Extension of Marginal Complementary Log–Log Model and Separations of Marginal Homogeneity for Ordinal Categorical Data

Kengo Fujisawa¹ · Kouu Mitomi¹ · Kouji Tahata²

Accepted: 29 April 2021 / Published online: 21 May 2021
© The Author(s) 2021

Abstract

The present paper considers a model that indicates the structure of inhomogeneity for marginal distributions for ordinal categorical data. The model is based on the complementary log–log transformation of marginal cumulative probability. A theorem that the marginal homogeneity model holds if and only if the proposed model and the marginal mean and variance equality model holds is given. Also, a model based on the conditional marginal distribution is considered under certain condition.

Keywords Decomposition · Moment · Square contingency table

1 Introduction

Two-way contingency tables with the same row and column categories occur frequently in, for example, panel studies, occupational studies and longitudinal studies. The data in Table 1 illustrate occupational status for Japanese father–son pairs, which is taken directly from Tominaga [19, p. 131]. Each observation pairs father’s occupation with son’s occupation. These variables have four levels with ordinal categories. The smaller category number means higher status. For the data in Table 1, many observations concentrate on main-diagonal cells. Therefore, the independence

This work was supported by JSPS KAKENHI Grant Number 20K03756.

✉ Kouji Tahata
kouji_tahata@rs.tus.ac.jp

Kengo Fujisawa
kfujisawa09@gmail.com

¹ Department of Information Sciences, Graduate School of Science and Technology, Tokyo University of Science, Noda City, Chiba 278-8510, Japan

² Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Noda City, Chiba 278-8510, Japan

Table 1 Occupational status for Japanese father–son pairs; from Tominaga [19, p. 131]

Son's status	Father's status				Total
	(1)	(2)	(3)	(4)	
(1)	80 (80.00)	44 (47.74)	26 (26.03)	69 (71.22)	219 (224.99)
(2)	72 (67.92)	155 (155.00)	73 (68.67)	156 (151.02)	456 (442.61)
(3)	37 (36.95)	61 (65.81)	218 (218.00)	166 (170.41)	482 (491.17)
(4)	19 (18.21)	31 (32.00)	45 (43.01)	614 (614.00)	709 (707.22)
Total	208 (203.09)	291 (300.55)	362 (355.71)	1005 (1006.65)	1866 (1866.00)

The parenthesized values are the MLEs of expected frequencies under the CEMCLL model

Status (1) is Professional and Managers, (2) Clerical and Sales, (3) Skilled manual and semiskilled manual, (4) Unskilled manual and Farmers

between two variables does not hold in the contingency table which is constructed from matched-pairs data. Then, we are interested in considering various symmetry (or asymmetry) instead of independence. This paper treats methods for analyzing square tables.

For a $r \times r$ square contingency table with ordered categories, let X and Y be the row and column random variables, respectively. Also, let p_{ij} denotes the probability that an observation will fall in the i th row and j th column of the table for $i = 1, \dots, r; j = 1, \dots, r$. We assume $p_{ij} > 0$ for all i and j . Bowker [4] considered the symmetry model defined by $p_{ij} = p_{ji}$ for $i < j$. The marginal homogeneity (MH) model is defined by

$$p_{i\cdot} = p_{\cdot i} \quad (i = 1, \dots, r),$$

where $p_{i\cdot} = \sum_{t=1}^r p_{it}$ and $p_{\cdot i} = \sum_{s=1}^r p_{si}$ (Stuart [15]). This model indicates the structure that satisfies the identity of marginal distributions of row and column. Also, Caussinus [5] showed the theorem that the symmetry model holds if and only if both the quasi-symmetry model, which indicates a symmetric structure for odds ratios, and the MH model holds. We shall refer this theorem as separation of symmetry. For the details of method for joint distribution, please see Bishop et al. [3] and Agresti [2].

Let F_i^X and F_i^Y denote the marginal cumulative probabilities of X and Y , respectively, namely $F_i^X = \sum_{s=1}^i p_{s\cdot}$ and $F_i^Y = \sum_{t=1}^i p_{\cdot t}$ for $i = 1, \dots, r-1$. The MH model may also be expressed as

$$F_i^X = F_i^Y \quad (i = 1, \dots, r-1).$$

When the MH model fits poorly for a real data set, we are interested in applying some extension of the MH model. Indeed, the MH model fits the data in Table 1 poorly (see Sect. 6.1). McCullagh [11] considered the marginal cumulative logistic (ML) model defined as follows:

$$L_i^X = \Delta + L_i^Y \quad (i = 1, \dots, r-1),$$

where L_i^X and L_i^Y denote the logit transformations of F_i^X and F_i^Y , respectively. That is,

$$L_i^X = \log \frac{F_i^X}{1 - F_i^X}, \quad L_i^Y = \log \frac{F_i^Y}{1 - F_i^Y}.$$

Also, see Agresti [1, p. 205]. The ML model with $\Delta = 0$ is the MH model, namely the ML model is the extension of the MH model. Moreover, Miyamoto et al. [12] proposed the conditional ML (CML) model, and Saigusa et al. [13] proposed the marginal complementary log–log (MCLL) model. The ML (CML) model states that one (conditional) marginal distribution is a location shift of the other (conditional) marginal distribution on a logistic scale. The MCLL model states that one marginal distribution is a location shift of the other marginal distribution in terms of a function $1 - \exp(-\exp(x))$. In this paper, we consider a model that is more relaxed than these models. The extensions of the ML and CML models exist (Kurakami et al. [9]). Therefore, an extension of the MCLL model is proposed in Sect. 2.

Caussinus [5] gave the separation of symmetry. The separation may be useful to see a reason for the poor fit of the symmetry model when the symmetry model fits the data poorly. Thus, we are interested in considering a necessary and sufficient condition of the MH model. We give the separation of marginal homogeneity in Sect. 3.

This paper is organized as follows. Section 2 proposes an extension of the MCLL model. Section 3 shows two theorems. Section 4 extends the proposed model into multi-way contingency table. Section 5 discusses the goodness-of-fit test. Section 6 gives examples. Section 7 concludes this paper.

2 Extension of MCLL

Let C_i^X and C_i^Y denote the marginal cumulative complementary log–log transforms of F_i^X and F_i^Y , respectively. That is,

$$C_i^X = \log(-\log(1 - F_i^X)), \quad C_i^Y = \log(-\log(1 - F_i^Y)) \quad (i = 1, \dots, r-1).$$

The extended MCLL (EMCLL) model is defined by

$$C_i^X = C_i^Y + i \log(\Delta_1) + \log(\Delta_2) \quad (i = 1, \dots, r-1). \quad (1)$$

The EMCLL model is an extension of the MCLL model. We would like to note that (i) the EMCLL model with $\Delta_1 = \Delta_2 = 1$ is reduced to the MH model and (ii) the EMCLL model with $\Delta_1 = 1$ is reduced to the MCLL model. This model indicates

that the probability that X is $i + 1$ or above, is equal to the probability that Y is $i + 1$ or above to the power of $\Delta_1^i \Delta_2$, for $i = 1, \dots, r - 1$, namely,

$$(1 - F_i^X) = (1 - F_i^Y)^{\Delta_1^i \Delta_2} \quad (i = 1, \dots, r - 1).$$

Also, in Eq. (1), $i \log(\Delta_1) + \log(\Delta_2)$ is the difference between the two random variables on complementary log–log scale, namely (i) the MCLL model (i.e., $\Delta_1 = 1$) indicates that one marginal distribution is a location shift of the other marginal distribution, and (ii) the EMCLL model indicates that the difference between two marginal distributions depends on the value of category i .

The Gumbel (minimum) distribution function is

$$G(x) = 1 - \exp \left(- \exp \left(\frac{x - \alpha}{\beta} \right) \right), \quad -\infty < x < \infty$$

where α is the location parameter and $\beta (> 0)$ is the scale parameter. Let $G^{\tilde{X}}(x)$ be the Gumbel distribution function with parameter (α_1, β_1) and $G^{\tilde{Y}}(x)$ be the Gumbel distribution function with parameter (α_2, β_2) . Then, the difference between two complementary log–log transforms of $G^{\tilde{X}}(x)$ and $G^{\tilde{Y}}(x)$ is expressed as

$$\log \left(-\log \left(1 - G^{\tilde{X}}(x) \right) \right) - \log \left(-\log \left(1 - G^{\tilde{Y}}(x) \right) \right) = \left(\frac{1}{\beta_1} - \frac{1}{\beta_2} \right) x + \frac{\alpha_2}{\beta_2} - \frac{\alpha_1}{\beta_1}.$$

This structure is similar to the EMCLL model. If we set $\beta_1 = \beta_2 = 1$ without loss of generality, then

$$\log \left(-\log \left(1 - G^{\tilde{X}}(x) \right) \right) - \log \left(-\log \left(1 - G^{\tilde{Y}}(x) \right) \right) = \alpha_2 - \alpha_1.$$

This is similar to the Eq. (1) with $\Delta_1 = 1$, that is, the location shift model. Also, if we set $\alpha_1 = \alpha_2 = 0$ without loss of generality, then

$$\log \left(-\log \left(1 - G^{\tilde{X}}(x) \right) \right) - \log \left(-\log \left(1 - G^{\tilde{Y}}(x) \right) \right) = \left(\frac{1}{\beta_1} - \frac{1}{\beta_2} \right) x.$$

The difference depends only on the scale parameter. This is similar to the Eq. (1) with $\Delta_2 = 1$. For ordinal categorical data, if it is reasonable to assume an underlying Gumbel distribution, then the proposed model may be appropriate for the square contingency tables.

Miyamoto et al. [12], Tahata et al. [18], and Shinoda et al. [14] considered the models depending on the conditional marginal cumulative probability. These models indicate the structure of marginal inhomogeneity on the condition that an observation will fall in one of off-diagonal cells of the table.

Let $F_i^{X(c)}$ and $F_i^{Y(c)}$ denote the conditional marginal cumulative probabilities of X and Y , respectively, given that $X \neq Y$. That is, for $i = 1, \dots, r - 1$,

$$F_i^{X(c)} = \Pr(X \leq i | X \neq Y) = \sum_{k=1}^i p_{k\cdot}^c,$$

$$F_i^{Y(c)} = \Pr(Y \leq i | X \neq Y) = \sum_{k=1}^i p_{\cdot k}^c,$$

where

$$p_{i\cdot}^c = \frac{p_{i\cdot} - p_{ii}}{\delta} = \Pr(X = i | X \neq Y),$$

$$p_{\cdot i}^c = \frac{p_{\cdot i} - p_{ii}}{\delta} = \Pr(Y = i | X \neq Y),$$

with $\delta = \sum_{s \neq t} p_{st} = \Pr(X \neq Y)$.

We shall consider the conditional EMCLL (CEMCLL) model which is defined by

$$C_i^{X(c)} = C_i^{Y(c)} + i \log(\Delta_1^*) + \log(\Delta_2^*) \quad (i = 1, \dots, r-1),$$

where

$$C_i^{X(c)} = \log \left(-\log \left(1 - F_i^{X(c)} \right) \right), \quad C_i^{Y(c)} = \log \left(-\log \left(1 - F_i^{Y(c)} \right) \right).$$

Similarly, (i) the CEMCLL model with $\Delta_1^* = 1$ reduces to the conditional MCLL (CMCLL) model proposed by Shinoda et al. [14]. The CMCLL model indicates that one conditional marginal distribution is a location shift of the other conditional marginal distribution. The CEMCLL model indicates that the difference between two conditional marginal distributions depends on the value of category i . In a similar manner to the EMCLL model, the CEMCLL model may be appropriate if it is reasonable to assume underlying Gumbel distribution for the conditional marginal distribution.

Both the EMCLL and CEMCLL models show the inhomogeneity of two marginal distributions. The EMCLL model shows the difference between the complementary log-log transformation of the row and column marginal distribution function is linear with respect to category i . On the other hand, the CEMCLL model shows that of row and column conditional marginal distribution function is linear with respect to category i .

3 A Necessary and Sufficient Condition for the MH

The MH model does not fit for a given dataset because the MH model is very restrictive. Marginal inhomogeneity models such as the ML and MCLL models were proposed. We are interested in detecting the probabilistic structure for the poor fit of the MH model. When the MH model can be separated into two or more models, analyzing these models may be helpful to elucidate the reason for the poor fit of the MH model.

Consider a model defined by

$$E(X) = E(Y),$$

where $E(X) = \sum_i i p_{i\cdot}$ and $E(Y) = \sum_i i p_{\cdot i}$. We shall refer to the model as the mean equality (ME) model. Also, we consider a model defined by

$$E(X) = E(Y) \quad \text{and} \quad E(X^2) = E(Y^2), \quad (2)$$

where $E(X^2) = \sum_i i^2 p_{i\cdot}$ and $E(Y^2) = \sum_i i^2 p_{\cdot i}$. We would like to note that Eq. (2) is equivalent to

$$E(X) = E(Y) \quad \text{and} \quad V(X) = V(Y), \quad (3)$$

where $V(X) = E(X^2) - E(X)^2$ and $V(Y) = E(Y^2) - E(Y)^2$. Thus, we shall refer to the Eq. (3) as the mean and variance equality (MVE) model.

We give the following lemma.

Lemma 1 *The MVE model is equivalent to the following restrictions:*

$$\sum_{i=1}^{r-1} (F_i^X - F_i^Y) = 0 \quad \text{and} \quad \sum_{i=1}^{r-1} i(F_i^X - F_i^Y) = 0. \quad (4)$$

Proof We can see that

$$\sum_{i=1}^r i^t (p_{i\cdot} - p_{\cdot i}) = \sum_{i=1}^r i^t \{ (F_i^X - F_i^Y) - (F_{i-1}^X - F_{i-1}^Y) \} \quad (t = 1, 2),$$

because $p_{i\cdot} = F_i^X - F_{i-1}^X$ and $p_{\cdot i} = F_i^Y - F_{i-1}^Y$ with $F_0^X = F_0^Y = 0$ and $F_r^X = F_r^Y = 1$. Since the right-hand side in the above equation is expressed as

$$\sum_{i=1}^{r-1} i^t (F_i^X - F_i^Y) - \sum_{i=2}^r ((i-1) + 1)^t (F_{i-1}^X - F_{i-1}^Y),$$

we can obtain

$$\sum_{i=1}^r i^t (p_{i\cdot} - p_{\cdot i}) = \sum_{i=1}^{r-1} \{ i^t - (i+1)^t \} (F_i^X - F_i^Y) \quad (t = 1, 2). \quad (5)$$

If the MVE model holds, then

$$\sum_{i=1}^r i^t (p_{i\cdot} - p_{\cdot i}) = 0 \quad (t = 1, 2) \quad (6)$$

from Eq. (2). Thus, we can obtain Eq. (4) from Eq. (5).

Conversely, if Eq. (4) holds, then we can obtain Eq. (6) from Eq. (5), namely the MVE model holds. The proof is completed. $\square$

We obtain the following theorem.

Theorem 1 *The MH model holds if and only if both the EMCLL and MVE models hold.*

Proof If the MH model holds, then (i) the EMCLL model with $\Delta_1 = \Delta_2 = 1$ holds and (ii) the MVE model holds because

$$\sum_{i=1}^r t^i(p_{i\cdot} - p_{\cdot i}) = \sum_{i=1}^r t^i(p_{i\cdot} - p_{i\cdot}) = 0 \quad (t = 1, 2).$$

Next, we assume that both the EMCLL and MVE models hold, and then we show that the MH model holds. If the MVE model holds, then

$$\sum_{i=1}^{r-1} (F_i^X - F_i^Y) = 0 \quad \text{and} \quad \sum_{i=1}^{r-1} i(F_i^X - F_i^Y) = 0,$$

from Lemma 1. So, we have

$$\sum_{i=1}^{r-1} (i \log(\Delta_1) + \log(\Delta_2))(F_i^X - F_i^Y) = 0. \quad (7)$$

Also, if the EMCLL model holds, we have

$$C_i^X - C_i^Y = i \log(\Delta_1) + \log(\Delta_2) \quad (i = 1, \dots, r-1). \quad (8)$$

Equations (7) and (8) lead to

$$\sum_{i=1}^{r-1} (C_i^X - C_i^Y)(F_i^X - F_i^Y) = 0.$$

As complementary log–log function increases monotonically, we see

$$(C_i^X - C_i^Y)(F_i^X - F_i^Y) \geq 0 \quad (i = 1, \dots, r-1).$$

Thus, we can obtain

$$F_i^X = F_i^Y \quad (i = 1, \dots, r-1).$$

This is the MH model. The proof is completed. $\square$

The MVE model can be expressed as

$$\sum_{i=1}^{r-1} (F_i^{X(c)} - F_i^{Y(c)}) = 0 \quad \text{and} \quad \sum_{i=1}^{r-1} i(F_i^{X(c)} - F_i^{Y(c)}) = 0.$$

In a similar manner to proof of Theorem 1, this leads to the equation

$$\sum_{i=1}^{r-1} (C_i^{X(c)} - C_i^{Y(c)})(F_i^{X(c)} - F_i^{Y(c)}) = 0.$$

Thus, we can obtain

$$F_i^{X(c)} = F_i^{Y(c)} \quad (i = 1, \dots, r-1).$$

This is the MH model. Therefore, we can obtain the following theorem.

Theorem 2 *The MH model holds if and only if both the CEMCLL and MVE models hold.*

From Theorems 1 and 2, we can obtain the following corollary.

Corollary 1 *If the MVE model holds, then the EMCLL model holds if and only if the CEMCLL model holds.*

Saigusa et al. [13] and Shinoda et al. [14] gave the separation of marginal homogeneity. We would like to note that Theorems 1 and 2 include these results.

4 Extension for Multi-Way Table

Consider a multi-way r^T contingency table of same classification having ordered categories. Let X_t denotes the t -th random variable for $t = 1, \dots, T$, and let $\Pr(X_1 = i_1, \dots, X_T = i_T) = p_{i_1 \dots i_T}$ for $i_t = 1, \dots, r$. The MH[T] model is defined by

$$p_i^{(1)} = \dots = p_i^{(T)} \quad (i = 1, \dots, r),$$

where

$$p_i^{(t)} = \Pr(X_t = i).$$

Let $F_i^{X_t}$ denotes the marginal cumulative probability of X_t , namely, $F_i^{X_t} = \sum_{s=1}^i p_s^{(t)}$ for $i = 1, \dots, r-1$, and let $C_i^{X_t}$ denotes the complementary log-log transform of $F_i^{X_t}$. The EMCLL[T] model is defined by

$$C_i^{X_1} = C_i^{X_t} + i \log(\Delta_{1t}) + \log(\Delta_{2t}) \quad (i = 1, \dots, r-1; t = 2, \dots, T).$$

Also, the ME[T] model is defined by

$$E(X_1) = \dots = E(X_T),$$

and the MVE[T] model is defined by

$$E(X_1) = \dots = E(X_T) \quad \text{and} \quad V(X_1) = \dots = V(X_T).$$

We obtain the following theorem.

Theorem 3 *The MH[T] model holds if and only if both the EMCLL[T] and the MVE[T] models hold.*

Proof If the MH[T] model holds, then the EMCLL[T] and MVE[T] models hold.

Next, we assume that both the EMCLL[T] and MVE[T] models hold, and then we show that the MH[T] model holds. In a similar manner to the proof of Theorem 1, if both the EMCLL[T] and MVE[T] models hold, for any $t = 2, \dots, T$,

$$\sum_{i=1}^{r-1} (C_i^{X_1} - C_i^{X_t})(F_i^{X_1} - F_i^{X_t}) = 0.$$

We can obtain

$$F_i^{X_1} = F_i^{X_t} \quad (i = 1, \dots, r-1; t = 2, \dots, T),$$

because the complementary log–log function is a strictly monotonically increasing function. This is the MH[T] model. The proof is completed. $\square$

Let $F_i^{X_t(c)}$ denotes the conditional marginal cumulative probability of X_t , given that there is at least one set of unequal random variables. That is, for $i = 1, \dots, r-1; t = 1, \dots, T$,

$$F_i^{X_t(c)} = \Pr(X_t \leq i | (X_1, \dots, X_T) \neq (s, \dots, s), s = 1, \dots, r) = \sum_{k=1}^i p_k^{(t)c},$$

where

$$p_i^{(t)c} = \frac{p_i^{(t)} - p_{i \dots i}}{\delta_T} = \Pr(X_t = i | (X_1, \dots, X_T) \neq (s, \dots, s), s = 1, \dots, r),$$

with

$$\delta_T = 1 - \sum_{i=1}^r p_{i \dots i} = \Pr((X_1, \dots, X_T) \neq (s, \dots, s), s = 1, \dots, r).$$

The CEMCLL[T] model is defined by,

$$C_i^{X_t(c)} = C_i^{X_t} + i \log(\Delta_{1t}^*) + \log(\Delta_{2t}^*) \quad (i = 1, \dots, r-1; t = 2, \dots, T),$$

where

$$C_i^{X_t(c)} = \log \left(-\log \left(1 - F_i^{X_t(c)} \right) \right).$$

We would like to note that the MVE[T] model can be expressed as

$$\sum_{i=1}^{r-1} (F_i^{X_1(c)} - F_i^{X_t(c)}) = 0 \quad \text{and} \quad \sum_{i=1}^{r-1} i(F_i^{X_1(c)} - F_i^{X_t(c)}) = 0 \quad (t = 2, \dots, T).$$

Then, we can obtain the following theorem.

Theorem 4 *The MH[T] model holds if and only if both the CEMCLL[T] and MVE[T] models hold.*

From Theorems 3 and 4, we can obtain the following corollary.

Corollary 2 *If the MVE[T] model holds, then the EMCLL[T] model holds if and only if the CEMCLL[T] model holds.*

5 Goodness-of-Fit Test

Let $n_{i_1 \dots i_T}$ denote the observed frequency in the $(i_1, \dots, i_T)$ cell of the r^T table with $n = \sum \dots \sum n_{i_1 \dots i_T}$, and let $m_{i_1 \dots i_T}$ denote the corresponding expected frequency, that is, $m_{i_1 \dots i_T} = np_{i_1 \dots i_T}$. We assume that a multinomial distribution applies to the table. The maximum likelihood estimates (MLEs) of expected frequencies under each model can be obtained using the Newton–Raphson method in the log-likelihood equation. The likelihood ratio chi-squared statistic for testing the goodness-of-fit of model M is given by

$$G^2(M) = 2 \sum_{i_1=1}^r \dots \sum_{i_T=1}^r n_{i_1 \dots i_T} \log \left(\frac{n_{i_1 \dots i_T}}{\hat{m}_{i_1 \dots i_T}} \right),$$

where $\hat{m}_{i_1 \dots i_T}$ are the MLE of $m_{i_1 \dots i_T}$ under the model. If zero cells occur, it may be desirable to aggregate a table. On the other hand, Goodman [6] and Grizzle et al. [8] recommended replacing n_{ij} by $n_{ij} + (1/2)$ and $n_{ij} + (1/r^2)$, respectively, when zero cells occur. Both suggestions are used in practice.

The MH[T] model has $(T-1)(r-1)$ restrictions. This implies the number of degrees of freedom (df) of statistic for testing the goodness-of-fit of the MH[T] model, namely the df for the MH[T] model is $(T-1)(r-1)$. Similarly, the df for the ME[T] model is $T-1$ and that for the MVE[T] model is $2(T-1)$. The MCLL[T] (CMCLL[T]) and EMCLL[T] (CEMCLL[T]) models have $T-1$ and $2(T-1)$ additional parameters than the MH[T] model, respectively. Then, those for testing the goodness-of-fit of the MCLL[T] (CMCLL[T]) and EMCLL[T] (CEMCLL[T]) models are $(T-1)(r-2)$ and $(T-1)(r-3)$, respectively, because $(T-1)(r-1) - (T-1) = (T-1)(r-2)$ and $(T-1)(r-1) - 2(T-1) = (T-1)(r-3)$. We would like to note that the df for the MH[T] model is equal to the sum of the df for the EMCLL[T] (CEMCLL[T]) model and the MVE[T] model.

Consider two models, say M_1 and M_2 , such that if model M_1 holds, then model M_2 holds. For testing the goodness-of-fit of model M_1 assuming that model M_2 holds, the conditional likelihood ratio statistic is given by $G^2(M_1|M_2) = G^2(M_1) - G^2(M_2)$. The number of df for the conditional test is the difference between the numbers of df for models M_1 and M_2 . The conditional test statistics are more powerful because they are based on fewer degrees of freedom.

As an example, we consider the MLEs of the expected frequencies $\{m_{ij}\}$ under the CEMCLL model for a square contingency table. Those under the EMCLL, MCLL (CMCLL), MH, ME, and MVE models can be obtained in a similar manner to this case, although those are omitted here. To obtain the MLEs under the CEMCLL model, we must maximize the Lagrangian

$$L = \sum_{i=1}^r \sum_{j=1}^r n_{ij} \log p_{ij} - \lambda \left(\sum_{i=1}^r \sum_{j=1}^r p_{ij} - 1 \right) - \sum_{i=1}^{r-1} \mu_i \left(\log \left(1 - F_i^{X(c)} \right) - \Delta_1^{*i} \Delta_2^* \log \left(1 - F_i^{Y(c)} \right) \right),$$

with respect to $\{p_{ij}\}$, λ , $\{\mu_i\}$, Δ_1^* and Δ_2^* . Setting the partial derivatives of L equal to zero, we obtain the equations

$$p_{ij} = n_{ij} \left[n + I_{(i \neq j)} \sum_{k=1}^{r-1} \frac{\mu_k}{\delta} \left\{ \frac{F_k^{X(c)} - I_{(k \geq i)}}{1 - F_k^{X(c)}} - \Delta_1^{*i} \Delta_2^* \frac{F_k^{Y(c)} - I_{(k \geq j)}}{1 - F_k^{Y(c)}} \right\} \right]^{-1}$$

for $i = 1, \dots, r; j = 1, \dots, r$, as well as

$$1 - F_i^{X(c)} = \left(1 - F_i^{Y(c)} \right)^{\Delta_1^{*i} \Delta_2^*} \quad (i = 1, \dots, r-1),$$

$$\sum_{i=1}^{r-1} \mu_i \log \left(1 - F_i^{Y(c)} \right)^{i \Delta_1^{*(i-1)} \Delta_2^*} = 0 \quad \text{and}$$

$$\sum_{i=1}^{r-1} \mu_i \log \left(1 - F_i^{Y(c)} \right)^{\Delta_1^{*i}} = 0,$$

where $I_{(\cdot)}$ is the indicator function. Using the Newton–Raphson method, we can solve the equations with respect to $\{p_{ij}\}$, $\{\mu_i\}$, Δ_1^* and Δ_2^* . Therefore, we can obtain the MLEs of $\{m_{ij}\}$, Δ_1^* and Δ_2^* under the CEMCLL model.

6 Examples

6.1 Occupational Status for Japanese

Consider the data in Table 1 taken from Tominaga [19, p. 131] again. Table 2 gives the values of likelihood ratio chi-square statistic G^2 for testing the goodness-of-fit

of models. We analyze the data using the new model and the properties of the MH model.

First, we want to see whether the marginal distribution of father's status is equal to that of his son. Since the MH model fits poorly from Table 2, we can infer that the marginal distribution of father's status is different from that of his son. Then, the extended models (i.e., the MCLL and EMCLL models) are applied to the data, and neither fit well.

Second, we want to see whether the conditional marginal distribution of father's status is equal to that of his son on the condition that his status is different from that of his son. Then, we shall apply the extended models (i.e., the CMCLL and CEMCLL models) based on the conditional marginal cumulative distributions. The CEMCLL model only fits well. Under the CEMCLL model, the MLEs of Δ_1^* and Δ_2^* are $\hat{\Delta}_1^* = 1.59$ and $\hat{\Delta}_2^* = 0.75$ with the standard errors 0.099 and 0.127, respectively. We want to see whether $\Delta_1^* = 1$ and $\Delta_2^* = 1$. Consider the hypothesis that the MH model holds under the assumption that the CEMCLL model holds. According to the test based on the difference between G^2 values for the MH and CEMCLL models, this hypothesis is rejected at the 0.05 significance level because $203.55 - 1.67 = 201.88$ with 2 df. Therefore, the CEMCLL model may be preferable to the MH model. Hence, under the CEMCLL model, the probability that the son's status in a pair is $i + 1$ or above, is estimated to be equal to the probability that his father's status in a pair is $i + 1$ or above to the power of 0.75×1.59^i , for $i = 1, 2, 3$, on condition that the father's status is different from his son's status. Since $\hat{\Delta}_1^* \hat{\Delta}_2^* > 1$ for $i = 1, 2, 3$ under the CEMCLL model, $\hat{F}_i^{X(c)} > \hat{F}_i^{Y(c)}$ and that difference tends to be greater as i increases, where $\hat{F}_i^{X(c)}$ and $\hat{F}_i^{Y(c)}$ are MLEs of the conditional marginal cumulative probabilities of X and Y for $i = 1, 2, 3$. Therefore, the occupational distribution for son is stochastically lower than the occupational distribution for father on the condition that father's status is different from his son's status. The difference becomes greater as the status category number increases. Please see Fig. 1.

Last, according to Theorem 2, we can see that the poor fit of the MH model is caused by the influence of the lack of structure of the MVE model rather than the CEMCLL model.

Table 2 Likelihood ratio chi-square values G^2 for models to the data in Table 1

Models	df	G^2
MH	3	203.55*
MCLL	2	23.82*
CMCLL	2	54.61*
MVE	2	202.09*
EMCLL	1	7.23*
CEMCLL	1	1.67
ME	1	107.18*

*Significant at 5% level

Fig. 1 Observed and estimated conditional marginal distribution functions (cmF): the solid line is for son's status and the dashed line is for father's status where the estimated cmF is red and the observed cmF is black

6.2 Occupational Status for British

Table 3 is taken directly from Agresti [2, p. 448]. The table relates father's and son's occupational status category for a British sample. Social mobility data in Britain have been analyzed many authors, for example, Bishop et al. [3], Goodman [7], Agresti [1], Xie [20], Lang and Scott [10], Sobel et al. [16], and Tahata [17]. These articles mainly focused on the structure of joint distributions. On the other hand, we focus on the structure of marginal distributions.

A cursory glance at the data reveals that the MH model is inappropriate. Indeed, $G^2(MH) = 32.80$ for testing its fit, with $df = 4$. For the population represented by this sample, we analyze whether the (conditional) occupational distribution for sons differs from the (conditional) occupational distribution for fathers.

First, we apply the MCLL and EMCLL models to compare two marginal distributions. These models fit well, having $G^2(MCLL) = 4.26$ ($df = 3$) and $G^2(EMCLL) = 3.04$ ($df = 2$), respectively. Their parameter estimates are $\hat{\Delta}_2 = 0.88$ with the standard error 0.020 under the MCLL model and $\hat{\Delta}_1 = 0.98$ and $\hat{\Delta}_2 = 0.96$ with the standard errors 0.022 and 0.079, respectively, under the EMCLL model. Consider the hypothesis that the MCLL model holds under the assumption that the EMCLL model holds, that is, $\Delta_1 = 1$. According to the test based on the difference between G^2 values for the MCLL and EMCLL models, this hypothesis is accepted at the 0.05 significance level because $4.26 - 3.04 = 1.22$ with 1 df. Therefore, the MCLL model may be preferable to the EMCLL model. The occupational distribution for son is stochastically higher than the occupational distribution for father. As the reference, we give the observed and estimated marginal distribution functions in Fig. 2.

Last, we apply the CMCLL and CEMCLL models to compare two conditional marginal distributions on the condition that father's status is different from his son's status. These models fit well, having $G^2(CMCLL) = 5.35$ ($df = 3$)

Table 3 Occupational status for British father–son pairs; from Agresti [2, p. 448]

Son's status	Father's status					Total
	(1)	(2)	(3)	(4)	(5)	
(1)	50	28	11	14	3	106
	(50.25)	(31.07)	(11.30)	(14.39)	(3.16)	(110.17)
	(50.00)	(31.37)	(11.48)	(14.41)	(3.21)	(110.47)
(2)	45	174	78	150	42	489
	(40.88)	(172.82)	(72.29)	(139.05)	(39.86)	(464.90)
	(40.56)	(174.00)	(72.42)	(137.56)	(39.91)	(464.45)
(3)	8	84	110	185	72	459
	(7.83)	(90.59)	(110.07)	(185.16)	(73.91)	(467.57)
	(7.75)	(89.88)	(110.00)	(182.51)	(73.80)	(463.94)
(4)	18	154	223	714	320	1429
	(17.62)	(166.05)	(223.10)	(714.46)	(328.43)	(1449.66)
	(17.69)	(167.45)	(226.34)	(714.00)	(333.10)	(1458.58)
(5)	8	55	96	447	411	1017
	(7.65)	(57.80)	(93.79)	(436.79)	(411.67)	(1007.70)
	(7.60)	(57.59)	(94.07)	(432.30)	(411.00)	(1002.56)
Total	129	495	518	1510	848	3500
	(124.24)	(518.33)	(510.56)	(1489.84)	(857.03)	(3500.00)
	(123.60)	(520.29)	(514.31)	(1480.78)	(861.02)	(3500.00)

The upper and lower parenthesized values are the MLEs of expected frequencies under the MCLL and CMCLL models, respectively

1, lowest status; 5, highest status

and $G^2(\text{CEMCLL}) = 3.53$ ($\text{df} = 2$), respectively. Their parameter estimates are $\hat{\Delta}_2^* = 0.82$ with the standard error 0.031 under the CMCLL model and $\hat{\Delta}_1^* = 0.95$ and $\hat{\Delta}_2^* = 0.97$ with the standard errors 0.035 and 0.130, respectively, under the CEMCLL model. Consider the hypothesis that the CMCLL model holds under the assumption that the CEMCLL model holds, that is, $\Delta_1^* = 1$. According to the test based on the difference between G^2 values for the CMCLL and CEMCLL models, this hypothesis is accepted at the 0.05 significance level because $5.35 - 3.53 = 1.82$ with 1 df. Therefore, the CMCLL model may be preferable to the CEMCLL model. The occupational distribution for son is stochastically higher than the occupational distribution for father on the condition that father's status is different from his son's status. As the reference, we give the observed and estimated conditional marginal distribution functions in Fig. 3.

Fig. 2 Observed and estimated marginal distribution functions (mF): the solid line is for son's status and the dashed line is for father's status where the estimated mF is red and the observed mF is black

Fig. 3 Observed and estimated conditional marginal distribution functions (cmF): the solid line is for son's status and the dashed line is for father's status where the estimated cmF is red and the observed cmF is black

7 Concluding Remarks

In this paper, the EMCLL[T] and CEMCLL[T] models have been proposed for $T \geq 2$. The EMCLL[T] model should be applied when we want to see the inhomogeneity between cumulative marginal distributions. On the other hand, the CEMCLL[T] model should be applied when we want to see the inhomogeneity between conditional cumulative marginal distributions on the condition that there is at least one variable unequal to others. As described in Sect. 2, if it is reasonable to assume an underlying Gumbel distribution with the location parameter and the scale parameter, then the proposed models may be appropriate for the square contingency

tables with ordinal categories. We would like to note that these models should not apply to the data with nominal categories.

In comparison with the $MCLL[T]$ ($CMCLL[T]$) model, additional parameters Δ_{1t} (Δ_{1t}^*) for $t = 2, \dots, T$ allow us to consider degrees of inhomogeneity proportional to category number as it shown in Sect. 6. The conditional test statistics are more powerful because they are based on fewer degrees of freedom. Thus, the proposed models enable more detailed analysis for the contingency tables with the same ordinal categories.

Theorems 1, 2, 3, and 4 have been shown in this paper. These results may be useful to see the reason why the poor fit of the $MH[T]$ model when the $MH[T]$ model fits poorly for the real dataset. If the $MCLL[T]$ model and the $EMCLL[T]$ model (or the $CMCLL[T]$ model and the $CEMCLL[T]$ model) both fit well,

$$G^2(MCLL[T]|EMCLL[T]) = G^2(MCLL[T]) - G^2(EMCLL[T])$$

or

$$G^2(CMCLL[T]|CEMCLL[T]) = G^2(CMCLL[T]) - G^2(CEMCLL[T])$$

with $(T - 1)$ df is useful to test the hypothesis that $\Delta_{1t} = 1$ (or $\Delta_{1t}^* = 1$) for $t = 2, \dots, T$ under the assumption that the $EMCLL[T]$ ($CEMCLL[T]$) model holds. Indeed, we have used these properties for the occupational status of British father–son pairs.

Acknowledgements We would like to thank the Associate Editor and the three reviewers for their helpful and valuable comments to improve our manuscript.

Declarations

Conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Agresti A (1984) Analysis of ordinal categorical data. Wiley, New York
2. Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken, New Jersey
3. Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. The MIT Press, Cambridge, Massachusetts
4. Bowker AH (1948) A test for symmetry in contingency tables. J Am Stat Assoc 43:572–574

5. Caussinus H (1965) Contribution à l'analyse statistique des tableaux de corrélation. *Ann Facult Sci l'Univ Toulouse Série 4*(29):77–182
6. Goodman LA (1970) The multivariate analysis of qualitative data: interaction among multiple classifications. *J Am Stat Assoc* 65:226–256
7. Goodman LA (1979) Multiplicative models for square contingency tables with ordered categories. *Biometrika* 66:413–418
8. Grizzle JE, Starmer CF, Koch GG (1969) Analysis of categorical data by linear models. *Biometrics* 25:489–504
9. Kurakami H, Tahata K, Tomizawa S (2010) Extension of the marginal cumulative logistic model and decompositions of marginal homogeneity for multi-way tables. *J Stat Adv Theo Appl* 3:135–152
10. Lang JB, Scott RE (1997) Application of association-marginal models to the study of social mobility. *Sociol Methods Res* 26:183–212
11. McCullagh P (1977) A logistic model for paired comparisons with ordered categorical data. *Biometrika* 64:449–453
12. Miyamoto N, Niibe K, Tomizawa S (2005) Decompositions of marginal homogeneity model using cumulative logistic models for square contingency tables with ordered categories. *Aust J Stat* 34:361–373
13. Saigusa Y, Maruyama T, Tahata K, Tomizawa S (2018) Extended marginal homogeneity model based on complementary log–log transform for square tables. *Int J Stat Prob* 7:27–31
14. Shinoda S, Tahata K, Iki K, Tomizawa S (2019) Extended marginal homogeneity models based on complementary log–log transform for multi-way contingency tables. *SUT J Math* 55:25–37
15. Stuart A (1955) A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42:412–416
16. Sobel ME, Becker MP, Minick SM (1998) Origins, destinations, and association in occupational mobility. *Am J Sociol* 104:687–721
17. Tahata K (2020) Separation of symmetry for square tables with ordinal categorical data. *Jpn J Stat Data Sci* 3:469–484
18. Tahata K, Katakura S, Tomizawa S (2007) Decompositions of marginal homogeneity model using cumulative logistic models for multi-way contingency tables. *Revstat Stat J* 5:163–176
19. Tominaga K (1979) *Nippon no Kaisou Kouzou* (Japanese Hierarchical Structure). University of Tokyo Press, Tokyo
20. Xie Y (1992) The log-multiplicative layer effect model for comparing mobility tables. *Am Sociol Rev* 57:380–395

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.