Measure of asymmetric association for ordinal contingency tables via the bilinear extension copula

Zheng Wei and Daeyoung Kim

University of Maine and University of Massachusetts-Amherst

Supplementary Material

The supplementary material consists of three sections. Section S1 provides a running example to illustrate the underlying idea of the proposed BECCR. The finite-sample performance of the BECCR is evaluated with the simulation study in Section S2.1 and a real data set in Section S2.2. Section S3 compares the proposed BECCR with the subcopula correlation ratio ρ_S^2 by Wei and Kim (2017) and the MCD measure μ_t by Shan et al. (2015).

S1. Running Example

Table S1 shows a 9×3 table for two ordinal variables X and Y with I=9 and J=3 ordered categories, $\{x_1 < \ldots < x_9\}$ and $\{y_1 < y_2 < y_3\}$, respectively. The joint p.m.f P=$\{p_{ij}, i = 1, \ldots, 9, j = 1, 2, 3\}$ of X and Y is constructed based on the completely dependent copula $C_a^{cd}(u^*, v^*)$ of U^* and V^* (left plot in Figure S1) such that $V^* = aU^*(mod\ 1)$, a=3, $U^* \sim uniform[0, 1]$ and $p_{ij} = Volume_{C_a^{cd}}\left(\left[\frac{i-1}{9}, \frac{i}{9}\right] \times \left[\frac{j-1}{3}, \frac{j}{3}\right]\right) = C_a^{cd}\left(\frac{i}{9}, \frac{j}{3}\right) - C_a^{cd}\left(\frac{i-1}{9}, \frac{j}{3}\right) - C_a^{cd}\left(\frac{i}{9}, \frac{j-1}{3}\right) + C_a^{cd}\left(\frac{i-1}{9}, \frac{j-1}{3}\right)$.

Table S1 reflects the completely dependent relationship shown in the left plot of Figure S1 in that Y is a function of X with probability 1 but not vice versa. The right plot of Figure S1 visualizes the bilinear extension copula density $c^+(u, v)$ in Eq. (3) associated with X and Y in Table S1, which well delineates the dependence structure of Table S1.

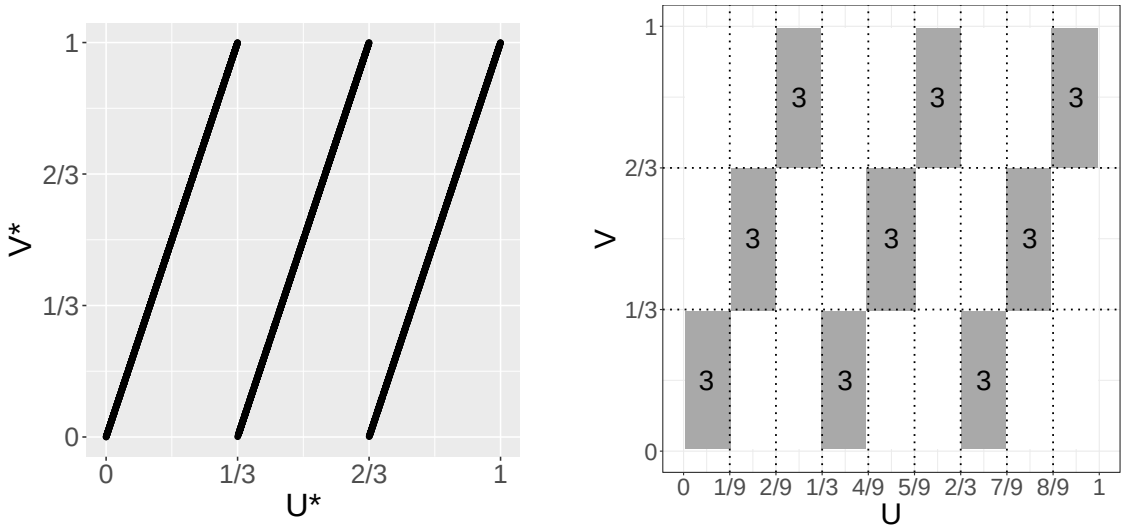

Fig. S1: Completely dependence copula C_a^{cd} (left) and bilinear extension copula density $c^+(u, v)$ for X and Y in Table S1 (right).

Y \ X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	$p_{\cdot j}$ (D_1)
y_1	1/9	0	0	1/9	0	0	1/9	0	0	1/3 (1/3)
y_2	0	1/9	0	0	1/9	0	0	1/9	0	1/3 (2/3)
y_3	0	0	1/9	0	0	1/9	0	0	1/9	1/3 (1)
$p_{i\cdot}$ (D_0)	1/9 (1/9)	1/9 (2/9)	1/9 (3/9)	1/9 (4/9)	1/9 (5/9)	1/9 (6/9)	1/9 (7/9)	1/9 (8/9)	1/9 (1)	

Table S1: Joint p.m.f for X and Y (p_{ij}), its marginal p.m.f.s ($p_{i\cdot}$, $p_{\cdot j}$) and their ranges (D_0, D_1).

For Table S1, we compute the bilinear extension copula regression functions of V given U and of U given V using Eq. (5), and visualize them in Figure S2: $r_{V|U}(u)$=1/6 for $u \in [0, 1/9] \cup (3/9, 4/9] \cup (6/9, 7/9]$, 1/2 for $u \in (1/9, 2/9] \cup (4/9, 5/9] \cup (7/9, 8/9]$, 5/6 for $u \in (2/9, 3/9] \cup (5/9, 6/9] \cup (8/9, 1]$, and $r_{U|V}(v) = 7/18$ for $v \in [0, 1/3]$, 1/2 for $v \in (1/3, 2/3]$, and $11/18$ for $v \in (2/3, 1]$. As shown in Figure S2, $r_{V|U}(u)$ detects the functional relationship between X and Y, unlike $r_{U|V}(v)$.

For the data in Table S1 we also calculate the BECCR of Y on X and X on Y in Definitions 2

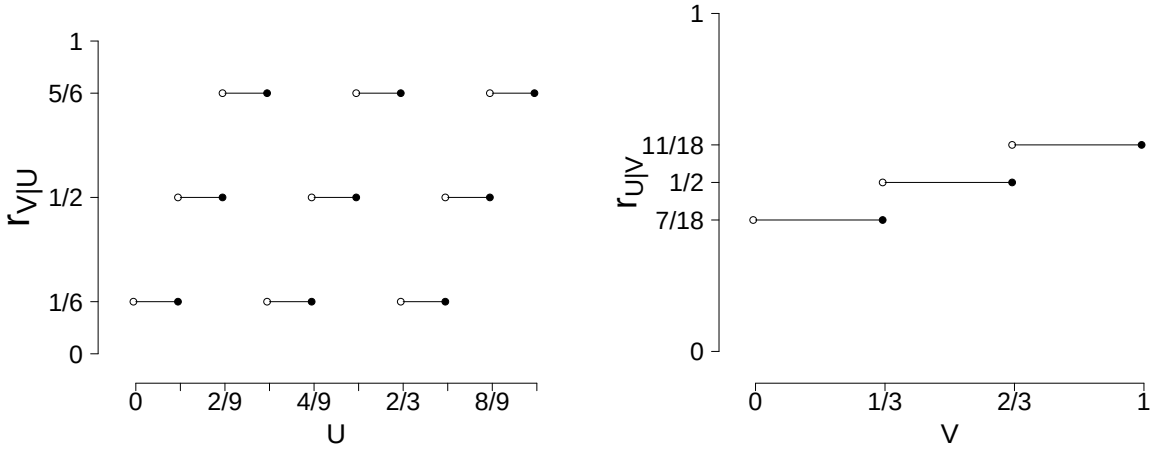

Fig. S2: Bilinear extension regressions for Table S1 - $r_{V|U}(u)$ (left) and $r_{U|V}(v)$ (right).

and 3 : $(\rho^2_{(X\to Y)}, \rho^2_{(Y\to X)})$=(8/9, 8/81) and $(\rho^{*2}_{(X\to Y)}, \rho^{*2}_{(Y\to X)})$=(1, 1/10). Note that the upper bounds of $\rho^2_{(X\to Y)}$ and $\rho^2_{(Y\to X)}$ are 8/9 and 80/81, respectively. This result supports the fact shown in Table S1 that Y is a function of X with probability 1 but not versa.

S2. Numerical example

We examine the finite-sample performance of the proposed BECCRs via Monte Carlo simulations and illustrate its utility using a real data set. In real data analysis, we use the bootstrap method to evaluate the uncertainty in the estimates of the proposed measures and their pairwise differences: compute the bootstrap bias-corrected and accelerated (BCa) confidence intervals for $\rho^2_{(X\to Y)}$, $\rho^2_{(Y\to X)}$, $\rho^{*2}_{(X\to Y)}$, $\rho^{*2}_{(Y\to X)}$, and $\rho^{*2}_{(X\to Y)} - \rho^{*2}_{(Y\to X)}$ from 1000 bootstrap replicates of size n which were simulated from the saturated log-linear model fitted to the observed contingency table.

S2.1. Simulation study

We consider the following simulation factors: type of asymmetric association, sample size n, and number of categories (I,J) for two ordinal variables (X,Y). Regrading the type of asymmetric association, three bivariate continuous copulas are employed: Pareto-Liouville copula (McNeil and Nešlehová, 2010) for monotonic association, Gaussian Mixture copula (Kasa et al., 2020; Rajan and Bhattacharya, 2016) for non-monotonic association, and the Completely Dependent copula (Junker et al., 2021) for completely functional association.

- A Pareto-Liouville (PL) copula C^{PL} is induced by the distribution of two continuous random variables $(R_\theta D_{\alpha_1}, R_\theta D_{\alpha_2})$, where R_θ follows the Pareto distribution with the distribution function $F_{R_\theta}(r) = 1 - r^{-\theta}$ for $r \geq 1$, $\theta > 0$, and $(D_{\alpha_1}, D_{\alpha_2})$ denotes a vector with Dirichlet distribution with parameters (α_1, α_2) . To generate the data showing a clear monotonic association, we consider the following parameter values in C^{PL}, numbers of categories and sample size (I, J, n) : $(\theta, \alpha_1, \alpha_2)$=(5,8,2), $(I, J) \in \{(5, 5), (10, 10)\}$, $n \in \{125, 250, 500, 1000, 2000, 2500, 5000\}$ for (I, J)=$(5, 5)$, and $n \in \{500, 1000, 2000, 4000, 8000, 10000, 20000\}$ for (I, J)=$(10, 10)$. The scatter plot simulated from C^{PL} with n=5000 is given in the first row of Figure S3.
- The Gaussian Mixture (GM) copula C^{GM} is derived by the bivariate Gaussian mixture distribution:

$$C^{GM}(u^*, v^*; \mathbf{\Psi}) = H_{12}^{GM}(F_1^{-1}(u^*; \pi, \mu_{11}, \mu_{21}), F_2^{-1}(v^*; \pi, \mu_{12}, \mu_{22}); \mathbf{\Psi}),$$

where H_{12}^{GM}, F_1 and F_2 denote the joint and marginal distributions of a 2-component bivariate Gaussian mixture random vector whose density is

$$h^{GM}(x_1, x_2; \mathbf{\Psi}) = \pi\phi_2(x_1, x_2; (\mu_{11}, \mu_{12})^\top, \theta_1) + (1 - \pi)\phi_2(x_1, x_2; (\mu_{21}, \mu_{22})^\top, \theta_2),$$

and $\mathbf{\Psi}$=(π, $\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \theta_1, \theta_2$). Note that $\phi_2(x_1, x_2; (\mu_1, \mu_2)^\top, \theta)$ denotes the bivariate normal density with mean vector $(\mu_1, \mu_2)^\top$, common variance equal to 1, and correlation θ. For the data with nonlinear non-monotonic association, we used the following parameter values in C^{GM} and values of (I, J, n) : $\mathbf{\Psi}$=$(0.5, 2, 1, -2, 1, -0.8, 0.8)$, $(I, J) \in \{(5, 5), (10, 10)\}$, $n \in \{125, 250, 500, 1000, 2000, 2500, 5000\}$ for (I, J)=$(5, 5)$ and $n \in \{500, 1000, 2000, 4000, 8000, 10000, 20000\}$ for (I, J)=$(10, 10)$. The first row of Figure S4 gives the scatter plot simulated from C^{GM} with n=5000.

- A Completely Dependent (CD) copula C_a^{CD} is the distribution of bivariate uniform vector (U^*,V^*) such that V^*=$aU^*(mod\ 1)$. The following parameter value in C_a^{CD} and values of (I, J, n) are employed: a=$(2, 5)$, $(I, J) \in \{(4, 2), (10, 5)\}$ for a=2, $(I, J) \in \{(10, 2), (15, 3)\}$ for a=5, and $n \in \{250, 500, 1000, 2000, 4000, 5000, 10000\}$. The scatter plots simulated from C_2^{CD} and C_5^{CD} are given in the first row of Figure S5.

For each combination of the simulation factors described above, we simulate a sample of size n, create an $I \times J$ contingency table by classifying n data points by the equal-width quantiles of two discrete uniform variables X and Y with I and J categories, respectively, and compute the estimators of the standardized BECCR in Eq. (7), $\hat{\rho}^{*2}_{X\to Y}$ and $\hat{\rho}^{*2}_{Y\to X}$. These steps are repeated for $B = 1000$ times. Note that the population value of $\rho^{*2}_{(X\to Y)}$ and $\rho^{*2}_{(Y\to X)}$ are obtained by generating the data of large sample size $N = 10^7$ for a given copula and a value of (I, J).

To examine the empirical behavior of the estimators, we construct the boxplots of the 1000 estimates of $\hat{\rho}^{*2}_{X\to Y}$ (blue) and $\hat{\rho}^{*2}_{Y\to X}$ (red) over each value of n and the log-log plots of the root mean square error (RMSE) versus n along with least-squares linear fits. Note that RMSEs were

computed using 1000 estimates of the proposed measure for the dominant direction, $\hat{\rho}^{*2}_{X\to Y}$.

The boxplots in the 2nd row of Figures S3 (obtained from C^{PL} copula) and S4 (obtained from C^{GM} copula) show that the sampling variability decreases as n increases and the bias is small even when n is small in that the means of the estimates are close to the population values (blue and red dashed lines) of $\rho^{*2}_{X\to Y}$ and $\rho^{*2}_{Y\to X}$. The log-log plots of the RMSE in the 3rd row of Figure S3 and S4 show a downward linear pattern with slope of approximately -1. This indicates that the empirical convergence rate of $\hat{\rho}^{*2}_{X\to Y}$ is of order n^{-1}.

From the boxplots in the 2nd and 3rd rows of Figure S5 obtained from the C^{CD}_a copula, we observe that $\hat{\rho}^{*2}_{X\to Y}$ (blue) has almost no sampling variability and is concentrated around 1 regardless of the values of (n, I, J, a), while $\hat{\rho}^{*2}_{Y\to X}$ (red) is concentrated around a lower value (zero for a=5). This observation results from the fact that the contingency tables simulated from the C^{CD}_a copula has complete and asymmetric dependence in that Y is a function of X with probability 1 but not vice versa, as shown in Table S1 in Section S1.

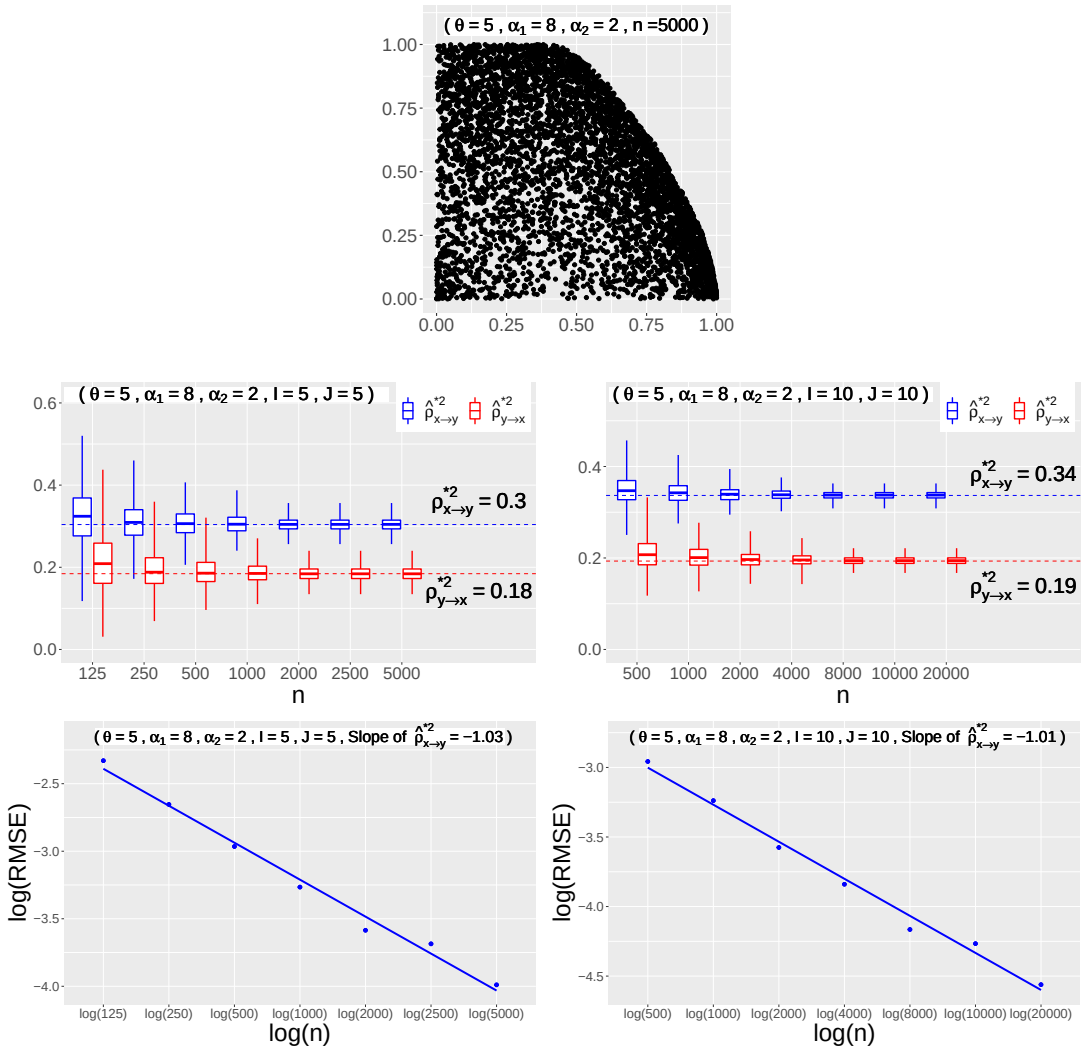

Fig. S3: 1st row - scatter plot of a sample of size n=5000 generated from Pareto-Liouville copula; **2nd row** - boxplots for the estimates of $\hat{\rho}^{*2}_{X\to Y}$ (blue) and $\hat{\rho}^{*2}_{Y\to X}$ (red) computed from 1000 simulated 5×5 contingency tables (left) and 1000 simulated 10×10 simulated contingency tables (right). Note that the dash lines represent $\rho^{*2}_{X\to Y}$ (blue) and $\rho^{*2}_{Y\to X}$ (red), respectively. **3rd row** - log-log plots of the RMSE of $\rho^{*2}_{X\to Y}$ and the sample size computed from 1000 simulated 5×5 contingency tables (left) and 1000 simulated 10×10 contingency tables (right).

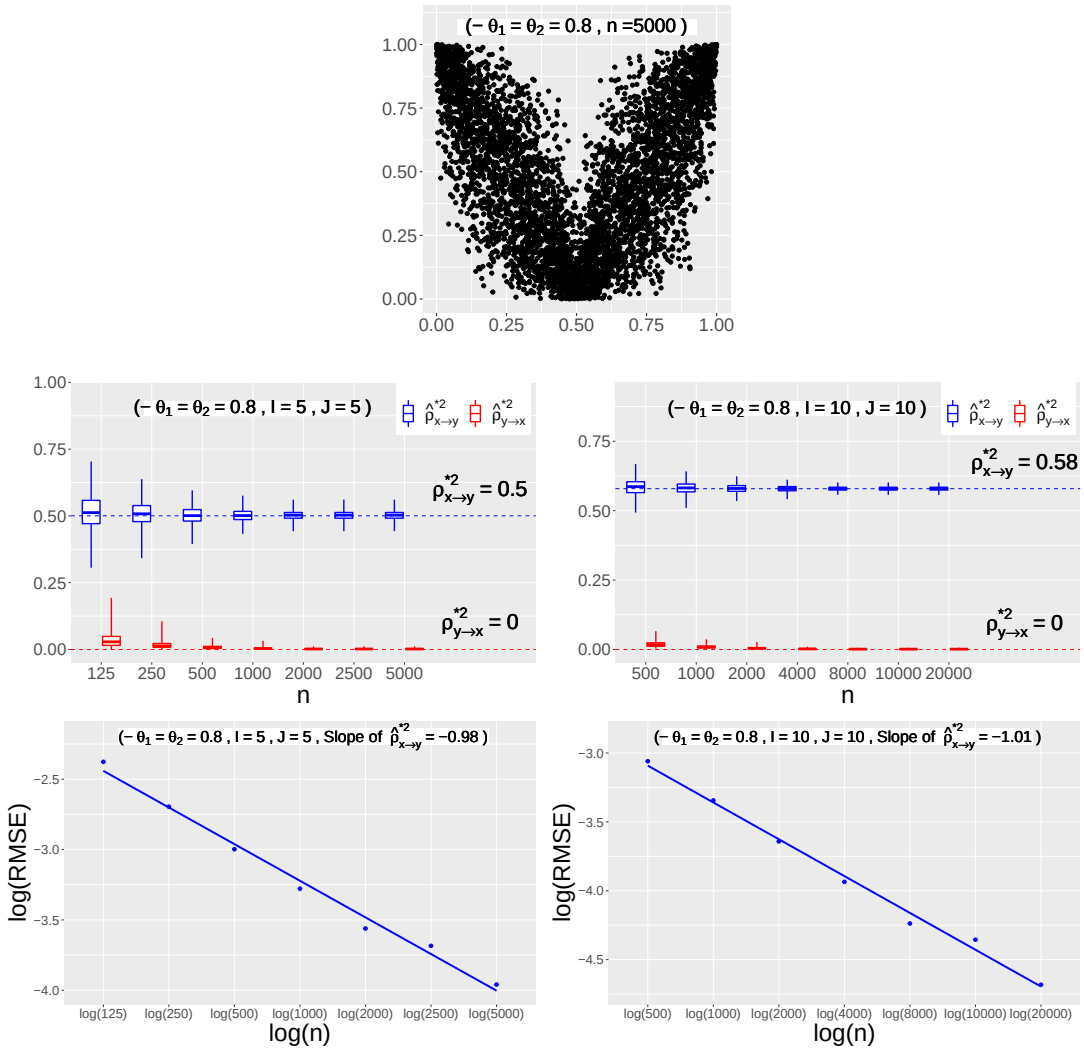

Fig. S4: 1st row : scatter plot of a sample of size n=5000 generated from Gaussian Mixture copula; **2nd row**: boxplots for the estimates of $\hat{\rho}^{*2}_{X\to Y}$ (blue) and $\hat{\rho}^{*2}_{Y\to X}$ (red) computed from 1000 simulated 5×5 contingency tables (left) and 1000 simulated 10×10 simulated contingency tables (right). Note that the dash lines represent $\rho^{*2}_{X\to Y}$ (blue) and $\rho^{*2}_{Y\to X}$ (red), respectively. **3rd row** : log-log plots of the RMSE of $\rho^{*2}_{X\to Y}$ and the sample size computed from 1000 simulated 5×5 contingency tables (left) and 1000 simulated 10×10 contingency tables (right).

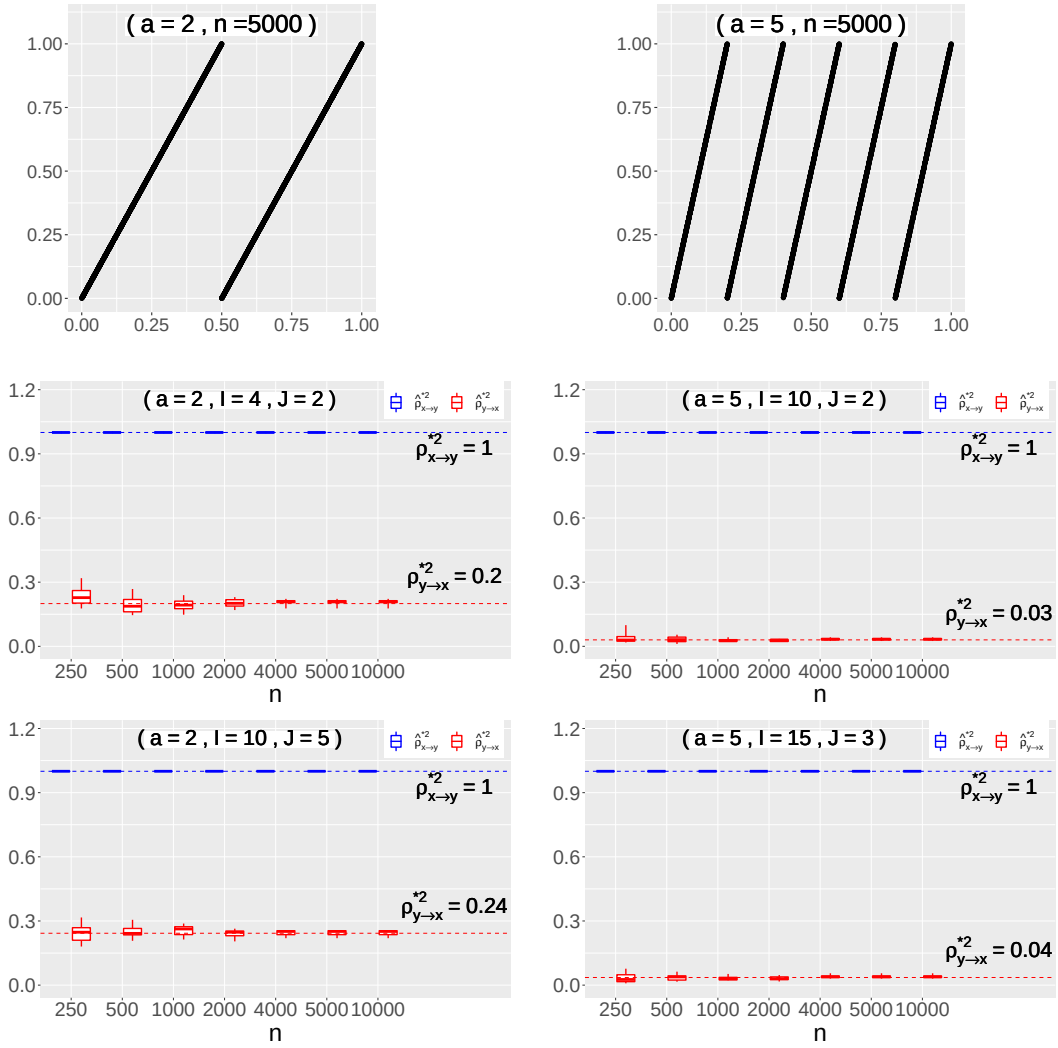

Fig. S5: 1st row - scatter plots of a sample of size n=5000 generated from the completely dependent copula C_a^{cd} with $a = 2$ (left) and $a = 5$ (right); **2nd row** - boxplots for the estimates of $\hat{\rho}^{*2}_{X\to Y}$ (blue) and $\hat{\rho}^{*2}_{Y\to X}$ (red) computed using 1000 4×2 contingency tables simulated from C_a^{cd} with $a = 2$ (left) and 1000 10×2 contingency tables simulated from C_a^{cd} with $a = 5$ (right); **3rd row** - boxplots for the estimates of $\hat{\rho}^{*2}_{X\to Y}$ (blue) and $\hat{\rho}^{*2}_{Y\to X}$ (red) computed using 1000 10×5 contingency tables simulated from C_a^{cd} with $a = 2$ (left) and 1000 15×3 contingency tables simulated from C_a^{cd} with $a = 5$ (right); Note that the dash lines shown in the 2nd and 3rd row represent $\rho^{*2}_{X\to Y}$ (blue) and $\rho^{*2}_{Y\to X}$ (red), respectively.

	Partner (Non-partner)			
Organizational	Relationship with users			
aspects	Unsatisfied	Little satisfied	Satisfied	Very satisfied
Unsatisfied	3 (15)	7 (4)	4 (2)	8 (4)
Little satisfied	5 (4)	9 (8)	10 (14)	12 (11)
Satisfied	1 (2)	9 (5)	22 (22)	30 (53)
Very satisfied	3 (2)	5 (4)	59 (14)	73 (39)

Table S2: Workers' satisfaction for organizational aspects by workers' satisfaction for relationships with users by workers' status

S2.2. Real data analysis

The data in Table S2, obtained from a survey concerning the job satisfaction of 426 workers in social enterprises of Caserta in Italy, has three variables: the satisfaction with respect to the organizational aspects (O), the satisfaction towards the relationships with users (R) and the workers' status with two levels ('Partner', 'Non-partner'). Both satisfaction variables are ordinal.

A question of interest is whether or not workers who are satisfied with organizational aspects are satisfied with the relationships with users (denoted as O→R) or vice versa (denoted as R→O), depending on the worker's status. To examine two types of association structure, we compute in Table S3 the (standardized) BECCR for the two types of association and their pairwise differences with the 95% BCa bootstrap intervals. We observe that for the partner's status, (R→O) is at least twice stronger than (O→R) ($\hat{\rho}^{*2}_{(R\to O)}$=0.1336> $\hat{\rho}^{*2}_{(O\to R)}$=0.0644), but for the non-partner's status (O→R) is a little larger than (R→O) ($\hat{\rho}^{*2}_{(R\to O)}$=0.2137< $\hat{\rho}^{*2}_{(O\to R)}$=0.2494).

From the pairwise differences and the corresponding confidence intervals, we see that statistically dominant reciprocal association structure is not the same between two workers's sta-

Partner status	$\hat{\rho}^2_{(R\to O)}$ (Upper bound)	$\hat{\rho}^2_{(O\to R)}$ (Upper bound)	$\hat{\rho}^{*2}_{(R\to O)}$	$\hat{\rho}^{*2}_{(O\to R)}$	$\hat{\rho}^{*2}_{(R\to O)} - \hat{\rho}^{*2}_{(O\to R)}$
Estimate	0.1104	0.0544	0.1336	0.0644	0.0691
	(0.827)	(0.844)			
95% BCa bootstrap CI	(0.0428, 0.1762)	(0.0119, 0.1092)	(0.0515, 0.2134)	(0.0139, 0.1278)	(0.0122, 0.1487)
Non-partner status	$\hat{\rho}^2_{(R\to O)}$ (Upper bound)	$\hat{\rho}^2_{(O\to R)}$ (Upper bound)	$\hat{\rho}^{*2}_{(R\to O)}$	$\hat{\rho}^{*2}_{(O\to R)}$	$\hat{\rho}^{*2}_{(R\to O)} - \hat{\rho}^{*2}_{(O\to R)}$
Estimate	0.1926	0.2080	0.2137	0.2494	-0.0357
	(0.9016)	(0.8341)			
95% BCa bootstrap CI	(0.0863, 0.2965)	(0.1162, 0.3229)	(0.0960, 0.3278)	(0.1357, 0.3780)	(-0.1147, 0.0270)

Table S3: Analysis of asymmetric association for Table S2.

tuses. For the partner's status, (R→O) is a more dominant reciprocal association structure than (O→R) because $\hat{\rho}^{*2}_{(R\to O)} - \hat{\rho}^{*2}_{(O\to R)} = 0.0691$ is positive and the corresponding 95% confidence intervals do not include 0. But, for the non-partner's status, the 95% confidence intervals for $\rho^{*2}_{(R\to O)} - \rho^{*2}_{(O\to R)}$ include 0 and thus two association structures are statistically equivalent. Therefore, we conclude that for the parter status, the satisfaction for the relationships with users (R) exerts more influence on the satisfaction for the organizational aspects (O) than the other way around. On the other hand, for the non-parter status, both satisfaction variables are symmetrically associated with each other.

S3. Comparison of the proposed BECCR with the subcopula correlation ratio (Wei and Kim, 2017) and the MCD measure (Shan et al., 2015)

When X and Y are the putatively independent and dependent ordinal variables, respectively, in an ordinal contingency table, two existing asymmetric association measures $\mu_t^{(1)}(X, Y)$ (Shan et al., 2015) and $\rho^2_{S(X\to Y)}$ (Wei and Kim, 2017) can be compared with $\rho^{*2}_{(X\to Y)}$ in Eq. (7) in several aspects.

First, $\rho^2_{S(X\to Y)}$ and $\rho^{*2}_{(X\to Y)}$ are the correlation ratios based on the (sub/bilinear extension) copula regression. On the other hand, $\mu_t^{(1)}(X, Y)$ is a normalized expectation of weighted distance between the marginal distribution of Y and its conditional distribution given X evaluated at two consecutive categories of Y which is defined through subcopula. The weight (denoted as t) determines the relative contributions of two consecutive categories of Y in the computation of $\mu_t^{(1)}(X, Y)$.

Second, $\rho^2_{S(X\to Y)}$ and $\rho^{*2}_{(X\to Y)}$ can be interpreted as the proportion of the variation associated with Y explained by X. However, $\mu_t^{(1)}(X, Y)$ does not enjoy the ease of interpretation that the other two indices have.

Third, all three indices range from 0 to 1 and 1 indicates that Y is completely dependent on X. While a zero value of $\mu_t^{(1)}(X, Y)$ implies independence between X and Y, a zero value of $\rho^2_{S(X\to Y)}$ and $\rho^{*2}_{(X\to Y)}$ does not imply independence. Instead, it represents no contribution of X in explaining the variation of Y through the (sub/bilinear extension) copula regression. Table S4 shows two illustrative examples where two ordinal random variables X and Y are not independent but $\rho^2_{S(X\to Y)}$ and $\rho^{*2}_{(X\to Y)} = 0$ are both zero. Note that both examples in Table S4 are the discretized versions of the examples considered in Dette et al. (2013).

Y \ X	x_1	x_2	x_3	x_4	x_5
y_1	0.1	0	0	0	0.1
y_2	0	0.1	0	0.1	0
y_3	0	0	0.2	0	0
y_4	0	0.1	0	0.1	0
y_5	0.1	0	0	0	0.1

(a) diagonally crossed distribution

Y \ X	x_1	x_2	x_3	x_4	x_5
y_1	0	0	0.2	0	0
y_2	0	0.1	0	0.1	0
y_3	0.1	0	0	0	0.1
y_4	0	0.1	0	0.1	0
y_5	0	0	0.2	0	0

(b) circular distribution

Table S4: Examples of dependent ordinal variables X and Y with zero value of $\rho^2_{S(X\to Y)}$ and $\rho^{*2}_{(X\to Y)}$

Fourth, while $\rho^2_{S(X\to Y)}$ and $\rho^{*2}_{(X\to Y)}$ are completely data-dependent, the computation of $\mu_t^{(1)}(X,Y)$ requires the determination of the weight t and in general the value of $\mu_t^{(1)}(X,Y)$ may change over t.

Last, $\rho^2_{S(X\to Y)}$ and its estimator are constructed based on the true subcopula associated with the true distribution function and the empirical subcopula obtained from the empirical distribution function of the data, respectively. Since the empirical subcopula is a consistent estimator for the true subcopula (Rachasingho and Tasena, 2020; Tasena, 2021a,b), the consistency of the estimator of $\rho^2_{S(X\to Y)}$ can be established if one can show that $\rho^2_{S(X\to Y)}$ is continuous with respect to the distance over the subcopula space (Tasena, 2021b). However, this was not studied in Wei and Kim (2017) and requires further investigation.

References

Dette, H., Siburg, K., and Stoimenov, P. A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics*, 40:21–41, 2013.

Junker, R., Griessenberger, F., and Trutschnig, W. Estimating scale-invariant directed dependence of bivariate distributions. *Computational Statistics & Data Analysis*, 153:107058, 2021. doi: https://doi.org/10.1016/j.csda.2020.107058.

Kasa, S., Bhattacharya, S., and Rajan, V. Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping. *Bioinformatics*, 36(2):621–628, 2020.

McNeil, A. and Nešlehová, J. From archimedean to liouville copulas. *Journal of Multivariate Analysis*, 101(8):1772–1790, 2010.

Rachasingho, J. and Tasena, S. A metric space of subcopulas – an approach via hausdorff distance. *Fuzzy Sets and Systems*, 378(1):144–156, 2020.

Rajan, V. and Bhattacharya, S. Dependency clustering of mixed data with gaussian mixture copulas. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1967–1973, 2016.

Shan, Q., Wongyang, T., Wang, T., and Tasena, S. A measure of mutual complete dependence in discrete variables through subcopula. *International Journal of Approximate Reasoning*, 65:11–23, 2015.

Tasena, S. On a distribution form of subcopulas. *International Journal of Approximate Reasoning*, 128:1–19, 2021a.

Tasena, S. On metric spaces of subcopulas. *Fuzzy Sets and Systems*, 415(15):76–88, 2021b.

Wei, Z. and Kim, D. Subcopula-based measure of asymmetric association for contingency tables. *Statistics in medicine*, 36(24):3875–3894, 2017.