相関のあるランダム行列に基づく時系列データのクラスタリング手法

理学専攻・情報科学コース　伊藤香織

1　はじめに

近年, 計算機の発達によりビックデータと呼ばれる大規模データの解析が注目されている. その中でも我々の生活に身近にあるものとして時系列データがある. また, ランダム行列理論に基づく最近の研究がデータ解析の分野でもその応用を期待されている. データ解析手法の一つとしてクラスタリングがある. データの集まりをデータ間の類似度に従って幾つかのグループに分けることでデータの特徴を把握しようというもので, その類似度をどのように設定するかが問題となる. 本研究ではランダム行列理論を応用した従来とは異なる時系列データのクラスタリング手法の提案を行う. 一般的に時系列データのクラスタリングではモデルパラメータの推定を行うが, 本研究ではパラメータの推定を行わず, モーメント値によって時系列のスペクトルの代用とし, クラスタリングを行う. パラメータの推定が必要ないため, 計算量を抑えることが可能であり大規模データの解析において提案手法が有用であると考える. 実際の時系列データに対して提案手法を適用し, そのクラスタリング結果も紹介する.

2　時系列モデル

時間の経過とともに形成されていくデータのことを一般に時系列データと呼ぶ. 時系列データの規則を記述する関係式を時系列モデルと呼び, その中に MA モデル（Moving Average model）がある. 次数 q の離散時間移動平均モデルを表し, 以下の数式で表わされる.

$$Y_n = \sum_{k=0}^{q} \theta_k Z_{n-k}$$

ここで $\theta_0,\cdots,\theta_q$ は MA モデルのパラメータで Z_i はホワイトノイズ $Z_j \sim N(0,1)$ *i.i.d.* である.

すなわち, MA モデルは時刻を現在から過去にさかのぼった撹乱項についての移動加重和として表わすモデルである.

3　ランダム行列理論

ランダム行列とは, 確率変数を要素にもつ行列である. また, ランダム行列が持つ統計的性質に関する理論をランダム行列理論という. 各要素が独立に標準正規分布に従う変数をもつ $N \times M$ ランダム行列を G とすると

$$S = \frac{1}{N} G^T G$$

で与えられる $N \times N$ 対象ランダム行列 S を *Wishart* 行列という. ここで M, N が漸近的に $M/N \to \lambda$, M, $N \to \infty$ となるような極限をとると, *Wishart* 行列 S の固有値経験分布は

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

ただし

$$\lambda_{max} = \left(1 + \sqrt{\lambda}\right)^2, \lambda_{min} = \left(1 - \sqrt{\lambda}\right)^2,$$

に収束することが知られている. これを *Marhenko-Pastur* 則と呼ぶ.

4　時系列の定常性

不規則に変動する時系列データに対して, その統計的な特性をより的確にとらえる試みとして確率過程に基づく分析がある. この場合, 時系列の統計的な特性は時間とともに変化しないといった定常性を仮定することが多い. 一般に時系列モデルが (弱) 定常性であるとは, 任意の時刻 t,s に対して

(I) $E(Y_t) = \mu \quad (< \infty)$

(II) $E(Y_t^2) = \sigma^2 \quad (< \infty)$

(III) $E[(Y_t - \mu)(Y_s - \mu)]$ が時差 $|t-s|$ のみに依存

が成り立つとき, この時系列モデルは定常であるという.

5　MA モデルを要素間の相関とするランダム行列

先に述べた *Wishart* 行列に関する結果は, データ行列の成分がすべて独立な場合であり, その共分散行列の *Marhenko-Pastur* 型の極限に対する極限分布を記述したものである. しかし, 最近 Hasegawa らにより [1] において行方向に関して MA モデルで相関が与えられるようなデータ行列, すなわち $X = (X_{i,j})_{N\times M}$ で, 各 i 行は

$$X_{i,j} = \sum_{k=0}^{q} \theta_k Z_{i,j-k}, \quad Z_{i,j} \sim \mathcal{N}(0,1) \; i.i.d.$$

と *MA* モデルで与えられる. これに対する共分散行列

$$\frac{1}{N} X^t X$$

の *Marhenko-Pastur* 型極限における極限分布は厳密に求められ, その極限分布のモーメントは MA パラメータのみに依存して定まることが示された. すなわち, 時系列の自己相関関数に依存して定まることになる.

したがって, 時系列データを行方向に適当に大きいサイズで並べて構成されるデータ行列の標本共分散行列の標本モーメント列には時系列データに内在する情報を含んでいるとみなせる.

そこで, 時系列データの MA モデル表現を用いて, このモデルのパラメータの推定をすることなく, これら時系列データを単純に並べることにより構成される行列のモーメント列を用いることで, 各時系列データの特徴を表すベクトルとみなせることが本研究における着眼点である.

また, 実データにおいては, 自己相関関数は時間と共に大きく減衰するため, 低次のモーメントらで時系列データの特徴を十分に捉えているものと考えられる.

6　提案手法

与えられた各時系列データ $\{X_i\}_{i=1}^{L}$ を適当な長さに区切り $N \times M$ 行列を構成する. ただし $NM \leqq L$

$$A = \begin{pmatrix} X_1 & \cdots & X_M \\ X_{M+1} & \cdots & X_{2M} \\ \vdots & \ddots & \vdots \\ X_{(N-1)M+1} & \cdots & X_{NM} \end{pmatrix}$$

これを相関のあるランダム行列の標本とみなして行列の k 次モーメント

$$m_k = tr\left[\left(\frac{1}{N}\,{}^tA_iA_i\right)^k\right]$$

を求められたモーメント値らを特徴ベクトルとして, クラスタリングを行う.

時系列の定常化

現実に存在する時系列データは非定常な時系列データであることが多い. そのひとつである気温データのような時系列データは季節による周期性を持つ. したがって時系列データに内在する周期性を除去し, 定常な時系列データに変換することを考える.

各時系列データ $\{X_i\}_{i=1}^{L}$ をそれぞれフーリエ変換し, パワースペクトルを見る. パワースペクトルの大きいものを時系列データの周期成分とみなし, 閾値を設定. 閾値より大きい値を 0 と設定し逆フーリエ変換によって周期成分を抜いた各時系列データ $\{X_i^{'}\}_{i=1}^{L}$ に対して本研究の手法を適用しクラスタリングを行う.

7　実験

7.1　実データによる実験

実際の時系列データとして日本の各地点における日別気温データを扱う. ある地域のある地点における L 日間の気温データを, その地域内で地点を変えて s 個用意する. この時系列データを D_i とする. 異なる地域における時系列データ群 $T = D_1, D_2, \cdots$ に対して提案手法を行い地域の違いによりクラスタリングされるかを調べる.

まず D_i を $N \times M$ 行列に区切り, この行列の k 次モーメント m_k を求める. k 次モーメントのベクトルを PCA 法によって次元削減することで可視化可能な空間にプロットしクラスタリング結果を見る.

7.2　実験例

気象庁 [2] が保有する日本の沖縄諸島, 関東内陸, 北海道の 3 つの地域における長さ 4018 の日別最高気温データをそれぞれ 10 個用意した. これを D_1, D_2, D_3 とする. この時系列データ D_i に対し, N=60, M=66 の $N \times M$ 行列を構成し, 第 5 次モーメントまで求めた. 得られたモーメントのベクトルから第 5 次キュムラント値までを求め, PCA 法により次元削減されたプロット結果を図 1 に示す.

7.3　定常化を行った実験例

各地点の時系列データ $\{X_i\}_{i=1}^{L}$ に対してフーリエ変換を行い, 周期性を除去した定常な時系列データ $\{X_i^{'}\}_{i=1}^{L}$

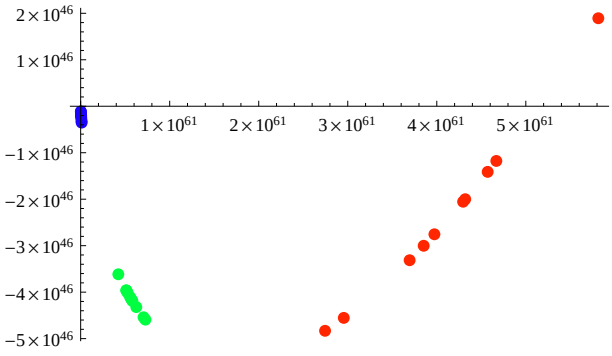

図 1: クラスタリング結果

に対して同様の操作を行った. プロット結果を図 2 に示す.

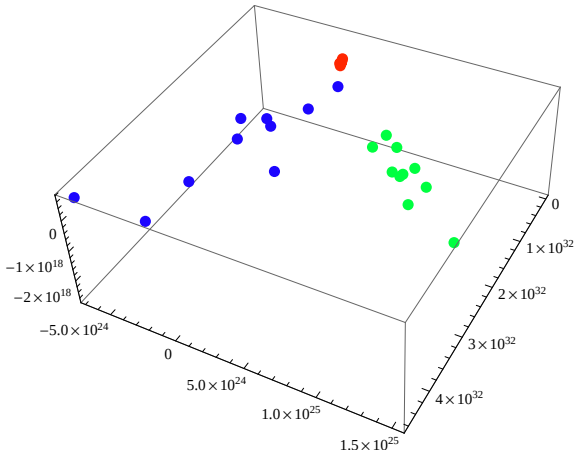

図 2: 定常化を行ったクラスタリング結果

図 1, 図 2 を比較すると図 1 はすべての点がある曲線に沿って分布することに対して, 図 2 は地域別に分布の仕方が異なっていることが見てとれる. 以上より, 周期成分をとることにより時系列データの特性をより捉えたクラスタリング結果になったといえる.

8　まとめ

本研究では, 要素間に相関が与えられるような MA モデルに対してランダム行列を構成し, その自己共分散行列の標本モーメント列が時系列データの特徴をあらわすことを応用した時系列データのクラスタリング手法を提案した. 実データを代表する気温データに対して, 提案手法を用いてデータに内在する特徴が抽出されたクラスタリング結果を見た. また, 実データの解析手段として時系列の定常化を行い, 定常化された時系列データに対して提案手法を適用したクラスタリング結果を併せて示した. 後者の方がより的確に特徴抽出できることを確認し, その実用性を示した.

参考文献

[1] A.Hasegawa,N.Sakumaand H.Yoshida,Randam Matrices by MA Models and Compound Free Poisson Laws,Probab.MAth.Statist,33(2)(2013),pp.243-254.

[2] http://www.jma.go.jp/ujma/index.html