順位情報に基づく事前分布を用いた多項確率の推定

東京理科大学　田畑　耕治
Kouji Tahata
Tokyo University of Science
E-mail address: kouji_tahata@rs.tus.ac.jp

東京理科大学大学院　岸村　遼
Ryo Kishimura
Tokyo University of Science

統計数理研究所　柳本　武美
Takemi Yanagimoto
The Institute of Statistical Mathematics

1　はじめに

確率ベクトル $\boldsymbol{p}=(p_1,\ldots,p_K)$ をパラメータとしてもつ多項分布 $Multi(n,\boldsymbol{p})$ から，標本 $\boldsymbol{x}=(x_1,\ldots,x_K)$ が得られたときの $\boldsymbol{p}$ の推定問題について考える．ただし，$\sum_{i=1}^{K}p_i=1$，$n=\sum_{i=1}^{K}x_i$ である．$\boldsymbol{p}$ の推定量として，最尤推定量 (MLE) である標本割合 $\hat{\boldsymbol{p}}=\boldsymbol{x}/n$ がよく用いられる．しかしながら，総観測度数 n が小さくカテゴリ数 K が大きい場合（$n<K$）には，しばしば度数 0 のセルが生じる．例えば，Tuyl (2017, 2019) は二つの極端な例を紹介している．

1. Berger *et al.* (2015) の例
   $K=1000$，$n=3$，$x_1=2$，$x_2=1$，$x_i=0$（$i=3,\ldots,1000$）

2. Berger の例
   $K=1000$，$n=x_1=100$，$x_i=0$（$i=2,\ldots,1000$）

本稿では，このような場合に適切な推定量について考える．

度数 0 のセルに対するセル確率を MLE により推定すると，0.0 となる．そのため，Agresti and Hitchcock (2005) はこのような場合には MLE は好ましくない推定量であり，ベイズ法を用いたより好ましい推定量の検討がなされてきたと述べている．ベイズ法を用いた $\boldsymbol{p}$ の推定問題は，Good (1967), Fienberg and Holland (1972, 1973), Bishop *et al.* (1975) などがある．

ベイズ法を用いた $\boldsymbol{p}$ の推定問題においては，平均母数を用いて議論される場合が多い．その一方で，Yanagimoto and Ohnishi (2009) は自然母数のベイズ推定量に関する最適

性について議論している．また，Ogura and Yanagimoto (2020) はポアソン分布の平均母数の推定において，共役事前分布を用いた自然母数の推定量の逆変換を用いた推定量を提案している．さらに，Ogura and Yanagimoto (2022) は多項分布の平均母数の推定において，Zeta 事前分布を用いた自然母数の推定量の逆変換を用いた推定量を提案している．Zeta 事前分布は，データセットから求められる順位情報を用いた事前分布であり，Ogura and Yanagimoto (2022) はハイパーパラメータの値を経験に基づき与えている．

本稿では，Zeta 事前分布のハイパーパラメータを未知母数として考え，二つの方法でその値を決定する方法を提案する．一つは，周辺尤度を最大化する方法であり，もう一つは，リスク関数を最小化する方法である．また，シミュレーションにより提案方法と従来の推定量について比較と考察を行う．

2　提案推定量

多項分布の母数 $\boldsymbol{p}$ に対する（共役）事前分布として，母数 $\boldsymbol{a} = (a_1, \ldots, a_K)$ のディリクレ分布を用いる：

$$\pi(\boldsymbol{p}; \boldsymbol{a}) = \frac{\Gamma(M)}{\prod_{i=1}^{K} \Gamma(a_i)} \prod_{i=1}^{K} p_i^{a_i - 1}.$$

ただし，$a_i > 0$，$M = \sum_{i=1}^{K} a_i$ である．ベイズの定理により $\boldsymbol{p}$ の事後分布は $\boldsymbol{\lambda} = \boldsymbol{a}/M$ とおくと，母数 $\boldsymbol{x} + M\boldsymbol{\lambda}$ のディリクレ分布に従う．このとき, 母数 $\boldsymbol{p}$ のベイズ推定量

$$\hat{p}_i(\boldsymbol{x}; \boldsymbol{\lambda}, M) = \frac{x_i + M\lambda_i}{n + M} \quad (i = 1, \ldots, K) \tag{1}$$

はよく知られている．式 (1) において $\boldsymbol{\lambda} = (1/K, \ldots, 1/K)$ として，$M \to 0$ のとき Haldane's prior を用いた推定量となる．同様に，$M = 1$ のとき Perks' prior，$M = K/2$ のとき Jeffreys' prior，$M = K$ のとき Bayes-Laplace's prior，$M = \sqrt{n}$ のとき Trybula's prior にそれぞれ対応する (Alvares *et al.* 2018)．

Ogura and Yanagimoto (2022) は，ディリクレ分布の母数として

$$\boldsymbol{r} = (1/r_1^s, \ldots, 1/r_K^s) \tag{2}$$

を用いた．ただし，r_i は第 i カテゴリのセル度数の降順の順位であり，s は実数である．同順位の場合には，小さい方の順位を割り当てることにする．例えば，$K = 5$ のとき $\boldsymbol{x} = (1, 5, 3, 5, 6)$ が得られたならば，$\boldsymbol{r} = (1/5^s, 1/2^s, 1/4^s, 1/2^s, 1/1^s)$ である．式 (2) を母数としてもつディリクレ分布を Zeta 事前分布と呼ぶ．Ogura and Yanagimoto (2022) は，$s = 1$ または $s = 1/2$ を用いることを推奨している．$\boldsymbol{p}$ の推定量として，Ogura and Yanagimoto (2022) は自然母数のベイズ推定量の逆変換を用いた推定量として

$$\hat{p}_i(\boldsymbol{x}; \boldsymbol{r}, s) = \frac{\exp\{\psi(x_i + 1/r_i^s)\}}{\sum_{h=1}^{K} \exp\{\psi(x_h + 1/r_h^s)\}} \quad (i = 1, \ldots, K), \tag{3}$$

を提案した．ただし，$\psi(\cdot)$ はディガンマ関数である．

式 (3) の s を未知母数として，最適な s を選択する方法を二つ提案する．一つは，周辺尤度を最大化する方法であり，もう一つは，Fienberg and Holland (1973) で提案された pseudo Bayes 推定量と類似の方法によって導出するものである．また，式 (2) において，同順位の場合には，大きい方の順位を割り当てることにする．例えば，$K = 5$ のとき $\boldsymbol{x} = (1, 5, 3, 5, 6)$ が得られたならば，$\boldsymbol{r} = (1/5^s, 1/3^s, 1/4^s, 1/3^s, 1/1^s)$ である．

2.1 周辺尤度最大化

式 (2) の s を以下のように決定する．$\boldsymbol{x} = (x_1, \ldots, x_K)$ が $Multi(n, \boldsymbol{p})$ からの標本であり，母数 $\boldsymbol{p}$ に対する事前分布として母数 $\boldsymbol{r} = (1/r_1^s, \ldots, 1/r_K^s)$ のディリクレ分布を仮定する．このとき，周辺尤度 $m(\boldsymbol{x}; \boldsymbol{r}, s)$ に対数をとった $\log m(\boldsymbol{x}; \boldsymbol{r}, s)$ は次のように与えられる：

$$\begin{aligned}\log m(\boldsymbol{x}; \boldsymbol{r}, s) = {} & \log \Gamma(n+1) + \log \Gamma(M(s)) + \sum_{i=1}^{K} \log \Gamma(x_i + 1/r_i^s) \\ & - \sum_{i=1}^{K} \log \Gamma(x_i + 1) - \sum_{i=1}^{K} \log \Gamma(1/r_i^s) - \log \Gamma(n + M(s)).\end{aligned} \tag{4}$$

ただし，$M(s) = \sum_{i=1}^{K} 1/r_i^s$ である．式 (4) を最大にする s を $\hat{s}$ とする．すなわち，

$$\hat{s} = \arg\max_{s} \log m(\boldsymbol{x}; \boldsymbol{r}, s).$$

以上から，式 (3) の s に $\hat{s}$ を代入した

$$\hat{p}_i(\boldsymbol{x}; \boldsymbol{r}, \hat{s}) = \frac{\exp\{\psi(x_i + 1/r_i^{\hat{s}})\}}{\sum_{h=1}^{K} \exp\{\psi(x_h + 1/r_h^{\hat{s}})\}} \quad (i = 1, \ldots, K) \tag{5}$$

を提案推定量とする．

2.2 リスク関数最小化

$\boldsymbol{x} = (x_1, \ldots, x_K)$ が $Multi(n, \boldsymbol{p})$ からの標本であり，母数 $\boldsymbol{p}$ に対する事前分布として母数 $\boldsymbol{r} = (1/r_1^s, \ldots, 1/r_K^s)$ のディリクレ分布を仮定する．$\boldsymbol{p}$ のベイズ推定量は次のように与えられる．

$$\tilde{\boldsymbol{p}}(\boldsymbol{x}; \boldsymbol{r}, s) = \frac{n}{n + M(s)} \frac{\boldsymbol{X}}{n} + \frac{1}{n + M(s)}\, \boldsymbol{r}.$$

ただし，$\boldsymbol{X}$ は $Multi(n, \boldsymbol{p})$ に従う確率変数である．Fienberg and Holland (1973) は，pseudo Bayes 推定量を次のように提案した：

$$\hat{p}_i(\boldsymbol{x}; \boldsymbol{\lambda}, M^*) = \left(\frac{n}{n+M^*}\right)\frac{x_i}{n} + \left(\frac{M^*}{n+M^*}\right)\lambda_i \quad (i = 1, \ldots, K).$$

ただし, $\boldsymbol{\lambda} = (1/K)\mathbf{1}$, $M^* = (n^2 - \sum x_k^2)/(\sum x_k^2 - n^2/K)$ である．

Fienberg and Holland (1973) で提案された pseudo Bayes 推定量の導出と類似の方法により，s を決定する．s と $\boldsymbol{r}$ を定数，損失関数を平均二乗損失として，$\tilde{\boldsymbol{p}}$ のリスク関数を求めると

$$R(\tilde{\boldsymbol{p}}, \boldsymbol{p}; s) = \left(\frac{n}{n+M(s)}\right)^2 \left(1 - \sum_{i=1}^{K} p_i^2\right) + n\left(\frac{1}{n+M(s)}\right)^2 \sum_{i=1}^{K}\left(\frac{1}{r_i^s} - M(s)p_i\right)^2.$$

となる．推定された $R(\tilde{\boldsymbol{p}}, \boldsymbol{p}; s)$ として，$\boldsymbol{p}$ に MLE である $\hat{\boldsymbol{p}}$ を代入した $R(\tilde{\boldsymbol{p}}, \hat{\boldsymbol{p}}; s)$ を考える．$(r_1, \ldots, r_K)$ を固定して，$R(\tilde{\boldsymbol{p}}, \hat{\boldsymbol{p}}; s)$ を最小にする s を $\tilde{s}$ とする．すなわち，

$$\tilde{s} = \arg\min_{s} R(\tilde{\boldsymbol{p}}, \hat{\boldsymbol{p}}; s).$$

以上から，式 (3) の s に $\tilde{s}$ を代入した

$$\hat{p}_i(\boldsymbol{x}; \boldsymbol{r}, \tilde{s}) = \frac{\exp\{\psi(x_i + 1/r_i^{\tilde{s}})\}}{\sum_{h=1}^{K} \exp\{\psi(x_h + 1/r_h^{\tilde{s}})\}} \quad (i = 1, \ldots, K) \tag{6}$$

を提案推定量とする．

3　シミュレーション

K $(K = 50, 300)$ と b $(b = 0, 1)$ を用いて，真の確率分布を

$$p_i = \frac{\exp\{bi\}}{\sum_{h=1}^{K} \exp\{bh\}} \quad (i = 1, \ldots, K)$$

によって与える．総観測度数を $n = 20, 60, 100$ とする．多項分布 $Multi(n, \boldsymbol{p})$ からの標本 $\boldsymbol{x}$ に対して，次の損失関数の値を計算する．

$$\text{KLD}: \ L(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \sum_{i=1}^{K} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i}\right).$$

これを 100 回繰り返した結果をまとめたものが表 1 である．

表 1: KLD 損失の平均値

b	n	$K=50$						$K=300$					
		$\hat{\boldsymbol{p}}$	$\hat{\boldsymbol{p}}_Z^{(1)}$	$\hat{\boldsymbol{p}}_Z^{(1/2)}$	$\hat{\boldsymbol{p}}_Z^{(\hat{s})}$	$\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$	$\hat{\boldsymbol{p}}_S$	$\hat{\boldsymbol{p}}$	$\hat{\boldsymbol{p}}_Z^{(1)}$	$\hat{\boldsymbol{p}}_Z^{(1/2)}$	$\hat{\boldsymbol{p}}_Z^{(\hat{s})}$	$\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$	$\hat{\boldsymbol{p}}_S$
0	20	1.153	1.234	1.077	0.940	1.141	**0.086**	2.752	2.760	2.258	2.778	2.791	**0.023**
	60	0.477	0.598	0.543	0.220	0.454	**0.116**	1.740	1.827	1.761	1.704	1.786	**0.057**
	100	0.277	0.371	0.338	0.151	0.280	**0.107**	1.309	1.428	1.371	0.932	1.308	**0.079**
1	20	0.127	0.566	4.430	**0.118**	0.121	1.028	0.126	19.438	91.267	**0.119**	0.137	6.553
	60	0.054	0.094	1.302	**0.050**	**0.050**	0.208	0.050	2.499	44.126	**0.049**	0.081	1.150
	100	0.035	0.470	0.694	**0.033**	**0.033**	0.106	0.033	0.949	27.671	**0.032**	0.064	0.498

表 1 における各記号を説明する．MLE を $\hat{\boldsymbol{p}}$ とし，式 (3) の推定量を $\hat{\boldsymbol{p}}_Z^{(s)}$ と記す．つまり，$\hat{\boldsymbol{p}}_Z^{(1)}$ と $\hat{\boldsymbol{p}}_Z^{(1/2)}$ は Ogura and Yanagimoto (2022) で提案された推定量である．また，式 (5) は $\hat{\boldsymbol{p}}_Z^{(\hat{s})}$ であり，式 (6) は $\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$ である．Tuyl (2019) は，度数 0 のセルが多い場合に適切とされる Spike and Slab prior をモデルセレクションアプローチを用いて構築した．その推定量を $\hat{\boldsymbol{p}}_S$ と記す．

$b=0$ のとき，真の確率分布 $\boldsymbol{p}$ は一様である．そのため，K と n がどの場合においても $\hat{\boldsymbol{p}}_S$ のパフォーマンスが良いことを確認できる．その理由は，Spike and Slab prior が事前分布として一様性を有しているためと考えられる．一方，$b=1$ のとき，真の確率分布 $\boldsymbol{p}$ は $p_1 < p_2 < \cdots < p_K$ となり，K が大きくなるにつれて第 K セル確率 p_K が他のセル確率に比べて大きくなる．つまり，本稿で検討している度数 0 のセルが多く生じる設定となっている．このとき，提案推定量である $\hat{\boldsymbol{p}}_Z^{(\hat{s})}$ と $\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$ のパフォーマンスが良いことを確認できる．この理由は，Zeta 事前分布が順位情報を用いるため，度数 0 のセルにおいては事前分布のパラメータへの縮小を小さくし，その他のセルについては度数の順位に応じてパラメータへの縮小がコントロールされているためだと考えられる．

統計解析ソフトウェア R を用いた計算時間の比較について述べる．パッケージ tictoc の関数 tic と関数 toc を用いて，提案推定量と Tuyl (2019) の推定量の計算時間を比較する．カテゴリ数 $K=10,50,100,500$，総観測度数 $n=10,100$ に対して，各推定量の計算時間を計測した結果をまとめたものが表 2 である．

表 2: Tuyl の推定量との計算時間の比較（単位：秒）

n	$K=10$			$K=50$			$K=100$			$K=500$		
	$\hat{\boldsymbol{p}}_Z^{(\hat{s})}$	$\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$	$\hat{\boldsymbol{p}}_S$	$\hat{\boldsymbol{p}}_Z^{(\hat{s})}$	$\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$	$\hat{\boldsymbol{p}}_S$	$\hat{\boldsymbol{p}}_Z^{(\hat{s})}$	$\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$	$\hat{\boldsymbol{p}}_S$	$\hat{\boldsymbol{p}}_Z^{(\hat{s})}$	$\hat{\boldsymbol{p}}_Z^{(\tilde{s})}$	$\hat{\boldsymbol{p}}_S$
10	0.03	0.05	0.03	0.03	0.04	0.51	0.03	0.03	3.51	0.03	0.03	386.47
100	0.03	0.04	0.03	0.02	0.03	0.43	0.03	0.04	3.66	0.03	0.04	412.08

表 2 から，総観測度数 n を固定し K を大きくするとき，提案推定量の計算時間にはほとんど変化が見られない．一方で，Tuyl の推定量は K が大きくなるにつれて，計算時間が長くなることを確認できる．また，K を固定して n を大きくしても，各推定量の計算時間には大きな差がないことを確認できる．

4　おわりに

本稿では，総観測度数 n が小さくカテゴリ数 K が大きい場合（$n < K$）に有用な推定量を提案した．この推定量は，得られたデータの順位情報を事前分布に用いており，客観的というよりもむしろ積極的に事前情報を取り入れたものである．そのため，3 節で述べたように従来法よりも提案法の方が良いパフォーマンスを示している．また，計算時間が短いことも利点の一つと考えられる．

参考文献

Agresti, A. and Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications* **14**, 297–330.

Alvares, D., Armero, C. and Forte, A. (2018). What does objective mean in a dirichlet-multinomial process? *International Statistical Review* **86**, 106–118.

Berger, J. O., Bernardo, J. M. and Sun, D. (2015). Overall objective priors. *Bayesian Analysis* **10**, 189–221.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, Massachusetts.

Fienberg, S. E. and Holland, P. W. (1972). On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis* **2**, 127–134.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association* **68**, 683–691.

Good, I. J. (1967). A bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **29**, 399–431.

Ogura, T. and Yanagimoto, T. (2020). Bayesian estimator of multiple poisson means assuming two different priors. *Communications in Statistics - Simulation and Computation*, 1–9.

Ogura, T. and Yanagimoto, T. (2022). Estimation of highly heterogeneous multinomial probabilities observed at the beginning of covid-19. *Biostatistics and Epidemiology*, to appear.

Tuyl, F. (2017). A note on priors for the multinomial model. *The American Statistician* **71**, 298–301.

Tuyl, F. (2019). A method to handle zero counts in the multinomial model. *The American Statistician* **73**, 151–158.

Yanagimoto, T. and Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference* **139**, 3064–3075.