

正方分割表における対称性のモデリング

東京理科大学・理工学部情報科学科 田畑 耕治

Kouji Tahata

Department of Information Sciences, Faculty of Science and Technology

Tokyo University of Science

Noda City, Chiba, 278-8510

JAPAN

E-mail address: kouji_tahata@rs.tus.ac.jp

1. はじめに

行と列に順序のある同じ分類からなる正方分割表においては、多くの観測値が分割表の主対角付近に集中する傾向がある。そのため、行変数と列変数の独立性が成り立たないことが多く、対称性に関するモデルが提案されている。例えば、Bowker (1948) の対称 (S) モデル、Caussinus (1965) の準対称 (QS) モデル、Stuart (1955) の周辺同等 (MH) モデルなどがある。また、Kateri and Papaioannou (1997) は f -divergence を用いて準対称性を一般化し (QS[f] モデルと記す)、Kateri and Agresti (2007) は順序カテゴリ分割表に対する準対称性 (OQS[f] モデルと記す) を提案した。本講演では、OQS[f] モデルと QS[f] モデルのギャップを補う f -divergence に基づく非対称モデルを提案する。このモデルは、Ireland, Ku and Kullback (1969) の情報理論的アプローチと関連が深く、ある条件を与えた下で f -divergence に関して S モデルにもっとも近いモデルであることが示される。また、提案モデルは、多くの非対称性のモデルを特別な場合を含む。

Caussinus (1965) は“S モデルが成り立つための必要十分条件は QS モデルと MH モデルの両方が成り立つことである”を証明した。S モデルがデータに適合しない場合に、この結果はその原因を探ることに有用である。本講演では、 f -divergence に基づく非対称モデルを用いた S モデルの分解も与える。さらに、適合度検定統計量の関係についても考察する。

2. f -divergence に基づく非対称モデル

順序カテゴリ $r \times r$ 正方分割表において、 (i, j) セル確率を π_{ij} とし ($i = 1, \dots, r; j = 1, \dots, r$)、 $\pi_{ij}^S = (\pi_{ij} + \pi_{ji})/2$ とする。また、周辺分布を $\pi_{i+} = \sum_{j=1}^r \pi_{ij}$ 、 $\pi_{+i} = \sum_{j=1}^r \pi_{ji}$ ($i = 1, \dots, r$) とし、 $\{u_i\}$ を $u_1 < u_2 < \dots < u_r$ を満たす既知のスコアとする。

確率ベクトル (π_{ij}^S) は、制約条件

$$\pi_{ij} + \pi_{ji} = t_{ij} = t_{ji} \quad (i = 1, \dots, r; j = 1, \dots, r), \quad (1)$$

の下で対称性を満たし、固定されることに注意する。 $I^C(\pi : \pi^S)$ を (π_{ij}) と (π_{ij}^S) の f -divergence とする。すなわち

$$I^C(\pi : \pi^S) = \sum_{i=1}^r \sum_{j=1}^r \pi_{ij}^S f\left(\frac{\pi_{ij}}{\pi_{ij}^S}\right)$$

である。ここに、 f は $(0, +\infty)$ で二階微分可能な狭義凸関数で $f(1) = 0$ を満たし、 $f(0) = \lim_{t \rightarrow 0} f(t)$ 、 $0 \cdot f(0/0) = 0$ 、 $0 \cdot f(a/0) = a \lim_{t \rightarrow \infty} [f(t)/t]$ とする。制約条件 (1)

と制約条件

$$\sum_{i=1}^r u_i^h \pi_{i+} = \mu^h \quad (h = 1, \dots, k), \quad (2)$$

の下で, $I^C(\pi : \pi^S)$ を最小化することを考える. このとき,

$$I^C(\pi : \pi^S) + \sum_{h=1}^k \lambda_h \left(\sum_{i=1}^r u_i^h \pi_{i+} - \mu^h \right) + \sum_{i=1}^r \sum_{j=1}^r \phi_{ij} (\pi_{ij} + \pi_{ji} - t_{ij})$$

を π_{ij} で偏微分して 0 と等しいとした方程式は

$$f' \left(\frac{\pi_{ij}}{\pi_{ij}^S} \right) + \sum_{h=1}^k \lambda_h u_i^h + \phi_{ij} + \phi_{ji} = 0 \quad (3)$$

である. f' を F とし, (3) の解を π_{ij}^* とすると,

$$F \left(\frac{\pi_{ij}^*}{\pi_{ij}^S} \right) = - \sum_{h=1}^k \lambda_h u_i^h - (\phi_{ij} + \phi_{ji})$$

である. ただし, π_{ij}^* は制約条件 (1) と (2) を満たすことに注意する. f が狭義凸関数であることから, F の逆関数が存在するので,

$$\frac{\pi_{ij}^*}{\pi_{ij}^S} = F^{-1} \left(- \sum_{h=1}^k \lambda_h u_i^h - (\phi_{ij} + \phi_{ji}) \right)$$

である. $-\lambda_h$ と $-(\phi_{ij} + \phi_{ji})$ をそれぞれ α_h と γ_{ij} とすると,

$$\pi_{ij}^* = \pi_{ij}^S F^{-1} \left(\sum_{h=1}^k \alpha_h u_i^h + \gamma_{ij} \right),$$

ただし, $\gamma_{ij} = \gamma_{ji}$ である. 以上から, 制約条件 (1) と (2) の下で, $I^C(\pi : \pi^S)$ は π_{ij}^* において最小値に達する.

f -divergence に基づく非対称 ($\text{AS}_k[f]$) モデルを次のように定義する: 任意に与えられた k ($k = 1, \dots, r-1$) に対して,

$$\pi_{ij} = \pi_{ij}^S F^{-1} \left(\sum_{h=1}^k u_i^h \alpha_h + \gamma_{ij} \right) \quad (i = 1, \dots, r; j = 1, \dots, r),$$

ただし, $\gamma_{ij} = \gamma_{ji}$, $\pi_{ij}^S = (\pi_{ij} + \pi_{ji})/2$, $F(t) = f'(t)$ である. 先の議論から, このモデルは $\{\pi_{ij} + \pi_{ji}\}$ と $h = 1, \dots, k$ について $\{\sum_i u_i^h \pi_{i+}\}$ (または $\{\sum_i u_i^h \pi_{+i}\}$) を与えた条件の下で f -divergence に関して S モデルにもっとも近いモデルである. 特に $\alpha_1 = \dots = \alpha_k = 0$ のとき, S モデルである.

$\text{AS}_{r-1}[f]$ モデルは, Kateri and Papaioannou (1997) の $\text{QS}[f]$ モデルであり, $\text{AS}_1[f]$ モデルは, Kateri and Agresti (2007) の $\text{OQS}[f]$ モデルである. このことから, $\text{AS}_k[f]$ モデル ($k = 2, \dots, r-1$) は, f -divergence に基づく $\text{OQS}[f]$ モデルの拡張であり, $\text{OQS}[f]$ モデルと $\text{QS}[f]$ モデルのギャップを埋める.

$f(t) = t \log(t)$ ($t > 0$) とした $AS_k[f]$ モデルは

$$\pi_{ij} = \pi_{ij}^S \exp \left[\sum_{h=1}^k u_i^h \alpha_h + \gamma_{ij} - 1 \right] \quad (i = 1, \dots, r; j = 1, \dots, r), \quad (4)$$

ただし, $\gamma_{ij} = \gamma_{ji}$ である. $f(t) = t \log(t)$ とした f -divergence は Kullback-Leibler distance である. 条件付き確率を $\pi_{ij}^c = \pi_{ij}/(\pi_{ij} + \pi_{ji})$ とする. 式 (4) の下で

$$\frac{\pi_{ij}^c}{\pi_{ji}^c} = \prod_{h=1}^k \beta_h^{u_i^h - u_j^h} \quad (i < j),$$

ただし, $\beta_h = \exp[\alpha_h]$ である. したがって, このモデルはセル確率 (条件付き確率) の比に関する非対称構造を示す. Tahata and Tomizawa (2011) は $u_i = i$ ($i = 1, \dots, r$) としたモデルを提案し, 多元分割表への拡張を行った. また, $f(t) = (1-t)^2$ とした $AS_k[f]$ モデルは

$$\pi_{ij} = \pi_{ij}^S \left(\frac{\sum_{h=1}^k u_i^h \alpha_h + \gamma_{ij}}{2} + 1 \right) \quad (i = 1, \dots, r; j = 1, \dots, r), \quad (5)$$

ただし, $\gamma_{ij} = \gamma_{ji}$ である. $f(t) = (1-t)^2$ とした f -divergence は Pearsonian distance である. 式 (5) の下で

$$\pi_{ij}^c - \pi_{ji}^c = \sum_{h=1}^k (u_i^h - u_j^h) \beta_h \quad (i < j),$$

ただし, $\beta_h = \alpha_h/4$, $\sum_h (u_j^h - u_i^h) \beta_h < 1$ である. したがって, このモデルは条件付き確率の差に関する非対称構造を示す. このように divergence に依存して, 条件付き確率に関する非対称構造が異なることに注意する.

3. 対称性の分解

行変数を X , 列変数を Y とする. 既知のスコア $\{u_i\}$ は $u_1 < u_2 < \dots < u_r$ を満たすとする. ここで $g(i) = u_i$ ($i = 1, \dots, r$) とするとき, 周辺 k 次積率一致 (ME_k) モデルを次のように考える: 任意に与えられた k ($k = 1, \dots, r-1$) に対して,

$$E(g(X)^h) = E(g(Y)^h) \quad (h = 1, \dots, k),$$

ただし,

$$E(g(X)^h) = \sum_{s=1}^r u_s^h \pi_{s+}, \quad E(g(Y)^h) = \sum_{s=1}^r u_s^h \pi_{+s}$$

である. このモデルは, $h = 1, \dots, k$ に対して, X と Y のスコアに関する周辺 h 次積率が一致する構造を示す. Tahata and Tomizawa (2008) は ME_{r-1} モデルが MH モデルと同値であることを示した. つまり, ME_{r-1} モデルは行変数と列変数の周辺分布に関する同等性を示す. このとき, 次の定理を得る.

定理 1. 任意に与えられた k ($k = 1, \dots, r-1$) に対して, S モデルが成り立つための必要十分条件は $AS_k[f]$ モデルと ME_k モデルの両方が成り立つことである.

定理 1 は Caussinus (1965), Kateri and Papaioannou (1997), Tahata and Tomizawa (2011) などの結果を含む.

モデル M の適合度検定は, たとえば, 尤度比カイ二乗統計量

$$G^2(M) = 2 \sum_{i=1}^r \sum_{j=1}^r n_{ij} \log \left(\frac{n_{ij}}{\hat{m}_{ij}} \right)$$

を用いる. ここに, n_{ij} は (i, j) セル観測度数, $\hat{m}_{ij}$ は (i, j) セル期待度数 m_{ij} のモデル M の下での最尤推定量である. このとき, 次の定理を得る.

定理 2. 任意に与えられた k ($k = 1, \dots, r-1$) に対して, S モデルの下で $G^2(S)$ は $G^2(AS_k[f])$ と $G^2(ME_k)$ の和と漸近的に同等である.

定理 1 と定理 2 の詳細については, Tahata (2019) を参照されたい. また, 定理 1 と定理 2 の $k = 1$ とした結果は, Saigusa, Tahata and Tomizawa (2015) を参照されたい.

定理 1 より $AS_k[f]$ モデルが成り立つとき, ME_k モデルと S モデルは同値である. したがって, $AS_k[f]$ モデルが成り立つという条件の下での ME_k モデルの適合度検定は,

$$G^2(ME_k|AS_k[f]) = G^2(S|AS_k[f]) = G^2(S) - G^2(AS_k[f])$$

与えられる. 一方, 定理 2 は S モデルの下で $G^2(S)$ が $G^2(AS_k[f])$ と $G^2(ME_k)$ の和と漸近的に同等であることを主張する. 以上から, S モデルの下で条件付き検定 $G^2(ME_k|AS_k[f])$ は $G^2(ME_k)$ と漸近的に同等である.

参考文献

- [1] Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**, 572-574.
- [2] Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, Série 4, **29**, 77-182.
- [3] Ireland, C. T., Ku, H. H. and Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *Journal of the American Statistical Association*, **64**, 1323-1341.
- [4] Kateri, M. and Agresti, A. (2007). A class of ordinal quasi-symmetry models for square contingency tables. *Statistics and Probability Letters*, **77**, 598-603.
- [5] Kateri, M. and Papaioannou, T. (1997). Asymmetry models for contingency tables. *Journal of the American Statistical Association*, **92**, 1124-1131.
- [6] Saigusa, Y., Tahata, K. and Tomizawa, S. (2015). Orthogonal decomposition of symmetry model using the ordinal quasi-symmetry model based on f -divergence for square contingency tables. *Statistics and Probability Letters*, **101**, 33-37.
- [7] Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412-416.
- [8] Tahata, K. (2019). Separation of symmetry for square tables with ordinal categorical data. *Japanese Journal of Statistics and Data Science*, <https://doi.org/10.1007/s42081-019-00066-8>.
- [9] Tahata, K. and Tomizawa, S. (2008). Generalized marginal homogeneity model and its relation to marginal equimoments for square contingency tables with ordered categories. *Advances in Data Analysis and Classification*, **2**, 295-311.
- [10] Tahata, K. and Tomizawa, S. (2011). Generalized linear asymmetry model and decomposition of symmetry for multiway contingency tables. *Biometrics and Biostatistics*, **2**, 1-6.