数理解析研究所講究録
1127巻 2000年 129-133



# マルコフ情報源から生成されるパタンの頻度分布について

九大システム情報科学研究科　香田　徹　(Tohru Kohda)
九大システム情報科学研究科　藤崎 礼志　(Hiroshi Fujisaki)

**Abstract:** In this note, we discuss how to identify a Markov information source by observing a sequence of symbols generated by the source. We theoretically discuss empirical distribution functions of random vectors generated by its sequences of finite-length.

## 1　はじめに

自然現象や社会現象は確率過程でしばしばモデル化される [1]. このとき，独立同分布（以下，i.i.d. と略称する）モデルとマルコフ・モデルが最も頻繁に用いられる．特に，情報理論においては，情報源が i.i.d. やマルコフ連鎖でモデル化される [2]–[5]. 現象を i.i.d. とみなすか，マルコフとみなすかは，何らかの物理量の時系列を観測することによって決まる．この場合，系列のどのような性質を基準として，系列が i.i.d もしくは，マルコフであると判断するかは，基本的ではあるが難しい問題である．考えられる全ての検定に合格したとしても，系列が i.i.d. であると判定することは難しい．推移確率を決定しなければならないマルコフの場合はさらに難しいといえる．

系列が2値 i.i.d. のとき，ある長さ $m$ のパタンの頻度に関する頻度分布はガウス分布に従うという中心極限定理が成立する．したがって，系列の i.i.d. 性評価法として，系列のパタンの頻度の期待値と分散を調べる評価法が考えられる．このテストは $m$ 次および $2m$ 次の相関関数を同時に調べることになる．すなわち，2値系列から生成される確率ベクトルのパタン，およびクラスの頻度に関する分布を調べ，それらの期待値と分散がどれだけガウス分布からずれるかを評価する．もし，期待値と分散が独立同分布の場合のそれらと一致すれば，系列の i.i.d. 性の必要条件が満たされたことになる

筆者の一人は，これまで，系列の i.i.d. 性の評価法として，系列のパタンの頻度の期待値と分散の理論的評価 [6]–[7] および実験による検証 [8] を行なってきた．小文では，系列のマルコフ性の評価法として，系列がマルコフ連鎖の場合のパタンの頻度の期待値と分散の理論的評価を行なう*.

## 2　マルコフ連鎖とマルコフ情報源

まず，マルコフ連鎖とマルコフ情報源について復習しよう [3]. 状態空間を $S=\{1,2,\cdots,N\}$，確率推移行列を $P=\{p_{ij}\}_{i,j=1}^{N}$ とする．ただし，任意の $i,j$ に対し，$p_{ij}\geq 0$ ; 任意の $i$ に対し，$\sum_{j=1}^{N}p_{ij}=1$ である．$S$ に値を取る確率変数列を $X_0,X_1,\cdots$ とする．任意の分布を有する $X_0$ に対して，

$$\mathrm{Prob}\{X_{n+1}=s_k|X_0=s_{i_0},\cdots,X_n=s_{i_n}\}=p_{i_n,k},$$

*2値の場合については，研究会「確率論と計算数学」(九州大学，1998年1月）において発表した.

$$\text{where } s_{i_j}(1 \leq j \leq N), s_k \in \mathcal{S} \tag{1}$$

のとき，確率変数列 $X_0, X_1, \cdots$ を $N$ 状態 マルコフ連鎖という (Prob($A$) は，事象 $A$ の起こる確率を表す)．マルコフ連鎖 $X_0, X_1, \cdots$ に対して， 定義域が $\mathcal{S}$ であり，値域がアルファベットの集合$\Gamma = \{u_1, \cdots, u_M\}$ である関数 $f$ を考える．初期状態 $X_0$ が定常分布 $\boldsymbol{p} = (p_1, \cdots, p_N)$ に一致するように与えられたとする．すなわち，

$$\text{Prob}\{X_0 = s_j\} = p_j, \quad \text{for all states } s_j, \tag{2}$$

であるとき，定常系列 $U_n = f(X_n)$, $n = 0, 1, 2, \cdots$ はマルコフ情報源と呼ばれる†．小文では，簡単のため，$\Gamma = \mathcal{S}$，$N = M$，さらに $f$ を恒等写像とする．

## 3 マルコフ情報源から生成される系列のパタンに関する分布

長さ $m$ の任意の系列を

$$\boldsymbol{U} = U_0 U_1 \cdots U_{m-1}, \quad U_k \in \mathcal{S}, \quad (0 \leq k \leq m-1). \tag{3}$$

とすると，$\boldsymbol{U}$ は $N^m$ 種類存在する．その $r$ 番目の系列パタンを

$$\boldsymbol{u}^{(r)} = u_0^{(r)} u_1^{(r)} \cdots u_{m-1}^{(r)} \quad (0 \leq r \leq N^m - 1) \tag{4}$$

とする．ただし，$u_k^{(r)} \in \mathcal{S}$, $(0 \leq k \leq m-1)$．マルコフ情報源から生成される系列 $\{X_n\}_{n=0}^{\infty}$ におけるパタン $\boldsymbol{u}^{(r)}$ の発生頻度を考察するために，2 値確率変数

$$Y_n(\boldsymbol{u}^{(r)}) = \begin{cases} 1 & (X_n X_{n+1} \cdots X_{n+m-1} = \boldsymbol{u}^{(r)}) \\ 0 & (X_n X_{n+1} \cdots X_{n+m-1} \neq \boldsymbol{u}^{(r)}) \end{cases}. \tag{5}$$

を導入する．$\{X_n\}_{n=0}^{T+m-2}$における $\boldsymbol{u}^{(r)}$ の個数は

$$M_T(\boldsymbol{u}^{(r)}) = \sum_{n=0}^{T-1} Y_n(\boldsymbol{u}^{(r)}), \tag{6}$$

で与えられる．ここで確率変数

$$Z_T(\boldsymbol{u}^{(r)}) = \frac{M_T(\boldsymbol{u}^{(r)}) - T\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]}{\sqrt{T}}. \tag{7}$$

を導入すると，$Z_T(\boldsymbol{u}^{(r)})$ の分散は次式で与えられる．

$$\begin{aligned} &Var(Z_T(\boldsymbol{u}^{(r)})) \\ &= \frac{1}{T}\mathbf{E}[M_T^2(\boldsymbol{u}^{(r)})] - T\{\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]\}^2, \end{aligned} \tag{8}$$

†文献 [4] で述べられているように，情報理論におけるマルコフ情報源は隠れマルコフ連鎖を含む．しかしながら，小文では，単純マルコフ情報源のみを議論する．

ただし，$\mathbf{E}[\cdot]$ は期待値（集合平均）を表す．連鎖が既約で非周期的であるならば，$T \to \infty$ のとき，(8) の極限が存在する [9]．その極限を

$$\sigma^2(\boldsymbol{u}^{(r)}) = \lim_{T-\infty} Var(Z_T(\boldsymbol{u}^{(r)})) \tag{9}$$

と表す．さらに，中心極限定理が成立し，$Z_T(\boldsymbol{u}^{(r)})$ の分布は，平均値 0 ，分散 $\sigma^2(\boldsymbol{u}^{(r)})$ のガウス分布

$$\phi(x; \boldsymbol{u}^{(r)}) = \frac{1}{\sqrt{2\pi}\sigma(\boldsymbol{u}^{(r)})} e^{-\frac{x^2}{2\sigma^2(u^{(r)})}} \tag{10}$$

に近付く．これより，系列のマルコフ性の評価法として，i.i.d. の場合 [6]–[8] と同様に，系列のパタンの頻度に関する分布を調べる評価法が考えられる．以下，分散を具体的に求めよう．

## 4 分散の評価

我々が観測できるのは，$\frac{1}{T}M_T(\boldsymbol{u}^{(r)})$ であるが，$T$が十分大きいとき，

$$\frac{1}{T}M_T(\boldsymbol{u}^{(r)}) \to \mathbf{E}[Y_n(\boldsymbol{u}^{(r)})] \tag{11}$$

となる．$\frac{1}{T}M_T(\boldsymbol{u}^{(r)})$ のヒストグラムを作成するためには，$\frac{1}{T}M_T(\boldsymbol{u}^{(r)})$ を十分多く観測する．$i$ 番目の観測量を $\frac{1}{T}M_T^i(\boldsymbol{u}^{(r)})$ と表すと，$\frac{1}{T}M_T^i(\boldsymbol{u}^{(r)})$ $(0 \le i \le L-1)$ が統計的に独立で，$T$ と $L$ が十分大きいとき，ヒストグラムの分散は，

$$\frac{1}{L}\sum_{i=0}^{L-1}\frac{1}{T}\left(M_T^i(\boldsymbol{u}^{(r)}) - T\frac{1}{L}\sum_{i=0}^{L-1}\frac{1}{T}M_T^i(\boldsymbol{u}^{(r)})\right)^2 \to Var[Z_T(\boldsymbol{u}^{(r)})]. \tag{12}$$

となる．したがって，以下，時間平均を集合平均で議論する．

$$\begin{aligned}\mathbf{E}[M_T^2(\boldsymbol{u}^{(r)})] &= T\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})^2] \\ &+ \sum_{i=1}^{m-1} 2(T-i)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})Y_{n+i}(\boldsymbol{u}^{(r)})] \\ &+ \sum_{i=m}^{T-1} 2(T-i)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})Y_{n+i}(\boldsymbol{u}^{(r)})].\end{aligned} \tag{13}$$

(13) の右辺第 2 項は

$$\begin{aligned}&\sum_{i=1}^{m-1} 2(T-i)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})Y_{n+i}(\boldsymbol{u}^{(r)})] \\ &= \sum_{i=1}^{m-1} 2(T-i)\left(p_{u_0^{(r)}}\prod_{k=1}^{i-1}p_{u_{k-1}^{(r)},u_k^{(r)}}\prod_{k=0}^{m-i-1}p_{u_{k+i-1}^{(r)},u_{k+i}^{(r)}}\delta_{u_{k+i}^{(r)},u_k^{(r)}}\prod_{k=0}^{i-1}p_{u_{k+m-i-1}^{(r)},u_{k+m-i}^{(r)}}\right)\end{aligned} \tag{14}$$

となる．ただし，$\delta_{i,j}$ は Dirac のデルタ記号である．(13) の右辺第 3 項は

$$\begin{aligned}
&\sum_{i=m}^{T-1} 2(T-i)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})Y_{n+i}(\boldsymbol{u}^{(r)})] \\
&= \sum_{i=m}^{T-1} 2(T-i)\left(p_{u_0^{(r)}} \prod_{k=1}^{m-1} p_{u_{k-1}^{(r)},u_k^{(r)}} (P^i)_{u_{m-1}^{(r)},u_0^{(r)}} \prod_{k=1}^{m-1} p_{u_{k-1}^{(r)},u_k^{(r)}}\right) \\
&= (T-m)(T-m+1)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]^2 \\
&+\frac{1}{p_{u_0}}\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]^2 \sum_{j=2}^{N}\sum_{i=m}^{T-1}(T-i)\lambda_j^i h_{u_{m-1}^{(r)},j} g_{u_0^{(r)},j}
\end{aligned} \tag{15}$$

となる．ただし，$\boldsymbol{g}_i = (g_{1,i}, g_{2,i}, \cdots, g_{n,i})^t$，$\boldsymbol{h}_i = (h_{1,i}, h_{2,i}, \cdots, h_{n,i})^t$ は各々，$P$ の $i$ 番目の固有値 $\lambda_i$ に対する左および右固有ベクトルを表し，次式を満たす．

$$\boldsymbol{g}_i^t \boldsymbol{h}_j = \delta_{i,j}, \quad 1 \le i,j \le N. \tag{16}$$

ここで，右肩の $t$ は転置を表す．さらに，$\lambda_1 = 1$ とした．このとき，$\boldsymbol{g}_1 = \boldsymbol{p}$，$\boldsymbol{h}_1 = \overbrace{(1,\cdots,1)}^{N}$ である．以上より，

$$\begin{aligned}
&Var(Z_T(\boldsymbol{u}^{(r)})) \\
&= \mathbf{E}[Y_n(\boldsymbol{u}^{(r)})] \\
&+\frac{1}{T}\sum_{i=1}^{m-1} 2(T-i)\left(p_{u_0^{(r)}} \prod_{k=1}^{i-1} p_{u_{k-1}^{(r)},u_k^{(r)}} \prod_{k=0}^{m-i-1} p_{u_{k+i-1}^{(r)},u_{k+i}^{(r)}} \delta_{u_{k+i}^{(r)},u_k^{(r)}} \prod_{k=0}^{i-1} p_{u_{k+m-i-1}^{(r)},u_{k+m-i}^{(r)}}\right) \\
&+\left(\frac{(T-m)(T-m+1)}{T} - T\right)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]^2 \\
&+\frac{1}{p_{u_0}}\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]^2 \sum_{j=2}^{N}\frac{1}{T}\sum_{i=m}^{T-1}(T-i)\lambda_j^i h_{u_{m-1}^{(r)},j} g_{u_0^{(r)},j}.
\end{aligned} \tag{17}$$

$T \to \infty$ のとき，

$$\begin{aligned}
&\lim_{T-\infty} Var(Z_T(\boldsymbol{u}^{(r)})) = \sigma(\boldsymbol{u}^{(r)})^2 \\
&= \mathbf{E}[Y_n(\boldsymbol{u}^{(r)})] \\
&+2\sum_{i=1}^{m-1}\left(p_{u_0^{(r)}} \prod_{k=1}^{i-1} p_{u_{k-1}^{(r)},u_k^{(r)}} \prod_{k=0}^{m-i-1} p_{u_{k+i-1}^{(r)},u_{k+i}^{(r)}} \delta_{u_{k+i}^{(r)},u_k^{(r)}} \prod_{k=0}^{i-1} p_{u_{k+m-i-1}^{(r)},u_{k+m-i}^{(r)}}\right) \\
&+(1-2m)\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]^2 \\
&+\frac{1}{p_{u_0}}\mathbf{E}[Y_n(\boldsymbol{u}^{(r)})]^2 \sum_{j=2}^{N}\frac{\lambda_j^m}{(1-\lambda_j)} h_{u_{m-1}^{(r)},j} g_{u_0^{(r)},j}.
\end{aligned} \tag{18}$$

## 5 まとめ

小文では，系列がマルコフ連鎖の場合，系列から生成される確率ベクトルのパタンの頻度分布が形成するガウス分布の分散を理論値を与えた．これより，実際に観測される系列から生成される確率ベクトルのパタンに関する分布を調べ，マルコフのそれと比較することによって，系列がどれだけマルコフ連鎖に近いかを評価することができる．

## References


[1] W. Feller, *An Introduction to Probability Theory and Its Applications*, Volume 1, Second Edition, John Wiley & Sons, Inc., 1957.

[2] F. M. Reza, *An Introduction to Information Theory*, Dover, 1961.

[3] R. B. Ash, *Information Theory*, Dover, 1965.

[4] C. M. Goldie and R. G. E. Pinch, "Communication Theory", London Mathematical Society Student Texts 20, Cambridge University Press, 1991.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[6] 香田 徹，大賀崇弘，常田明夫，"2値系列の頻度分布について"，信学技報，CAS96-10, VLD96-10, DSP96-30, pp. 65–68, 1996.

[7] 香田 徹，藤崎礼志，大賀崇弘，"不等確率2値系列の頻度分布について"，信学技報，NLP97-39, pp. 1–4, 1997.

[8] 香田 徹，宗 広和，"BBS 生成器と離散カオス生成器の i.i.d. 性に基づくランダムネステスト" SCIS '98-5.1.A, 1998.

[9] P. Billingsley, *Probability and Measure* Third Edition, John Wiley & Sons, Inc., 1995.