

Pooling incomplete samples による統計解析

関東学院大経済 布能 英一郎 (Eiichiro Funo)

1. Introduction

Asano(1965) は、On estimating multinomial probabilities by pooling incomplete samples という paper を発表した (Ann. Inst. Stat. Math., **17**, 1-17)。Pooling incomplete samples という名前は、これに由来する。後に、Johnson, Kotz, Balakrishnan(1997) は、分布論を網羅した書籍 *Discrete Multivariate Distributions* (Wiley) にて、第 35 章 Multinomial distributions の中で、Incomplete and modified multinomial distributions という節を設け (第 10 節, 71-72 ページ)、ここで Asano の成果を解説している。ここでは、Johnson らの記述に従うが、記号等は、適宜変更する。

$\mathbf{X} = (X_1, X_2, \dots, X_k)$ は、 $\text{multinomial}(N_x; p_1, \dots, p_k)$ に従い、 $\mathbf{Y} = (Y_1, \dots, Y_m)$ ($m < k$) は、 $\mathbf{X}$ とは独立で、 $\text{multinomial}(N_y; \frac{p_1}{\sum_{i=1}^m p_i}, \dots, \frac{p_m}{\sum_{i=1}^m p_i})$ に従うものとする。このとき、 p_i の最尤推定量 $\hat{p}_i$ は

$$\hat{p}_i = \begin{cases} \frac{x_i + y_i}{N_x \left(1 + \frac{N_y}{\sum_{j=1}^m x_j}\right)} & \text{if } i \leq m, \\ \frac{x_i}{N_x} & \text{if } i > m. \end{cases} \quad (1)$$

である。

更に、次のことも知られている。

- (i) $\hat{p}_i$ は p_i の minimum variance unbiased estimator である (Asano, 1965)
- (ii) $\hat{p}_i$ は、自乗損失下で許容的 (布能, 2000)
- (iii) Pooling incomplete samples による多項分布からの 2 標本問題において、Total information = Within information + Between information (以後、これを Kullback 情報量の直和分解と略記) が成り立つ。 (布能, 2011)

注 Kullback(1959, 1968) は、2 標本問題において、仮説 H_1 を「2 つの母集団は同じ」、仮説 H_2 を「2 つの母集団は異なる」に設定したとき、Between Information を H_2 の下における「(最良) 推定量とパラメータの間の Kullback 距離」、Within Information

を「 H_1 の下での推定量と、仮説 H_2 の下での推定量の間の Kullback 距離」、Total Information を「 H_2 の下での推定量と、 H_1 の下でのパラメータ間の Kullback 距離」と定めた。そして、(通常の) 多項分布において、Kullback 情報量の直和分解 が成り立つことを示した。

本稿は、これらの内容を概観したのち、更なる発展を考察する。

2. 既存結果の概要

MLE を求める 尤度方程式を直接解いて求められるが、次のようなパラメータ変換を行うと容易。 $u_i = p_i / \sum_{j=1}^m p_j$, $v_i = p_i / \sum_{j=m+1}^k p_j$, $s = \sum_{j=1}^m p_j$. そうすると

$$L \propto \prod_{i=1}^m (su_i)^{x_i} \prod_{i=m+1}^k ((1-s)v_i)^{x_i} \prod_{i=1}^m u_i^{y_i} = s^{\sum_{i=1}^m x_i} (1-s)^{\sum_{i=m+1}^k x_i} \prod_{i=1}^m u_i^{x_i+y_i} \prod_{i=m+1}^k v_i^{x_i}$$

これより $\hat{s} = \sum_{j=1}^m x_j / N_x$, $\hat{u}_i = (x_i + y_i) / (\sum_{j=1}^m x_j + N_y)$, $\hat{v}_i = x_i / (\sum_{j=m+1}^k x_j)$. これを $\hat{p}_i = \hat{s}\hat{u}_i$ ($i = 1, \dots, m$), $\hat{p}_i = (1-\hat{s})\hat{v}_i$ ($i = m+1, \dots, k$), に代入することで (1) が得られる。

2.1 $\hat{p}_i$ は unbiased $k = 3, m = 2$ の場合に示す。一般の k, m でも、指針は同じ。 $\hat{p}_3$ が unbiased であることは、通常の多項分布の期待値計算と全く同じ。 $\hat{p}_i, i = 1, 2$, が unbiased であることを示すには、同時分布が

$$L = \frac{N_x!}{x_3! (x_1 + x_2)!} \frac{(N_y + x_1 + x_2)!}{(x_1 + y_1)! (x_2 + y_2)!} \frac{p_3^{x_3}}{(p_1 + p_2)^{N_y}} p_1^{x_1+y_1} p_2^{x_2+y_2}.$$

と書けることを利用する。これにより、 $E(\hat{p}_1)$ は

$$\begin{aligned} E(\hat{p}_1) &= \sum_{x_1+x_2=0}^{N_x} \sum_{x_1+y_1=0}^{N_y+x_1+x_2} \frac{x_1+x_2}{N_x} \frac{x_1+y_1}{N_y+x_1+x_2} \frac{N_x!}{x_3! (x_1+x_2)!} \\ &\quad \times \frac{(N_y+x_1+x_2)!}{(x_1+y_1)! (x_2+y_2)!} \frac{p_3^{x_3}}{(p_1+p_2)^{N_y}} p_1^{x_1+y_1} p_2^{x_2+y_2} \\ &= \sum_{x_1+x_2=1}^{N_x-1} \frac{(N_x-1)!}{x_3! (x_1+x_2-1)!} \frac{p_3^{x_3}}{(p_1+p_2)^{N_y}} \\ &\quad \times \left(p_1 \sum_{x_1+y_1=1}^{N_y+x_1+x_2-1} \frac{(N_y+x_1+x_2-1)!}{(x_1+y_1-1)! (x_2+y_2)!} p_1^{x_1+y_1-1} p_2^{x_2+y_2} \right) \\ &= p_1 \sum_{x_1+x_2=1}^{N_x-1} \frac{(N_x-1)!}{x_3! (x_1+x_2-1)!} \frac{p_3^{x_3}}{(p_1+p_2)^{N_y}} \times (p_1+p_2)^{N_y+x_1+x_2-1} \\ &= p_1 \sum_{x_1+x_2=1}^{N_x-1} \frac{(N_x-1)!}{x_3! (x_1+x_2-1)!} p_3^{x_3} (p_1+p_2)^{x_1+x_2-1} = p_1. \end{aligned}$$

$\hat{p}_2$ も全く同様の計算。

2.2 $\hat{p}_i$ は 自乗損失下で許容的 離散分布で、sample space が有限、母数空間がコンパクトのとき、許容性を示すには、stepwise Bayesian procedure を用いるのが簡潔である。

標本空間 $\mathcal{X}$, 母数空間 Θ , 損失関数 L 対して (i) 各 $x \in \mathcal{X}$ に対し $P(x|\theta) > 0$ を満たす $\theta \in \Theta$ が少なくとも 1 つ存在 (ii) L は δ に関して strictly convex を仮定する。 $\mathcal{X}$ の空でない部分集合 $\mathcal{X}(i)$ に対し、 $\Theta(\mathcal{X}(i)) = \{\theta \in \Theta | g_i(\theta) = \sum_{x \in \mathcal{X}} P(x|\theta) > 0\}$ と定める。そうすると、標本空間を $\mathcal{X}(i)$, 母数空間を $\Theta(\mathcal{X}(i))$ とする restricted probability distribution $P_{\mathcal{X}(i)}(x|\theta)/g_i(\theta)$ が well-defined である。 Θ の空でない部分集合 Θ^* に対して、 Θ^* 上で定義されている事前確率 $d\tau(\theta)$ に対し、 $\Theta - \Theta^*$ 上で zero mass を持つと定める。これにより、 $d\tau(\theta)$ は Θ 上で定義された事前確率となる。 $g(x:\tau) = \int P(x|\theta)d\tau(\theta)$ と定める。

定理 (Hsuan(1979), Meeden and Ghosh(1981)) $\mathcal{X}$ の空でない部分集合の列 $\{\mathcal{X}(i) | i \in I\}$ と、事前確率の列 $\{d\tau_i(\theta) | i \in I\}$ が (a) $\mathcal{X}(1) = \{x \in \mathcal{X} | g(x:\tau_1) > 0\}, \dots, \mathcal{X}(j) = \{x \in \mathcal{X} - \mathcal{X}(2) - \dots - \mathcal{X}(j-1) | g(x:\tau_j) > 0\}$ を満たし、かつ $\mathcal{X}(i) \neq \emptyset, \mathcal{X} = \bigcup_{i \in I} \mathcal{X}(i)$. (b) $\Theta(i) = \{\theta \in \Theta(\mathcal{X}(i)) : d\tau_i(\theta) \text{ は positive mass を持つ}\}$ と置くと、 $\{\Theta(i) | i \in I\}$ は disjoint (c) 推定量 $\delta(x)$ が各 $(\Theta(i), \mathcal{X}(i))$ 上で事前確率 $d\tau_i(\theta)$ より一意に定まる Bayes 解 これら 3 つの条件が満たされれば、 $\delta(x)$ は $(\Theta, \mathcal{X})$ で許容的。

例 二項分布 $P(x|\theta) \propto \theta^x(1-\theta)^{n-x}$, $\Theta = [0, 1]$, $\mathcal{X} = \{0, 1, \dots, n\}$ にて、 $\hat{\theta} = x/n$ が自乗損失下で許容的であることを stepwise Bayes 法で示す。 $d\tau_1(\theta) = (dI_{\{\theta=0\}}(\theta) + dI_{\{\theta=1\}}(\theta))/2$, $d\tau_2(\theta) \propto (1-\theta^n - (1-\theta)^n)/(\theta(1-\theta))d\theta$ に選ぶと、 $\Theta(1) = \{0, 1\}$, $\Theta(2) = [0, 1]$, $\mathcal{X}(1) = \{0, n\}$. $\mathcal{X}(2) = \{1, \dots, n-1\}$. あとは、各 $(\Theta(i), \mathcal{X}(i))$ 上で $d\tau_i(\theta)$ より定まる θ の Bayes 解を計算すれば x/n となり、自乗損失下で許容的。

Pooling Complete Samples の下での推定量の許容性 $k=3, m=2$ の場合に、推定量 (1) の自乗損失下での示す。一般の k, m に対しても、同じ指針で示せる。

標本分割	X_1	X_2	X_3	Y_1	Y_2	Marginal が positive となるような母数 (の範囲)
$\mathcal{X}(1, 1)$	N_x	0	0	N_y	0	$\theta_1 = 1$
$\mathcal{X}(1, 2)$	0	N_x	0	0	N_y	$\theta_2 = 1$
$\mathcal{X}(1, 3)$	0	0	N_x	—	—	$\theta_3 = 1$
$\mathcal{X}(2, 1)$	0	x_2	x_3	0	N_y	$\theta_1 = 0, 0 < \theta_2, \theta_3 < 1$
$\mathcal{X}(2, 2)$	x_1	0	x_3	N_y	0	$\theta_2 = 0, 0 < \theta_1, \theta_3 < 1$
$\mathcal{X}(2, 3)$	x_1	x_2	0	y_1	y_2	$\theta_3 = 0, 0 < \theta_1, \theta_2 < 1$
$\mathcal{X}(3)$	x_1	x_2	x_3	y_1	y_2	$0 < \theta_1, \theta_2, \theta_3 < 1$

$d\tau_{1j}(\theta) = dI_{\{\theta_j=1\}}(\theta)$, $(j = 1, 2, 3)$ $d\tau_{2j}(\theta) \propto I_{\{\theta_j=1\}}(\theta) \frac{\text{restriction}}{\theta_\alpha \theta_\beta} d\theta_\alpha d\theta_\beta$, $(\alpha, \beta \neq j, j = 1, 2, 3)$ $d\tau_3(\theta) \propto \frac{\text{restriction}}{\theta_1 \theta_2 \theta_3} d\theta_1 d\theta_2 d\theta_3$ に選ぶと、 $\mathcal{R}$ は、表のように直和に分割される。

Θ の分解 $\{\Theta(i, j)\}$ は、表の「Marginal が positive となるような母数の範囲」を見れば良い。

各 $\{(\Theta(i, j), \mathcal{R}(i, j))\}$ 上でベイズ解を計算すると、推定量 (1) に一致し、よって自乗損失下で許容的。

Pooling incomplete samples を伴う多項分布からの 2 標本問題 $i = 1, 2$ に対し、確率変数 $\mathbf{X}^{[i]} = (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]})$, $\mathbf{Y}^{[i]} = (Y_1^{[i]}, \dots, Y_m^{[i]})$ はすべて独立 (すなわち、 $\mathbf{X}^{[1]}$, $\mathbf{Y}^{[1]}$, $\mathbf{X}^{[2]}$, $\mathbf{Y}^{[2]}$ はすべて独立)。そして

$$\mathbf{X}^{[i]} \sim \text{Multi.}(N_x^{[i]}; p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}), \quad \mathbf{Y}^{[i]} \sim \text{Multi.}(N_y^{[i]}; \frac{p_1^{[i]}}{\sum_{j=1}^m p_j^{[i]}}, \dots, \frac{p_m^{[i]}}{\sum_{j=1}^m p_j^{[i]}}).$$

仮説 H_1 , H_2 を、 $H_2: p_i^{[1]} = p_i^{[2]} = p_i$, $(i = 1, \dots, k)$ $H_1: \text{not } H_2$ に選ぶ。

パラメータ変換 $u_i = p_i / \sum_{j=1}^m p_j$, $v_i = p_i / \sum_{j=m+1}^k p_j$, $s = \sum_{j=1}^m p_j$ により計算が簡単になるので、これを用いる。そうすると、上記の設定は

$$\begin{aligned} \mathbf{X}^{[i]} &\sim \text{Multi.}(N_x^{[i]}; s^{[i]}u_1^{[i]}, \dots, s^{[i]}u_m^{[i]}, (1-s^{[i]})v_m^{[i]}, \dots, (1-s^{[i]})v_k^{[i]}), \\ \mathbf{Y}^{[i]} &\sim \text{Multi.}(N_y^{[i]}; u_1^{[i]}, \dots, u_m^{[i]}). \end{aligned}$$

$H_2: s^{[1]} = s^{[2]} = s$, $u_i^{[1]} = u_i^{[2]} = u_i$, $v_i^{[1]} = v_i^{[2]} = v_i$, $H_1: \text{not } H_2$ と書きかえられる。

表記の簡略化の為に、 $T_m^{[i]} = \sum_{j=1}^m x_j^{[i]}$ と定める。そうすると、

$$\hat{s}^{[i]} = \frac{T_m^{[i]}}{N_x^{[i]}}, \quad \hat{u}_j^{[i]} = \frac{x_j^{[i]} + y_j^{[i]}}{T_m^{[i]} + N_y^{[i]}}, \quad \hat{v}_j^{[i]} = \frac{x_j^{[i]}}{\sum_{l=m+1}^k x_l^{[i]}}.$$

$$\hat{s} = \frac{T_m^{[1]} + T_m^{[2]}}{N_1^{[1]} + N_1^{[2]}}, \quad \hat{u}_j^{[i]} = \frac{x_j^{[i]} + y_j^{[i]} + x_j^{[i]} + y_j^{[i]}}{T_m^{[i]} + N_y^{[i]} + T_m^{[i]} + N_y^{[i]}}, \quad \hat{v}_j^{[i]} = \frac{x_j^{[i]} + x_j^{[i]}}{\sum_{l=m+1}^k (x_l^{[i]} + x_l^{[i]})}.$$

$$\begin{aligned} \frac{P(\mathbf{x}^{[1]}, \mathbf{y}^{[1]}, \mathbf{x}^{[2]}, \mathbf{y}^{[2]} \mid H_1)}{P(\mathbf{x}^{[1]}, \mathbf{y}^{[1]}, \mathbf{x}^{[2]}, \mathbf{y}^{[2]} \mid H_2)} &= \prod_{i=1}^2 \frac{\prod_{j=1}^m (s^{[i]}u_j^{[i]})^{x_j^{[i]}} \prod_{j=m+1}^k ((1-s^{[i]})v_j^{[i]})^{x_j^{[i]}} \prod_{j=1}^m (u_j^{[i]})^{y_j^{[i]}}}{\prod_{j=1}^m (su_j)^{x_j^{[i]}} \prod_{j=m+1}^k ((1-s)v_j)^{x_j^{[i]}} \prod_{j=1}^m (u_j)^{y_j^{[i]}}} \\ &= \prod_{i=1}^2 \left\{ \left(\frac{s^{[i]}}{s} \right)^{T_m^{[i]}} \left(\frac{1-s^{[i]}}{1-s} \right)^{N_x^{[i]} - T_m^{[i]}} \prod_{j=1}^m \left(\frac{u_j^{[i]}}{u_j} \right)^{y_j^{[i]} + y_j^{[i]}} \prod_{j=m+1}^k \left(\frac{v_j^{[i]}}{v_j} \right)^{x_j^{[i]}} \right\}, \end{aligned}$$

$$\log \frac{P(\mathbf{x}^{[1]}, \mathbf{y}^{[1]}, \mathbf{x}^{[2]}, \mathbf{y}^{[2]} | H_1)}{P(\mathbf{x}^{[1]}, \mathbf{y}^{[1]}, \mathbf{x}^{[2]}, \mathbf{y}^{[2]} | H_2)} = \sum_{i=1}^2 \left\{ T_m^{[i]} \log \frac{s^{[i]}}{s} + (N_x^{[i]} - T_m^{[i]}) \log \frac{1 - s^{[i]}}{1 - s} \right. \\ \left. + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{u_j^{[i]}}{u_j} + \sum_{j=m+1}^k (x_j^{[i]}) \log \frac{v_j^{[i]}}{v_j} \right\}.$$

ここで、 $j = 1, \dots, m$ に対して

$$\hat{E}_{H_1}(X_j^{[i]} + Y_j^{[i]}) = N_x^{[i]} \hat{s}^{[i]} \hat{u}_j^{[i]} + N_y^{[i]} \hat{u}_j^{[i]} = \left(N_x^{[i]} \frac{T_m^{[i]}}{N_x^{[i]}} + N_y^{[i]} \right) \frac{x_j^{[i]} + y_j^{[i]}}{T_m^{[i]} + N_y^{[i]}} = x_j^{[i]} + y_j^{[i]}$$

が成り立つので

$$\begin{aligned} & \text{Total information} \\ &= \sum_{i=1}^2 \left\{ \hat{E}_{H_1} T_m^{[i]} \log \frac{\hat{s}^{[i]}}{s} + (N_x^{[i]} - \hat{E}_{H_1} T_m^{[i]}) \log \frac{1 - \hat{s}^{[i]}}{1 - s} \right. \\ & \quad \left. + \sum_{j=1}^m \hat{E}_{H_1}(X_j^{[i]} + Y_j^{[i]}) \log \frac{\hat{u}_j^{[i]}}{u_j} + \sum_{j=m+1}^k \hat{E}_{H_1}(X_j^{[i]}) \log \frac{\hat{v}_j^{[i]}}{v_j} \right\} \\ &= \sum_{i=1}^2 \left\{ T_m^{[i]} \log \frac{\hat{s}^{[i]}}{s} + (N_x^{[i]} - T_m^{[i]}) \log \frac{1 - \hat{s}^{[i]}}{1 - s} + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{\hat{u}_j^{[i]}}{u_j} + \sum_{j=m+1}^k x_j^{[i]} \log \frac{\hat{v}_j^{[i]}}{v_j} \right\}, \\ & \text{Within information} \\ &= \sum_{i=1}^2 \left\{ T_m^{[i]} \log \frac{\hat{s}^{[i]}}{\hat{s}} + (N_x^{[i]} - T_m^{[i]}) \log \frac{1 - \hat{s}^{[i]}}{1 - \hat{s}} + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{\hat{u}_j^{[i]}}{\hat{u}_j} + \sum_{j=m+1}^k x_j^{[i]} \log \frac{\hat{v}_j^{[i]}}{\hat{v}_j} \right\}. \end{aligned}$$

同様な計算で

$$\begin{aligned} \text{Between information} &= \hat{E}_{H_2} \left(\log \frac{\hat{P}(\mathbf{X}^{[1]}, \mathbf{Y}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[2]} | H_2)}{P(\mathbf{X}^{[1]}, \mathbf{Y}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[2]} | H_2)} \right) \\ &= (T_m^{[1]} + T_m^{[2]}) \log \frac{\hat{s}}{s} + (N_1^{[1]} + N_1^{[2]} - T_m^{[1]} - T_m^{[2]}) \log \frac{1 - \hat{s}}{1 - s} \\ & \quad + \sum_{j=1}^m (x_j^{[1]} + y_j^{[1]} + x_j^{[2]} + y_j^{[2]}) \log \frac{\hat{u}_j}{u_j} + \sum_{j=m+1}^k (x_j^{[1]} + x_j^{[2]}) \log \frac{\hat{v}_j}{v_j} \end{aligned}$$

となり、Kullback 情報量の直和分解 が示せた。

3. その後の発展

Pooling incomplete samples によるサンプリングは、多項分布の場合に限らず、一般の分布の下で考えることが出来る。「カテゴリーが減少する場合、減少後のカテゴリーには、元の生起確率を比例配分したものを割り当てる」ものであるから、離散分布にはこの考え方をそのまま当てはめられる。ところが、多項分布の場合に成り立っていることが、他の分

布のとき、あるいは、多項分布の場合でも、Asano が設定した Pooling incomplete samples の枠組から外れると、成り立たないことが生じる。

3.1 unbiasedness 2.1 で言及した方法は、Pooling incomplete samples を伴う他の分布に適用できる場合もあれば、うまくゆかない場合もある。

例 1. $X_i \sim \text{Poisson}(\lambda_i)$, ($i = 1, 2, 3, \dots$), $Y_i \sim \text{Poisson}\left((\lambda_1 + \lambda_2 + \lambda_3) \frac{\lambda_i}{\lambda_1 + \lambda_2}\right)$, $i = 1, 2$.
にて

$$\begin{aligned}\tilde{\lambda}_i &= \frac{x_i + y_i}{x_1 + y_1 + x_2 + y_2} \frac{x_1 + x_2}{x_1 + x_2 + x_3} \frac{Total}{2}, \quad i = 1, 2, \\ \tilde{\lambda}_3 &= \frac{x_3}{x_1 + x_2 + x_3} \frac{Total}{2}, \quad \text{where} \quad Total = \sum_i \sum_j x_{i,j}.\end{aligned}$$

は unbiased である。これは、同時確率が

$$\begin{aligned}L &= \frac{2^{Total}}{Total!} \frac{(x_1 + x_2 + x_3)!}{(x_1 + x_2)! x_3!} \frac{(x_1 + y_1 + x_2 + y_2)!}{(x_1 + y_1)! (x_2 + y_2)!} \exp(-2(\lambda_1 + \lambda_2 + \lambda_3)) \\ &\quad \times \lambda_1^{x_1+y_1} \lambda_2^{x_2+y_2} \lambda_3^{x_3} (\lambda_1 + \lambda_2 + \lambda_3)^{y_1+y_2} \left(\frac{1}{\lambda_1 + \lambda_2}\right)^{y_1+y_2}.\end{aligned}$$

と書けることを利用する。実際

$$\begin{aligned}E(\hat{\lambda}_1) &= E\left(\frac{x_1 + y_1}{x_1 + y_1 + x_2 + y_2} \frac{x_1 + x_2}{x_1 + x_2 + x_3} \frac{Total}{2}\right) \\ &= \sum \sum \sum \frac{x_1 + y_1}{x_1 + x_2 + x_3} \frac{Total}{2} \frac{2^{Total}}{Total!} \frac{(x_1 + x_2 + x_3)!}{(x_1 + x_2)! x_3!} \\ &\quad \times (\lambda_1 + \lambda_2 + \lambda_3)^{y_1+y_2} \lambda_3^{x_3} \exp(-2(\lambda_1 + \lambda_2 + \lambda_3)) \left(\frac{1}{\lambda_1 + \lambda_2}\right)^{y_1+y_2} \\ &\quad \times \lambda_1 \sum \frac{(x_1 + y_1 + x_2 + y_2 - 1)!}{(x_1 + y_1 - 1)! (x_2 + y_2)!} \lambda_1^{x_1+y_1-1} \lambda_2^{x_2+y_2} \\ &= \sum \sum \frac{2^{Total-1}}{Total-1!} (\lambda_1 + \lambda_2 + \lambda_3)^{y_1+y_2} \frac{(x_1 + x_2 + x_3 - 1)!}{(x_1 + x_2 - 1)! x_3!} \lambda_3^{x_3} \\ &\quad \times \exp(-2(\lambda_1 + \lambda_2 + \lambda_3)) \left(\frac{1}{\lambda_1 + \lambda_2}\right)^{y_1+y_2} \lambda_1 (\lambda_1 + \lambda_2)^{x_1+y_1+x_2+y_2-1} \\ &= \lambda_1 \sum \frac{2^{Total-1}}{(Total-1)!} \exp(-2(\lambda_1 + \lambda_2 + \lambda_3)) (\lambda_1 + \lambda_2 + \lambda_3)^{y_1+y_2} \\ &\quad \times \sum \frac{(x_1 + x_2 + x_3 - 1)!}{(x_1 + x_2 - 1)! x_3!} \lambda_3^{x_3-1} (\lambda_1 + \lambda_2)^{x_1+x_2-1} \\ &= \lambda_1 \sum \frac{2^{Total-1}}{(Total-1)!} \exp(-2(\lambda_1 + \lambda_2 + \lambda_3)) (\lambda_1 + \lambda_2 + \lambda_3)^{y_1+y_2} \\ &\quad \times (\lambda_1 + \lambda_2 + \lambda_3)^{x_1+x_2+x_3-1} \\ &= \lambda_1 \sum \frac{(2(\lambda_1 + \lambda_2 + \lambda_3))^{Total-1}}{(Total-1)!} \exp(-2(\lambda_1 + \lambda_2 + \lambda_3)) = \lambda_1.\end{aligned}$$

$\hat{\lambda}_2, \hat{\lambda}_3$ についても同様。

例 2. X, Y は互いに独立, $X = (X_1, X_2) \sim NM(k_x, \theta_1, \theta_2)$, $Y \sim NB(k_y, \frac{\theta_2}{\theta_1 + \theta_2})$. このとき、自乗損失下で、improper prior $d\pi = d\theta_1 d\theta_2 / (\theta_0^2 \theta_1 \theta_2)$ に対する (一般化) ベイズ解は

$$\tilde{\theta}_0 = \frac{k_x - 1}{x_1 + x_2 + k_x - 1}, \quad \tilde{\theta}_i = \frac{x_1 + x_2}{x_1 + x_2 + k_x - 1} \frac{x_i + k_y}{x_1 + k_y + x_2 + y}, \quad (i = 1, 2)$$

である。 $\tilde{\theta}_0$ が unbiased であることは、straightfoward. 先ほどと同じように 同時確率が

$$L = \frac{(k_x + x_1 + x_2 - 1)! (k_y + x_1 + x_2 + y)!}{(k_x - 1)! (x_1 + x_2)! (k_y + x_1)! (x_2 + y)!} \frac{\theta_0^{k_x}}{(\theta_1 + \theta_2)^{k_y + y}} \theta_1^{k_y + x_1} \theta_2^{x_2 + y}.$$

と書きあらわせるので

$$\begin{aligned} E(\tilde{\theta}_1) &= E\left(\frac{X_1 + X_2}{X_1 + X_2 + k_x - 1} \frac{X_1 + k_y}{X_1 + k_y + X_2 + Y}\right) \\ &= \sum \sum \frac{x_1 + x_2}{x_1 + x_2 + k_x - 1} \frac{x_1 + k_y}{x_1 + k_y + x_2 + y} \frac{(k_x + x_1 + x_2 - 1)!}{(k_x - 1)! (x_1 + x_2)!} \\ &\quad \times \frac{(k_y + x_1 + x_2 + y)!}{(k_y + x_1)! (x_2 + y)!} \frac{\theta_0^{k_x}}{(\theta_1 + \theta_2)^{k_y + x_2}} \theta_1^{k_y + x_1} \theta_2^{x_2 + y} \\ &= \sum \frac{(k_x + x_1 + x_2 - 2)!}{(k_x - 1)! (x_1 + x_2 - 1)!} \frac{\theta_0^{k_x}}{(\theta_1 + \theta_2)^{k_y + x_2}} \\ &\quad \times \theta_1 \sum \frac{(k_y + x_1 + x_2 + y - 1)!}{(k_y + x_1 - 1)! (x_2 + y)!} \theta_1^{k_y + x_1 - 1} \theta_2^{x_2 + y} \\ &= \theta_1 \sum \frac{(k_x + x_1 + x_2 - 2)!}{(k_x - 1)! (x_1 + x_2 - 1)!} \frac{\theta_0^{k_x} (\theta_1 + \theta_2)^{k_y + x_1 + x_2 + y - 1}}{(\theta_1 + \theta_2)^{k_y + x_2}} \\ &= \theta_1 \sum \frac{(k_x + (x_1 + x_2 - 1) - 1)!}{(k_x - 1)! (x_1 + x_2 - 1)!} \theta_0^{k_x} (\theta_1 + \theta_2)^{x_1 + x_2 - 1} = \theta_1. \end{aligned}$$

と計算できるように思える。しかしながら、上記では $(\theta_1 + \theta_2)^{k_y + x_1 + x_2 + y - 1}$ を二項展開して $\sum \frac{(k_y + x_1 + x_2 + y - 1)!}{(k_y + x_1 - 1)! (x_2 + y)!} \theta_1^{k_y + x_1 - 1} \theta_2^{x_2 + y}$ としており、この部分が正しくない。

3.2 許容性の証明 Pooling incomplete samples を伴う多項分布の場合、Stepwise Bayes 法は、の許容性の証明法として「相性が良い」と言える。それは、Brown(1981) が「sample space が有限、母数空間がコンパクトのとき、任意の許容的推定量に対して、sequence of priors(あるいは、1つの prior) が存在し、これに対する Stepwise Bayes 解 (あるいは、1つの prior に対する Bayes 解) となっている」という定理を示しているからである。他方、Stepwise Bayes 法は、母数空間がコンパクトでないときには、必ずしもうまくゆくとは限らない。

3.3 2 標本問題における Kullback 情報量の直和分解 多項分布以外の Pooling incomplete samples では、成り立たない場合の方が多いと言わざるを得ない。

Pooling incomplete samples を伴う負の多項分布の場合

例 3. $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で、各 $i = 1, 2$ に対し

$$\begin{aligned}\mathbf{X}^{[i]} &= (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{NM}(r_x^{[i]}, \theta_1^{[i]}, \dots, \theta_m^{[i]}, \dots, \theta_k^{[i]}), \\ \mathbf{Y}^{[i]} &= (Y_1^{[i]}, \dots, Y_m^{[i]}) \sim \text{NM}(r_y^{[i]}, \sum_{j=1}^k \theta_j^{[i]} \frac{\theta_1^{[i]}}{\sum_{j=1}^m \theta_j^{[i]}}, \dots, \sum_{j=1}^k \theta_j^{[i]} \frac{\theta_m^{[i]}}{\sum_{j=1}^m \theta_j^{[i]}})\end{aligned}\quad (2)$$

なる確率モデルを考える。これは、「負の多項分布で自然に考えられる Pooling incomplete samples」と言える。ところが、この場合、Kullback 情報量の直和分解が成り立たない。他方、次の場合には、Kullback 情報量の直和分解が成立する。

Proposition (負の多項分布と、Pooling incomplete sample が多項分布の 2 標本問題) 上記の $\mathbf{Y}^{[i]}$ を

$$\mathbf{Y}^{[i]} = (Y_1^{[i]}, \dots, Y_m^{[i]}) \sim \text{Multi.}(N_y^{[i]}, \frac{\theta_1^{[i]}}{\sum_{j=1}^m \theta_j^{[i]}}, \dots, \frac{\theta_m^{[i]}}{\sum_{j=1}^m \theta_j^{[i]}})$$

に変更すると、Kullback 情報量の直和分解が成立する。

上記 2 つと類似した確率モデルを考える。

例 4. $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で、各 $i = 1, 2$ に対し

$$\begin{aligned}\mathbf{X}^{[i]} &= (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{NM}(r_x^{[i]}, \theta_1^{[i]}, \dots, \theta_m^{[i]}, \dots, \theta_k^{[i]}), \\ \mathbf{Y}^{[i]} &= (Y_0^{[i]}, \dots, Y_m^{[i]}) \sim \text{M}(N_y^{[i]}, \frac{\theta_0^{[i]}}{\sum_{j=0}^m \theta_j^{[i]}}, \dots, \frac{\theta_m^{[i]}}{\sum_{j=0}^m \theta_j^{[i]}}),\end{aligned}$$

この場合、Total $\neq$ Between + Within

例 5. $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で、各 $i = 1, 2$ に対し

$$\begin{aligned}\mathbf{X}^{[i]} &= (X_1^{[i]}, X_2^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \\ &\sim \text{NM}(r_x^{[i]}, \theta_1^{[i]}, \theta_2^{[i]}, \dots, \theta_m^{[i]}, \dots, \theta_k^{[i]}), \\ \mathbf{Y}^{[i]} &= (Y_2^{[i]}, \dots, Y_m^{[i]}) \sim \text{NM}(r_y^{[i]}, \frac{\theta_2^{[i]}}{\sum_{j=1}^m \theta_j^{[i]}}, \dots, \frac{\theta_m^{[i]}}{\sum_{j=1}^m \theta_j^{[i]}}),\end{aligned}$$

この場合も、Total $\neq$ Between + Within

これらの現象に対する (1 つの) 解釈

負の多項分布 = 負の二項分布 $\times$ 多項分布 と分解されるが、Pooling incomplete samples が「多項分布」の部分で行われていれば Total = Between + Within が成り立つ。しかし、他のところで Pooling incomplete samples が行われた場合は、等号が成立するとは限らない。

参考 負の多項分布 = 負の二項分布 × 多項分布 であること :

$$\begin{aligned} & \frac{(x_1 + \cdots + x_k + r - 1)!}{x_1! \cdots x_k! (r - 1)!} p_0^r p_1^{x_1} \cdots p_k^{x_k} \\ &= \frac{(x_1 + \cdots + x_k + r - 1)!}{(x_1 + \cdots + x_k)! (r - 1)!} p_0^r (1 - p_0)^{x_1 + \cdots + x_k} \\ & \times \frac{(x_1 + \cdots + x_k)!}{x_1! \cdots x_m! \cdots x_k!} \left(\frac{p_1}{1 - p_0} \right)^{x_1} \cdots \left(\frac{p_m}{1 - p_0} \right)^{x_m} \cdots \left(\frac{p_k}{1 - p_0} \right)^{x_k} \end{aligned}$$

パラメータを分離して Kullback 情報量の直和分解が成立するようにする

(2) の確率モデルは、パラメータ変換により、次のように書きかえられる。

$$\begin{aligned} \mathbf{X}^{[i]} &= (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{Negative Multi.}(r_x^{[i]}, p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}), \\ \mathbf{Y}^{[i]} &= (Y_1^{[i]}, \dots, Y_m^{[i]}) \sim \text{Negative Multi.}(r_y^{[i]}, q_1^{[i]}, \dots, q_m^{[i]}), \quad \text{ここで} \end{aligned}$$

$$p_j^{[i]} = \begin{cases} 1 - s^{[i]} & j = 0, \\ s^{[i]} t^{[i]} u_j^{[i]} & 1 \leq j \leq m, \\ s^{[i]} (1 - t^{[i]}) v_j^{[i]} & j > m, \end{cases} \quad q_j^{[i]} = \begin{cases} 1 - s^{[i]} & j = 0, \\ s^{[i]} u_j^{[i]} & j \geq 1, \end{cases}$$

このときは、Kullback 情報量の直和分解が成立しなかった。そこで、パラメータ s を分離して

$$p_j^{[i]} = \begin{cases} 1 - s_1^{[i]} & j = 0, \\ s_1^{[i]} t^{[i]} u_j^{[i]} & 1 \leq j \leq m, \\ s_1^{[i]} (1 - t^{[i]}) v_j^{[i]} & j > m, \end{cases} \quad q_j^{[i]} = \begin{cases} 1 - s_2^{[i]} & j = 0, \\ s_2^{[i]} u_j^{[i]} & j \geq 1, \end{cases}$$

とすると Kullback 情報量の直和分解が成立。

今後の課題 カテゴリーが減少した際、「Kullback 情報量の直和分解が成立」するような Pooling incomplete samples 風の確率モデルを探すことも必要だが、Kullback 情報量の直和分解が成立しない原因を突き止めることも重要なことである。願わくば、Kullback 情報量の直和分解が成立しないとき、Total information と Within information + Between information の差が、何らかの意味を持つ量であることを突き止められたらと思う。

References

- [1] Brown, L.D. (1981). A complete class theorem for statistical problems with finite sample space. *Annals of Statistics* **9**, 1289-1300.
- [2] Brown, L.D. and Hwang, J.T. (1982). A unified admissibility proof. *Statistical Decision Theory and Related Topics III: vol 1*. 205-230 Academic Press.

- [3] Brown, L.D. and Farrell, R.H. (1985). Complete class theorems for estimation of multivariate Poisson means and related problems. *Annals of Statistics* **13**, 706-726.
- [4] Brown, L.D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics, Lecture notes monograph series volume 9.
- [5] 布能英一郎 (2000). Missing, Pooling を伴う離散サンプリングの統計的推測. 京都大学数理解析研究所講究録 No 1161, 141-158.
- [6] 布能英一郎 (2011). On estimating multivariate discrete probabilities by pooled incomplete samples and related topics. 京都大学数理解析研究所講究録 No 1758.
- [7] 布能英一郎 (2015). 負の多項分布における Kullback 情報量の直和分解—— Pooling incomplete samples の場合を含めた考察—— 京都大学数理解析研究所講究録 No 1954, 90-103.
- [8] Ghosh, M. and Meeden, G. (1997). *Bayesian methods for finite population sampling*. Chapman and Hall.
- [9] Hsuan, F. (1979). A stepwise Bayesian procedure. *Annals of Statistics* **7**, 860-868.
- [10] Kullback, S. (1959). *Information Theory and Statistics*. Wiley.
- [11] Kullback, S. (1968). *Information Theory and Statistics*, Revised edition. Dover.
- [12] Meeden, G. and Ghosh, M. (1981). Admissibility in finite problems. *Annals of Statistics* **9**, 846-852.

文中で十分な検索情報を与えた文献は、割愛した。