On exploratory analytic method for multi-way contingency tables with an ordinal response variable and categorical explanatory variables

Zheng Wei[a], Daeyoung Kim[b,*]

[a]Department of Mathematics and Statistics, University of Maine, Orono, ME 04469-5752, USA
[b]Department of Mathematics and Statistics, University of Massachusetts-Amherst, Amherst, MA 01003-9305, USA

Abstract

In this paper, we propose a new model-free exploratory method for descriptive modeling that identifies and measures the regression dependence between an ordinal response variable and categorical (ordinal or nominal) explanatory variables in a multi-way contingency table. The proposed methodology consists of three parts, checkerboard copula score, checkerboard copula regression, and checkerboard copula association measure. The checkerboard copula score is a new type of score for ordinal variables that preserves the natural ordering of the categorical scale and it will be exploited for developing the methods measuring the association between the variables of interest. The checkerboard copula regression identifies the regression dependence between an ordinal response variable and categorical explanatory variables. It enables delineating the identified dependence in an exploratory manner. The checkerboard copula association measure quantifies the strength of the dependence identified by the checkerboard copula regression. We investigate the properties of checkerboard copula scores, checkerboard copula regression, its association measure, and their estimators. Finally, the performance of the proposed method is illustrated with simulation and real data.

Keywords: Association measure, Copula, Multivariate categorical data, Regression.
2020 MSC: Primary 62H20, Secondary 62H17

1. Introduction

The analysis of data with ordinal categorical responses has been an important task in various research fields such as social, medical, and public health sciences. The use of the ordering information in the categories of ordinal variables results in more powerful inferences and various methods designed for the ordinal response data have been developed (see [1, 2, 10, 22, 25, 31, 32, 34, 43, 47, 48] and references therein).

A primary step for properly analyzing data with ordinal responses is to explore the data and discover different types of dependence structures among the variables in as many ways as possible for exploratory and descriptive modeling. This step enables the identification of dependency patterns that are perhaps unknown or informally formulated, to summarize the identified dependencies, and to obtain clues for potential variables that may affect the ordinal response variable [6, 46]. This information would be beneficial toward the goals of formal modeling such as explanatory modeling to test underlying causal hypotheses and predictive modeling to predict new or future observations [40].

One of the dependence structures that has attracted the interest of researchers in the statistical modeling process is the regression dependence, where the main interest is to identify and measure the relationship of one variable designated as a response variable on the remaining variables designated as explanatory variables. Many model-based approaches explicitly modeling regression dependence in ordinal response data have been proposed, especially for the formal modeling (explanatory modeling or predictive modeling). Commonly used methods include proportional odds model (cumulative logit model) [35], other cumulative link models (with probit, log-log, and complementary log-log links) [3, 9], adjacent-categories logit model [17, 41], continuation ratio logit model [29], stereotype model [4], latent variable models, association models [16, 18], correspondence analysis models, and canonical correlation

*Corresponding author. Email address: daeyoung@math.umass.edu

models [13, 14, 19, 20]. On the other hand, one may employ a variety of non-model-based approaches for ordinal response data, including (but not limited to) generalized Cochran-Mantel-Haenszel methods [30, 33], ordinal odds ratios [7, 16, 17, 35], and rank-based methods, for example, Kendall's tau [28] and its variants [26, 45], Goodman and Kruskal's gamma [21], Spearman's rank-based correlation [1], Somers' D [44], Kendall's partial tau and its extension [23, 27]. These non-model based methods provide description and inference for ordinal data without explicit specification of the underlying dependence structure among the variables.

However, it is not straightforward to use the non-model-based methods listed above when the goal is to learn and discover the regression dependence in the multivariate ordinal data with an ordinal response variable. This is because these methods are mainly designed for bivariate (marginal/conditional) association between two ordinal variables of interest (unconditional/conditional on the other variables), and some of these methods treat two ordinal variables symmetrically (i.e., no distinction between the response variable and the explanatory variable).

To this end, we propose a novel non-model based, data-driven exploratory method to identify and quantify the regression dependence in a multi-way contingency table with an ordinal response variable and categorical (ordinal or nominal) explanatory variables. The principal mathematical tool employed in the proposed method is the checkerboard copula, also called multilinear extension copula [11, 37, 39]. The checkerboard copula, constructed by multilinear interpolation of the joint distribution function of discrete marginal distributions, uniquely links the marginal distributions of discrete random variables to their joint distribution function. Recent researches show that the checkerboard copula best represents the dependence structure among multivariate discrete variables and it plays an important role in developing new statistical methods for ordinal contingency tables [5, 11, 12, 38].

The proposed methodology consists of three parts: checkerboard copula score, checkerboard copula regression, and checkerboard copula association measure. First, the checkerboard copula score is a new type of score for ordinal variables accounting for the ordinal nature of the categories. To exploit the fact that the ordering of the categories of ordinal variables is informative, the proposed scores will be utilized in the methods developed in this paper. Second, the checkerboard copula regression is a model-free approach to identify the regression dependence between an ordinal response variable and categorical explanatory variables. It also enables delineating the identified dependence in an exploratory manner via the prediction of an ordinal response variable for each combination of categories of explanatory variables. Third, the checkerboard copula regression association measure is an index to quantify the strength of the association identified by the checkerboard copula regression. It is the average proportion of variance in the response variable (with respect to its checkerboard copula score and its marginal distribution) attributable to the checkerboard copula regression. Note that the proposed methods, designed for contingency tables where all variables are ordinal, are also applicable for the tables with nominal explanatory variables.

A previous work relevant to the method proposed in this paper is the one introduced in [31], where a set of test statistics was developed to examine the association between two ordinal categorical variables after adjusting for continuous and/or categorical covariates. In their method, both ordinal variables are treated symmetrically, meaning that it does not distinguish which ordinal variable is to be considered the dependent variable and which is to be considered the explanatory variable. The first step of their approach is to separately fit the two ordinal variables on the covariates using ordinal categorical response models (requiring explicit modeling of the relationship between each ordinal variable and covariates) and obtain the predicted probability distributions for these two ordinal variables. The next step is to construct test statistics using the predicted probability distributions and bivariate association measures (such as Goodman and Kruskal's gamma) and compute p-values for testing the null of conditional independence. The main features of the proposed method that differs from the work of [31] are four-fold: (i) the proposed method focuses on the regression dependence leading to distinguishing between a response variable and explanatory variables; (ii) it is designed for measuring the strength of the regression dependence, though it can be used to test the null of no regression association based on permutation test; (iii) it does not require an explicit model for regression dependence; (iv) it is applicable only when explanatory variables are categorical (ordinal or nominal).

The remainder of this article is organized as follows. In Section 2, we briefly review the checkerboard copula in a multi-way ordinal contingency table. We introduce the checkerboard copula score for the ordinal categorical variable, and propose the checkerboard copula regression and its association measure for a multi-way contingency table with an ordinal response variable and categorical explanatory variables. Their theoretical properties are also investigated. Section 3 introduces the estimation of the methods proposed in Section 2 and presents their asymptotic properties. To demonstrate the fundamental ideas of the proposed methodology, we consider an artificial two-way contingency table as a running example through Sections 2-3. In Section 4, we perform simulation and data analysis to evaluate the performance of the proposed method. Finally, Section 5 closes the paper with a discussion.

2. Checkerboard copula score, checkerboard copula regression and its association measure for the multi-way contingency table with an ordinal response

2.1. Review on checkerboard copula

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ be a d-dimensional random vector with the joint cumulative distribution function (c.d.f.) H and arbitrary marginal c.d.f.s $F_1, \ldots, F_d$. Then, by Sklar [42], there exists at least one copula $C : [0,1]^d \to [0,1]$ such that

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \quad x_1, \ldots, x_d \in \mathbb{R}.$$

If $F_1, \ldots, F_d$ are continuous, C is unique and is the c.d.f. of $\boldsymbol{F}(\boldsymbol{X}) = (F_1(X_1), \ldots, F_d(X_d))$; otherwise, C is uniquely determined on the domain $\prod_{j=1}^d D_j$, where D_j is the range of F_j. Copula is a powerful tool to model the dependence of multivariate data. For an overview of copula theory and applications, see [24, 37].

Specifically, suppose each X_j in $\boldsymbol{X}$ is an ordinal variable with I_j categories $\{x_1^j < \cdots < x_{i_j}^j < \cdots < x_{I_j}^j\}$. Note that "<" represents the natural ordering of categories of X_j. We can form a d-way contingency table that cross-classifies subjects with respect to d variables. Let the joint probability mass function (p.m.f.) of $\boldsymbol{X}$ in the d-way contingency table be the array with size $\prod_{j=1}^d I_j$,

$$P = \left\{ p_{\boldsymbol{i}} = p_{i_1,\ldots,i_d} = \Pr(X_1 = x_{i_1}^1, \ldots, X_d = x_{i_d}^d) \,\middle|\, i_j \in \{1, \ldots, I_j\},\ j \in \{1, \ldots, d\},\ \boldsymbol{i} = (i_1, \ldots, i_d)^\top \in \prod_{j=1}^d \mathbb{I}_j \right\},$$

where $\mathbb{I}_j = \{1, \ldots, I_j\}$ denotes the index set for X_j, $\mathbf{1} = (1, \ldots, 1)^\top$ and $\boldsymbol{I} = (I_1, \ldots, I_d)^\top$ denote index vectors of length d, and $\sum_{i_1=1}^{I_1} \cdots \sum_{i_d=1}^{I_d} p_{i_1,\ldots,i_d} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{I}} p_{\boldsymbol{i}} = 1$. The marginal p.m.f. of i_j-th entry in X_j is denoted as

$$p_{+i_j+} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_{j-1}=1}^{I_{j-1}} \sum_{i_{j+1}=1}^{I_{j+1}} \cdots \sum_{i_d=1}^{I_d} p_{i_1,\cdots,i_d} = \sum_{\boldsymbol{i}_{-j}=\mathbf{1}_{-j}}^{\boldsymbol{I}_{-j}} p_{\boldsymbol{i}}, \tag{1}$$

where $\boldsymbol{i}_{-j}$, $\mathbf{1}_{-j}$, and $\boldsymbol{I}_{-j}$ denote index vectors of $\boldsymbol{i}$, $\mathbf{1}$, and $\boldsymbol{I}$ without j-th entry, respectively. The marginal p.m.f. for $\boldsymbol{X}_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d)^\top$, the $(d-1)$-dimensional random vector without X_j, is denoted by

$$p_{i_1,\ldots,+_j,\ldots,i_d} = \sum_{i_j=1}^{I_j} p_{i_1,\ldots,i_d}.$$

Furthermore, the conditional p.m.f. of X_j given $\boldsymbol{X}_{-j}$ is

$$p_{i_j|\boldsymbol{i}_{-j}} = \frac{p_{i_1,\ldots,i_d}}{p_{i_1,\ldots,+_j,\ldots,i_d}}. \tag{2}$$

We denote the range of the marginal distribution of X_j to be $D_j = \{u_0^j, \ldots, u_{i_j}^j, \ldots, u_{I_j}^j\}$, where $u_0^j = 0$, $u_{I_j}^j = 1$, and

$$u_{i_j}^j = \sum_{k_j=1}^{i_j} p_{+k_j+}. \tag{3}$$

Note that if the superscript in $u_{i_j}^j$ can be deduced trivially from the subscript, the superscript may be omitted. Otherwise, the superscript will be kept.

Then, by Sklar [42], the unique subcopula C^s associated with d-dimensional random vector $\boldsymbol{X}$ over $\prod_{j=1}^d D_j$ is given by

$$H(x_{i_1}^1, \ldots, x_{i_d}^d) = \sum_{k_1 \le i_1} \cdots \sum_{k_d \le i_d} p_{k_1,\ldots,k_d} = C^s(u_{i_1}^1, \ldots, u_{i_d}^d). \tag{4}$$

As noted in [11, 37], any subcopula C^s on $\prod_{j=1}^d D_j$ can be extended to a copula C on $[0,1]^d$ by multilinear interpolation [39], and the constructed copula, called the checkerboard copula, links the marginal distributions of discrete random variables to their joint distribution function in a unique way. The definition of the checkerboard copula and its density function are given below.

Definition 1. [11, 37] Let C^s be a subcopula on $\prod_{j=1}^d D_j$ satisfying (4). For any $\boldsymbol{u} = (u_1, \dots, u_d) \in [0,1]^d$, let u_j^l and u_j^u be, respectively, the least and greatest elements of $\overline{D_j}$, the closure of set D_j satisfying $u_j^l \leq u_j \leq u_j^u$. Note that if u_j is in $\overline{D_j}$, then $u_j^l = u_j = u_j^u$. Furthermore, for any $S \subseteq \{1, \dots, d\}$, let

$$\lambda_j(u_j) = \begin{cases} \frac{u_j - u_j^l}{u_j^u - u_j^l}, & \text{if } u_j^l < u_j^u, \\ 1, & \text{if } u_j^l = u_j^u, \end{cases} \quad \text{and} \quad \lambda_S(u_1, \dots, u_d) = \prod_{\ell \in S} \lambda_\ell(u_\ell) \prod_{\ell \notin S} (1 - \lambda_\ell(u_\ell)).$$

Then, the checkerboard copula C^+ of the ordinal random vector $\boldsymbol{X}$ is defined as

$$C^+(\boldsymbol{u}) = C^+(u_1, \dots, u_d) = \sum_{S \subseteq \{1, \dots, d\}} \lambda_S(u_1, \dots, u_d) C^s(u_{s_1}, \dots, u_{s_d}), \tag{5}$$

where $u_{s_j} = u_j^u$ if $j \in S$ and $u_{s_j} = u_j^l$, otherwise. By taking the derivatives of C^+ with respect to $u_1, \dots, u_d$, the checkerboard copula density function is defined to be

$$c^+(\boldsymbol{u}) = c^+(u_1, \dots, u_d) = \frac{p_{i_1, \dots, i_d}}{\prod_{j=1}^{d} p_{+i_j+}}, \qquad u_{i_j-1}^j < u_j \leq u_{i_j}^j. \tag{6}$$

It follows from Definition 1 that C^+ coincides with the unique subcopula C^s of (4) on $\prod_{j=1}^d \overline{D_j}$ associated with the ordinal random vector $\boldsymbol{X}$. Genest et al. [12] obtained the stochastic representation of $C^+(\boldsymbol{u})$ in (5) : $C^+(\boldsymbol{u})$ is the joint distribution function of the standard uniform random vector $\boldsymbol{U}=(U_1, \dots, U_d)^\top$ where

$$U_j = F_j(X_j-) + \{F_j(X_j) - F_j(X_j-)\}V_j \sim \mathcal{U}(0,1), \tag{7}$$

$F(x-)$ refers to the left limit of F and $V_1, \dots, V_d$ denote independent uniform random variables on $[0,1]$ independent of $\boldsymbol{X}$. Note that the expression for U_j in (7) can be viewed as a distributional transform extending the distribution of X_j to the continuous distribution over $[0,1]$.

For the better illustration of the checkerboard copula and the proposed methods in the following sections, we present an artificial example.

Example 1. Suppose that X_1 and X_2 are two ordinal variables representing the dose of a treatment drug for acute migraine and the severity of migraine pain recorded after treatment with $I_1 = 5$ and $I_2 = 3$ ordered categories, respectively: $(x_1^1, x_2^1, x_3^1, x_4^1, x_5^1)$=(very low, low, medium, high, very high) and (x_1^2, x_2^2, x_3^2)=(mild, moderate, severe).

Table 1: Joint p.m.f of X_1 and X_2, $P = \{p_{i_1 i_2}\}$.

X_1 \ X_2	x_1^2	x_2^2	x_3^2
x_1^1	0	0	2/8
x_2^1	0	1/8	0
x_3^1	2/8	0	0
x_4^1	0	1/8	0
x_5^1	0	0	2/8

Table 1 shows the joint p.m.f of X_1 and X_2, $P = \{p_{i_1 i_2}, i_1 \in \{1, \dots, 5\}, i_2 \in \{1,2,3\}\}$. Note that X_2 has a quadratic relationship with X_1 in that the level of X_2 decreases and then increases as the level of X_1 increases. We also observe that X_2 is a function of X_1 with probability 1, but not vice versa: for a given category of X_1, there is one and only one category of X_2 whose corresponding joint probability is non-zero. The marginal p.m.f.s of X_1 and X_2 are $p_{i_1+} \in$ {2/8, 1/8, 2/8, 1/8, 2/8} and $p_{+i_2} \in$ {2/8, 2/8, 4/8}, respectively. The ranges of the marginal c.d.f.s of X_1 and X_2 in (3) are $D_1 = \{u_0^1, u_1^1, u_2^1, u_3^1, u_4^1, u_5^1\}$ = {0,2/8,3/8,5/8,6/8,1} and $D_2 = \{u_0^2, u_1^2, u_2^2, u_3^2\}$ = {0,2/8,4/8,1}, respectively. Fig. S1 of the supplementary material shows that the checkerboard copula density of X_1 and X_2 in (6) inherits the dependence between X_1 and X_2.

2.2. *Checkerboard copula score*

Ordinal variables have ordered categories whose distances are unknown and their categories do not necessarily represent actual magnitudes. In order to exploit the fact that the ordering of categories of ordinal variables is informative in measuring association in an ordinal contingency table, we propose a new type of scores for ordinal variables obtained from the checkerboard copula.

As shown in Definition 1 of Section 2.1, the checkerboard copula C^+ is a smooth version of the subcopula associated with the ordinal random vector $\boldsymbol{X}$ in that it spreads the (zero or non-zero) mass uniformly over each of all $\prod_{j=1}^{d} I_j$ d-dimensional hyperrectangles in $[0,1]^d$, $\prod_{j=1}^{d}[u^j_{i_j-1}, u^j_{i_j}]$, where $u^j_{i_j}$ is defined in (3). In addition, C^+ is the joint distribution function of $\boldsymbol{U}$ in (7), which is a distributional transform of $\boldsymbol{X}$. Motivated by these properties of the checkerboard copula, we define a new random variable S_j to be a transformation of X_j via U_j in (7): $S_j = E[U_j \mid X_j]$ where $j \in \{1, \ldots, d\}$. As shown in Proposition 1 (i)-(ii), S_j is an ordinal random variable with numerical support values $\{s^j_1, \ldots, s^j_{i_j}, \ldots, s^j_{I_j}\}$ where $s^j_{i_j} = (u^j_{i_j-1} + u^j_{i_j})/2$, and S_j has the same p.m.f. as X_j.

We propose the support values of S_j above as a new type of scores for X_j, named as checkerboard copula scores.

Definition 2. The checkerboard copula scores of ordinal variable X_j are

$$\{s^j_1, \ldots, s^j_{i_j}, \ldots, s^j_{I_j}\}, \quad s^j_{i_j} = (u^j_{i_j-1} + u^j_{i_j})/2, \tag{8}$$

for $i_j \in \{1, \ldots I_j\}$ and $u^j_{i_j}$ are given in (3).

The checkerboard copula scores in (8) is a set of the average of the marginal distributions evaluated at every two consecutive categories of X_j. The following proposition shows properties of the proposed checkerboard copula scores.

Proposition 1.

(i) *The checkerboard copula scores in* (8) *have the same ordering as the categories of* X_j*:* $0 < s^j_1 < \cdots < s^j_{i_j} < \cdots < s^j_{I_j} < 1$.

(ii) *The conditional expectation of the stochastic representation in* (7) *given* $X_j = x_j$ *with respect to* U_j *(which is also with respect to* V_j *independent of* X_j*) is equal to the* i_j*-th checkerboard score for* X_j*:* $\mathrm{E}\left(U_j | X_j = x_{i_j}\right) = s^j_{i_j}$.

(iii) *The mean and variance of* S_j *are* $\mu_{S_j} = 0.5$ *and* $\sigma^2_{S_j} = \frac{1}{4}\sum_{i_j=1}^{I_j} u^j_{i_j-1} u^j_{i_j} p_{+i_j+}$. *The maximum value of the variance of* S_j *is achieved when* $p_{+i_j+} = 1/I_j$ *(i.e., the marginal distribution of* X_j *is discrete uniformly distributed).*

Proof. See Appendix. □

Example 1 (continued). For the table given in Example 1 of Section 2.1, we obtain the checkerboard copula scores for X_1 and X_2: (2/16,5/16,8/16,11/16,14/16) for X_1 and (2/16,6/16,12/16) for X_2. We also compute the means and variances of $S_1 = E(U_1 \mid X_1)$ and $S_2 = E(U_2 \mid X_2)$ (transformations of X_1 and X_2): (0.5, 81/1024) for S_1 and (0.5, 9/128) for S_2.

2.3. *Checkerboard copula regression*

In this section, we propose the checkerboard copula regression function for a multi-way contingency table with an ordinal response variable and a set of explanatory variables.

Let $\boldsymbol{U}$ in (7) be a uniform random vector on $[0,1]^d$ associated with the checkerboard copula C^+ in (5) for a d-way ordinal contingency table. The (d-1)-marginal density for $\boldsymbol{U}_{-j} = (U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_d)^\top$ and the conditional density of U_j given $\boldsymbol{U}_{-j}$ are given as follows: for any $\boldsymbol{u}_{-j} = (u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_d)^\top$ in $[0,1]^{d-1}$ and $u^k_{i_k-1} < u_k \leq u^k_{i_k}$,

$$c^+(\boldsymbol{u}_{-j}) = \frac{p_{i_1,\ldots,+_j,\ldots,i_d}}{\prod\limits_{k=1,k\neq j}^{d} p_{+i_k+}}, \qquad c^+(u_j|\boldsymbol{u}_{-j}) = \frac{c^+(\boldsymbol{u})}{c^+(\boldsymbol{u}_{-j})} = \frac{p_{i_j|\boldsymbol{i}_{-j}}}{p_{+i_j+}},$$

where $k \in \{1, \ldots, j\text{-}1, j\text{+}1, \ldots, d\}$, $j \in \{1, \ldots, d\}$, and p_{+i_j+} and $p_{i_j|\boldsymbol{i}_{-j}}$ are defined in (1) and (2), respectively.

Definition 3. The **checkerboard copula regression function** of U_j on $\boldsymbol{U}_{-j}$ is defined as follows: for $u^k_{i_k-1} < u_k \le u^k_{i_k}$ and $k \in \{1, \ldots, j-1, j+1, \ldots, d\}$,

$$r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j}) \equiv \mathrm{E}_{c^+}(U_j|\boldsymbol{U}_{-j} = \boldsymbol{u}_{-j}) = \int_0^1 u_j c^+(u_j|\boldsymbol{u}_{-j}) du_j = \sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}} s^j_{i_j}. \tag{9}$$

The checkerboard copula regression function defined in Definition 3 can be interpreted as the mean checkerboard score of X_j with respect to the conditional distribution at the category $\boldsymbol{i}_{-j}$ of $(d-1)$ explanatory variables $\boldsymbol{X}_{-j}$.

Example 1 (continued). Table 2(a) shows the conditional p.m.f. of X_2 given X_1, $p_{i_2|i_1}$, computed from Table 1 in Section 2.1. We see that X_1 is quadratically associated with X_2 as shown in Table 1 and, for each category of X_1, there exists only one category of X_2 whose conditional probability is exactly 1. Using the checkerboard copula score of X_2 and the conditional p.m.f. of X_2 given X_1 in Table 2(a), we obtain the checkerboard copula regression of U_2 on U_1, $r_{U_2|U_1}(u_1)$, in Table 2(b) and visualize it in Fig. S2 of the supplementary material.

Table 2: (a) Conditional p.m.f. of X_2 given X_1; (b) Checkerboard copula regression of U_2 on U_1; (c) Conditional p.m.f. of X_1 given X_2

X_1 \ X_2	x^2_1	x^2_2	x^2_3
x^1_1	0	0	1
x^1_2	0	1	0
x^1_3	1	0	0
x^1_4	0	1	0
x^1_5	0	0	1

u_1	$r_{U_2\|U_1}(u_1)$
[0,2/8]	12/16
(2/8,3/8]	6/16
(3/8,5/8]	2/16
(5/8,6/8]	6/16
(6/8,1]	12/16

X_1 \ X_2	x^2_1	x^2_2	x^2_3
x^1_1	0	0	1/2
x^1_2	0	1/2	0
x^1_3	1	0	0
x^1_4	0	1/2	0
x^1_5	0	0	1/2

We observe from Table 2(b) and Fig. S2 that $r_{U_2|U_1}(u_1)$ can capture the quadratic dependence given in Table 1 as it reflects the changes in the D_2 associated with X_2 according to the changes in the D_1 associated with X_1, and it is equal to one and only one of the checkerboard score of X_2 for each interval in U_1. On the other hand, we find that the checkerboard copula regression of U_1 on U_2, $r_{U_1|U_2}(u_2)$, equals 0.5 for all $u_2 \in [0,1]$. This is due to the facts that i) checkerboard copula scores for X_1 are equally spaced, centered at the middle category x^1_3 whose score is equal to 1/2(=8/16), and ii) the conditional p.m.f. of X_1 given X_2 in Table 2(c), $p_{i_1|i_2}$, is symmetrically distributed in terms of X_1, centered at x^1_3, and for each of second and third categories (x^2_2 and x^2_3) of X_2, the corresponding conditional probabilities are equally and symmetrically distributed over two categories of X_1, (x^1_2,x^1_4) for x^2_2 and (x^1_1,x^1_5) for x^2_3.

The checkerboard copula regression in Definition 3 can be used for predicting the category of a response variable for a given combination of categories of explanatory variables and describing the dependence structure between them. Assume the j-th variable X_j is a response variable and the remaining variables in $\boldsymbol{X}_{-j}$ are explanatory variables. For a given combination categories of $\boldsymbol{X}_{-j}$, we find the corresponding $\boldsymbol{u}^*_{-j}$ from the range $\prod_{k=1,k\neq j}^d D_k$ associated with $\boldsymbol{X}_{-j}$ in (3), and obtain the estimated value of the checkerboard copula regression in (9), $u^*_j = r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}^*_{-j})$. From the range D_j of the marginal distribution of X_j, we get i^*_j and $u^j_{i^*_j}$ such that $u^j_{i^*_j-1} < u^*_j \le u^j_{i^*_j}$ and obtain the predicted category of a response variable X_j, $x^j_{i^*_j}$.

As shown in Proposition 2 below, prediction of response category using the checkerboard copula regression is invariant under permutation of categories of explanatory variables. Thus, the proposed prediction method can also be employed for nominal explanatory variables.

Proposition 2. *For the d-way ordinal contingency table of* $X_1, \ldots, X_d$*, let* $f_{X_j|\boldsymbol{X}_{-j}}$ *be the prediction of a category of* X_j *for a given combination of categories in* $\boldsymbol{X}_{-j}$ *using the checkerboard copula regression. If* $\tilde{X}_k = g_k(X_k)$ *where* g_k *is an injective function of* X_k *and* $k = 1, \ldots, j-1, j+1, \ldots, d$*, the prediction of a category of* X_j *is invariant over the injective transformation on* $\boldsymbol{X}_{-j}$*. That is,* $f_{X_j|\boldsymbol{X}_{-j}}(\boldsymbol{x}^*_{\boldsymbol{i}_{-j}}) = f_{X_j|\tilde{\boldsymbol{X}}_{-j}}(\boldsymbol{x}^*_{\boldsymbol{i}_{-j}})$*, where* $\boldsymbol{x}^*_{\boldsymbol{i}_{-j}}$ *is denoted as a combination of categories in* $\boldsymbol{X}_{-j}$ *and* $\tilde{\boldsymbol{X}}_{-j} = (\tilde{X}_1, \ldots, \tilde{X}_{j-1}, \tilde{X}_{j+1}, \ldots, \tilde{X}_d)^\top$.

Proof. See Appendix. □

Example 1 (continued). Using the checkerboard copula regression $r_{U_2|U_1}(u_1)$ in Table 2(b), we predict a category of X_2 for each category of X_1. For example, given the third category of X_1 (i.e., $x^{1*}_{i_1=3} = x^1_3$), the corresponding $u^{1*}_3 \in D_1 = \{0,2/8,3/8,5/8,6/8,1\}$ is 5/8 and the predicted value of the checkerboard copula regression is $u^*_2 = r_{U_2|U_1}(5/8) = 1/8$. Then we see that there exists $i^*_2 = 1$ and $u^2_{i^*_2=1} = 2/8$ from $D_2 = \{0,2/8,4/8,1\}$ such that $u^2_0 = 0 < u^*_2 = 1/8 \leq u^2_1 = 2/8$, and thus the predicted category of X_2 for $X_1 = x^1_3$ is the first category of X_2, $f_{X_2|X_1}(x^1_3) = x^2_1$. Applying the same prediction method to the other categories of X_1, we obtain the predicted categories of X_2: $f_{X_2|X_1}=(x^2_3,x^2_2,x^2_1,x^2_2,x^2_3)$ for $X_1=(x^1_1,x^1_2,x^1_3,x^1_4,x^1_5)$, respectively. The prediction result reflects the quadratic relationship shown in Table 1. Note that we provide an example in Table S1 of the supplementary material to illustrate the invariance property in Proposition 2.

Remark 1. The property in Proposition 2 does not imply that the ordering information of ordinal explanatory variables is sacrificed. First, an ordinal variable has an intrinsic ordering to its categories, and thus one cannot change the order of categories without changing the meaning of the variable. Second, the checkerboard copula regression is defined over (d-1)-dimensional hyperrectangles in $[0,1]^{d-1}$, $\prod_{k=1,k\neq j}^{d}[u^k_{i_k-1},u^k_{i_k}]$, which includes information on the order of an ordinal explanatory variable. For example, in Table 2(b) of Example 1, the checkerboard copula regression $r_{U_2|U_1}(u_1)$ was computed over an interval associated with the i_1-th category of X_1, $[u^1_{i_1-1},u^1_{i_1}]$ where $i_1 \in \{1,\dots,I_1\}$ and $u^1_{i_1}=\sum_{\ell_1=1}^{i_1} p_{+\ell_1+}$ is the cumulative probability at the i_1-th category of X_1.

2.4. Checkerboard copula regression association measure

Using the checkerboard copula regression proposed in Section 2.3, we propose the checkerboard copula regression based association measure for a multi-way contingency table with an ordinal response variable and categorical (ordinal or nominal) explanatory variables.

Definition 4. For the ordinal contingency table of $\boldsymbol{X}$, the checkerboard copula regression association measure of X_j on $\boldsymbol{X}_{-j}=(X_1,\dots,X_{j-1},X_{j+1},\dots,X_d)^\top$ is

$$\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} \equiv \frac{\mathrm{Var}\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})\}}{\mathrm{Var}(U_j)} = \frac{\mathrm{E}\left[\left\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})-1/2\right\}^2\right]}{1/12} = 12\sum_{\boldsymbol{i}_{-j}=\boldsymbol{1}_{-j}}^{\boldsymbol{I}_{-j}}\left(\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}} s^j_{i_j} - 1/2\right)^2 p_{i_1,\dots,+_j,\dots,i_d}, \quad (10)$$

where $j \in \{1,\dots,d\}$, and U_j and $\boldsymbol{U}_{-j} = (U_1,\dots,U_{j-1},U_{j+1},\dots,U_d)^\top$ are the random variables on $[0,1]^d$ associated with the checkerboard copula density $c^+(\boldsymbol{u})$ in (6).

In the following, we study properties of the proposed association measure in (10).

Proposition 3.

(i) $0 \leq \rho^2_{(\boldsymbol{X}_{-j}\to X_j)} \leq 12\sigma^2_{S_j} < 1$ *where* $\sigma^2_{S_j}$ *is the variance of* $S_j = E(U_j \mid X_j)$ *(a transformation of* X_j *via* U_j *in (7)), as given in Proposition 1(iii).*

(ii) *If* X_j *and* $\boldsymbol{X}_{-j}$ *are independent, then* $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 0$.

(iii) *If* $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 0$, *then the mean checkerboard copula score for* X_j *with respect to the conditional distribution of* X_j *given* $\boldsymbol{X}_{-j}$ *is constant over every combination of categories of* $\boldsymbol{X}_{-j}$.

(iv) $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 12\sigma^2_{S_j}$ *if and only if* $X_j = g(\boldsymbol{X}_{-j})$ *almost surely for some measurable function of* g.

(v) $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} < \mathrm{Var}\left\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})\right\}\Big/\sigma^2_{S_j}$.

(vi) *Let* $\tilde{X}_k=g_k(X_k)$, *where* g_k *is an injective function of* X_k *for* $k=1,\dots,j\text{-}1,j\text{+}1,\dots,d$, *and* $\widetilde{\boldsymbol{X}}_{-j} = (\tilde{X}_1, \dots, \tilde{X}_{j-1}, \tilde{X}_{j+1}, \dots, \tilde{X}_d)$. *Then,* $\rho^2_{(\widetilde{\boldsymbol{X}}_{-j}\to X_j)} = \rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$. *Particularly,* $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ *is invariant over the permutation on the categories of* X_k *in* $\boldsymbol{X}_{-j}$.

(vii) $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ *is invariant over the permutation on the categories of* X_j *only when* X_j *is binary.*

Proof. See Appendix. □

From the Proposition 3 (i)-(iv), we can see that the proposed measure $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ can identify linear/nonlinear relationship between a response variable X_j and explanatory variables $\boldsymbol{X}_{-j}$. The Proposition 3 (i) means that $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ ranges from 0 to $12\sigma^2_{S_j}$. According to Proposition 3 (iii) and (iv), the zero value of $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ means no contribution of explanatory variables $\boldsymbol{X}_{-j}$ to the construction of the checkerboard copula regression function, and $12\sigma^2_{S_j}$ is a sharp upper bound for $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$. The Proposition 3 (v) means that $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ is the lower bound on the average proportion of variance for X_j with respect to its checkerboard copula scores and its marginal distribution explained by the checkerboard copula regression. The Proposition 3 (vi) and (vii) imply that $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ can also be applied when any explanatory variables in $\boldsymbol{X}_{-j}$ are nominal and/or a response variable X_j is binary.

If one wishes to assess whether the association measured by $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ in (10) is strong or weak for the data at hand, it is necessary to take into account the fact that the upper bound of $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ depends on the marginal distribution of X_j. Thus, we propose the “scaled" version of the checkerboard copula regression based association measure,

$$\rho^{2\,*}_{(\boldsymbol{X}_{-j}\to X_j)} = \frac{\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}}{12\sigma^2_{S_j}}. \tag{11}$$

Note that the range of $\rho^{2\,*}_{(\boldsymbol{X}_{-j}\to X_j)}$ is between 0 and 1.

Example 1 (continued). Using the checkerboard copula regressions $r_{U_2|U_1}(u_1)$ and $r_{U_1|U_2}(u_2)$ obtained in Example 1 of Section 2.3, we calculate the proposed association measures in (10), their upper bounds and their corresponding scaled versions in (11): $(\rho^2_{(X_1\to X_2)},\ 12\sigma^2_{S_2},\ \rho^{2\,*}_{(X_1\to X_2)}) = (27/32, 27/32, 1)$ and $(\rho^2_{(X_2\to X_1)},\ 12\sigma^2_{S_1},\ \rho^{2\,*}_{(X_2\to X_1)}) = (0, 243/256, 0)$. $\rho^{2\,*}_{(X_1\to X_2)}$=1 implies that X_1 perfectly explains the variation in X_2 (induced by its checkerboard copula score and its marginal distribution) and this result stems from the observation that $r_{U_2|U_1}(u_1)$ equals one and only one of the checkerboard score of X_2 for each interval in U_1. Note that this result also supports the fact that X_2 is functionally dependent on X_1 with probability 1 given in Table 1 and also verifies Proposition 3 (iv). On the other hand, $\rho^{2\,*}_{(X_2\to X_1)}$=0 means that $r_{U_1|U_2}(u_2)$ is equal to $E(U_1) = 1/2$ for all values of u_2 due to the checkerboard copula score of X_1 and the conditional p.m.f. of X_1 given X_2 (as shown in the example of Section 2.3). Thus, X_2 has no contribution in explaining the variation in X_1 (arising from its score and its marginal distribution).

3. Estimation

We present estimators of the proposed checkerboard copula score, checkerboard copula regression and its association measure, and examine their asymptotic properties. The proofs of theorems are given in Appendix.

Let $\{n_{i_1,\ldots,i_d}\}$, $i_j \in \{1,\ldots,I_j\}$, $j \in \{1,\ldots,d\}$, denote counts in a d-way contingency table obtained by classifying $n(=\sum_{i_1=1}^{I_1}\cdots\sum_{i_d=1}^{I_d} n_{i_1,\ldots,i_d})$ cases according to categories of the d variables, $X_1,\ldots,X_d$. The marginal sums of i_j-th category in X_j are denoted as $n_{+i_j+} = \sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d} n_{i_1,\ldots,i_d}$, and $(d-1)$-variate marginal frequencies of $\boldsymbol{X}_{-j}$ are denoted as $n_{i_1,\ldots,+_j,\ldots,i_d} = \sum_{i_j=1}^{I_j} n_{i_1,\ldots,i_d}$. Then, estimators for $p_{i_1,\ldots,i_d}$, p_{+i_j+}, $p_{i_1,\ldots,+_j,\ldots,i_d}$, $p_{i_j|\boldsymbol{i}_{-j}}$ in (1)-(2) are given by

$$\hat{p}_{i_1,\ldots,i_d} = \frac{n_{i_1,\ldots,i_d}}{n},\ \ \hat{p}_{+i_j+} = \frac{n_{+i_j+}}{n},\ \ \hat{p}_{i_1,\ldots,+_j,\ldots,i_d} = \frac{n_{i_1,\ldots,+_j,\ldots,i_d}}{n},\ \ \hat{p}_{i_j|\boldsymbol{i}_{-j}} = \frac{\hat{p}_{i_1,\ldots,i_d}}{\hat{p}_{i_1,\ldots,+_j,\ldots,i_d}}.$$

Furthermore, the range of marginal c.d.f. of X_j, D_j in (3), is estimated by $\hat{D}_j = \{\hat{u}^j_0,\ldots,\hat{u}^j_{i_j},\ldots,\hat{u}^j_{I_j}\}$ with $\hat{u}^j_0 = 0$ and $\hat{u}^j_{i_j} = \sum_{k_j=1}^{i_j}\hat{p}_{+k_j+}$.

3.1. Estimation of checkerboard copula score and checkerboard copula regression

By utilizing the estimators above, the checkerboard copula score for X_j in (8), the variance of S_j and the checkerboard copula regression in (9) are estimated by $\{\hat{s}_1^j, \ldots, \hat{s}_{I_j}^j\}$ with $\hat{s}_{i_j}^j = (\hat{u}_{i_j-1}^j + \hat{u}_{i_j}^j)/2$, $\hat{\sigma}^2_{\hat{S}_j} = \sum_{i_j=1}^{I_j} \hat{u}_{i_j-1}^j \hat{u}_{i_j}^j \hat{p}_{+i_j+}/4$, and, for $k \in \{1, \ldots, j-1, j+1, \ldots, d\}$,

$$\hat{r}_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j}) = \sum_{i_j=1}^{I_j} \hat{p}_{i_j|\boldsymbol{i}_{-j}} \hat{s}_{i_j}^j \quad \text{for } \hat{u}_{i_k-1}^k < u_k \leq \hat{u}_{i_k}^k. \tag{12}$$

Using the estimated checkerboard copula regression and the prediction procedure in Section 2.3, we can obtain the predicted category of a response variable for each combination of categories of explanatory variables. That is, for a given combination categories of the $(d-1)$ explanatory variables $\boldsymbol{X}_{-j}$, we find the corresponding $\hat{\boldsymbol{u}}^*_{-j}$ from the estimated ranges $\prod_{k=1,k\neq j}^{d} \hat{D}_k$ associated with $\boldsymbol{X}_{-j}$ and then obtain the estimated value of the checkerboard copula regression in (12), $\hat{u}_j^* = \hat{r}_{U_j|\boldsymbol{U}_{-j}}(\hat{\boldsymbol{u}}^*_{-j})$. From the estimated range $\hat{D}_j$ of a response variable X_j, we get i_j^* and $\hat{u}_{i_j^*}^j$ such that $\hat{u}_{i_j^*-1}^j < \hat{u}_j^* \leq \hat{u}_{i_j^*}^j$ and finally obtain the predicted category of X_j, $\hat{x}_{i_j^*}^j$.

Theorem 1 establishes the asymptotic distribution for the estimator of the proposed regression in (12).

Theorem 1. *Let $\hat{r}_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})$ be the estimator for the checkerboard copula regression $r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})$ in (12). Then,*

$$\sqrt{n}\left(\hat{r}_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j}) - r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})\right) \xrightarrow{D} \mathcal{N}_1\left(0, \nabla h_{r_j}(\boldsymbol{p})^\top \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right) \nabla h_{r_j}(\boldsymbol{p})\right),$$

where $\boldsymbol{p}$ is the vectorized version of the joint p.m.f. P in the d-way contingency table, i.e., $\boldsymbol{p} = (p_{1,1,\ldots,1}, \ldots, p_{I_1,1,\ldots,1}, \ldots, p_{I_1,\ldots,I_{d-1},1}, \ldots, p_{I_1,\ldots,I_{d-1},I_d})^\top$, $\nabla h_{r_j}(\boldsymbol{p}) = \left(\partial r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})/\partial p_{1,1,\ldots,1}, \ldots, \partial r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})/\partial p_{I_1,I_2,\ldots,I_d}\right)^\top$, and $\partial r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})/\partial p_{k_1,k_2,\ldots,k_d}$ is given in (A.1) of Appendix.

Proof. See Appendix. □

Using Theorem 1, one can obtain the asymptotic confidence interval of the estimated checkerboard copula regression $\hat{r}_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})$ for a given combination categories of explanatory variables, denoted as $[\hat{u}_j^L, \hat{u}_j^U]$, and apply these lower and upper limits to the proposed prediction method for obtaining the corresponding range of the estimated category of a response variable X_j. Alternatively, to better quantify uncertainty of the predicted category of X_j, we propose using bootstrap resampling [8]: simulate B bootstrap samples of size n from the saturated log-linear model fitted to the observed multi-way contingency table and compute the proportion of each category of X_j obtained by the estimated checkerboard copula regression.

3.2. Estimation of the checkerboard copula regression based association measure

The estimators for the checkerboard copula regression based association measure $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$, $j \in \{1, \ldots, d\}$, of (10) and its scaled version in (11) are given below:

$$\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)} = 12 \sum_{\boldsymbol{i}_{-j}=\boldsymbol{1}_{-j}}^{\boldsymbol{I}_{-j}} \left(\sum_{i_j=1}^{I_j} \hat{p}_{i_j|\boldsymbol{i}_{-j}} \hat{s}_{i_j}^j - \frac{1}{2}\right)^2 \hat{p}_{i_1,\ldots,+_j,\ldots,i_d}, \qquad \hat{\rho}^{2\,*}_{(\boldsymbol{X}_{-j}\to X_j)} = \frac{\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)}}{12\hat{\sigma}^2_{S_j}}. \tag{13}$$

Once we compute the value of $\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)}$ in (13), it is nature to ask if $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ equals zero. Here, we propose using the permutation test [8, 15, 36] for a hypothesis testing $H_0 : \rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 0$. In the data analysis, we will report the estimated permutation p-value, denoted as p_{est}, and its associated relative error $1/\sqrt{R \times p_{est}}$ where R is the number of permutation samples.

When the proposed association measure in (10) is statistically greater than zero, Theorem 2 below presents the asymptotic distribution of the estimator in (13).

Theorem 2. *Assume that* $0 < \rho^2_{(\boldsymbol{X}_{-j}\to X_j)} < 12\sigma^2_{S_j}$. *Let* $\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)}$ *denote the estimator in* (13) *for* $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$. *Then,*

$$\sqrt{n}\left(\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)} - \rho^2_{(\boldsymbol{X}_{-j}\to X_j)}\right) \xrightarrow{D} \mathcal{N}_1\left(0, \nabla h_{X_j}(\boldsymbol{p})^\top \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right) \nabla h_{X_j}(\boldsymbol{p})\right),$$

where $\boldsymbol{p} = (p_{1,1,\ldots,1}, \ldots, p_{I_1,1,\ldots,1}, \ldots, p_{I_1,\ldots,I_{d-1},1}, \ldots, p_{I_1,\ldots,I_{d-1},I_d})^\top$, $\nabla h_{X_j}(\boldsymbol{p}) = \left(\partial\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}/\partial p_{1,1,\ldots,1},\right.$ $\left.\ldots\ \partial\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}/\partial p_{I_1,I_2,\ldots,I_d}\right)^\top$, *and* $\partial\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}/\partial p_{k_1,k_2,\ldots,k_d}$ *is given in (A.2) of Appendix.*

Proof. See Appendix. □

In the real data analysis, we will obtain the estimator for the asymptotic variance of $\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)}$ in Theorem 2 by the plug-in principle and then construct the asymptotic confidence interval of $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$.

Example 1 (continued). To illustrate the estimation of the proposed measure, we consider a two-way ordinal table with the counts $\{n_{ij}\}$ where n_{ij}=80$\times p_{ij}$ and p_{ij} is given in Table 1. We compute the proposed association measures in (13) : $(\hat{\rho}^2_{(X_1\to X_2)}, \hat{\rho}^{2\,*}_{(X_1\to X_2)})$=(27/32, 1) and $(\hat{\rho}^2_{(X_2\to X_1)}, \hat{\rho}^{2\,*}_{(X_2\to X_1)})$=(0, 0). Using the permutation test with 10^6 permutations, the p-values (relative errors) for $H_0 : \rho^2_{(X_1\to X_2)}$=0 and $H_0 : \rho^2_{(X_2\to X_1)}$=0 are $<$ 0.001 (0.0316) and $>$ 0.999 (0.001), respectively.

We also predict the category of X_2 given a category of X_1 using the estimated checkerboard copula regression as follows : $\hat{f}_{X_2|X_1} = (x_3^2, x_2^2, x_1^2, x_2^2, x_3^2)$ for $X_1 = (x_1^1, x_2^1, x_3^1, x_4^1, x_5^1)$, respectively. Using 1000 bootstrap resampling, we quantify uncertainty of the predicted category of X_2, and it turns out that the proportion that the category of X_2 predicted by the estimated copula regression for each category of X_1 selected by the bootstrap method is 100%. These results support that the proposed method successfully identified the functional dependence in Table 1. For detailed results, please see Table S2 of the supplementary material.

4. Numerical examples

In this section, the performance of the proposed methods is assessed with simulation studies and real data.

4.1. Simulation study

We conduct a simulation study to evaluate the proposed association measure in (13) under different types of association scenarios in a $I_1 \times I_2$ ordinal table with the explanatory variable X_1 and the response variable X_2. We consider the five simulation factors : (i) the type of association between X_1 and X_2 (linear pattern, monotone nonlinear pattern and nonmonotone nonlinear pattern), (ii) the magnitude of association, (iii) the marginal distributions of X_1 and X_2, (iv) the sample size n=(500,1000,2000), and (v) the table size (I_1=(3,5) and I_2=(3,5)). To simulate the contingency table with the intended association, we employed the proportional odds cumulative logit model (CLM), one of the widely used ordinal response models :

$$\begin{aligned}
\text{logit}[P(X_2 \le i_2|X_1)] &= \alpha_{i_2}, \ \text{for no association,}\\
&= \alpha_{i_2} - \beta X_1, \ \text{for linear and monotone nonlinear pattern,}\\
&= \alpha_{i_2} - \beta_1 X_1 - \beta_2 X_1^2, \ \text{for nonmonotone nonlinear pattern,}
\end{aligned}$$

where $i_2 \in \{1, \ldots, I_2 - 1\}$, β, β_1 and β_2 are regression coefficients, and α_{i_2}s are intercepts satisfying $\alpha_1 < \cdots < \alpha_{I_2-1}$. Note that regression coefficients and intercepts will determine the second and third simulation factors, the magnitude of association and the marginal distribution of X_2, respectively. Due to the limited space, the detailed information on the marginal distributions of X_1, X_2, and the magnitude of association (the parameter values of the proportional odds CLM) used in the simulation is given in Section S2.1 of the supplementary material.

For each combination of simulation factors, we simulated 1000 tables, and calculated $\hat{\rho}^2_{X_1\to X_2}$ in (13). We presented simulation results using the boxplots. For the sake of space, we provide them in Section S2.3 of the supplementary material. Here, we briefly summarize the simulation results. First, the variability of the distribution for $\hat{\rho}^2_{X_1\to X_2}$ decreases as n increases and the center of the distribution is stable over different sample sizes, regardless of the table size and association pattern. Second, for a case of no association, the values of $\hat{\rho}^2_{X_1\to X_2}$ are very close to zero (their range is less than 0.05), regardless of the table size and the sample size. The sampling distributions of $\hat{\rho}^2_{X_1 \to X_2}$ are right-skewed, but the amount of skewness decreases as n increases. Third, for tables with linear and monotone nonlinear pattern, as the magnitude of association increases from weak to very strong, the magnitude of $\hat{\rho}^2_{(X_1 \to X_2)}$ also increases. Last, for tables with linear and (monotone/nonmonotone) nonlinear patterns, the proposed measure $\hat{\rho}^2_{(X_1 \to X_2)}$ increases as the table size increases and the increment of I_1 leads to larger value of $\hat{\rho}^2_{(X_1 \to X_2)}$ than the increment of I_2 : $\hat{\rho}^2_{(X_1 \to X_2)}$ (3×3 table) < $\hat{\rho}^2_{(X_1 \to X_2)}$ (3×5 table) < $\hat{\rho}^2_{(X_1 \to X_2)}$ (5×3 table) < $\hat{\rho}^2_{(X_1 \to X_2)}$ (5×5 table).

Remark 2. We also carried out a simulation study to assess the performance of the proposed association measure in a two-contingency table with a nominal explanatory variable and an ordinal response variable. According to the simulation results obtained from various combinations of simulation factors (sample size, table size and magnitude of association), we found that the proposed measure performed very similarly as in the simulation study of Section 4.1 (i.e., the table with an ordinal explanatory variable). For more details, see Sections S2.2 and S2.3.2 of the supplementary material.

4.2. Real data analysis

In this section, we demonstrate the utility of the proposed methods using a multidimensional contingency table data. Table 3 is a 2 × 3 × 2 × 6 table from a study of 101 patients suffering from back pain [4].

Table 3: 101 back pain patients cross classified by three prognostic variables (X_1, X_2, X_3) and progress on pain levels X_4.

			Back Pain (X_4)					
Length of previous attack (X_1)	Pain change (X_2)	Lordosis (X_3)	W	S	SI	MODI	MARI	CR
short	better	absent/decreasing	0	1	0	0	2	4
short	better	present/increasing	0	0	0	1	3	0
short	same	absent/decreasing	0	2	3	0	6	4
short	same	present/increasing	0	1	0	2	0	1
short	worse	absent/decreasing	0	0	0	0	2	2
short	worse	present/increasing	0	0	1	1	3	0
long	better	absent/decreasing	0	0	3	0	1	2
long	better	present/increasing	0	1	0	0	3	0
long	same	absent/decreasing	0	3	4	5	6	2
long	same	present/increasing	1	4	4	3	0	1
long	worse	absent/decreasing	2	2	1	5	2	0
long	worse	present/increasing	2	0	2	3	0	0

Patients suffering from back pain received a treatment and after three weeks, their progress (X_4) was assessed on six ordered categories : Worse (W), Same (S), Slight Improvement (SI), Moderate improvement (MODI), Marked improvement (MARI), Complete Relief (CR). The three explanatory variables were measured at the beginning of the treatment period: X_1=length of previous attack (two levels : S=Short, L=Long), X_2=pain change (three levels: B=getting Better, S=Same, W=Worse), and X_3=lordosis (two levels: AD=Absent/Decreasing, PI=Present/Increasing).

Table 3 was analyzed by Anderson [4] using the stereotype model that is nested between the proportional odds adjacent-categories logit model and the general adjacent-categories logit model. The analyses in [4, 49] showed that the three explanatory variables X_1, X_2, X_3 in the main effect stereotype model (with no interactions among the explanatory variables) are statistically significant and patients get worse as the level of each of the three explanatory variables increases.

In order to assess the explanatory power of three explanatory variables on back pain assessment X_4 in a model-free fashion, we first compute the checkerboard copula regression based association measure in (13), $\hat{\rho}^2_{(X_1,X_2,X_3 \to X_4)}$. The value of $\hat{\rho}^2_{(X_1,X_2,X_3 \to X_4)}$ is 0.257 and the corresponding 95% confidence interval is (0.1099, 0.405). Note that the permutation test p-value (its relative error) for $H_0 : \rho^2_{X_1,X_2,X_3 \to X_4} = 0$ using 10^6 permutations were 0.0018 (0.0235). The estimated association measure means that the lower bound on the average proportion of variance for the checkerboard copula score of the back pain assessment (X_4) explained by the checkerboard copula regression using three explanatory variables is 25.7%. Since the estimated upper bound of $\rho^2_{(X_1,X_2,X_3 \to X_4)}$ is 0.9585 (=$12\hat{\sigma}^2_{S_4}$), the scaled association measure is $\hat{\rho}^{2\,*}_{(X_1,X_2,X_3 \to X_4)}$=0.257/0.9585 =0.2681. From the results above, we confirm that three explanatory variables (X_1, X_2, X_3) are statistically important explanatory variables for the back pain assessment of patients.

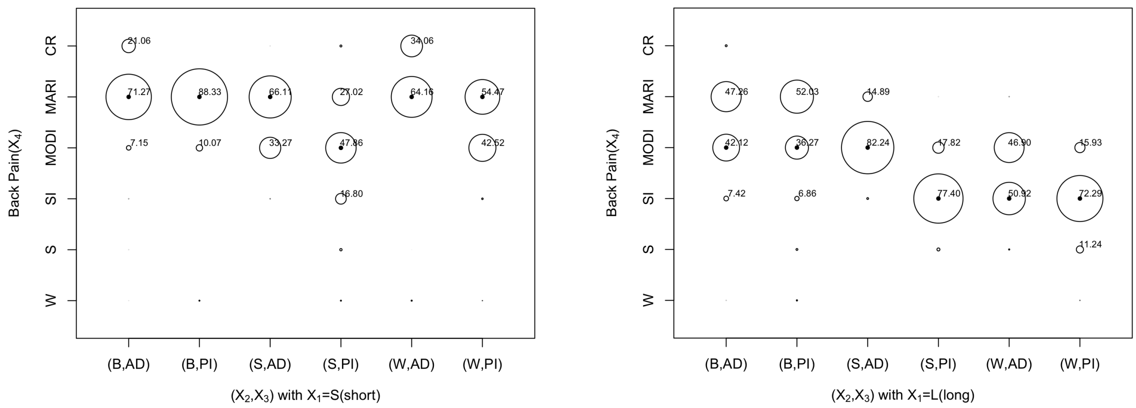

Fig. 1: Predicted category of X_4 (dark dots) by the checkerboard copula regression for each combination of categories of X_1, X_2, X_3 and the proportion of each category of X_4 (circles whose sizes are proportional to the numbers) estimated by the copula regression in the 1000 bootstrap resampling.

To identify the potential interactions among three explanatory variables, we present in Fig.1 the prediction of the category of X_4 (dark dots) for a combination of categories of the three explanatory variables using the proposed prediction method. Note that the size of the circles represents the proportion of each category of X_4 obtained from 1000 bootstrap resampling and we report the proportions larger than 5%.

From Fig.1, we see potential interactions between three explanatory variables, particularly interaction between the length of previous attack (X_1) and pain change (X_2). The left plot in Fig.1 shows that for the patients whose length of previous attack is short (X_1=S), their assessment tends to be the marked improvement (X_4=MARI) regardless of pain change (X_2) and lordosis (X_3) except when pain change is same and lordosis is present/increasing, i.e., (X_2, X_3)=(S,PI). On the other hand, the right plot in Fig.1 shows that when the previous attack is long (X_1=L), the back pain tends to be the markedly/moderately improved or moderately/slightly improved, depending on pain change (X_2) and lordosis (X_3). When the pain change is “getting better" (X_2=B), the back pain is markedly/moderately improved (X_4=MARI/MODI) irrespective of the status of lordosis (X_3). When there is no change in pain (X_2=S), the back pain is improved from “slightly" (SI) to “moderately" (MODI) as the status of lordosis (X_3) changes from “present/increasing" (PI) to “absent/decreasing" (AD). For the “worse" pain change (X_2=W), the back pain stays at “slightly" (SI).

5. Discussion

In this paper, we proposed a model-free exploratory tool to identify and quantify the regression dependence between an ordinal response variable and a set of categorical (ordinal or nominal) explanatory variables in a multi-way contingency table. The performance and utility of the proposed method, consisting of the checkerboard copula regression and its association measure, were demonstrated through simulation studies and a real data analysis. The proposed association measure represents the magnitude of the explanatory power of explanatory variables in the checkerboard copula regression. However, this does not indicate either the goodness of fit of parametric ordinal response models with the same explanatory variables or their statistical significance.

In Section 2.3, we proposed the method to predict a category of an ordinal response variable. As illustrated in the real data analysis, the prediction method is useful for delineating the regression dependence identified by the checkerboard copula regression by visualizing the patterns of change in the predicted category of a response variable across combinations of categories of explanatory variables. Note that the prediction method was developed to serve the purpose of exploratory modeling, not predictive modeling focusing on predictive power.

As a reviewer pointed out, the predictive accuracy of the proposed prediction method is also an important topic to investigate. We computed the prediction accuracy of the proposed method using leave-one-out cross-validation (LOOCV) for two data sets, the artificial data (n=80) in Example 1 of Section 2.1 and the back pain data (n=101) in

Section 4.2. For Example 1 where the estimated proposed association measure is $\hat{\rho}^{2\,*}_{(X_1 \to X_2)}$=1, the LOOCV prediction accuracy (=1-the error rate) was 1. On the other hand, the LOOCV prediction accuracy was 0.2376 for the back pain data with $\hat{\rho}^{2\,*}_{(X_1, X_2, X_3 \to X_4)}$ =0.268. Note that the accuracy for 101 training data sets (each of size 100) in the back pain data ranges from 0.28 to 0.35. This limited investigation indicates that the performance of the proposed prediction method may be related to several factors, including (i) the sample size relative to the size of the contingency table (i.e., both the number of variables and the number of categories of each variable), and (ii) the magnitude of the regression dependence (i.e., the quality of the explanatory variables and their association with an ordinal response variable).

If prediction is the main purpose of statistical modeling for multidimensional contingency tables with an ordinal response variable, the proposed association measure would still be a useful index. This is because a low value of $\rho^{2\,*}_{(\mathbf{X}_{-j} \to X_j)}$ indicates that the checkerboard copula regression based on given explanatory variables may not account for much of the variance associated with an ordinal response variable and it may introduce more error in the prediction. That is, a low value of $\rho^{2\,*}_{(\mathbf{X}_{-j} \to X_j)}$ can warn of low prediction accuracy. While a high value of the proposed association measure appears to be required for precise predictions, we believe that it is not sufficient by itself. A valuable future work would be a thorough investigation of the predictive power of the proposed prediction method to facilitate the predictive modeling process.

Some other future directions that we are interested in are as follows: First, we plan to develop the decomposition of the proposed association measure $\rho^{2\,*}_{(\mathbf{X}_{-j} \to X_j)}$ to quantify the contributions of every subset of explanatory variables to $\rho^{2\,*}_{(\mathbf{X}_{-j} \to X_j)}$, and investigate its application to the variable selection problem for a multi-way contingency table with an ordinal response variable. Second, we will propose an asymptotic approach for testing the hypothesis of no regression dependence using the proposed association measure.

Supplementary material

The supplementary material includes the additional results for Example 1, and the detailed information on the simulation design and simulation results of Section 4.1.

Acknowledgments

The authors would like to thank the Editor, Associate Editor and two anonymous referees for their insightful and constructive comments and suggestions, which helped us greatly in revising this work.

Appendix Proofs

Proof of Proposition 1. The proof of properties of the checkerboard copula scores is given below.

(i) : As $s^j_{i_j} - s^j_{i_j-1} = \left(u^j_{i_j} + u^j_{i_j-1}\right)/2 - \left(u^j_{i_j-1} + u^j_{i_j-2}\right)/2 = \left(p_{+i_j+} + p_{+i_j-1+}\right)/2 > 0$, we have $s^j_{i_j} > s^j_{i_j-1}$. In addition, we have $s^j_1 = \left(u^j_1 + u^j_0\right)/2 = u^j_1/2 > 0$ and $s^j_{I_j} = \left(u^j_{I_j} + u^j_{I_j-1}\right)/2 = \left(1 + u^j_{I_j-1}\right)/2 < 1$.

(ii) : $\mathrm{E}\left(U_j | X_j = x_{i_j}\right) = F_j(x_{i_j-1}) + (F_j(x_{i_j}) - F_j(x_{i_j-1}))E(V_j) = u^j_{i_j-1} + (u^j_{i_j} - u^j_{i_j-1}) \times \frac{1}{2} = s^j_{i_j}$.

(iii) :

$$\mu_{S_j} = \sum_{i_j=1}^{I_j} s^j_{i_j} p_{+i_j+} = \sum_{i_j=1}^{I_j} \left(\sum_{i'_j=1}^{i_j-1} p_{+i'_j+} + \frac{p_{+i_j+}}{2} \right) p_{+i_j+} = \frac{2 \sum\limits_{i_j=1}^{I_j} \sum\limits_{i'_j=1}^{i_j-1} p_{+i'_j+} p_{+i_j+} + \sum\limits_{i_j=1}^{I_j} p^2_{+i_j+}}{2} = \frac{\left(\sum\limits_{i_j=1}^{I_j} p_{+i_j+} \right)^2}{2} = \frac{1}{2}.$$

$$\sigma^2_{S_j} = \sum_{i_j=1}^{I_j} (s^j_{i_j} - \frac{1}{2})^2 p_{+i_j+} = \frac{1}{4} \sum_{i_j=1}^{I_j} (u^j_{i_j} + u^j_{i_{j-1}} - 1)^2 p_{+i_j+} = \frac{1}{4} \left(\sum_{i_j < i'_j} p_{+i_j+} p_{+i'_j+} u^j_{i_j} \right) = \frac{1}{4} \left(\sum_{i_j=1}^{I_j} u^j_{i_j-1} u^j_{i_j} p_{+i_j+} \right).$$

To maximize $\sigma^2_{S_j}$ under the constraint $\sum\limits_{i_j=1} p_{+i_j+} = 1$, we need to find the stationary point for the Lagrange function

$$f(p_{+1+}, \ldots, p_{+I_j+}, \lambda) = \sigma^2_{S_j} - \lambda (\sum_{i_j=1} p_{+i_j+} - 1).$$

Setting the first derivative of the function above with respect to p_{+i_j+} equal to zero implies that p_{+i_j+} are all same, and the constraint $\sum_{i_j=1} p_{+i_j+} = 1$ implies $\sigma^2_{S_j}$ is maximized at $p_{+i_j+} = 1/I_j$. □

Proof of Proposition 2. For $k \in \{1, \cdots, j-1, j+1, \cdots, d\}$, let g_k be an injective function of X_k and $\tilde{X}_k = g_k(X_k)$. Assume that $\tilde{p}_{i_1,\ldots,i_d}$ is the joint p.m.f. of $(X_1, \ldots, X_{k-1}, \tilde{X}_k, X_{k+1}, \ldots, X_d)^\top$ and C^s_k is the corresponding subcopula defined on $D_1 \times \cdots \times D_{k-1} \times \tilde{D}_k \times D_{k+1} \times \cdots \times D_d$, where $\tilde{D}_k = \{\tilde{u}^k_0, \tilde{u}^k_1, \ldots, \tilde{u}^k_{I_k}\}$ is the range of the marginal distribution of $\tilde{X}_k$. Suppose that C^+_k is the checkerboard copula of C^s_k and denote $\widetilde{\boldsymbol{U}} = (U_1, \ldots, U_{k-1}, \tilde{U}_k, U_{k+1}, \ldots, U_d)^\top$ be the uniform random vector associated with C^+_k.

Now, let $\boldsymbol{X}_{-j} = \boldsymbol{x}^*_{-j}$ with $x^*_{i_k}$ be the i_k-th category of X_k and it corresponds to the i'_k-th category $g_k(x^*_{i_k})$ of $\tilde{X}_k$. Let $\boldsymbol{u}^*_{-j}$ be the corresponding elements of $\boldsymbol{x}^*_{-j}$ in $\prod_{\ell=1,\ell\neq j}^d D_\ell$, and $\tilde{\boldsymbol{u}}^*_{-j}$ be the corresponding elements of $\tilde{\boldsymbol{x}}^*_{-j}$ in $D_1 \times \cdots \times D_{k-1} \times \tilde{D}_k \times D_{k+1} \times \cdots \times D_d$. Since g_k is an injective function of X_k, there is an one-to-one corresponds between $\tilde{p}_{i_1,\ldots,i_d}$ and $p_{i_1,\ldots,i_d}$. This implies that for each $\boldsymbol{i} = (i_1, \ldots, i_d)$, there exists $\boldsymbol{i'} = (i_1, \ldots, i_{k-1}, i'_k, i_{k+1}, \ldots, i_d)$ such that $p_{i_j|\boldsymbol{i'}_{-j}} = p_{i_j|\boldsymbol{i}_{-j}}$ for each $i_j \in \{1, \ldots, I_j\}$ where $\boldsymbol{i'}_{-j} = (i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_{k-1}, i'_k, i_{k+1}, \ldots, i_d)$ and $\boldsymbol{i}_{-j} = (i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_d)$. Therefore, we have

$$r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}^*{}_{-j}) = \sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}} \left(\frac{u^j_{i_j} + u^j_{i_j-1}}{2} \right) = \sum_{i_j=1}^{I_j} \tilde{p}_{i_j|\boldsymbol{i'}_{-j}} \left(\frac{u^j_{i_j} + u^j_{i_j-1}}{2} \right) = r_{U_j|\tilde{\boldsymbol{U}}_{-j}}(\tilde{\boldsymbol{u}}^*_{-j}).$$

Two regression functions $r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}^*_{-j})$ and $r_{U_j|\tilde{\boldsymbol{U}}_{-j}}(\tilde{\boldsymbol{u}}^*_{-j})$ will give the same value of u^*_j in D_j (the range of the marginal distribution of X_j) and in turn provide the same estimation category of X_j. □

Proof of Proposition 3. The following lemma is needed to prove Proposition 3.

Lemma 1. *If the subcopula C^s in* (4) *is an independent subcopula over $\prod_{j=1}^d D_j$ in that $C^s(\boldsymbol{u}) = \prod_{j=1}^d u_j$, the checkerboard copula C^+ in* (5) *is also an independent copula.*

Proof. For any $\boldsymbol{u} \in [0,1]^d$, the checkerboard copula in (5) is

$$C^+(\boldsymbol{u}) \quad = \quad \sum_{S \subseteq \{1,\ldots,d\}} \lambda_S(\boldsymbol{u}) C^s(u_{s_1}, \ldots, u_{s_d}) = \sum_{S \subseteq \{1,\ldots,d\}} \lambda_S(\boldsymbol{u}) \prod_{j=1}^d u_{s_j} = \prod_{j=1}^d u_j,$$

where the last equality is due to $u^l_j + \lambda(u_j)(u^u_j - u^l_j) = u_j$ for each j. Therefore, C^+ is an independent copula. □

We turn to the proof for the Proposition 3.

(i) By the variance decomposition formula $\mathrm{Var}(U_j) = \mathrm{E}\{\mathrm{Var}(U_j\ |\boldsymbol{U}_{-j})\} + \mathrm{Var}\{\mathrm{E}(U_j|\boldsymbol{U}_{-j})\} = \mathrm{E}\{\mathrm{Var}(U_j|\boldsymbol{U}_{-j})\}$ $+\mathrm{Var}\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})\}$, we can rewrite the association measure of X_j on $\boldsymbol{X}_{-j}$ as

$$\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 1 - \frac{\mathrm{E}\{\mathrm{Var}(U_j|\boldsymbol{U}_{-j})\}}{\mathrm{Var}(U_j)} = 1 - \frac{\mathrm{E}[\{U_j - \mathrm{E}(U_j|\boldsymbol{U}_{-j})\}^2]}{\mathrm{Var}(U_j)}.$$

Therefore, $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$ is between 0 and 1.

In addition, the following inequality holds for the numerator of $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$:

$$\mathrm{Var}\Big\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})\Big\} = \mathrm{E}\Big\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j}) - 1/2\Big\}^2 = \sum_{\boldsymbol{i}_{-j}=\boldsymbol{1}_{-j}}^{\boldsymbol{I}_{-j}} \left(\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}} \frac{u^j_{i_j} + u^j_{i_j-1}}{2} - 1/2 \right)^2 p_{i_1,\ldots,+_j,\ldots,i_d} \le \sigma^2_{S_j}.$$

By dividing $1/12$ from both side of the above inequality, we have $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} \le 12\sigma^2_{S_j}$. From the Proposition 1 (iii), we know the maximum value of $\sigma^2_{S_j}$ is $(1 - 1/I^2_j)/12$ which is less than 1/12. Thus, $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} \le 12\sigma^2_{S_j} < 1$.

(ii) If the discrete random variables X_j and $\boldsymbol{X}_{-j}$ are independent, and from the above Lemma, the checkerboard copula C^+ of X_j and $\boldsymbol{X}_{-j}$ is also independent. The checkerboard copula regression function $r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j}) = \mathrm{E}(U_j)$ which is a constant function of $\boldsymbol{u}_{-j}$. Thus, $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 0$.

(iii) $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 0$ implies $r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j}) = 1/2$ almost surely, which means that the mean checkerboard copula score for the categories of X_j with respect to the conditional distribution of X_j given $\boldsymbol{U}_{-j}$ is constant.
(iv) Since $\mathrm{Var}_{\boldsymbol{X}_{-j}}\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})\} = \sigma^2_{S_j} - \mathrm{E}_{\boldsymbol{X}_{-j}}\{\mathrm{Var}(S_j|\boldsymbol{S}_{-j})\}^2 = \sigma^2_{S_j} - \mathrm{E}_{\boldsymbol{X}}\{S_j - E(S_j|\boldsymbol{S}_{-j})\}^2$, $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 12\sigma^2_{S_j}$ if and only if $\mathrm{E}_{\boldsymbol{X}}[\{S_j - \mathrm{E}(S_j|\boldsymbol{S}_{-j})\}^2] = 0$, which is also equivalent to $S_j = \mathrm{E}_{X_j}(S_j|\boldsymbol{S}_{-j})$ almost surely. Here $\boldsymbol{S} = (S_1,\ldots,S_d)^\top$ denotes the checkboard scores in (8) equipped with the probability distribution of the ordinal random vector $\boldsymbol{X}$.
(v) By Proposition 1 (c), we know that the maximum value of $\sigma^2_{S_j}$ is $(1-1/I_j^2)/12$ and it is less than $1/12$. Thus, the property holds.
(vi) It is enough to show the part (f) holds for one X_k. Let g_k be an injective function of X_k, and $\tilde{X}_k = g_k(X_k)$. By following the same notations and argument as in the proof for Proposition 2, we know for each $\boldsymbol{i}$, there exist $\boldsymbol{i}'$ such that $p_{i_j|\boldsymbol{i}'_{-j}} = p_{i_j|\boldsymbol{i}_{-j}}$ and $p_{\boldsymbol{i}_{-j}} = \tilde{p}_{\boldsymbol{i}'_{-j}}$. Thus we have,

$$\mathrm{E}\{r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{U}_{-j})^2\} = \sum_{\boldsymbol{i}_{-j}}\left\{\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}}\frac{u^j_{i_j}+u^j_{i_j-1}}{2}\right\}^2 p_{\boldsymbol{i}_{-j}} = \sum_{\boldsymbol{i}'_{-j}}\left\{\sum_{i_j=1}^{I_j} \tilde{p}_{i_j|\boldsymbol{i}'_{-j}}\frac{u^j_{i_j}+u^j_{i_j-1}}{2}\right\}^2 \tilde{p}_{\boldsymbol{i}'_{-j}} = \mathrm{E}\{r_{U_j|\widetilde{\boldsymbol{U}}_{-j}}(\widetilde{\boldsymbol{U}}_{-j})^2\},$$

which in turn implies that $\rho^2_{(\widetilde{\boldsymbol{X}}_{-j}\to X_j)} = \rho^2_{(\boldsymbol{X}_{-j}\to X_j)}$.
(vii) Let X_j be a binary variable with two categories $\{x^j_1, x^j_2\}$ and the corresponding p.m.f., $\Pr(X_j = x^j_1) = p_{+1+}$ and $\Pr(X_j = x^j_2) = p_{+2+}$. The range of the marginal distribution of X_j is $D_j = \{0, u^j_1, 1\}$ where $u^j_1 = p_{+1+}$. Now, assume X'_j is the binary variable with two categories $\{x^j_2, x^j_1\}$ so that the range of the marginal distribution of X'_j is $D'_j = \{0, u'^j_1, 1\}$, where $u'^j_1 = p_{+2+} = 1 - u^j_1$. For each combination category $\boldsymbol{i}_{-j}$ of $\boldsymbol{X}_{-j}$, we have $p_{i_j=1|\boldsymbol{i}_{-j}}u^j_1 + p_{i_j=2|\boldsymbol{i}_{-j}}(1+u^j_1) - 1 = -\left(p_{i_j=2|\boldsymbol{i}_{-j}}u'^j_1 + p_{i_j=1|\boldsymbol{i}_{-j}}(1+u'^j_1) - 1\right)$. Thus,

$$\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = 3\sum_{\boldsymbol{i}_{-j}=\boldsymbol{1}_{-j}}^{\boldsymbol{I}_{-j}}\left(p_{i_j=2|\boldsymbol{i}_{-j}}u'^j_1 + p_{i_j=1|\boldsymbol{i}_{-j}}(1+u'^j_1) - 1\right)^2 p_{i_1,\ldots,+_j,\ldots,i_d} = \rho^2_{(\boldsymbol{X}_{-j}\to X'_j)}. \qquad \square$$

Proof of Theorem 1 and Theorem 2. We denote the relative frequencies of the $I_1\times\cdots\times I_d$ contingency table by $\hat{P} = \{\hat{p}_{\boldsymbol{i}} = n_{\boldsymbol{i}}/n \mid \boldsymbol{i}\in\prod_{j=1}^d \mathbb{I}_j\}$. Define the estimator $\hat{\boldsymbol{p}} = \mathrm{Vec}(\hat{P})$ to be the vector of length I for $\boldsymbol{p} = \mathrm{Vec}(P)$, where $I = \prod_{j=1}^d I_j$, and denote the expectation and the covariance matrix of $\hat{\boldsymbol{p}}$ to be $\mathrm{E}(\hat{\boldsymbol{p}}) = \boldsymbol{p}$ and $\mathrm{Cov}\,(\hat{\boldsymbol{p}})$. Let $(i_1,\ldots,i_d)^\top$ and $(i'_1,\ldots,i'_d)^\top$ be two index vectors with $i_j \neq i'_j$ for some $j\in\{1,\ldots,d\}$. Then $n\hat{p}_{i_1,\ldots,i_d}$ and $n\hat{p}_{i'_1,\ldots,i'_d}$ follow a trinomial distribution with parameters $(n,(p_{i_1,\ldots,i_d}, p_{i'_1,\ldots,i'_d}, 1-p_{i_1,\ldots,i_d}-p_{i'_1,\ldots,i'_d})^\top)$, and $\mathrm{Cov}(n\hat{p}_{i_1,\ldots,i_d}, n\hat{p}_{i'_1,\ldots,i'_d}) = -np_{i_1,\ldots,i_d}p_{i'_1,\ldots,i'_d}$. Thus, $\mathrm{Cov}\,(\hat{\boldsymbol{p}}) = \{\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\}/n$ and

$$\sqrt{n}\,(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{D} \mathcal{N}_I\left(\boldsymbol{0}, \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right)\right),$$

where $\mathcal{N}_I(\boldsymbol{\mu},\Sigma)$ denotes the multivariate normal distribution of size I with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

The regression function $r_{U_j|\boldsymbol{U}_{-j}} = h_{r_j}(\boldsymbol{p})$ and the association measure $\rho^2_{(\boldsymbol{X}_{-j}\to X_j)} = h_{X_j}(\boldsymbol{p})$ are functions of $\boldsymbol{p}$. We now calculate their gradients, $\nabla h_{r_j}(\boldsymbol{p}) = \left(\partial r_{U_j|\boldsymbol{U}_{-j}}/\partial p_{1,\ldots,1},\ldots,\partial r_{U_j|\boldsymbol{U}_{-j}}/\partial p_{I_1,\ldots,I_d}\right)^\top$ and $\nabla h_{X_j}(\boldsymbol{p}) = \left(\partial\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}/\partial p_{1,\ldots,1},\ldots,\partial\rho^2_{(\boldsymbol{X}_{-j}\to X_j)}/\partial p_{I_1,\ldots,I_d}\right)^\top$. For any $\boldsymbol{k} = (k_1,\ldots,k_d)\in\mathbb{I}_1\times\cdots\times\mathbb{I}_d$, we have

$$(i)\ \text{ for } i_j < k_j, \boldsymbol{i}_{-j}\neq\boldsymbol{k}_{-j},\quad \frac{\partial p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j}+u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = 0;$$

$$(ii)\ \text{ for } i_j = k_j, \boldsymbol{i}_{-j}\neq\boldsymbol{k}_{-j},\quad \frac{\partial p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j}+u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = p_{k_j|\boldsymbol{i}_{-j}};$$

(iii) for $i_j > k_j, \boldsymbol{i}_{-j} \neq \boldsymbol{k}_{-j}$, $\quad \dfrac{\partial p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = 2p_{k_j|\boldsymbol{i}_{-j}};$

(iv) for $i_j < k_j, \boldsymbol{i}_{-j} = \boldsymbol{k}_{-j}$, $\quad \dfrac{\partial p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = -\left(\dfrac{p_{k_1,\ldots,i_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\right)\left(u^j_{i_j} + u^j_{i_j-1}\right);$

(v) for $i_j = k_j, \boldsymbol{i}_{-j} = \boldsymbol{k}_{-j}$, $\quad \dfrac{\partial p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = p_{k_j|\boldsymbol{k}_{-j}} + \left(\dfrac{p_{k_1,\ldots,+_j,\ldots,k_d} - p_{k_1,\ldots,k_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\right)\left(u^j_{k_j} + u^j_{k_j-1}\right);$

(vi) for $i_j > k_j, \boldsymbol{i}_{-j} = \boldsymbol{k}_{-j}$, $\quad \dfrac{\partial p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = 2p_{i_j|\boldsymbol{k}_{-j}} - \left(\dfrac{p_{k_1,\ldots,i_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\right)\left(u^j_{i_j} + u^j_{i_j-1}\right).$

By utilizing (iv)-(vi), we obtain,

$$\frac{\partial r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})}{\partial p_{k_1,\ldots,k_d}} = \frac{1}{2}\frac{\partial \sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right)}{\partial p_{k_1,\ldots,k_d}} = \frac{1}{2}\left[\sum_{i_j<k_j} -\frac{p_{k_1,\ldots,i_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\left(u^j_{i_j} + u^j_{i_j-1}\right)\right. \tag{A.1}$$
$$\left. + \left(p_{k_j|\boldsymbol{k}_{-j}} + \frac{p_{k_1,\ldots,+_j,\ldots,k_d} - p_{k_1,\ldots,k_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\left(u^j_{k_j} + u^j_{k_j-1}\right)\right) + \sum_{i_j>k_j}\left\{2p_{i_j|\boldsymbol{k}_{-j}} - \frac{p_{k_1,\ldots,i_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\left(u^j_{i_j} + u^j_{i_j-1}\right)\right\}\right].$$

By the delta method, we have

$$\sqrt{n}\left\{\hat{r}_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j}) - r_{U_j|\boldsymbol{U}_{-j}}(\boldsymbol{u}_{-j})\right\} \xrightarrow{D} \mathcal{N}\left(0, \nabla h_{r_j}(\boldsymbol{p})^\top\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right)\nabla h_{r_j}(\boldsymbol{p})\right).$$

This concludes the proof of Theorem 1.

In addition, by utilizing (i)-(vi), we obtain,

$$\frac{\partial \rho^2_{(\boldsymbol{X}_{-j}\to X_j)}}{\partial p_{k_1,\ldots,k_d}} = \frac{\partial 3\sum_{\boldsymbol{i}_{-j}=\boldsymbol{1}_{-j}}^{\boldsymbol{I}_{-j}}\left(\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right) - 1\right)^2 p_{i_1,\ldots,+_j,\ldots,i_d}}{\partial p_{k_1,\ldots,k_d}} = 3\left\{\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{k}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right) - 1\right\}^2 \tag{A.2}$$
$$+ 6\sum_{\boldsymbol{i}_{-j}\neq\boldsymbol{k}_{-j}}\left(\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{i}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right) - 1\right)\left(p_{k_j|\boldsymbol{i}_{-j}} + \sum_{i_j\geq k_j+1} 2p_{i_j|\boldsymbol{i}_{-j}}\right)p_{i_1,\ldots,+_j,\ldots,i_d}$$
$$+ 6\left\{\sum_{i_j=1}^{I_j} p_{i_j|\boldsymbol{k}_{-j}}\left(u^j_{i_j} + u^j_{i_j-1}\right) - 1\right\} \times \left[\sum_{i_j<k_j} -\frac{p_{k_1,\ldots,i_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\left(u^j_{i_j} + u^j_{i_j-1}\right) + \left\{p_{k_j|\boldsymbol{k}_{-j}} + \left(\frac{p_{k_1,\ldots,+_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\right.\right.\right.$$
$$\left.\left.\left. - \frac{p_{k_1,\ldots,k_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\right)\left(u^j_{k_j} + u^j_{k_j-1}\right)\right\} + \sum_{i_j>k_j}\left\{2p_{i_j|\boldsymbol{k}_{-j}} - \frac{p_{k_1,\ldots,i_j,\ldots,k_d}}{p^2_{k_1,\ldots,+_j,\ldots,k_d}}\left(u^j_{i_j} + u^j_{i_j-1}\right)\right\}\right]p_{k_1,\ldots,+_j,\ldots,k_d}.$$

By the delta method, we have

$$\sqrt{n}\left(\hat{\rho}^2_{(\boldsymbol{X}_{-j}\to X_j)} - \rho^2_{(\boldsymbol{X}_{-j}\to X_j)}\right) \xrightarrow{D} \mathcal{N}(0, \nabla h_{X_j}(\boldsymbol{p})^\top\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right)\nabla h_{X_j}(\boldsymbol{p})).$$

This concludes the proof of Theorem 2. □

References

[1] A. Agresti, Analysis of Ordinal Categorical Data, Wiley, New York, 2 edition, 2010.

[2] A. Agresti, M. Kateri, Categorical Data Analysis, Springer, Berlin, 2011.

[3] J. Aitchison, S. Silvey, The generalization of probit analysis to the case of multiple responses, Biometrika 44 (1957) 131–140.

[4] J. A. Anderson, Regression and ordered categorical variables, Journal of the Royal Statistical Society: Series B (Methodological) 46 (1984) 1–22.

[5] M. Denuit, P. Lambert, Constraints on concordance measures in bivariate discrete data, Journal of Multivariate Analysis 93 (2005) 40–57.

[6] D. Donoho, 50 years of data science, Journal of Computational and Graphical Statistics 26 (2017) 745–766.

[7] R. Douglas, S. Fienberg, M.-L. Lee, A. Sampson, L. Whitaker, Positive dependence concepts for ordinal contingency tables, in: H. W. Block, A. R. Sampson, T. H. Savits (Eds.), Topics in Statistical Dependence, Institute of Mathematical Statistics, Hayward, CA, 1991, pp. 189–202.

[8] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[9] V. Farewell, A note on regression analysis of ordinal data with variability of classification, Biometrika 69 (1982) 533–538.

[10] A. S. Fullerton, J. Xu, Ordered Regression Models: Parallel, Partial, and Non-Parallel Alternatives, Chapman and Hall/CRC, New York, 2016.

[11] C. Genest, J. G. Nešlehová, B. Rémillard, On the empirical multilinear copula process for count data, Bernoulli 20 (2014) 1344–1371.

[12] C. Genest, J. G. Nešlehová, B. Rémillard, Asymptotic behavior of the empirical multilinear copula process under broad conditions, Journal of Multivariate Analysis 159 (2017) 82–110.

[13] Z. Gilula, S. Haberman, The analysis of multivariate contingency tables by restricted canonical and restricted association models, Journal of the American Statistical Association 83 (1988) 760–771.

[14] Z. Gilula, Y. Ritov, Inferential ordinal correspondence analysis: Motivation, derivation and limitations, International Statistical Review 58 (1990) 99–108.

[15] P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Springer, New York, 2000.

[16] L. Goodman, Simple models for the analysis of association in cross-classifications having ordered categories, Journal of the American Statistical Association 74 (1979) 537–552.

[17] L. Goodman, The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds, Biometrics 39 (1983) 149–160.

[18] L. Goodman, The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, Annals of Statistics 13 (1985) 10–69.

[19] L. Goodman, Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, International Statistical Review 54 (1986) 243–309.

[20] L. Goodman, A single general method for the analysis of cross-classified data: reconciliation and synthesis of some methods of pearson, yule, and fisher, and also some methods of correspondence analysis and association analysis, Journal of the American Statistical Association 91 (1996) 408–428.

[21] L. A. Goodman, W. H. Kruskal, Measures of association for cross classifications, Journal of the American statistical association 49 (1954) 732–764.

[22] F. E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer, New York, 2015.

[23] R. Hawkes, The multivariate analysis of ordinal measures, American Journal of Sociology 76 (1971) 908–926.

[24] H. Joe, Dependence Modeling with Copulas, Chapman and Hall/CRC, New York, 2014.

[25] M. Kateri, Contingency Table Analysis: Methods and Implementation using R, Springer, New York, 2014.

[26] M. Kendall, The treatment of ties in ranking problems, Biometrika 33 (1945) 239–251.

[27] M. Kendall, Rank Correlation Methods, C. Griffin, London, 1948.

[28] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

[29] E. Läärä, J. Matthews, The equivalence of two models for ordinal data, Biometrika 72 (1985) 206–207.

[30] J. Landis, E. Heyman, G. Koch, Average partial association in three-way contingency tables: A review and discussion of alternative tests, International Statistical Review 46 (1978) 237–254.

[31] C. Li, B. Shepherd, Test of association between two ordinal variables while adjusting for covariates, Journal of the American Statistical Association 105 (2010) 612–620.

[32] I. Liu, A. Agresti, The analysis of ordered categorical data: An overview and a survey of recent developments, Test 14 (2005) 1–73.

[33] N. Mantel, Chi-square tests with one degree of freedom: extensions of the mantel-haenszel procedure, Journal of the American Statistical Association 58 (1963) 690–700.

[34] A. Maydeu-Olivares, H. Joe, Assessing approximate fit in categorical data analysis, Multivariate Behavioral Research 49 (2014) 305–328.

[35] P. McCullagh, Regression models for ordinal data, Journal of the Royal Statistical Society, Ser. B 42 (1980) 109–142.

[36] P. W. Mielke, K. J. Berry, Permutation Methods: A Distance Function Approach, Springer, New York, 2001.

[37] R. B. Nelsen, An Introduction to Copulas, Springer, New York, 2006.

[38] J. Nešlehová, On rank correlation measures for non-continuous random variables, Journal of Multivariate Analysis 98 (2007) 544–567.

[39] B. Schweizer, A. Sklar, Operations on distribution functions not derivable from operations on random variables, Studia Mathematica 52 (1974) 43–52.

[40] G. Shmueli, To explain or to predict?, Statistical Science 25 (2010) 289–310.

[41] G. Simon, Alternative analyses for the singly-ordered contingency table, Journal of the American Statistical Association 69 (1974) 971–976.

[42] A. Sklar, Fonctions de r$\acute{e}$partition $\grave{a}$ n dimensions et leurs marges, Publications de l'Institut de Statistique de L'Universit$\acute{e}$ de Paris 8 (1959) 229–231.

[43] M. Smithson, E. C. Merkle, Generalized Linear Models for Categorical and Continuous Limited Dependent Variables, CRC Press, Boca Raton, FL, 2013.

[44] R. H. Somers, A new asymmetric measure of association for ordinal variables, American sociological review (1962) 799–811.
[45] A. Stuart, The estimation and comparison of strengths of association in contingency tables, Biometrika 40 (1953) 105–110.
[46] J. Tukey, Exploratory Data Analysis, Addison Wesley, Reading, MA, 1977.
[47] G. Tutz, Regression for Categorical Data, Cambridge University Press, Cambridge, 2011.
[48] L. Yang, E. Frees, Z. Zhang, Nonparametric estimation of copula regression models with discrete outcomes, Journal of the American Statistical Association 115 (2020) 707–720.
[49] T. Yee, The vgam package for categorical data analysis, Journal of Statistical Software 32 (2010) 1–34.