マルコフモデルの幾何学について

竹内純一
九州大学 大学院システム情報科学研究院

2009年 1月 24日

内容

目標 : Markovモデルの期待値パラメータに関するFisher情報量の行列式を求める

目次 :

- マルコフモデルの指数型分布族
- 確率的コンプレキシティ
- 最尤符号
- ベイズ符号
- Fisher情報量の行列式

定常マルコフモデル(単純マルコフ連鎖)

アルファベット

$$\mathcal{X} := \{0, 1, 2, ..., d\}$$

マルコフカーネル$p_{j|i}$ $(i \in \mathcal{X}, j \in \mathcal{X})$

$$\sum_{j=0}^{d} p_{j|i} = 1, \;\; p_{j|i} > 0$$

$x_t = i$の条件のもと$x_{t+1} = j$を観測する確率

$$q(j|i, \boldsymbol{p}) := p_{j|i}$$

列x_0^nに関する確率質量関数

$$q(x_0^n|\boldsymbol{p}) := q(x_0|\boldsymbol{p}) \prod_{t=1}^{n-1} q(x_t|x_{t-1}, \boldsymbol{p}) = q_0(x_0|\boldsymbol{p}) \prod_{ij \in \mathcal{X}^2} (p_{j|i})^{\tau_{ij}}$$

ただし,

$$x_m^n := x_m x_{m+1} ... x_n, \quad x^n := x_1 x_2 ... x_n, \quad \boldsymbol{p} := (p_{j|i})_{i \in \mathcal{X}, j \in \mathcal{X} \setminus \{0\}},$$

$$\tau_{ij} := \#\{t | x_t x_{t+1} = ij, 0 \leq t \leq n-1\}, \quad q_0(x|\boldsymbol{p})は\boldsymbol{p}で定まる定常確率分布$$

マルコフモデルの定常確率分布

$$q_0(x|\boldsymbol{p}) = \frac{\Delta_{xx}}{\sum_{i\in\mathcal{X}} \Delta_{ii}}$$

$$\Delta_{xy} := \det_{xy}(I - \Pi),\ \ \Pi = (\pi_{ij}),\ \pi_{ij} := p_{i|j}$$

Proof: $B = (b_{ij}) := I - \Pi$ とすると $\sum_{i\in\mathcal{X}} b_{ij} = 0$ より $\det B = 0$.

Adj$B$ で，その (i, j) 成分が Δ_{ji} である行列を表すと

$$(I - \Pi)\mathrm{Adj}B = B\mathrm{Adj}B = (\det B)I = 0$$

よって $\boldsymbol{v}(k) := (\Delta_{k0}, \Delta_{k1}, \ldots\Delta_{kd})^t (k \in \mathcal{X})$ は Π に関する固有値 1 の右固有ベクトル…… (1)

B の第2行から第 d 行までを第0行に加えると，第0行は第1行の符号を変えたものになる．よって $\Delta_{0j} = \Delta_{1j}$．対称性より

$$\forall ij \in \mathcal{X}^2,\ \ \Delta_{ij} = \Delta_{kj}.$$

また，Π の最大固有値は 1 で，Perron–Frobenius の定理より固有空間は 1 次元であり，(1) と同定理より $\boldsymbol{v}(k)$ の各成分は同符号*．ゆえに

$$q_0(i|\boldsymbol{p}) = \frac{\Delta_{ji}}{\sum_{k=0}^{d} \Delta_{jk}} = \frac{\Delta_{ii}}{\sum_{k=0}^{d} \Delta_{kk}}.$$

* Matrix-tree theorem (例えば [5]) を用いると，$\forall ij \in \mathcal{X}^2,\ \Delta_{ij} \geq d\epsilon^d$ $(\epsilon := \min_{ij} p_{j|i})$ を示せる．

指数型分布族 [2]

$x \in \mathcal{X} \subseteq \mathfrak{R}^k$, $\theta \in \mathfrak{R}^k$, ν: 基準測度

$$\begin{aligned} \psi(\theta) &:= \log \int_{\mathcal{X}} \exp(\theta^i x_i) \nu(dx) \text{ (和の規約使用)} \\ p(x|\theta) &:= \exp(\theta^i x_i - \psi(\theta)) \\ \Theta &:= \{\theta : \psi(\theta) < \infty\} \\ k\text{-次元指数型分布族} \quad S &:= \{p(\cdot|\theta) : \theta \in \Theta\} \end{aligned}$$

e.g. 正規分布、ポアソン分布、Bernoulli 分布、逆ガウス分布等

$$\text{期待値パラメータ} \quad \eta_i := E_\theta x_i = \frac{\partial \psi(\theta)}{\partial \theta^i}$$

$$\text{Fisher 情報量} \quad J_{ij}(\theta) = \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j}$$

η の Fisher 情報量は J^{-1}.

マルコフモデルの指数型分布族

[伊藤，甘利 88], [Merhav 89], [Küchler & Sorensen 97], [Takeuchi, Kawabata, & Barron 01], [Nagaoka 05]

[Nagaoka 05] より

$$\begin{aligned}
\frac{1}{n}\log q(x_0^n|\boldsymbol{p}) &= \sum_{j=1}^{d}\sum_{i=0}^{d}\theta^{ij}\nu_{ij}(x_0^n) - \psi(\theta) + \frac{1}{n}\big(\log q_0(x_0|\boldsymbol{p}) + \log p_{0|x_0} - \log p_{0|x_n}\big),\\
\theta^{ij} &:= \log\frac{p_{j|i}p_{0|j}}{p_{0|i}p_{0|0}},\\
\psi(\theta) &:= -\log p_{0|0}.
\end{aligned}$$

$$\begin{aligned}
\text{マルコフタイプ}\quad \nu_{ij}(x_0^n) &:= \frac{\tau_{ij}(x_0^n)}{n}\\
\text{自然パラメータ}\quad \theta &:= (\theta^{ij})_{i\in\mathcal{X},j\in\mathcal{X}\setminus\{0\}},\\
\text{期待値パラメータ}\quad \eta &:= (\eta_{ij})_{i\in\mathcal{X},j\in\mathcal{X}\setminus\{0\}},\\
\eta_{ij} &:= E_{\boldsymbol{p}}\nu_{ij} = q_0(i|\boldsymbol{p})p_{j|i} = \frac{\Delta_{ii}p_{j|i}}{\sum_{k\in\mathcal{X}}\Delta_{kk}}
\end{aligned}$$

期待値パラメータの範囲

$\mu_i := q_0(i|\boldsymbol{p})$ とおく.

$$\eta_i := \sum_j \eta_{ij} \text{ とすると } \eta_{ij} = p_{j|i}\mu_i \text{ より } \eta_i = \sum_j p_{j|i}\mu_i = \mu_i$$

$$\text{また, } \sum_j \eta_{ji} = \sum_j p_{i|j}\mu_j = \mu_i = \eta_i \text{ より } \sum_i \eta_{ji} = \sum_i \eta_{ij}$$

$$\eta \in H := \{\eta := (\eta_{ij})_{i\in\mathcal{X}, j\in\mathcal{X}\setminus\{0\}} | \quad \forall ij \in \mathcal{X}^2,\ \eta_{ij} > 0, \textstyle\sum_{ij\in\mathcal{X}^2} \eta_{ij} = 1, \tag{1}$$

$$\forall i \leq d-1, \textstyle\sum_{j\in\mathcal{X}} \eta_{ij} = \sum_{j\in\mathcal{X}} \eta_{ji}\} \tag{2}$$

Note: (2) から $\sum_{j\in\mathcal{X}} \eta_{dj} = \sum_{j\in\mathcal{X}} \eta_{jd}$ が出る.

c.f.

$$x_0 = x_n \ \Rightarrow\ \forall j \in \mathcal{X},\ \sum_i \tau_{ji}(x_0^n) = \sum_i \tau_{ij}(x_0^n)$$

期待値パラメータによる記述

$$q(x_0^n|\boldsymbol{p}) := \eta_{x_0} \prod_{ij\in\mathcal{X}^2} \left(\frac{\eta_{ij}}{\eta_i}\right)^{\tau_{ij}(x_0^n)} \tag{3}$$

$$q(x_1^n|x_0,\boldsymbol{p}) := \prod_{ij\in\mathcal{X}^2} \left(\frac{\eta_{ij}}{\eta_i}\right)^{\tau_{ij}(x_0^n)} \tag{4}$$

$$M := \{q(\cdot|\boldsymbol{p}(\eta))|\eta \in H\} \tag{5}$$

確率的コンプレキシティ

概念的定義: 「データがモデルに対してもつ情報量」「モデルの助けを借りてデータを出来るだけ短い符号長で記述するときの符号長」 [Rissanen '96]
情報量規準 (モデル選択), ユニバーサル予測, ユニバーサル符号に応用
具体的には最尤符号の符号長

$$-\log \hat{m}_n(x^n) = -\log \frac{p(x^n|\hat{\theta}(x^n))}{\sum_{x^n} p(x^n|\hat{\theta}(x^n))}$$

典型的なケース ([Rissanen 96], 中心極限定理による証明)

$$-\log \hat{m}_n(x^n) = -\log p(x^n|\hat{\theta}(x^n)) + \frac{\dim(\theta)}{2}\log\frac{n}{2\pi} + \log\int \sqrt{\det J(\theta)}d\theta + o(1)$$

c.f. AIC は $-\log p(x^n|\hat{\theta}(x^n)) + \dim(\theta)$

- $S = \{p(\cdot|\theta) : \theta \in \Theta\}$:モデル (確率過程の族)
- $J(\theta)$: θ に関する Fisher 情報行列
- $\hat{\theta}(x^n)$: データ列 $x^n = x_1x_2...x_n$ が与えられたもとでの θ の最尤推定値

Remark: $-\log p(x^n)$ は, 確率 $p(x^n)$ に基づく語頭符号の符号長. 確率過程 p 自体を符号と呼ぶ (c.f. Kraft の不等式).

Kraftの不等式

関数$\varphi : \mathcal{X} \to \bigcup_{n=0}^{\infty}\{0,1\}^n$を$\mathcal{X}$上の符号，$\varphi(x)$を$x$の符号語と呼ぶ．$|\varphi(x)|$で符号語$\varphi(x)$の長さを表す．次が成り立つ．

1. 符号φが瞬時複号可能であるとき，次のKraftの不等式が成り立つ．

$$\sum_{x\in\mathcal{X}} 2^{-|\varphi(x)|} \leq 1 \tag{6}$$

2. 関数$l : \mathcal{X} \to \boldsymbol{N}$が

$$\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq 1 \tag{7}$$

を満たすならば，$\forall x \in \mathcal{X}$, $|\varphi(x)| = l(x)$ なる瞬時復号可能な符号φを構成できる．

瞬時復号可能: 符号語を繋げた列$\varphi(x_1)\varphi(x_2)\ldots$を先頭から読み込む時，逐次的に復号できるという条件．これは符号φが実用的であるための条件である．

Kraftの不等式により，符号と確率分布の間に次の対応があることが分かる．

$$p(x) \Leftrightarrow 2^{-l(x)}$$

$-\log p(x)$を一般化符号長と呼ぶことがある(非整数だから)．

最尤符号の意味

“regret” (pointwise redundancy) の最悪値を最小にする瞬時復号可能な符号 [Shutarkov ’87]

$$\hat{m}_n = \arg\min_q \max_{x^n \in \mathcal{X}^n} \Big(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\hat{\theta}(x^n))} \Big).$$

すなわち

$$\mathsf{regret}(q, x^n) := \log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\hat{\theta}(x^n))}$$

とすると,

$$\mathsf{regret}(\hat{m}_n(x^n)) = \log \sum_{x^n \in \mathcal{X}^n} p(x^n|\hat{\theta}(x^n)) = (x^n \text{によらない定数})$$

このような “equalizing rule” が minimax 解となる.

例：ベルヌーイモデルの確率的コンプレキシティ

最尤符号の符号長.

$$-\log \frac{p(x^n|\hat{\eta}(x^n))}{\sum_{x^n} p(x^n|\hat{\eta}(x^n))}$$

$$p(x^n|\eta) = \eta^k (1-\eta)^{n-k}$$

タイプクラス $E(k, n-k)$(1 を k 個含む x^n の集合)

$$|E(n,k)| = \frac{n!}{k!(n-k)!}$$

よって，$\hat{\eta} = k/n$ とすると，

$$\sum_{x^n} p(x^n|\hat{\eta}) = \sum_{k=0}^{n} \hat{\eta}^k (1-\hat{\eta})^{n-k} \frac{n!}{k!(n-k)!}$$

Stirling の公式より，

$$\sum_{x^n} p(x^n|\hat{\eta}) \approx \sum_{k=0}^{n} \sqrt{\frac{1}{2\pi n \hat{\eta}(1-\hat{\eta})}} \approx \sqrt{\frac{n}{2\pi}} \int_0^1 \frac{1}{\sqrt{\hat{\eta}(1-\hat{\eta})}} d\hat{\eta}$$

Bayes 符号

$$m(x^n) = \int p(x^n|\theta)w(\theta)d\theta \quad w(\theta)\text{: } \Theta \text{ 上の確率密度 (事前分布)}$$

周辺確率が簡単に求められる

$$\int p(x^{n-1}|\theta)w(\theta)d\theta = \sum_{x_n \in \mathcal{X}} \int p(x^{n-1}x_n|\theta)w(\theta)d\theta$$

$\Rightarrow$ 条件付き確率 $m(x_n|x^{n-1}) = m(x^n)/m(x^{n-1})$ が容易に計算可

$$m(x_n|x^{n-1}) = \int p(x_n|\theta)w(\theta|x^{n-1})d\theta, \quad w(\theta|x^{n-1}) := \frac{p(x^{n-1}|\theta)w(\theta)}{\int p(x^{n-1}|\theta)w(\theta)d\theta} \text{ (事後分布)}$$

特に

$$\text{Jeffreys 事前分布: } w_J(\theta) := \frac{\sqrt{\det(J(\theta))}}{C_J}, \quad C_J := \int \sqrt{\det(J(\theta))}d\theta$$

のとき，$p(\cdot|\theta)$ が指数型ならば minimax regret を達成し，次のようになる.

$$-\log m(x^n) = -\log p(x^n|\hat{\theta}(x^n)) + \frac{\dim(\theta)}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det J(\theta)}d\theta + o(1)$$

逆に指数型でないならば，$m(x^n)$ では minimax regret を達成できない (埋め込み e-曲率の影響. [Takeuchi & Barron 98]).

Bayes符号の符号長評価1

$m_J : \int p(\cdot|\theta) w_J(\theta) d\theta$, B_n : $\hat{\theta}$ 周りの半径 $\log n/\sqrt{n}$ のボール, θ : d 次元
$\hat{J}(x^n, \theta) := -(1/n)\partial_i \partial_j \log p(x^n|\theta)$ は経験的Fisher情報量

$$
\begin{aligned}
& \frac{m_J(x^n)}{p(x^n|\hat{\theta})} \\
\sim \quad & \frac{\int_{B_n} p(x^n|\theta) w_J(\theta) d\theta}{p(x^n|\hat{\theta})} \\
= \quad & \int_{B_n} \exp(\log \frac{p(x^n|\theta)}{p(x^n|\hat{\theta})}) w_J(\theta) d\theta \\
\sim \quad & \int_{B_n} \exp(\frac{-n(\theta - \hat{\theta}) \cdot \hat{J}(x^n, \hat{\theta})(\theta - \hat{\theta})}{2}) w_J(\theta) d\theta \\
\sim \quad & \frac{w_J(\hat{\theta})(2\pi)^{d/2}}{n^{d/2}(\det(\hat{J}(x^n, \hat{\theta})))^{1/2}} \\
\sim \quad & \frac{(\det(J(\hat{\theta})))^{1/2}(2\pi)^{d/2}}{C_J n^{d/2}(\det(\hat{J}(x^n, \hat{\theta})))^{1/2}}
\end{aligned}
$$

Bayes符号の符号長評価2

$$-\log m_J(x^n) = -\log p(x^n|\hat{\theta}(x^n)) + \frac{d}{2}\log\frac{n}{2\pi} + \log C_J + \frac{1}{2}\log\frac{\det(\hat{J}(x^n,\hat{\theta}))}{\det(J(\hat{\theta}))} + o(1)$$

指数型ならば，$J(\hat{\theta}) = \hat{J}(x^n,\hat{\theta})$

$$\Downarrow$$

$$-\log m_J(x^n) = -\log p(x^n|\hat{\theta}(x^n)) + \frac{d}{n}\log\frac{n}{2\pi} + \log C_J + o(1)$$

Note: 実際の証明では，あらゆる系列x^nについて一様な収束を示している．

Markovモデルの Fisher情報量

The Fisher information with respect to $(\theta^{ij}, \theta^{i'j'})$ is defined as

$$J_{(ij)(i'j')}(\theta) \quad := \quad \lim_{n\to\infty} \frac{1}{n} E_{\boldsymbol{p}}(\partial_{ij} l_n)(\partial_{i'j'} l_n) \tag{8}$$

$$\text{or} \quad J_{(ij)(i'j')}(\theta) \quad := \quad -\lim_{n\to\infty} \frac{1}{n} E_{\boldsymbol{p}} \partial_{ij}\partial_{i'j'} l_n \tag{9}$$

where we let $l_n := \log q(x_0^n|\boldsymbol{p})$ and $\partial_{ij} := \partial/\partial\theta^{ij}$.

Empirical Fisher information w.r.t. $(p_{j|i}, p_{j'|i'})$

$$\hat{I}_{(ij)(i'j')}(x^n, \boldsymbol{p}) \quad := \quad \frac{-1}{n}\frac{\partial^2 \log q(x^n|\boldsymbol{p})}{\partial p_{j|i}\partial p_{j'|i'}} = \delta_{ii'}\hat{p}_i\Big(\frac{\delta_{jj'}\hat{p}_{j|i}}{(p_{j|i})^2} + \frac{\hat{p}_{0|i}}{(p_{0|i})^2}\Big)$$

$$\hat{I}_{(ij)(i'j')}(x^n, \hat{\boldsymbol{p}}) \quad = \quad \delta_{ii'}\hat{p}_i\Big(\frac{\delta_{jj'}}{\hat{p}_{j|i}} + \frac{1}{\hat{p}_{0|i}}\Big)$$

Fisher information : $I_{(ij)(i'j')}(\boldsymbol{p}) \quad := \quad \lim_{n\to\infty} E_{\boldsymbol{p}} \hat{I}_{(ij)(i'j')}(x^n, \boldsymbol{p}) = \delta_{ii'}\mu_i\Big(\frac{\delta_{jj'}}{p_{j|i}} + \frac{1}{p_{0|i}}\Big),$

where $\hat{p}_{j|i} := \tau_{ij}/\tau_i$ ($\tau_i := \sum_j \tau_{ij}$), $\hat{p}_i := \tau_i/n$, $\mu_i := E_p\hat{p}_i$ (stationary probability)

$$\mu_i(\hat{p}) = \hat{p}_i \Leftrightarrow \hat{J}(x^n, \hat{p}) = J(\hat{p})$$

$\Rightarrow \hat{J}(x^n, \hat{p}) - J(\hat{p})$ は埋め込みe-曲率に対応する量であり，上記からもMarkovモデルが指数型であることが分かる [Takeuchi, Kawabata, & Barron 2001].

Markovモデルの SC

[Takeuchi, Kawabata, & Barron 2001](Bayes符号による)

$$-\log q(x^n|x_0,\hat{p}) + \frac{d(d+1)}{2}\log\frac{n}{2\pi} + \log\int_P \sqrt{\det I(p)}dp + +o(1)$$

$$\det I(p) = \prod_{i\in\mathcal{X}} \frac{\eta_i^d}{\prod_{j\in\mathcal{X}} p_{j|i}} = \prod_{i\in\mathcal{X}} \Big(\frac{\Delta_{ii}}{\sum_{k\in\mathcal{X}}\Delta_{kk}}\Big)^d \frac{1}{\prod_{j\in\mathcal{X}} p_{j|i}}$$

Recall

$$I_{(ij)(i'j')} = \delta_{ii'}\eta_i\Big(\frac{\delta_{jj'}}{p_{j|i}} + \frac{1}{p_{0|i}}\Big).$$

It is easy to find

$$\det(I_{(ij)(i'j')}) = \prod_{i\in\mathcal{X}} \frac{\eta_i^d}{\prod_{j\in\mathcal{X}} p_{j|i}}. \tag{10}$$

最尤符号によるSC評価

[Jacquet and Szpankowski 2004] による

$$\log \frac{1}{\hat{m}_n(x^n)} - \log \frac{1}{q(x^n|x_0, \hat{\boldsymbol{p}})} = \log \sum_{x_0^n \in \mathcal{X}^{n+1}} q(x^n|\hat{u})$$

$$\begin{aligned} \sum_{x_0^n} q(x^n|\hat{\boldsymbol{p}}) &= \left(\frac{n}{2\pi}\right)^{d(d+1)/2} A_d(1 + O(1/n)) \\ A_d &= \int_H (d+1) \sum_x \Delta_{xx} \prod_{x \in \mathcal{X}} \frac{\sqrt{\eta_x}}{\prod_y \sqrt{\eta_{xy}}} d\eta. \end{aligned}$$

すなわち,

$$-\log q(x^n|x_0, \hat{p}) + \frac{d(d+1)}{2} \log \frac{n}{2\pi} + \log \int_H (d+1) \sum_x \Delta_{xx} \prod_{x \in \mathcal{X}} \frac{\sqrt{\eta_x}}{\prod_y \sqrt{\eta_{xy}}} d\eta + o(1)$$

Jacquet and Szpankowskiの証明

τ_{xy}：系列x_0^nに現れるパターンxyの回数を表す(x_0^nのMarkovタイプ).
$E(\boldsymbol{N})$：Markovタイプが$\boldsymbol{N} = (\tau_{xy})$に等しい系列の集合

$$q(x_0^n|\hat{\boldsymbol{p}}(x^n)) = \prod_{xy\in\mathcal{X}^2} (\hat{p}_{y|x})^{\tau_{xy}} = \Big(\frac{n_{xy}}{\sum_y \tau_{xy}}\Big)^{\tau_{xy}}$$

であり，Markovタイプのみで決まる.
すなわち，$x_{-1}^n \in E(\boldsymbol{N})$について$q(x_0^n|\hat{\boldsymbol{p}}(x^n))$は一定．その値を$L(\boldsymbol{N})$と書くと，

$$\sum_{x^n} q(x^n|\hat{p}(x^n)) = \sum_{\boldsymbol{N}} |E(\boldsymbol{N})| \cdot L(\boldsymbol{N})$$

となる．$|E(\boldsymbol{N})|$はWhittleの公式(ただし以下は$x_0 = x_n$をみたす系列だけを数える公式)

$$|E(\boldsymbol{N})| = \sum_z \Delta_{zz} \prod_x \frac{\tau_x!}{\prod_y \tau_{xy}!}$$

で与えられる．これをStiringの公式で近似し，積分で評価する．$\boldsymbol{N}$の成分は$|\mathcal{X}|^2$だけあるが，これは全てが独立ではない．今，任意の$xy \in \mathcal{X}^2$について$\sum_t \tau_{st} = \sum_t \tau_{ts} \pm 1$，$\sum_{t,s} \tau_{ts} = n$が成り立つ．逆に，これらが満たされれば，$\boldsymbol{N}$はMarkovタイプである．すなわち，$\boldsymbol{N}$は$|\mathcal{X}|^2$次元空間の中の超平面上の格子点に値を取る．これは一様に分布するので$\boldsymbol{N}$に関する和は積分で近似できる.

比較

$$\int_P \prod_i \left(\frac{\Delta_{ii}}{\sum_k \Delta_{kk}}\right)^{d/2} \frac{1}{\sqrt{\prod_j p_{j|i}}} d\boldsymbol{p} = (d+1) \int_H \sum_x \Delta_{xx} \prod_{x\in\mathcal{X}} \frac{\sqrt{\eta_x}}{\prod_y \sqrt{\eta_{xy}}} d\eta$$

より

$$\prod_i \left(\frac{\Delta_{ii}}{\sum_k \Delta_{kk}}\right)^{d/2} \frac{1}{\sqrt{\prod_j p_{j|i}}} d\boldsymbol{p} = (d+1) \sum_x \Delta_{xx} \prod_{x\in\mathcal{X}} \frac{\sqrt{\eta_x}}{\prod_y \sqrt{\eta_{xy}}} d\eta$$

が出る．すなわち,

$$\sqrt{\det J^{(ij)(i'j')}} = (d+1) \sum_x \Delta_{xx} \prod_{x\in\mathcal{X}} \frac{\sqrt{\eta_x}}{\prod_y \sqrt{\eta_{xy}}}$$

となる !? (実は $(d+1)$ が余分)

$$\Downarrow$$

直接 $\det J^{(ij)(i'j')}$ を計算したい.

$\det I_{(ij)(i'j')}$ よりも

$$\det J^{(ij)(i'j')} = \det\left(\frac{\partial\theta}{\partial\eta}\right)$$

が重要な量と思える.

方針

全ての η_{ij} が独立でないため $J^{(ij)(i'j')}$ は結構複雑.

$$\Downarrow$$

η_{ij} の制限を外して，拡大モデル $\bar{M}$ を利用したらどうか？

拡大モデル

$$
\begin{aligned}
\bar{M} &:= \{r(\cdot|\bar{\eta})|\bar{\eta} \in \bar{H}\} \\
r(x_0^n|\bar{\eta}) &:= \bar{\eta}_{x_0} \prod_{ij \in \mathcal{X}^2} \left(\frac{\bar{\eta}_{ij}}{\bar{\eta}_i}\right)^{\tau_{ij}(x_0^n)} \\
\bar{H} &:= \{\bar{\eta} = (\bar{\eta}_{ij})_{ij \in \mathcal{X}^2} | \forall ij \in \mathcal{X}^2,\ \bar{\eta}_{ij} > 0\}
\end{aligned}
$$

Recall

$$
\begin{aligned}
H = \{\eta &= (\eta_{ij})_{i \in \mathcal{X}, j \in \mathcal{X} \setminus \{0\}} | \forall ij \in \mathcal{X}^2,\ \bar{\eta}_{ij} > 0\} \\
&\wedge \sum_{ij \in \mathcal{X}^2} \eta_{ij} = 1 \ \wedge\ \forall i \in \mathcal{X}, \sum_{j \in \mathcal{X}} (\eta_{ij} - \eta_{ji}) = 1\}
\end{aligned}
$$

M は $\bar{M}$ の中で -1-自己平行
(c.f. α-表現の議論)

拡大モデルの Fisher 情報量 $\bar{J}$

$$\bar{J}^{(ij)(i'j')} := \lim_{n\to\infty} \frac{1}{n} \sum_{x_0^n \in \mathcal{X}^{n+1}} (\bar{\partial}^{ij} l_n)(\bar{\partial}^{i'j'} l_n) r(x_0^n|\bar{\eta})$$

where we let $l_n := \log r(x_0^n|\bar{\eta})$ and $\bar{\partial}^{ij} := \partial/\partial\bar{\eta}_{ij}$. Since effect of η_{x_0} in l_n vanishes when n goes to infinity, we have

$$\bar{J}^{(ij)(i'j')} = \sum_{kl\in\mathcal{X}^2} \bar{\eta}_{kl} \bar{\partial}^{ij} \log\Big(\frac{\bar{\eta}_{kl}}{\bar{\eta}_k}\Big) \cdot \bar{\partial}^{i'j'} \log\Big(\frac{\bar{\eta}_{kl}}{\bar{\eta}_k}\Big),$$

which yields

$$\bar{J}^{(ij)(i'j')} = \delta_{ii'} \Big(\frac{\delta_{jj'}}{\bar{\eta}_{ij}} - \frac{1}{\bar{\eta}_i} \Big). \tag{11}$$

$\bar{J}$は退化している

Let us see this utilising the fact that $\bar{J}$ is a block diagonal matrix, We focus on a miner of $\bar{J}$ for which i is fixed to k. We denote it by $\bar{J}_{(k)}$:

$$\bar{J}^{ij}_{(k)} := \frac{\delta_{ij}}{\bar{\eta}_{ki}} - \frac{1}{\bar{\eta}_k}.$$

Let $\boldsymbol{w}(k) = (\bar{\eta}_{k0}, \bar{\eta}_{k1}, ..., \bar{\eta}_{kd})$. Then we have $\bar{J}_{(k)}\boldsymbol{w}(k)^T = 0$. Hence for all $k \in \mathcal{X}$, $\bar{J}_{(k)}$'s rank is smaller than $d+1$. Since J's rank is $d(d+1)$, $\bar{J}$'s rank is $d(d+1)$, which equals the dimension of H.

これは，$l_n = \log r(x_0^n|\bar{\eta})$の中の$\log \bar{\eta}_{x_0}$の影響が消えることに起因する.

Fisher Information Determinant of η

Theorem 1. *Let $J(\eta)$ be the Fisher information matrix of expectation parameter η of a stationary Markov model on finite alphabet $\mathcal{X}$. Then, the following holds.*

$$\det(J^{(ij)(i'j')}(\eta)) = \Big(\sum_{\alpha\in\mathcal{X}} \Delta_{\alpha\alpha}\Big)^2 \prod_{\alpha\in\mathcal{X}} \frac{\eta_\alpha}{\prod_{\beta\in\mathcal{X}} \eta_{\alpha\beta}}. \tag{12}$$

Example 1. *Let $\mathcal{X} := \{0,1\}$. Then from Theorem 1, we have*

$$\sqrt{\det(J^{(ij)(i'j')})} = (p_{1|0}+p_{0|1})\sqrt{\frac{\eta_0\eta_1}{\eta_{01}\eta_{00}\eta_{10}\eta_{11}}}. \tag{13}$$

On the other hand, we have

$$\sqrt{\det(I_{(ij)(i'j')})} = \frac{1}{(p_{1|0}+p_{0|1})\sqrt{p_{1|0}+p_{0|1}}}, \tag{14}$$

since (10). Note that the Jacobian of the function h which maps $(p_{0|1}, p_{1|1})$ to

$$(\eta_{01},\eta_{11}) = \Big(\frac{p_{0|1}p_{1|0}}{p_{1|0}+p_{0|1}}, \frac{p_{1|0}p_{1|1}}{p_{1|0}+p_{0|1}}\Big)$$

is given as follows. Then, we can derive (13) from (14).

$$\det\Big(\frac{\partial h(p_{0|1},p_{1|1})}{\partial(p_{0|1},p_{1|1})}\Big) = \frac{p_{0|1}p_{1|0}}{(p_{0|1}+p_{1|0})^3}.$$

準備1

Define

$$\bar{T}_{\bar{\eta}} = \{d\bar{\eta} = (d\bar{\eta}_{ij}) | d\bar{\eta} \in \boldsymbol{R}^{(d+1)^2}\}. \text{ (the space of displacement of } \bar{\eta})$$

$$dH := \{d\eta = (d\eta_{ij}) | d\eta \in \boldsymbol{R}^{d(d+1)}\}. \text{(the space of displacement of } \eta)$$

Define a subspace of $\bar{T}_{\bar{\eta}}$ as

$$T_{\bar{\eta}} := \{d\bar{\eta} \in \bar{T}_{\bar{\eta}} : \sum_{ij \in \mathcal{X}^2} d\bar{\eta}_{ij} = 0 \ \wedge \ \forall i \in \mathcal{X} \setminus \{0\}, \sum_{j \in \mathcal{X}} (d\bar{\eta}_{ij} - d\bar{\eta}_{ji}) = 0\}.$$

Define a function $\phi : dH \to T_{\bar{\eta}}$, which maps $d\eta$ to $d\bar{\eta} = \phi(d\eta)$ such that

$$\forall i \in \mathcal{X}, \forall j \in \mathcal{X} \setminus \{0\}, \ \ d\bar{\eta}_{ij} = d\eta_{ij}.$$

The Fisher information $\bar{J}$ defines the Fisher metric on $\bar{T}_{\bar{\eta}}$ as

$$\bar{J}(d\bar{\eta}, d\bar{\eta}') := \sum_{ij \in \mathcal{X}^2} \sum_{i'j' \in \mathcal{X}^2} \bar{J}^{(ij)(i'j')}(\bar{\eta}) d\bar{\eta}_{ij} d\bar{\eta}'_{i'j'}.$$

準備2

Similarly, we let $J(d\eta, d\eta')$ denote the Fisher metric on dH. Note that

$$J(d\eta, d\eta') = \bar{J}(\phi(d\eta), \phi(d\eta'))$$

holds. Let $\{\boldsymbol{e}(ab)\}_{ab}$ be a basis of dH, i.e. we define $\boldsymbol{e}(ab)$ as $e_{ij}(ab) := \delta_{ai}\delta_{bj}$. Then, the following holds.

$$J^{(ij)(i'j')} = \bar{J}(\phi(\boldsymbol{e}(ij)), \phi(\boldsymbol{e}(i'j'))).$$

$$\left(\text{普通の表記では } \boldsymbol{e}(ab) = \frac{\partial}{\partial \eta_{ab}}\right)$$

Besides the Fisher metric, we also use the Euclidean metric on $\bar{T}_{\bar{\eta}}$:

$$\bar{g}(d\bar{\eta}, d\bar{\eta}') := \sum_{ij \in \mathcal{X}^2} d\bar{\eta}_{ij} d\bar{\eta}'_{ij}.$$

証明アウトライン

Let $K_{\bar{\eta}} \subset \bar{T}_{\bar{\eta}}$ denote the kernel of $\bar{J}(\bar{\eta})$, which is a $(d+1)$-dimensional space.
Let λ denote the angle between T_η and ${K_{\bar{\eta}}}^{\perp}$.
For a linear subspace U of $\bar{T}_{\bar{\eta}}$, $U^{\perp}$ denotes the orthogonal complement space of U with respect to $\bar{g}$.

Our approach is as follows:

1. Find the determinant D of restriction of $\bar{J}$ to ${K_{\bar{\eta}}}^{\perp}$.
2. Find the angle λ between $T_{\bar{\eta}}$ and ${K_{\bar{\eta}}}^{\perp}$.
3. The desired determinant is proportional to $D \cdot (\cos\lambda)^2$.
4. Determine the proportional constant.

Some notation:
$\bar{T}_{\bar{\eta}}$ 中の $d(d+1)$ 次元立方体 C について

$$V_F(C) \quad : \quad \bar{J} \text{で定まる} C \text{の} d(d+1) \text{次元体積}$$
$$V_{\bar{g}}(C) \quad : \quad \bar{g} \text{で定まる} C \text{の} d(d+1) \text{次元体積}$$

Determinant of restriction of $\bar{J}$

Lemma 1. *Let C be a cube in ${K_{\bar{\eta}}}^{\perp}$. Assume that C's volume in ${K_{\bar{\eta}}}^{\perp}$ induced by $\bar{g}$ is 1. Then, the volume of C induced by $\bar{J}$ is given by*

$$\sqrt{\prod_{\alpha\in\mathcal{X}} \frac{1}{\prod_{\beta\in\mathcal{X}} \bar{\eta}_{\alpha\beta}} \cdot \frac{\sum_{\gamma\in\mathcal{X}} \bar{\eta}_{\alpha\gamma}^2}{\bar{\eta}_{\alpha}}}.$$

Angle between T_η and ${K_{\bar{\eta}}}^{\perp}$

Define vectors $\boldsymbol{m}(k)$ $(k = 0, .., d)$ as:

$$\begin{aligned} \forall ij \in \mathcal{X}^2,\ m_{ij}(0) &= 1, \\ \forall ij \in \mathcal{X}^2,\ m_{ij}(k) &= \delta_{ki} - \delta_{kj}\ \ (k \geq 1). \end{aligned}$$

The vector $\boldsymbol{m}(0)$ is a normal (in terms of $\bar{g}$) vector of the plane $\sum_{ij \in \mathcal{X}^2} d\bar{\eta}_{ij} = 0$ in $\bar{T}_{\bar{\eta}}$, and for each $k \geq 1$, $\boldsymbol{m}(k)$ is a normal vector (in terms of $\bar{g}$) of the plane $\sum_{j \in \mathcal{X}} d\bar{\eta}_{kj} = \sum_{j \in \mathcal{X}} d\bar{\eta}_{jk}$ in $\bar{T}_{\bar{\eta}}$. Note that $\{\boldsymbol{m}(k)\}_k$ spans ${T_{\bar{\eta}}}^{\perp}$.

The following holds.

Lemma 2. Let λ denote the angle between T_η and ${K_{\bar{\eta}}}^{\perp}$ with respect to the Euclidean metric. Let Q denote the parallelepiped spanned by $\{\boldsymbol{m}(k)\}_k$. Then, the following holds:

$$\cos\lambda = \frac{1}{V_{\bar{g}}(Q)} \frac{(\prod_{\beta \in \mathcal{X}} \bar{\eta}_\beta) \sum_{\alpha \in \mathcal{X}} \tilde{\Delta}_{\alpha\alpha}(\bar{\eta})}{\prod_{\beta \in \mathcal{X}} \sqrt{\sum_{\alpha \in \mathcal{X}} \bar{\eta}_{\beta\alpha}^2}},$$

where we let $\tilde{\Delta}_{\alpha\alpha}(\bar{\eta}) := \det_{\beta\beta}(\delta_{ij} - \bar{\eta}_{ij}/\bar{\eta}_i)$.

Proportional Constant

From Lemmas 1 and 2, we can evaluate the volume (by $\bar{J}$) of the unit cube C in $T_{\bar{\eta}}$:

$$V_F(C) = \frac{\sum_{\alpha\in\mathcal{X}} \tilde{\Delta}_{\alpha\alpha}(\bar{\eta})}{V_{\bar{g}}(Q)} \sqrt{\prod_{\alpha\in\mathcal{X}} \frac{\bar{\eta}_\alpha}{\prod_{\beta\in\mathcal{X}} \bar{\eta}_{\alpha\beta}}}. \tag{15}$$

Let $\tilde{C}$ denote the unit cube in dH. Then, the square of $V_{\bar{g}}(\phi(\tilde{C}))V_F(C)$ $(= V_F(\phi(\tilde{C})))$ gives the required determinant. Hence, the remaining task is to find $V_{\bar{g}}(\phi(\tilde{C}))$. It is given by $1/\cos\varphi$, where we let φ denote the angle (by $\bar{g}$) between $T_{\bar{\eta}}$ and the subspace W of $\bar{T}_{\bar{\eta}}$ defined by $d\bar{\eta}_{k0} = 0$ for all $k \in \mathcal{X}$. Below, we will find $\cos\varphi$ similarly as the proof of Lemma 2.

Let $\{\boldsymbol{r}(k)\}_k$ denote an orthonormal basis of the orthogonal complement of the subspace W. Concretely, we let $r_{ij}(k) := \delta_{ik}\delta_{j0}$, for all $k \in \mathcal{X}$. (The vector $\boldsymbol{r}(k)$ is a unit normal vector of the hyper plane $d\bar{\eta}_{k0} = 0$.) Recalling that $\{\boldsymbol{m}(k)\}_k$ spans ${T_{\bar{\eta}}}^\perp$, and that Q denotes the parallelepiped spanned by $\{\boldsymbol{m}(k)\}_k$, we have

$$\cos\varphi = \frac{\det(\bar{g}(\boldsymbol{r}(i), \boldsymbol{m}(j))}{V_{\bar{g}}(Q)}.$$

Since $\bar{g}(\boldsymbol{r}(i), \boldsymbol{m}(j)) = \delta_{ij}$, when $j \geq 1$, and $\bar{g}(\boldsymbol{r}(i), \boldsymbol{m}(0)) = 1$, otherwise, hold, we have $\det(\bar{g}(\boldsymbol{r}(i), \boldsymbol{m}(j)) = 1$. Hence $\cos\varphi = 1/V_{\bar{g}}(Q)$. Together with (15), we have completed the proof of Theorem 1.

What is $\sum_\alpha \Delta_{\alpha\alpha}$?

1. $K_{\bar{\eta}}^{\perp}$ と $T_{\bar{\eta}}$ のなす角の余弦の中に現れる．(Lemma 2)
2. Whittle の公式に現れる因子 (下記)

Bilingsley[1] のバージョン

$$\#\{x_0^n : \tau_{ij}(x_0^n) = F_{ij}, x_0 = a, x_n = b\} = \det_{ab}(\tilde{F}_{ij}) \prod_\alpha \frac{F_\alpha!}{\prod_{\beta\in\mathcal{X}} F_{\alpha\beta}!}, \tag{16}$$

where F_{ij} is a non-negative integer matrix such that

$$\sum_{ij\in\mathcal{X}^2} F_{ij} = n, \ \wedge \forall i \in \mathcal{X}, \ \sum_{j\in\mathcal{X}} (F_{ij} - F_{ji}) = \delta_{bi} - \delta_{ai}.$$

$$\tilde{F}_{ij} := \begin{cases} \delta_{ij} - F_{ij} / \sum_{j\in\mathcal{X}} F_{ij}, & \text{when } \sum_{j\in\mathcal{X}} F_{ij} \neq 0 \\ \delta_{ij}, & \text{otherwise.} \end{cases} \tag{17}$$

Note that

$$\det_{ab}(\tilde{F}_{ij}) \to \det_{aa}(\tilde{F}_{ij}) \ (n \to \infty).$$

References

[1] P. Billingsley, "Statistical Methods in Markov Chains," *Ann. Math. Statist.* Volume 32, Number 1, 12-40 1961.

[2] L. Brown, *Fundamentals of statistical exponential families*, Institute of Mathematical Statistics, 1986.

[3] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE trans. Inform. Theory*, Vol. 50, No. 7, July 2004.

[4] U. Küchler and M. Sorensen, *Exponential Families of Stochastic Processes*, Springer-Verlag, 1997.

[5] J. W. Moon, "Some determinant expansions and the matrix-tree theorem," *Discrete Mathematics*, 124, pp. 163-171, 1994.

[6] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. on Inform. Theory*, vol.35, no.5, 1109.1114, 1989.

[7] H. Nagaoka, "The exponential family of Markov chains and its information geometry," *Proc. of the 28th Symposium on Information Theory and its Applications* (SITA2005), 2005.

[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. Inform. Theory*, vol. 40, pp. 40-47, 1996.

[9] Yu M. Shtar'kov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1987.

[10] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of IEEE International Symposium on Inform. Theory*, 1998.

[11] J. Takeuchi, T. Kawabata, & A. R. Barron, "Properties of Jeffreys mixture for Markov sources," *Proc. of Workshop on Information Based Induction Sciences* (*IBIS2001*), pp. 327-332, 2001. Full paper was accepted for publication, *IEEE trans. Inform. Theory*

[12] P. Whittle, "Some distribution and moment formula for Markov chain," *J. Roy. Statist. Soc.*, Ser. B, vol. 17, pp. 235-242, 1955.

[13] 伊藤, 甘利, "情報源の幾何学," *Proc. of SITA88*, pp. 57–60, 1988.