

負の多項分布における Kullback 情報量の直和分解

— Pooling incomplete samples を含めた考察 —

関東学院大学 経済学部 布能 英一郎

Eiichiro Funo

School of Economics, Kanto Gakuin University

はじめに (本稿の目的)

離散分布の 2 標本問題において、Between information と Within information の和が Total information に等しい、すなわち、Kullback 情報量の直和分解が成り立つことが多くみられる。このことは、離散解析においても、分散分析と似た解析ができ、有益であると言える。他方、Asano(1965) は、多項分布において、pooling incomplete samples の下での統計的推測問題を論じた。近年、Funo(2012) によって、多項分布における pooling incomplete samples の下での 2 標本問題において、直和分解が成り立つことが示された。では、負の多項分布の 2 標本問題に関してはいかがであろうか？ 本稿は、負の多項分布の場合の Kullback 情報量の直和分解に関して、pooling incomplete samples の場合を含めて議論したものである。

1. Introduction

1.1 Kullback 情報量

本稿では、離散型分布のみ取り扱う。未知母数を $(\theta_1, \theta_2, \dots)$ として、仮説 H_1, H_2 を $H_1 : \theta_j = p_j, H_2 : \theta_j = q_j, (j = 1, \dots)$ に選ぶ。このとき、Kullback 情報量は、

$$I(H_1 : H_2) = E_{H_1} \left(\log \frac{P(X | H_1)}{P(X | H_2)} \right) \quad \text{であるから}$$

多項分布 $\frac{N!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}$ では $\log \frac{P(X | H_1)}{P(X | H_2)} = \log \frac{p_1^{x_1} \cdots p_k^{x_k}}{q_1^{x_1} \cdots q_k^{x_k}} = \sum_{j=1}^k x_j \log \frac{p_j}{q_j}$ より

$$I(H_1 : H_2) = \sum_{j=1}^k E_{H_1} X_j \log \frac{p_j}{q_j} = N \sum_{j=1}^k p_j \log \frac{p_j}{q_j},$$

負の多項分布 $\frac{(r + x_1 + \cdots + x_k - 1)!}{(r - 1)! x_1! \cdots x_k!} \theta_0^r \theta_1^{x_1} \cdots \theta_k^{x_k}, \theta_0 = 1 - \sum_{j=1}^k \theta_j$ では

$$\log \frac{P(X | H_1)}{P(X | H_2)} = \log \frac{p_0^r p_1^{x_1} \cdots p_k^{x_k}}{q_0^r q_1^{x_1} \cdots q_k^{x_k}} = r \log \frac{p_0}{q_0} + \sum_{j=1}^k x_j \log \frac{p_j}{q_j} \quad \text{により}$$

$$I(H_1 : H_2) = r \log \frac{p_0}{q_0} + \sum_{j=1}^k E_{H_1}(X_j) \log \frac{p_j}{q_j} = r \left\{ \log \frac{p_0}{q_0} + \frac{1}{p_0} \sum_{j=1}^k p_j \log \frac{p_j}{q_j} \right\}$$

である。

1.2 多項分布の2標本問題における Kullback 情報量の直和分解

多項分布の2標本問題を考える。すなわち、 $\mathbf{X}^{[1]} = (X_1^{[1]}, \dots, X_k^{[1]})$ と $\mathbf{X}^{[2]} = (X_1^{[2]}, \dots, X_k^{[2]})$ は互いに独立で、各 $i = 1, 2$ に対して $\mathbf{X}^{[i]} \sim \text{Multinomial}(N^{[i]}; p_1^{[i]}, \dots, p_k^{[i]})$, $\sum_{j=1}^k X_j^{[i]} = N^{[i]}$, なる状況下で考える。

H_1 を異なる母集団 *i.e.*, $H_1 : (p_1^{[1]}, \dots, p_k^{[1]}) = \mathbf{p}^{[1]} \neq \mathbf{p}^{[2]} = (p_1^{[2]}, \dots, p_k^{[2]})$,

H_2 を同じ母集団からの標本 *i.e.*, $H_2 : p_j^{[1]} = p_j^{[2]} = p_j, j = 1, \dots, k$

に選ぶ。このとき、

$$I(H_1 : H_2) = N^{[1]} \sum_{j=1}^k p_j^{[1]} \log \frac{p_j^{[1]}}{p_j} + N^{[2]} \sum_{j=1}^k p_j^{[2]} \log \frac{p_j^{[2]}}{p_j}$$

である。

2標本問題にて **Total Information** とは $\hat{I}(H_1, \mathbf{p})$, すなわち、仮説 H_1 : 「2つの母集団は異なる」の下での(最良)推定量と、 $\mathbf{p} = (p_1, \dots)$ 間の Kullback 情報量, **Between Information** とは $\hat{I}(H_2, \mathbf{p})$, すなわち、仮説 H_2 : 「2つの母集団は同じ」の下での推定量と、 $\mathbf{p} = (p_1, \dots)$ 間の Kullback 情報量, **Within Information** とは $\hat{I}(H_1, H_2)$, すなわち、仮説 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 情報量のことを指す。

多項分布の2標本問題にて、これらは

$$\text{Between} = \hat{I}(\hat{\mathbf{p}}, \mathbf{p}) = (N^{[1]} + N^{[2]}) \sum_{j=1}^k \hat{p}_j \log \frac{\hat{p}_j}{p_j}$$

$$\text{Within} = \sum_{i=1}^2 \hat{I}(\hat{\mathbf{p}}^{[i]}, \hat{\mathbf{p}}) = N^{[1]} \sum_{j=1}^k \hat{p}_j^{[1]} \log \frac{\hat{p}_j^{[1]}}{\hat{p}_j} + N^{[2]} \sum_{j=1}^k \hat{p}_j^{[2]} \log \frac{\hat{p}_j^{[2]}}{\hat{p}_j}$$

$$\text{Total} = \sum_{i=1}^2 \hat{I}(\hat{\mathbf{p}}^{[i]}, \mathbf{p}) = N^{[1]} \sum_{j=1}^k \hat{p}_j^{[1]} \log \frac{\hat{p}_j^{[1]}}{p_j} + N^{[2]} \sum_{j=1}^k \hat{p}_j^{[2]} \log \frac{\hat{p}_j^{[2]}}{p_j}$$

であるが、これに推定量 $\hat{p}_j = \frac{x_j^{[1]} + x_j^{[2]}}{N^{[1]} + N^{[2]}}$, $\hat{p}_j^{[1]} = \frac{x_j^{[1]}}{N^{[1]}}$, $\hat{p}_j^{[2]} = \frac{x_j^{[2]}}{N^{[2]}}$ を代入することで

due to	Information
Between $\hat{I}(\hat{\mathbf{p}}, \mathbf{p})$	$\sum_{j=1}^k (x_j^{[1]} + x_j^{[2]}) \log \frac{x_j^{[1]} + x_j^{[2]}}{(N^{[1]} + N^{[2]})p_j}$
Within $\sum_{i=1}^2 \hat{I}(\hat{\mathbf{p}}^{[i]}, \hat{\mathbf{p}})$	$\sum_{j=1}^k \left(x_j^{[1]} \log \frac{(N^{[1]} + N^{[2]})x_j^{[1]}}{N^{[1]}(x_j^{[1]} + x_j^{[2]})} + x_j^{[2]} \log \frac{(N^{[1]} + N^{[2]})x_j^{[2]}}{N^{[2]}(x_j^{[1]} + x_j^{[2]})} \right)$
Total $\sum_{i=1}^2 \hat{I}(\hat{\mathbf{p}}^{[i]}, \mathbf{p})$	$\sum_{j=1}^k \left(x_j^{[1]} \log \frac{x_j^{[1]}}{N^{[1]}p_j} + x_j^{[2]} \log \frac{x_j^{[2]}}{N^{[2]}p_j} \right)$

を得る。

Proposition 1 多項分布の2標本問題にて、Between information と Within information の和は、Total information に等しい。

実際 $\frac{x_j^{[i]}}{N^{[i]}p_j} = \frac{x_j^{[1]} + x_j^{[2]}}{(N^{[1]} + N^{[2]})p_j} \frac{(N^{[1]} + N^{[2]})x_j^{[i]}}{N^{[i]}(x_j^{[1]} + x_j^{[2]})}$ を用いることで

$$\begin{aligned} x_j^{[i]} \log \frac{x_j^{[i]}}{N^{[i]}p_j} &= x_j^{[i]} \log \frac{x_j^{[1]} + x_j^{[2]}}{(N^{[1]} + N^{[2]})p_j} \frac{(N^{[1]} + N^{[2]})x_j^{[i]}}{N^{[i]}(x_j^{[1]} + x_j^{[2]})} \\ &= x_j^{[i]} \log \frac{x_j^{[1]} + x_j^{[2]}}{(N^{[1]} + N^{[2]})p_j} + x_j^{[i]} \log \frac{(N^{[1]} + N^{[2]})x_j^{[i]}}{N^{[i]}(x_j^{[1]} + x_j^{[2]})} \end{aligned}$$

が各 $i = 1, 2$ に対して成り立つ。あとは単純計算で、Total = Between + Within が示せる。

2. 負の多項分布の2標本問題における Kullback 情報量の直和分解

$\mathbf{X}^{[1]}$ と $\mathbf{X}^{[2]}$ は独立で

$$\mathbf{X}^{[1]} = (X_1^{[1]}, X_2^{[1]}, \dots, X_k^{[1]}) \sim \text{Negative Multinomial}(r^{[1]}, p_1^{[1]}, p_2^{[1]}, \dots, p_k^{[1]})$$

$$\mathbf{X}^{[2]} = (X_1^{[2]}, X_2^{[2]}, \dots, X_k^{[2]}) \sim \text{Negative Multinomial}(r^{[2]}, p_1^{[2]}, p_2^{[2]}, \dots, p_k^{[2]})$$

であるとき、 $H_1 : (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2 : p_j^{[1]} = p_j^{[2]} = p_j$, $j = 1, \dots, k$, の下で

$$\hat{p}_0^{[i]} = \frac{r^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}, \quad \hat{p}_j^{[i]} = \frac{x_j^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}} \quad (1 \leq j \leq k), \quad (i = 1, 2),$$

$$\hat{p}_0 = \frac{r^{[1]} + r^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}, \quad \hat{p}_j = \frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})} \quad (1 \leq j \leq k),$$

であるから、

$$\text{Total} = \sum_{i=1}^2 \left\{ r^{[i]} \log \frac{\frac{r^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}}{\hat{p}_0} + \sum_{j=1}^k x_j^{[i]} \log \frac{\frac{x_j^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}}{\hat{p}_j} \right\},$$

$$\begin{aligned} \text{Between} &= (r^{[1]} + r^{[2]}) \log \frac{\frac{r^{[1]} + r^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}}{\hat{p}_0} \\ &\quad + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]}) \log \frac{\frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}}{\hat{p}_j}, \end{aligned}$$

$$\text{Within} = \sum_{i=1}^2 \left\{ r^{[i]} \log \frac{\frac{r^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}}{\frac{r^{[1]} + r^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}} \right\}$$

$$+ \sum_{j=1}^k x_j^{[i]} \log \frac{\frac{x_j^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}}{\frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}} \Bigg\}.$$

Proposition 2 負の多項分布の2標本問題にて、Between information と Within information の和は、Total information に等しい。

Proposition 2 は

$$\begin{aligned} \frac{r^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}} &= \frac{r^{[1]} + r^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})} \times \frac{\frac{r^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}}{\frac{r^{[1]} + r^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}} \\ \frac{x_j^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}} &= \frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})} \times \frac{\frac{x_j^{[i]}}{r^{[i]} + \sum_{j=1}^k x_j^{[i]}}}{\frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + \sum_{j=1}^k (x_j^{[1]} + x_j^{[2]})}} \end{aligned}$$

および、対数の性質 $\log AB = \log A + \log B$ を用いれば、容易に導ける。

3. Pooling incomplete samples を伴う多項分布の2標本問題

k, m は $m < k$ なる自然数。確率変数 $\mathbf{X}, \mathbf{Y}$ は互いに独立で

$$\begin{aligned} \mathbf{X} &= (X_1, \dots, X_m, \dots, X_k) \sim \text{Multinomial}(N_1; \theta_1, \dots, \theta_m, \dots, \theta_k) \\ \mathbf{Y} &= (Y_1, \dots, Y_m) \sim \text{Multinomial}(N_2; \frac{\theta_1}{\sum_{l=1}^m \theta_l}, \dots, \frac{\theta_m}{\sum_{l=1}^m \theta_l}) \end{aligned}$$

とする。このようなモデルを、Asano(1965) は pooling incomplete samples と言った。この場合、 θ_j の MVUE $\hat{\theta}_j$ は

$$\hat{\theta}_j = \frac{x_j + y_j}{N_1 \left(1 + \frac{N_2}{\sum_{j=1}^m x_j}\right)} \quad \text{if } j \leq m, \quad \hat{\theta}_j = \frac{x_j}{N_1} \quad \text{if } j > m$$

である。なお、上記の推定量 $\hat{\theta}_j$ は、 θ_j の MLE でもある。

Pooling incomplete samples を伴う場合、 $H_1: \theta_j = p_j, \quad H_2: \theta_j = q_j, \quad (j = 1, \dots, k)$ に対する Kullback 情報量 $I(H_1: H_2)$ を計算すると

$$N_1 \sum_{j=1}^k p_j \log \frac{p_j}{q_j} + \frac{N_2}{\sum_{j=1}^m p_j} \sum_{j=1}^m p_j \log \frac{p_j}{q_j} + N_2 \log \frac{\sum_{l=1}^m q_l}{\sum_{l=1}^m p_l}$$

である。

さて、Pooling incomplete samples を伴う場合の 2 標本問題を考える。すなわち、2 つの独立な多項分布 $i = 1, 2$ からの確率変数 $\mathbf{X}^{[i]} = (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]})$, $\mathbf{Y}^{[i]} = (Y_1^{[i]}, \dots, Y_m^{[i]})$, がある。そして、各 $i = 1, 2$ にて、 $\mathbf{X}^{[i]}$ と $\mathbf{Y}^{[i]}$ は独立。そして

$$\mathbf{X}^{[i]} \sim \text{Multinomial}(N_1^{[i]}; p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}),$$

$$\mathbf{Y}^{[i]} \sim \text{Multinomial}(N_2^{[i]}; \frac{p_1^{[i]}}{\sum_{l=1}^m p_l^{[i]}}, \dots, \frac{p_m^{[i]}}{\sum_{l=1}^m p_l^{[i]}}),$$

$H_1 : (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2 : p_i^{[1]} = p_i^{[2]} = p_i$ のとき、Between information, Within information, Total information は、以下の表の通り：

	Information
Between	$(N_1^{[1]} + N_1^{[2]}) \sum_{j=1}^k \hat{p}_j \log \frac{\hat{p}_j}{p_j} + (N_2^{[1]} + N_2^{[2]}) \sum_{j=1}^m \frac{\hat{p}_j}{\sum_{l=1}^m \hat{p}_l} \log \frac{\hat{p}_j / \sum_{l=1}^m \hat{p}_l}{p_j / \sum_{l=1}^m p_l} \}$
Within	$\sum_{i=1}^2 \left\{ N_1^{[i]} \sum_{j=1}^k \hat{p}_j^{[i]} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_j} + N_2^{[i]} \sum_{j=1}^m \frac{\hat{p}_j^{[i]}}{\sum_{l=1}^m \hat{p}_l^{[i]}} \log \frac{\hat{p}_j^{[i]} / \sum_{l=1}^m \hat{p}_l^{[i]}}{\hat{p}_j / \sum_{l=1}^m \hat{p}_l} \right\}$
Total	$\sum_{i=1}^2 \left\{ N_1^{[i]} \sum_{j=1}^k \hat{p}_j^{[i]} \log \frac{\hat{p}_j^{[i]}}{p_j} + N_2^{[i]} \sum_{j=1}^m \frac{\hat{p}_j^{[i]}}{\sum_{l=1}^m \hat{p}_l^{[i]}} \log \frac{\hat{p}_j^{[i]} / \sum_{l=1}^m \hat{p}_l^{[i]}}{p_j / \sum_{l=1}^m p_l} \right\}$

H_1 の下での MVUE $\hat{p}_j^{[i]}$ および H_2 の下での MVUE $\hat{p}_j$ は

$$\hat{p}_j^{[i]} = \begin{cases} \frac{x_j^{[i]} + y_j^{[i]} \frac{T_x^{[i]}}{T_x^{[i]} + N_2^{[i]}}}{\frac{T_x^{[i]} + N_2^{[i]}}{N_1^{[i]}}} & \text{if } j \leq m, \\ x_j^{[i]} / N_1^{[i]} & \text{if } j > m, \end{cases}$$

$$\hat{p}_j = \begin{cases} \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]} \frac{T_x^{[1]} + T_x^{[2]}}{T_x^{[1]} + T_x^{[2]} + N_2^{[1]} + N_2^{[2]}}}{\frac{T_x^{[1]} + T_x^{[2]} + N_2^{[1]} + N_2^{[2]}}{N_1^{[1]} + N_1^{[2]}}} & \text{if } j \leq m, \\ (x_j^{[1]} + x_j^{[2]}) / (N_1^{[1]} + N_1^{[2]}) & \text{if } j > m. \end{cases}$$

である。但し、 $T_x^{[i]} = \sum_{j=1}^m x_j^{[i]}$. これを代入して書き下すと

$$\begin{aligned} \text{Between} &= \sum_{j=1}^m (x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}) \log \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{(T_x^{[1]} + T_x^{[2]} + N_2^{[1]} + N_2^{[2]}) p_j} \\ &+ \sum_{j=m+1}^k (x_j^{[1]} + x_j^{[2]}) \log \frac{x_j^{[1]} + x_j^{[2]}}{(N_1^{[1]} + N_1^{[2]}) p_j} + (T_x^{[1]} + T_x^{[2]}) \log \frac{T_x^{[1]} + T_x^{[2]}}{N_1^{[1]} + N_1^{[2]}} \end{aligned}$$

$$\begin{aligned}
& +(N_2^{[1]} + N_2^{[2]}) \log(\sum_{l=1}^m p_l) \\
\text{Within} = & \sum_{i=1}^2 \left\{ \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{(x_j^{[i]} + y_j^{[i]})(T_x^{[1]} + T_x^{[2]} + N_2^{[1]} + N_2^{[2]})}{(x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]})(T_x^{[i]} + N_2^{[i]})} \right. \\
& \left. + \sum_{j=m+1}^k x_j^{[i]} \log \frac{(N_1^{[1]} + N_1^{[2]})x_j^{[i]}}{N_1^{[i]}(x_j^{[1]} + x_j^{[2]})} + T_x^{[i]} \log \frac{T_x^{[i]}}{N_1^{[i]}} \right\} + (T_x^{[1]} + T_x^{[2]}) \log \frac{N_1^{[1]} + N_1^{[2]}}{T_x^{[1]} + T_x^{[2]}} \\
\text{Total} = & \sum_{i=1}^2 \left\{ \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{x_j^{[i]} + y_j^{[i]}}{(T_x^{[i]} + N_2^{[i]})p_j} + \sum_{j=m+1}^k x_j^{[i]} \log \frac{x_j^{[i]}}{N_1^{[i]}p_j} \right. \\
& \left. + T_x^{[i]} \log \frac{T_x^{[i]}}{N_1^{[i]}} + N_2^{[i]} \log(\sum_{l=1}^m p_l) \right\}
\end{aligned}$$

である。

Proposition 3 上記の状況下において、すなわち、Pooling incomplete samples を伴う多項分布の 2 標本問題において、Between information と Within information の和は、Total information に等しい。

Funo(2012) は Proposition 3 を、Total, Between, Within の各式に推定量を代入して、式変形を行うことで示した。しかしながら、この式変形には膨大な計算が必要であった。ところがその後、直和分解を示すのに次のような簡便な方法が見つかったので、これを記載する。

$q_j = p_j / \sum_{l=1}^m p_l$ とおくと、

$$\begin{aligned}
\text{Between} = & (N_1^{[1]} + N_1^{[2]}) \sum_{j=1}^k \hat{p}_j \log \frac{\hat{p}_j}{p_j} + (N_2^{[1]} + N_2^{[2]}) \sum_{j=1}^m \hat{q}_j \log \frac{\hat{q}_j}{q_j} \} \\
\text{Within} = & \sum_{i=1}^2 \left\{ N_1^{[i]} \sum_{j=1}^k \hat{p}_j^{[i]} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_j} + N_2^{[i]} \sum_{j=1}^m \hat{q}_j^{[i]} \log \frac{\hat{q}_j^{[i]}}{\hat{q}_j} \right\} \\
\text{Total} = & \sum_{i=1}^2 \left\{ N_1^{[i]} \sum_{j=1}^k \hat{p}_j^{[i]} \log \frac{\hat{p}_j^{[i]}}{p_j} + N_2^{[i]} \sum_{j=1}^m \hat{q}_j^{[i]} \log \frac{\hat{q}_j^{[i]}}{q_j} \right\}
\end{aligned}$$

と書き表せる。ここで

$$\sum_j \hat{p}_j^{[i]} \log \frac{\hat{p}_j^{[i]}}{p_j} = \sum_j \hat{p}_j^{[i]} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_i} + \sum_j \hat{p}_j^{[i]} \log \frac{\hat{p}_i}{p_j}, \quad \sum_j \hat{q}_j^{[i]} \log \frac{\hat{q}_j^{[i]}}{q_j} = \sum_j \hat{q}_j^{[i]} \log \frac{\hat{q}_j^{[i]}}{\hat{q}_i} + \sum_j \hat{q}_j^{[i]} \log \frac{\hat{q}_i}{q_j}$$

は常に成り立つ。よって、

$$(N_1^{[1]} + N_1^{[2]})\hat{p}_j + (N_2^{[1]} + N_2^{[2]})\hat{q}_j = N_1^{[1]}\hat{p}_j^{[1]} + N_1^{[2]}\hat{p}_j^{[2]} + N_2^{[1]}\hat{q}_j^{[1]} + N_2^{[2]}\hat{q}_j^{[2]}$$

が成り立つことを示せば良い。この計算は、比較的簡単にできる。実際、

$$\sum_{j=1}^m \hat{p}_j = \frac{T_x^{[1]} + T_x^{[2]}}{N_1^{[1]} + N_1^{[2]}}, \quad \hat{q}_j = \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{T_x^{[1]} + T_x^{[2]} + N_2^{[1]} + N_2^{[2]}}, \quad \sum_{j=1}^m \hat{p}_j^{[i]} = \frac{T_x^{[i]}}{N_1^{[i]}}, \quad \hat{q}_j^{[i]} = \frac{x_j^{[1]} + y_j^{[1]}}{T_x^{[1]} + N_2^{[1]}}$$

であるから、これらを代入することで

$$\begin{aligned}(N_1^{[1]} + N_1^{[2]})\hat{p}_j + (N_2^{[1]} + N_2^{[2]})\hat{q}_j &= x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]} \\ N_1^{[1]}\hat{p}_j^{[1]} + N_1^{[2]}\hat{p}_j^{[2]} + N_2^{[1]}\hat{q}_j^{[1]} + N_2^{[2]}\hat{q}_j^{[2]} &= x_j^{[1]} + y_j^{[1]} + x_j^{[2]} + y_j^{[2]}\end{aligned}$$

が得られ、Proposition 3 が示せた。

4. Pooling incomplete sample を伴う負の多項分布の 2 標本問題

これより後、各 $i = 1, 2$ に対して

$$T_m^{[i]}(x) = \sum_{j=1}^m x_j^{[i]}, \quad T_m^{[i]}(y) = \sum_{j=1}^m y_j^{[i]}, \quad T_k^{[i]}(x) = \sum_{j=1}^k x_j^{[i]}, \quad T^{[i]}(y) = \sum_{j=2}^m y_j^{[i]}$$

なる記号を用いる。また、パラメーター $(p_0, p_1, \dots, p_k)$ は $0 \leq p_l \leq 1, \sum_{l=0}^k p_l = 1$ を満たすものとする。

Proposition 4 $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で

$$\begin{aligned}\mathbf{X}^{[i]} &= (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{Negative Multinomial}(r_1^{[i]}, p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}), \\ \mathbf{Y}^{[i]} &= (Y_1^{[i]}, \dots, Y_m^{[i]}) \sim \text{Negative Multinomial}(r_2^{[i]}, \frac{p_1^{[i]}}{\sum_{j=0}^m p_j^{[i]}}, \dots, \frac{p_m^{[i]}}{\sum_{j=0}^m p_j^{[i]}}).\end{aligned}$$

であるとき、 $H_1 : (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2 : p_j^{[1]} = p_j^{[2]} = p_j, j = 1, \dots, k$, の下で Total information, Between information, Within information は

$$\begin{aligned}\text{Total} &= \sum_{i=1}^2 \left\{ r_1^{[i]} \frac{r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]}} \log \frac{r_1^{[i]} + T_m^{[i]}(x)}{r_1^{[i]} + T_k^{[i]}(x)} \right. \\ &\quad + (r_1^{[i]} + r_2^{[i]}) \log \frac{r_1^{[i]} + r_2^{[i]}}{(r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y))p_0} \\ &\quad + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{x_j^{[i]} + y_j^{[i]}}{(r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y))p_j} \\ &\quad + r_1^{[i]} \frac{r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]}} \sum_{j=m+1}^k \frac{x_j^{[i]}}{r_1^{[i]} + T_m^{[i]}(x)} \log \frac{x_j^{[i]}}{(r_1^{[i]} + T_k^{[i]}(x))p_j} \\ &\quad \left. + r_2^{[i]} \frac{r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]}} \log(p_0 + \dots + p_m) \right\},\end{aligned}$$

$$\begin{aligned}\text{Between} &= (r_1^{[1]} + r_1^{[2]}) \frac{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}} \\ &\quad \times \log \frac{r_1^{[1]} + r_1^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)}{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}\end{aligned}$$

$$\begin{aligned}
& + (r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}) \\
& \quad \times \log \frac{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}}{(r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y))p_0} \\
& + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{x_j^{[i]} + y_j^{[i]}}{(r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y))p_j} \\
& + (r_1^{[1]} + r_1^{[2]}) \frac{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}} \\
& \quad \times \sum_{j=m+1}^k \frac{x_j^{[i]}}{r_1^{[1]} + r_1^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)} \log \frac{x_j^{[i]}}{(r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x))p_j} \\
& + (r_2^{[1]} + r_2^{[2]}) \frac{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}} \\
& \quad \times \log(p_0 + \cdots + p_m), \\
\text{Within} = & \sum_{i=1}^2 \left\{ r_1^{[i]} \frac{r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]}} \log \frac{r_1^{[i]} + T_m^{[i]}(x)}{r_1^{[i]} + T_k^{[i]}(x)} \right. \\
& + (r_1^{[i]} + r_2^{[i]}) \left(\log \frac{r_1^{[i]} + r_2^{[i]}}{(r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y))} + \log \frac{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}{r_1^{[1]} + r_1^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)} \right. \\
& \left. \left. + \log \frac{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}} \right) \right. \\
& + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \left(\log \frac{x_j^{[i]} + y_j^{[i]}}{(r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y))} + \log \frac{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}{r_1^{[1]} + r_1^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)} \right. \\
& \left. \left. + \log \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}} \right) \right. \\
& + r_1^{[i]} \frac{r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]}} \sum_{j=m+1}^k \frac{x_j^{[i]}}{r_1^{[i]} + T_m^{[i]}(x)} \\
& \quad \times \left(\log \frac{x_j^{[i]}}{(r_1^{[i]} + T_k^{[i]}(x))\hat{p}_j} + \log \frac{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}{x_j^{[1]} + x_j^{[2]}} \right) \\
& \left. + r_2^{[i]} \frac{r_1^{[i]} + r_2^{[i]} + T_m^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]}} \log \frac{r_1^{[1]} + r_1^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)}{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \right\}
\end{aligned}$$

である。

Remark 上記の結果より、Total = Between + Within が成り立たない。

Proposition 4 の仮定条件は、Pooling incomplete sample を伴う負の多項分布の 2 標本問題として、きわめて自然なものと思えるものの、Total = Between + Within(直和分解) が成立しないのはどうしてなのか、最初は戸惑いを感じた。モデルの設定条件をいろいろ変更してみたところ、次のような場合に直和分解が成立することがわかった。

Proposition 5 (負の多項分布と、Pooling incomplete sample が多項分布の場合の 2 標本問題)

$\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で

$$\mathbf{X}^{[i]} = (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{Negative Multinomial}(r_1^{[i]}, p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}),$$

$$\mathbf{Y}^{[i]} = (Y_1^{[i]}, \dots, Y_m^{[i]}) \sim \text{Multinomial}(N_2^{[i]}, \frac{p_1^{[i]}}{\sum_{j=1}^m p_j^{[i]}}, \dots, \frac{p_m^{[i]}}{\sum_{j=1}^m p_j^{[i]}})$$

であるとき、 $H_1 : (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2 : p_j^{[1]} = p_j^{[2]} = p_j$, $j = 1, \dots, k$, の下で Total information, Between information, Within information は

$$\begin{aligned} \text{Total} = & \sum_{i=1}^2 \left(r^{[i]} \log \frac{r^{[i]}}{(r^{[i]} + T_k^{[i]}(x))p_0} + (T_m^{[i]}(x) + N_2^{[i]}) \log \frac{T_m^{[i]}(x)}{r^{[i]} + T_k^{[i]}(x)} \right. \\ & + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{x_j^{[i]} + y_j^{[i]}}{(T_m^{[i]}(x) + N_2^{[i]})p_j} + \sum_{j=m+1}^k x_j^{[i]} \log \frac{x_j^{[i]}}{(r^{[i]} + T_k^{[i]}(x))p_j} \\ & \left. + N_2^{[i]} \log \frac{\sum_{j=1}^m p_j}{T_m^{[i]}(x)/(r^{[i]} + T_k^{[i]}(x))} \right), \end{aligned}$$

$$\begin{aligned} \text{Between} = & (r^{[1]} + r^{[2]}) \log \frac{r^{[1]} + r^{[2]}}{(r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x))p_0} \\ & + (T_m^{[1]}(x) + T_m^{[2]}(x) + N_2^{[1]} + N_2^{[2]}) \log \frac{T_m^{[1]}(x) + T_m^{[2]}(x)}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \\ & + \sum_{j=1}^m (x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}) \log \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{(T_m^{[1]}(x) + T_m^{[2]}(x) + N_2^{[1]} + N_2^{[2]})p_j} \\ & + \sum_{j=m+1}^k (x_j^{[1]} + x_j^{[2]}) \log \frac{x_j^{[1]} + x_j^{[2]}}{(r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x))p_j} \\ & + (N_2^{[1]} + N_2^{[2]}) \log \frac{\sum_{j=1}^m p_j}{(T_m^{[1]}(x) + T_m^{[2]}(x))/(r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x))}, \end{aligned}$$

$$\begin{aligned} \text{Within} = & \sum_{i=1}^2 \left(r^{[i]} \log \frac{\frac{r^{[i]}}{r^{[i]} + T_k^{[i]}(x)}}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} + (T_m^{[i]}(x) + N_2^{[i]}) \log \frac{T_m^{[i]}(x)}{r^{[i]} + T_k^{[i]}(x)} \right. \\ & + \sum_{j=1}^m (x_j^{[i]} + y_j^{[i]}) \log \frac{\frac{x_j^{[i]} + y_j^{[i]}}{T_m^{[i]}(x) + N_2^{[i]}}}{\frac{T_m^{[1]}(x) + T_m^{[2]}(x)}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{T_m^{[1]}(x) + T_m^{[2]}(x) + N_2^{[1]} + N_2^{[2]}}} \end{aligned}$$

$$+ \sum_{j=m+1}^k x_j^{[i]} \log \frac{\frac{x_j^{[i]}}{r^{[i]} + T_k^{[i]}(x)}}{\frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}} + N_2^{[i]} \log \frac{\frac{T_m^{[1]}(x) + T_m^{[2]}(x)}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}}{\frac{T_m^{[i]}(x)}{r^{[i]} + T_k^{[i]}(x)}} \Bigg).$$

よって、Total = Between + Within, すなわち、直和分解が成り立つ。

この現象に関する考察は、次節で行う。

5. 考察および今後の課題

負の多項分布は、負の二項分布 × 多項分布 と分解できる。実際

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

このことと Proposition 5 により、pooling incomplete samples が「(分布の分解後の) 多項分布」の部分で行われていれば Total = Between + Within すなわち、直和分解が成り立つ、といえる。しかし、これ以外の箇所で pooling incomplete samples が行われた場合は、Total = Between + Within が成り立つとは限らない。たとえば、次のような例 (例 1 ~ 例 3) を見つけることができた。

例 1 (Proposition 5 とは別のモデル) $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で

$$\begin{aligned} \mathbf{X}^{[i]} &= (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{Negative Multinomial}(r_1^{[i]}, p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}), \\ \mathbf{Y}^{[i]} &= (Y_0^{[i]}, \dots, Y_m^{[i]}) \sim \text{Multinomial}(N_2^{[i]}, \frac{p_0^{[i]}}{\sum_{j=0}^m p_j^{[i]}}, \dots, \frac{p_m^{[i]}}{\sum_{j=0}^m p_j^{[i]}}) \end{aligned}$$

であるとき、 $H_1: (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2: p_j^{[1]} = p_j^{[2]} = p_j$, $j = 1, \dots, k$, の下で Total information, Between information, Within information は

$$\text{Total} = \sum_{i=1}^2 \left[r_1^{[i]} \left\{ \log \frac{\hat{p}_0^{[i]}}{p_0} + \sum_{j=1}^k \frac{\hat{p}_j^{[i]}}{\hat{p}_0^{[i]}} \log \frac{\hat{p}_j^{[i]}}{p_j} \right\} + N_2^{[i]} \left\{ \sum_{j=0}^m \frac{\hat{p}_j^{[i]}}{\sum_{j=0}^m \hat{p}_j^{[i]}} \log \frac{\hat{p}_j^{[i]} / \sum_{l=0}^m \hat{p}_l^{[i]}}{p_j / \sum_{l=0}^m p_l} \right\} \right]$$

$$\text{Between} = (r_1^{[1]} + r_1^{[2]}) \left\{ \log \frac{\hat{p}_0}{p_0} + \sum_{j=1}^k \frac{\hat{p}_j}{\hat{p}_0} \log \frac{\hat{p}_j}{p_j} \right\} + (N_2^{[1]} + N_2^{[2]}) \left\{ \sum_{j=0}^m \frac{\hat{p}_j}{\sum_{j=0}^m \hat{p}_j} \log \frac{\hat{p}_j / \sum_{l=0}^m \hat{p}_l}{p_j / \sum_{l=0}^m p_l} \right\}$$

$$\text{Within} = \sum_{i=1}^2 \left[r_1^{[i]} \left\{ \log \frac{\hat{p}_0^{[i]}}{\hat{p}_0} + \sum_{j=1}^k \frac{\hat{p}_j^{[i]}}{\hat{p}_0^{[i]}} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_j} \right\} + N_2^{[i]} \left\{ \sum_{j=0}^m \frac{\hat{p}_j^{[i]}}{\sum_{j=0}^m \hat{p}_j^{[i]}} \log \frac{\hat{p}_j^{[i]} / \sum_{l=0}^m \hat{p}_l^{[i]}}{\hat{p}_j / \sum_{l=0}^m \hat{p}_l} \right\} \right]$$

に、推定量

$$\hat{p}_j^{[i]} = \begin{cases} \frac{r^{[i]} + T_m^{[i]}(x)}{r^{[i]} + T_k^{[i]}(x)} \frac{r^{[i]} + y_0^{[i]}}{r^{[i]} + T_m^{[i]}(x) + N_2^{[i]}}, & \text{if } j = 0, \\ \frac{r^{[i]} + T_m^{[i]}(x)}{r^{[i]} + T_k^{[i]}(x)} \frac{x_j^{[i]} + y_j^{[i]}}{r^{[i]} + T_m^{[i]}(x) + N_2^{[i]}}, & \text{if } 1 \leq i \leq m, \\ \frac{x_j^{[i]}}{r^{[i]} + T_k^{[i]}(x)}, & \text{if } i > m, \end{cases}$$

$$\hat{p}_j = \begin{cases} \frac{r^{[1]} + r^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{r^{[1]} + r^{[2]} + y_0^{[1]} + y_0^{[2]}}{r^{[1]} + r^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + N_2^{[1]} + N_2^{[2]}}, & \text{if } j = 0, \\ \frac{r^{[1]} + r^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x)}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{r^{[1]} + r^{[2]} + T_m^{[1]}(x) + T_m^{[2]}(x) + N_2^{[1]} + N_2^{[2]}}, & \text{if } 1 \leq i \leq m, \\ \frac{x_j^{[1]} + x_j^{[2]}}{r^{[1]} + r^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{r^{[1]} + r^{[2]} + y_0^{[1]} + y_0^{[2]}}{r^{[1]} + r^{[2]}}, & \text{if } i > m, \end{cases}$$

を代入したもの。これを計算してみたところ、Total = Between + Within が成立しないことがわかった。

例 2 $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で

$$\mathbf{X}^{[i]} = (X_1^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{Negative Multinomial}(r_1^{[i]}, p_1^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}),$$

$$\mathbf{Y}^{[i]} = (Y_1^{[i]}, \dots, Y_m^{[i]}) \sim \text{Negative Multinomial}(r_2^{[i]}, \frac{(1 - p_0^{[i]})p_1^{[i]}}{\sum_{j=1}^m p_j^{[i]}}, \dots, \frac{(1 - p_0^{[i]})p_m^{[i]}}{\sum_{j=1}^m p_j^{[i]}})$$

であるとき、 $H_1 : (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2 : p_j^{[1]} = p_j^{[2]} = p_j$, $j = 1, \dots, k$, の下で Total information, Between information, Within information は

$$\text{Total} = \sum_{i=1}^2 \left\{ (r_1^{[i]} + r_2^{[i]}) \log \frac{\hat{p}_0^{[i]}}{p_0} + r_1^{[i]} \sum_{j=1}^k \frac{\hat{p}_j^{[i]}}{p_0} \log \frac{\hat{p}_j^{[i]}}{p_j} \right. \\ \left. + r_2^{[i]} \frac{1 - \hat{p}_0^{[i]}}{\hat{p}_0^{[i]}} \left(\sum_{j=1}^m \frac{\hat{p}_j^{[i]}}{\sum_{l=1}^m \hat{p}_l^{[i]}} \log \frac{\hat{p}_j^{[i]}}{p_j} + \log \frac{1 - \hat{p}_0^{[i]}}{1 - p_0} + \log \frac{\sum_{j=1}^m p_j}{\sum_{j=1}^m \hat{p}_j^{[i]}} \right) \right\}$$

$$\text{Between} = (r_1^{[1]} + r_2^{[1]} + r_1^{[2]} + r_2^{[2]}) \log \frac{\hat{p}_0}{p_0} + (r_1^{[1]} + r_2^{[1]}) \sum_{j=1}^k \frac{\hat{p}_j}{p_0} \log \frac{\hat{p}_j}{p_j} \\ + (r_1^{[2]} + r_2^{[2]}) \frac{1 - \hat{p}_0}{\hat{p}_0} \left(\sum_{j=1}^m \frac{\hat{p}_j}{\sum_{l=1}^m \hat{p}_l} \log \frac{\hat{p}_j}{p_j} + \log \frac{1 - \hat{p}_0}{1 - p_0} + \log \frac{\sum_{j=1}^m p_j}{\sum_{j=1}^m \hat{p}_j} \right)$$

$$\text{Within} = \sum_{i=1}^2 \left\{ (r_1^{[i]} + r_2^{[i]}) \log \frac{\hat{p}_0^{[i]}}{\hat{p}_0} + r_1^{[i]} \sum_{j=1}^k \frac{\hat{p}_j^{[i]}}{\hat{p}_0} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_j} \right. \\ \left. + r_2^{[i]} \frac{1 - \hat{p}_0^{[i]}}{\hat{p}_0^{[i]}} \left(\sum_{j=1}^m \frac{\hat{p}_j^{[i]}}{\sum_{l=1}^m \hat{p}_l^{[i]}} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_j} + \log \frac{1 - \hat{p}_0^{[i]}}{1 - \hat{p}_0} + \log \frac{\sum_{j=1}^m \hat{p}_j}{\sum_{j=1}^m \hat{p}_j^{[i]}} \right) \right\}$$

に、推定量

$$\hat{p}_j^{[i]} = \begin{cases} \frac{r_1^{[i]} + r_2^{[i]}}{r_1^{[i]} + r_2^{[i]} + T_k^{[i]}(x) + T_m^{[i]}(y)}, & \text{if } j = 0, \\ \frac{T_k^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]} + T_k^{[i]}(x) + T_m^{[i]}(y)} \frac{T_m^{[i]}(x)}{T_k^{[i]}(x)} \frac{x_j^{[i]} + y_j^{[i]}}{T_m^{[i]}(x) + T_m^{[i]}(y)}, & \text{if } 1 \leq j \leq m, \\ \frac{T_k^{[i]}(x) + T_m^{[i]}(y)}{r_1^{[i]} + r_2^{[i]} + T_k^{[i]}(x) + T_m^{[i]}(y)} \frac{x_j^{[i]}}{T_k^{[i]}(x)}, & \text{if } j > m, \end{cases}$$

$$\hat{p}_j = \begin{cases} \frac{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]}}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)} & \text{if } j = 0, \\ \frac{T_k^{[1]}(x) + T_k^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)} \\ \times \frac{T_m^{[1]}(y) + T_m^{[2]}(y)}{T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{T_m^{[1]}(x) + T_m^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)} & \text{if } 1 \leq j \leq m, \\ \frac{T_k^{[1]}(x) + T_k^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)}{r_1^{[1]} + r_1^{[2]} + r_2^{[1]} + r_2^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x) + T_m^{[1]}(y) + T_m^{[2]}(y)} \\ \times \frac{x_j^{[1]} + x_j^{[2]}}{T_k^{[1]}(x) + T_k^{[2]}(x)} & \text{if } j > m, \end{cases}$$

を代入したもの。これを計算してみたところ、Total = Between + Within が成立しないことがわかった。

例 3 $\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ は互いに独立で

$$\mathbf{X}^{[i]} = (X_1^{[i]}, X_2^{[i]}, \dots, X_m^{[i]}, \dots, X_k^{[i]}) \sim \text{Negative Multinomial}(r_1^{[i]}, p_1^{[i]}, p_2^{[i]}, \dots, p_m^{[i]}, \dots, p_k^{[i]}),$$

$$\mathbf{Y}^{[i]} = (Y_2^{[i]}, \dots, Y_m^{[i]}) \sim \text{NM}(r_2^{[i]}, \frac{p_2^{[i]}}{\sum_{j=1}^m p_j^{[i]}}, \dots, \frac{p_m^{[i]}}{\sum_{j=1}^m p_j^{[i]}})$$

であるとき、 $H_1: (p_1^{[1]}, \dots, p_k^{[1]}) \neq (p_1^{[2]}, \dots, p_k^{[2]})$, $H_2: p_j^{[1]} = p_j^{[2]} = p_j$, $j = 1, \dots, k$, の下で Total information, Between information, Within information は

$$\text{Total} = \sum_{i=1}^2 \left\{ r_1^{[i]} \sum_{j=0}^k \frac{\hat{p}_j^{[i]}}{\hat{p}_0^{[i]}} \log \frac{\hat{p}_j^{[i]}}{p_j} + r_2^{[i]} \sum_{j=1}^m \frac{\hat{p}_j^{[i]}}{\hat{p}_1^{[i]}} \log \frac{\hat{p}_j^{[i]} / \sum_{l=1}^m \hat{p}_l^{[i]}}{p_j / \sum_{l=1}^m p_l} \right\}$$

$$\text{Between} = (r_1^{[1]} + r_1^{[2]}) \sum_{j=0}^k \frac{\hat{p}_j}{\hat{p}_0} \log \frac{\hat{p}_j}{p_j} + (r_2^{[1]} + r_2^{[2]}) \sum_{j=1}^m \frac{\hat{p}_j}{\hat{p}_1} \log \frac{\hat{p}_j / \sum_{l=1}^m \hat{p}_l}{p_j / \sum_{l=1}^m p_l}$$

$$\text{Within} = \sum_{i=1}^2 \left\{ r_1^{[i]} \sum_{j=0}^k \frac{\hat{p}_j^{[i]}}{\hat{p}_0^{[i]}} \log \frac{\hat{p}_j^{[i]}}{\hat{p}_j} + r_2^{[i]} \sum_{j=1}^m \frac{\hat{p}_j^{[i]}}{\hat{p}_1^{[i]}} \log \frac{\hat{p}_j^{[i]} / \sum_{l=1}^m \hat{p}_l^{[i]}}{\hat{p}_j / \sum_{l=1}^m p_l} \right\}$$

に、推定量

$$\hat{p}_j^{[i]} = \begin{cases} \frac{r_1^{[i]}}{r_1^{[i]} + T_k^{[i]}(x)}, & \text{if } j = 0, \\ \frac{T_m^{[i]}(x)}{r_1^{[i]} + T_k^{[i]}(x)} \frac{x_1^{[i]} + r_2^{[i]}}{T_m^{[i]}(x) + r_2^{[i]} + T^{[i]}(y)}, & \text{if } j = 1, \\ \frac{T_m^{[i]}(x)}{r_1^{[i]} + T_k^{[i]}(x)} \frac{x_j^{[i]} + y_j^{[i]}}{T_m^{[i]}(x) + r_2^{[i]} + T^{[i]}(y)}, & \text{if } 2 \leq j \leq m, \\ \frac{x_j^{[i]}}{r_1^{[i]} + T_k^{[i]}(x)}, & \text{if } j > m, \end{cases}$$

$$\hat{p}_j = \begin{cases} \frac{r_1^{[1]} + r_1^{[2]}}{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}, & \text{if } j = 0, \\ \frac{T_k^{[1]}(m) + T_m^{[2]}(x)}{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{x_1^{[1]} + x_1^{[2]} + r_2^{[1]} + r_2^{[2]}}{T_k^{[1]}(m) + T_m^{[2]}(x) + r_2^{[1]} + r_2^{[2]} + T^{[1]}(y) + T^{[2]}(y)}, & \text{if } j = 1, \\ \frac{T_k^{[1]}(m) + T_m^{[2]}(x)}{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)} \frac{x_j^{[1]} + x_j^{[2]} + y_j^{[1]} + y_j^{[2]}}{T_k^{[1]}(m) + T_m^{[2]}(x) + r_2^{[1]} + r_2^{[2]} + T^{[1]}(y) + T^{[2]}(y)}, & \text{if } 2 \leq j \leq m, \\ \frac{x_j^{[1]} + x_j^{[2]}}{r_1^{[1]} + r_1^{[2]} + T_k^{[1]}(x) + T_k^{[2]}(x)}, & \text{if } j > m, \end{cases}$$

を代入したもの。これを計算してみたところ、 $\text{Total} = \text{Between} + \text{Within}$ が成立しないことがわかった。

負の多項分布が 負の二項分布 $\times$ 多項分布 と分解されること、および Proposition 5 によれば、カテゴリー 0 と カテゴリー $1, 2, \dots, k$ をまず分けて、カテゴリー $1, 2, \dots, k$ の部分で「多項分布の Pooling incomplete sample」を行えば、2 標本問題で Total information が Between information と Within information の直和に分解ができた。ところが、Proposition 4 における確率モデルは、カテゴリー $0, 1, 2, \dots, m$ での「負の多項分布の Pooling incomplete sample」である。例 1 は、多項分布ではあるものの、カテゴリー $0, 1, 2, \dots, m$ での Pooling incomplete sample である。例 2 は、カテゴリー 0 と カテゴリー $1, 2, \dots, k$ をまず分けており、「多項分布の箇所での Pooling incomplete sample」のように思えるが、 $\mathbf{Y}^{[1]}, \mathbf{Y}^{[2]}$ の分布は負の多項分布である。更に、カテゴリー $1, \dots, m$ の確率の中に $1 - p_0 = \sum_{l=1}^k p_l$ の項が入っているため、

確率構造が tree structure(樹木図) で書かれていない。 Pooling incomplete sample の確率構造は tree structure で書かれるものゆえ、これも例 2 が Proposition 5 の仮定に反している点である。例 3 は、カテゴリ 0 と カテゴリ $1, 2, \dots, k$ を分けてはいるが、 $\mathbf{Y}^{[1]}$, $\mathbf{Y}^{[2]}$ の分布は負の多項分布。それゆえ、直和分解が保証されないのであるが、直和分解が成立しないことが示せたのである。

さて、上記のことから、「pooling incomplete samples が行われた箇所が、分布分解後の多項分布以外の箇所であれば、Total $\neq$ Between + Within となるであろう」と予想されるが、この予想の解決に関しては、今後の課題である。

References

- [1] Asano, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. *Annals of the Institute of Statistical Mathematics* **17**, 1-13.
- [2] Funo, E. (2012). Analysis of two independent samples from pooling incomplete multinomial distributions. *Quarterly Journal of Economics*, edited by The Society of Economics, Kanto Gakuin University, **253**, 1-14.
- [3] 稲垣宣生 (2003). 数理統計学 (改訂版), 裳華房.
- [4] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley.
- [5] Kullback, S. (1959). *Information Theory and Statistics*. Wiley.
- [6] Kullback, S. (1968). *Information Theory and Statistics*, Revised edition. Dover.