ランダム行列の時系列データ解析への応用

黒崎志帆　(指導教員：吉田裕亮)

1　はじめに

身の回りには気象情報や経済現象の記録, 医学データなど様々な時系列データが存在する. これら時系列データの解析方法として, 一般的には時系列を構成するパラメータの推定が知られているが, 計算量が大きい等の問題が挙げられる. ここで本研究では, 時系列データから得られる行列を Compound Wishart 行列の標本とみなし, ランダム行列の特性を応用することで時系列データを解析することを目的とする.

2　ランダム行列

ランダム行列とは確率変数を要素に持つ行列のことを指す. 代表的なものに Wishart 行列があり, 行列 G を各要素が独立に標準正規分布に従う変数を持つ $n \times m$ ランダム行列とすると,

$$S = \frac{1}{n} G^t G$$

で与えられる. $m/n = \lambda$ を有限に保ちながら $n \to \infty, m \to \infty$ の極限をとると, 固有値経験分布はある分布 $p(t)$ に収束することが知られている. これを Marchenko-Pastur 則といい, 密度関数は以下の式で表される.

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

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

3　時系列モデル

時間の経過と共に不規則に変動する現象の記録が時系列である. 平均, 分散, 共分散が時間をシフトしても変化しないことを定常性といい, 本研究は定常な時系列を対象とする.

ARMA モデル

時系列モデルの代表例として ARMA モデルがあり, 時系列 y_n を過去の観測値 y_{n-i} と白色雑音の現在および過去の値の線形和で表現したモデルのことである.

$$y_n = \sum_{i=1}^{m} a_i y_{n-i} + v_n - \sum_{i=1}^{l} b_i v_{n-i}$$

m, l はモデルの次数, a_i, b_iは はパラメータ, v_n は白色雑音（ノイズ）である. ARMA モデルにおいて $m = 0$ と置くと移動平均部分のみの MA モデルが得られ, 以下の式で表される.

$$y_n = v_n - \sum_{i=1}^{l} b_i v_{n-i}$$

4　モーメント理論値

$\{x_1, x_2, \cdots, x_p\}$ を定常な時系列データとするとき, 自己分散は

$$\gamma(h) = Cov(x_{n+|h|}, x_n)$$

で与えられる. この $\gamma(h)$ を用いて作られる行列

$$\Gamma = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(n) \\ \gamma(1) & \gamma(0) & \dots & \gamma(n-1) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n) & \gamma(n-1) & \dots & \gamma(0) \end{pmatrix}$$

を考える. ここで, 時系列 x_n を $n \times m$ に折りたたんで以下の行列 X を構成する. ($\lambda = m/n$)

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_m \\ x_{m+1} & x_{m+2} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(n-1)m+1} & \dots & \dots & x_{nm} \end{pmatrix}$$

このとき

$$D = \frac{1}{n} X^t X$$

は, ガウスランダム行列 G と先程の Γ によって表される行列

$$W = \frac{1}{n} G \Gamma^t G$$

と相似であることが知られている. この行列 W は Compound Wishart 行列と呼ばれる. W のモーメント列を μ_k とすると, 1〜4 次のモーメント列は

$$\mu_1 = \lambda m_1$$

$$\mu_2 = \lambda^2 m_1^2 + \lambda m_2$$

$$\mu_3 = \lambda m_3 + 3\lambda^2 m_1 m_2 + \lambda^3 m_1^3$$

$$\mu_4 = \lambda m_4 + 4\lambda^2 m_1 m_3 + 2\lambda^2 m_2^2 + 6\lambda^3 m_1^2 m_2 + \lambda^4 m_1^4$$

で与えられることが知られている.

5　実験

時系列データから得られる行列 D のモーメント列と W から得られるモーメント理論値との誤差と λ の関係を求めることを目的とする. (実験 1) また, 得られた最適な λ を用いて時系列データのクラスタリングを試みる. (実験 2)

5-1 実験 1

1. 時系列データを用意し, $n \times m$ に折りたたむ.(行列 X)
2. $D = \frac{1}{n} X^t X$ のモーメント列 (1〜3 次) を調べる.

3. μ_k の理論値との誤差を調べ (50 回の平均をとる), λ の値による誤差の変化を調べる

用意したデータ

モデル	パラメータ
MA(2)	$b_1 = 1, b_2 = 1$
ARMA(2,2)	$a_1 = 0.5, a_2 = 0.3, b_1 = 0.9, b_2 = 0.7$

実験結果

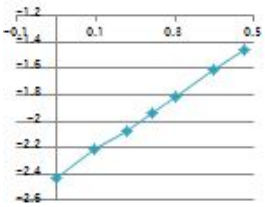

図 1: MA(2)

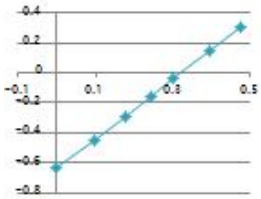

図 2: ARMA(2,2)

横軸：λ の値 (log)
縦軸：理論値との誤差二乗 (log)

グラフは 1 次モーメントの結果であり,2〜3 次も同様に比例関係がみられた.log の値が比例であることから誤差は λ の値に伴い指数関数的に増加することが分かる. また誤差が最も小さくなるのは $\lambda = 1$ の場合である.

5-2 実験 2

1. 実験 1 同様行列 X を作成し, モーメント列 (1〜3 次) を調べる.($\lambda = 1$)
2. (μ_1, μ_2, μ_3) を三次元空間にプロットし、時系列モデルごとにクラスタリングされるかを調べる.

用意したデータ

実験番号	モデル	パラメータ
[1]	ARMA(2,2)	$a_1 = 0.5, a_2 = 0.3$ $b_1 = 0.9, b_2 = 0.7$
	ARMA(1,3)	$a_1 = 0.5,$ $b_1 = 1, b_2 = 1, b_3 = 1$
	MA(2)	$b_1 = 1, b_2 = 1$
[2]	ARMA(2,2)	$a_1 = 0.5, a_2 = 0.3$ $b_1 = 0.8, b_2 = 0.5$
	ARMA(2,2)	$a_1 = 0.53, a_2 = 0.3$ $b_1 = 0.8, b_2 = 0.5$
	ARMA(2,2)	$a_1 = 0.56, a_2 = 0.3$ $b_1 = 0.8, b_2 = 0.5$

[1] では全くモデルの異なる三種類のデータ, [2] ではモデルは同じでパラメータのみ異なる類似したデータを用意. 20 組ずつ, 計 60 組でクラスタリングを試みる.

実験結果

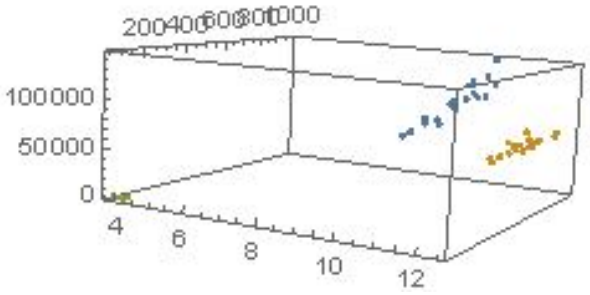

図 3: 実験 [1]

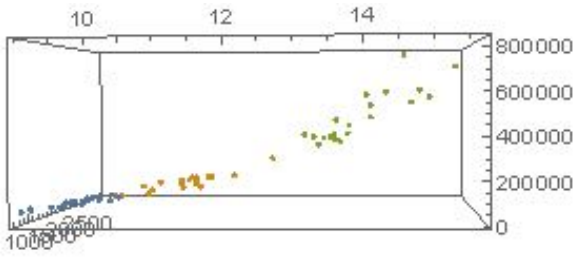

図 4: 実験 [2]

6 考察

時系列データから構成される行列 D のモーメント列と理論値との誤差は, λ の値に伴い指数関数的に増加するという関係性が得られた. さらに, 誤差を最も小さくする λ の値は $\lambda = 1$ であることが分かった.

クラスタリングの実験では, 実験 [1] のようにモデルの異なるデータでは明らかに 3 つにクラスタリングされた. また実験 [2] よりモデルは同じでパラメータのみ異なる類似したデータは, 分布は同一曲線状に連続して現れるという特徴があることが分かった.

7 おわりに

実験より, ランダム行列を利用し, モーメント列を推定することでパラメータ推定等の計算をせずにクラスタリングできることがわかった. 本研究では, 時系列のシミュレーションデータを作成してクラスタリング実験を行ったが, この結果は実データの解析結果の解釈に応用可能である.

8 参考文献

[1]Ayako Hasegawa, Noriyoshi Sakuma, Hiroaki Yoshida On limit spectral measures of Marchenko-Pastur limit of random matrices with dependent entries an application of fluctuations

[2] 北川源四郎著, 時系列解析プログラミング

[3] 渡辺澄夫, 永尾太郎, 樺島祥介, 田中利幸, 中島伸一 共著, ランダム行列の数理と科学