理論

一般に, ランダム行列とは確率変数を要素に持つ行列であり, その代表例として Wishart 行列が挙げられる.

2.1　Wishart 行列

各成分が独立に $N(0,1)$ の標準正規分布に従う変数をもつ $n \times p$ の行列を C とする. このランダム行列 C から

$$S = \frac{1}{n} C\ {}^T C$$

で求められる $n \times n$ 対称ランダム行列 S を Wishart 行列という. $p/n = \lambda$ を保ちながら, $n \to \infty$, $p \to \infty$ の極限をとると, Wishart 行列 S の固有値の経験分布は, $\lambda_{min} \leq t \leq \lambda_{max}$ のときに以下の確率密度関数に収束することが知られている.

$$p(t) = \frac{1}{2\pi} \frac{\sqrt{-(t-\lambda_{max})(t-\lambda_{min})}}{\lambda t} \quad ,$$

$$\lambda_{min}^{max} = (1 \pm \sqrt{\lambda})^2.$$

また, このような確率密度関数を持つ分布は Marcenko-Pastur 分布と呼ばれている.

2.2　ガウスカーネル

変数の集合の二つの要素 x, x' に対し、カーネル関数 $k(x, x')$ は x, x' それぞれの特徴ベクトルどうしの内積

$$k(x, x') = \phi(x)^T \phi(x')$$

として定義される. カーネルには様々なものがあるが, その中でもガウスカーネル

$$k(x, x') = \exp(-\beta ||x - x'||^2)$$

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. ここで, $||\ ||^2$ は通常のユークリッド 2 乗距離で, $\beta \in R$ は適当なパラメータである.

また, Wishart 行列の固有値分布と, ガウスカーネルで写像した特徴空間における内積行列の固有値分布は等価であると知られており, これによりガウスカーネル行列におけるノイズ部に相当する固有値分布も Marcenko-Pastur 分布と同様の性質を持つことがわかる.

3　モーメント法

$f(x)$ を連続確率変数 X の密度関数とするとき, 原点まわりの k 次モーメント m_k は,

$$m_k = \int_{-\infty}^{\infty} x^k f(x) dx$$

と表せ, これらの値は確率分布の特徴を与える. また, 先にランダム行列理論で述べた Marcenko-Pastur 分布のモーメントは,

$$m_k = \frac{2k!}{k!(k+1)!}\ m_1^k$$

で与えられる.

本研究では, 観測データからの標本モーメント列を理論値と比較することにより, 最適なノイズ部の推定を行う. 目視でスペクトルを比較するより, 定量的な推定が行える.

4　スペクトラルクラスタリング

スペクトラルクラスタリングは, サンプル点をグラフ構造として考え, 各頂点がサンプル点で, 枝にはサンプル点同士の近さを表す重みがついているとする. 例えば, サンプル点を 2 つのグループに分けると, それに伴いグラフも 2 分割される. 分割されたグループ間を結ぶ枝のことを分割のカットと呼び, このカットの重みの合計が小さくなるようにグループ分けを行う. 式で表すと, 以下のようになる.

$$\min_{\beta} \sum_{i,j} K_{ij} (\beta_i - \beta_j)^2 = 2\ {}^T \beta P \beta, \quad \beta_i = \pm 1$$

ここで, P は対角行列 Λ を $\Lambda_{ii} = \sum_{j=1}^{n} K_{ij}$ として, $P = \Lambda - K$ と書ける. β は 2 値ベクトルという制約がある. それは整数計画問題と呼ばれ, 一般には解くのが困難である. そこで, 整数という制約を取り払って任意の実数ベクトルに, ${}^T \beta \Lambda \beta = 1$ という条件の下, 制約を緩めることにより推定を行うことになる. この場合, 最小固有値 0 が存在するが, これはすべてのサンプルを 1 つにまとめてしまうという意味のない解のため, 実際には 2 番目以降の固有ベクトルの成分符号に基づいてクラスタリングを行う.

5　提案手法

本研究では、以下のようにスペクトラルクラスタリングを行うことを提案する.

1. サンプルデータのガウスカーネル行列 K

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

を構成し, そのスペクトル分布を求める.
2. Marcenko-Pastur 分布のモーメント列に最も適合するように, 1 で求めたスペクトルからモーメント法を用いてノイズ部と構造部を推定する.
3. 構造部のスペクトルのみを用いてカーネル行列 K' を再構成する.
4. この新たな K' を与えられたカーネル行列と見なし, スペクトラルクラスタリングを行う.

6 実験例

図 1 のような, 線形で分けることの出来ない 3 群のデータを用意する. 1 群, 2 群が各 100 点, 3 群が 200 点の合計 400 点のサンプルデータとなっている. これらをスペクトラルクラスタリングで, 3 つにクラス分けする.

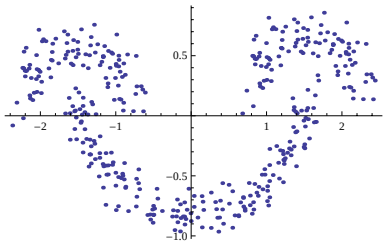

図 1: サンプルデータ

6.1 ノイズ部スペクトルの推定

	1乗	2乗	3乗	4乗	5乗	6乗
理論値	1	2	5	14	42	132
モーメント	1	2. 04523	5. 10791	14. 1827	42. 1473	131. 177

図 2: Marcenko-Pastur 分布のモーメント理論値 (上段) とスペクトル 338 個のモーメント値 (下段) の比較

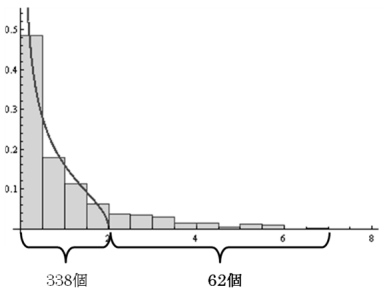

図 3: ヒストグラムと Marcenko-Pastur 分布

サンプルデータのガウスカーネル行列のスペクトルをもとにモーメントを計算していったところ, 値の小さい方から 338 個用いた時に, Marcenko-Pastur 分布のモーメント列に最も適合することが分かった (図 2). その 338 個のスペクトルを目安にして, 実際にサンプルデータのヒストグラムと Marcenko-Pastur 分布を重ね合わせてみると, 図 3 のように分布が合っていることも見て取れる. (今回, ガウスカーネル行列 K は, 正方行列なので λ は 1 となり, マルチェンコパスツール分布は図 3 の曲線のような形となっている.)

そこで, Marcenko-Pastur 分布から外れた残りのスペクトル 62 個がサンプルデータの構造部であると考え, このスペクトルのみを用いてカーネル行列を再編成し, スペクトラルクラスタリングを行った.

6.2 結果

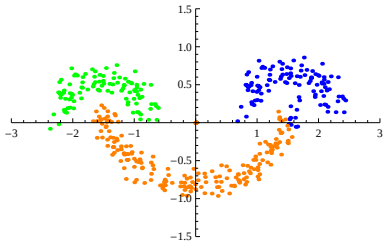

図 4: ノイズを除いた場合

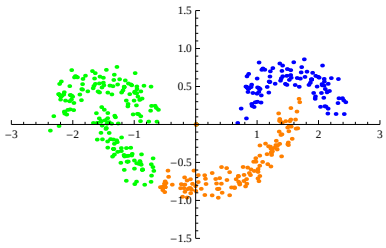

図 5: 除いていない場合

このとき, 図 4 の結果が得られた. 比較の為, 下に通常のスペクトラルクラスタリングの結果 (図 5) を並べておく. 図 4 は, ほぼ設定通りに 3 つにクラス分けが出来ている. ノイズを除いた場合の方が, クラスタリングが適切に行えていることが分かる.

7 まとめ

今回の実験で, データに含まれるノイズを除去することにより, スペクトラルクラスタリングの精度を高めることが可能である場合があることが, 確認出来た. また, ノイズ推定の際, モーメント法を用いて理論値と比較することによって, 目視でスペクトルを比較するより, 定量的な推定が行えることが分かった. また今後の課題として, Marcenko-Pastur 分布の適切なスケーリングパラメータの推定を行いたいと考えている.

参考文献

1. 赤穂昭太郎, カーネル多変量解析〜非線形データ解析の新しい展開, 岩波書店 (2008)
2. 伊藤里江, ランダム行列理論を用いた Gaussian カーネルにおける雑音の推定, お茶の水女子大学大学院理学専攻情報科学コース修士論文, 2009 年