GOODMAN AND KRUSKAL'S GAMMA COEFFICIENT FOR ORDINALIZED BIVARIATE NORMAL DISTRIBUTIONS

Abstract

We consider a bivariate normal distribution with linear correlation ρ whose random components are discretized according to two assigned sets of thresholds. On the resulting bivariate ordinal random variable, one can compute Goodman and Kruskal's gamma coefficient, which is a common measure of ordinal association. Given the known analytical monotonic relationship between Pearson's ρ and Kendall's rank correlation τ for the bivariate normal distribution, and since for a bivariate continuous variable, Kendall's τ coincides with Goodman and Kruskal's γ, the change of this association measure before and after discretization is worth being studied. We consider several experimental settings obtained by varying the two sets of thresholds, or, equivalently, the marginal distributions of the final ordinal variables; in particular, we examine the case of uniform, symmetrical unimodal or bimodal, and triangular distributions, with different number of categories. This study, confirming previous findings, shows how the gamma coefficient is in absolute value always larger than Kendall's rank correlation; this discrepancy lessens when the number of categories increases or, given the same number of categories, by using equally probable categories. Based on these results, a proposal is suggested to build a bivariate ordinal variable with assigned margins and association (expressed in terms of Goodman and Kruskal's γ), by ordinalizing a bivariate normal distribution. An application to statistical inference and an illustrative example employing real data are provided.

Key words: Bivariate normal distribution, Discretization, Gamma coefficient, Latent variable, Ordinal association

1. Introduction

The use of the multivariate normal distribution as a latent construct for modelling observed correlated or associated discrete or ordinal variables can be dated back to the seminal book by Lazarsfeld and Henry (1968) and to the later work by Muthen (1983), in the context of structural equation modeling. Latent variable modelling has then gradually become an integral part of mainstream statistics and is currently used for a multitude of applications in different subject areas (Beaujean, 2014).

It is also an indisputable fact that the ability to simulate artificial data resembling the main features of some observed dataset or following the specifications of a study design is necessary when comparing and investigating the behaviour of statistical procedures and exploring their robustness; such features or specifications can be often conveniently summarized by the empirical marginal distributions and pairwise measures of correlation or association.

In this work, we will focus our attention on the relationship between a concordance measure of the bivariate normal latent variable (Kendall's rank correlation, straight related to the more popular Pearson's correlation) and a correspondent association measure of the ordinalized variable (Goodman and Kruskal's gamma). On the one hand, we will analyze how the correlation of the latent variable and the marginal distributions of the final ordinal variables affect their correlation/association: we can call it the "direct problem". Specifically, we will investigate the effect of the number of categories and the probability distribution (uniform/symmetrical/skewed) on the distortion of the association measure. On the other hand, we will provide a procedure that allows constructing and simulating a random vector of discrete variables with assigned marginal distributions and association by discretizing a bivariate normal random vector with a pairwise correlation that is able to induce the desired association among the ordinal variables (the "inverse problem", i.e., finding the parameters of the continuous model that produce the (parameters of the) observed ordinal variables).

The paper is structured as follows. In the next section, we will recall the definition of Goodman and Kruskal's gamma coefficient and resume its main properties, also in comparison with alternative measures of ordinal association. In Section 3, we will analyze the change in magnitude of association before and after ordinalization of a bivariate normal variable, measured by Kendall's tau and Goodman and Kruskal's gamma, respectively, under a wide array of experimental conditions. Section 4 suggests and examines a procedure for building a bivariate ordinal variable with assigned marginal distributions and association. Section 5 proposes a possible application to inference of the algorithm of Section 4; Section 6 presents an illustrative example on real data. The last section is devoted to some final remarks and possible research prospects.

2. Measures of ordinal association: Goodman and Kruskal's gamma

Measures of ordinal association between two variables use the property that the categories of ordinal variables have a natural order; but the way in which these measures use this information may differ considerably. Some measures are based on the intuitive notion that ordinal association should have an interpretation analogous to that of metric association, say Pearson's correlation; in this group, we find Spearman's ρ, Kendall's τ, polychoric correlation, Somer's d, and

Goodman-Kruskal's γ. These coefficients allow us to make statements of the general form "if scores on X increase, than most probably scores on Y will increase" (Kampen and Swyngedouw, 2000). Other measures of ordinal association have a background in information theory, and have interpretations in terms of stochastic entropy (Bryson and Phillips, 1975; Laird, 1979; Gilula, 1988). Henceforth, we will focus on the former family of measures.

Let us consider a couple of ordinal variables (X, Y), with H and K ordered categories, respectively, and introduce the concept of concordance/discordance by considering two independent realizations (x_i, y_i) and (x_j, y_j). (x_i, y_i) and (x_j, y_j) will be said concordant if $x_i < x_j$ and $y_i < y_j$, or if $x_i > x_j$ and $y_i > y_j$. Conversely, (x_i, y_i) and (x_j, y_j) will be said discordant if $x_i < x_j$ and $y_i > y_j$, or if $x_i > x_j$ and $y_i < y_j$. Thus, one can define the probability of concordance as

$$\mathit{\Pi}_c = Pr\left\{X_i < X_j \text{ and } Y_i < Y_j\right\} + Pr\left\{X_i > X_j \text{ and } Y_i > Y_j\right\}$$

and similarly the probability of discordance as

$$\mathit{\Pi}_d = Pr\left\{X_i < X_j \text{ and } Y_i > Y_j\right\} + Pr\left\{X_i > X_j \text{ and } Y_i < Y_j\right\}.$$

The gamma coefficient belongs to a larger family of ordinal correlation measures (see e.g. Woods, 2009, for an exhaustive account); it is defined as the following ratio:

$$\gamma = \frac{\mathit{\Pi}_c - \mathit{\Pi}_d}{\mathit{\Pi}_c + \mathit{\Pi}_d} \tag{1}$$

$\mathit{\Pi}_c$ and $\mathit{\Pi}_d$ can be conveniently expressed in terms of the joint probabilities $p_{ij} = P(X = x_i, Y = y_j)$:

$$\mathit{\Pi}_c = 2\sum_{i<h}\sum\sum_{j<k}\sum p_{ij}p_{hk}, \quad \mathit{\Pi}_d = 2\sum_{i<h}\sum\sum_{j>k}\sum p_{ij}p_{hk}.$$

By defining the quantities

$$p_{hk}^{(c)} = \sum_{a<h}\sum_{b<k} p_{ab} + \sum_{a>h}\sum_{b>k} p_{ab}$$

and

$$p_{hk}^{(d)} = \sum_{a<h}\sum_{b>k} p_{ab} + \sum_{a>h}\sum_{b<k} p_{ab}$$

for each $h = 1, \ldots, H$, $k = 1, \ldots, K$, then $\mathit{\Pi}_c$ and $\mathit{\Pi}_d$ can be rewritten as

$$\Pi_c = \sum_{h=1}^{H}\sum_{k=1}^{K} p_{hk}p_{hk}^{(c)}, \; \Pi_d = \sum_{h=1}^{H}\sum_{k=1}^{K} p_{hk}p_{hk}^{(d)}.$$

For a bivariate continuous random variable, γ can be still computed through Equation (1): in this case, since the probability of concordance and the probability of discordance sum up to 1 (there is no probability of tied values for neither X nor Y), then $\gamma = \mathit{\Pi}_c - \mathit{\Pi}_d = 2\mathit{\Pi}_c - 1$, and it coincides with Kendall's rank correlation τ (Kendall, 1945), simply defined as the difference between the probability of concordance and the probability of discordance. We recall that for two continuous random variables X and Y, Kendall's τ depends only on the unique copula C of the bivariate random vector (X, Y), specifically:

$$\tau(X,Y) = 4\int_0^1\int_0^1 C(u_1,u_2)\mathrm{d}C(u_1,u_2) - 1,$$

and not on the marginal distributions of X and Y. τ ranges between -1 and $+1$; it attains the upper bound if and only if X and Y are comonotonic, whereas it attains the lower bound if and only if X and Y are countermonotonic.

Like Pearson's correlation and other correlation measures such as Spearman's rho and the aforementioned Kendall's tau, γ takes values in $[-1, +1]$. The values -1, 0, and $+1$ are attained when $\Pi_c = 0$, $\Pi_c = \Pi_d$, $\Pi_d = 0$, respectively. There are however some potential problems with the gamma coefficient, which were immediately recognized by Goodman and Kruskal (1954):

- γ is unstable over various "cutting points", that is to say, γ tends to increase as the categories of a contingency table are collapsed, since γ gives no consideration to tied pairs, as can be seen by Equation (1), and the number of tied pairs increases as the table is collapsed;
- γ also usually yields greater association values than other measures of ordinal association as it does not consider any of the tied pairs;
- Finally, γ is a weakly monotonic measure of ordinal association, i.e., it reaches $+1$ under a variety of cell frequency configurations, not only in case of strict perfect association (Berry and Johnston and Mielke, 2018). For example, the bivariate distribution of Table 1, where all the joint probabilities are zero except those labelled with $\times$, presents a value of γ equal to $+1$, although the relationship between X and Y is not perfectly monotonic (Kendall's τ and Spearman's ρ would be strictly smaller than 1).

With the objective of overcoming these pitfalls, some modifications to the gamma coefficient have been proposed (Rousson, 2007).

==========================

Insert Table 1 about here

==========================

A remark on the possible extension of the gamma coefficient to higher dimensions can be made. Goodman-Kruskal's gamma can be rewritten as

$$\gamma = \mathbb{E}\left[\mathrm{sgn}(X-Y)(\tilde{X}-\tilde{Y})|\mathrm{sgn}(X-Y)(\tilde{X}-\tilde{Y}) \neq 0\right]$$

where $(\tilde{X}, \tilde{Y})$ is an independent copy of (X, Y) and the function sgn is defined as

$$\mathrm{sgn}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$$

In dimension $d > 2$, let us consider the d-variate random vector $\boldsymbol{X} = (X_1 X_2 \dots X_d)^T$, then the γ association matrix remains defined as

$$\Gamma(\boldsymbol{X}) = [\gamma_{ij}]_{i=1,\dots,d; j=1,\dots,d}$$

with $\gamma_{ij} = \gamma(X_i, X_j)$. This matrix however is not necessarily positive semidefinite, as happens for Spearman's or Kendall's rank correlation matrices, which are both valid correlation matrices. An easy counterexample is provided as follows. Consider the trivariate distribution of Table 2:

=========================
Insert Table 2 about here
=========================

The values of γ for the bivariate distributions (X, Y), (X, Z), and (Y, Z) are $\gamma_{xy} = 1/3$, $\gamma_{xz} = 1$, and $\gamma_{yz} = -7/9$, respectively. The corresponding γ association matrix,

$$\Gamma(X, Y, Z) = \begin{bmatrix} 1 & 1/3 & 1 \\ 1/3 & 1 & -7/9 \\ 1 & -7/9 & 1 \end{bmatrix},$$

is clearly not positive semidefinite.

On n pairs of observations constituting a multinomial sample, the sample analog of γ is given by

$$\hat{\gamma} = \frac{C - D}{C + D},$$

with C the number of concordant pairs and D the number of discordant pairs. It is possible to prove that $\sqrt{n}(\hat{\gamma} - \gamma)$ is asymptotically normal with mean zero and variance given by

$$\sigma^2 = \sum_i \sum_j p_{ij} \phi_{ij}^2 / (\Pi_c + \Pi_d)^4 = \frac{16}{(\Pi_c + \Pi_d)^4} \sum_i \sum_j p_{ij} (\Pi_c p_{ij}^{(d)} - \Pi_d p_{ij}^{(d)})^2$$

being $\phi_{ij} = 4(\Pi_d p_{ij}^{(c)} - \Pi_c p_{ij}^{(d)})$ (see, e.g., Agresti, 2010).

In practice, replacing p_{ij}, Π_c, and Π_d by their sample values in σ^2 yields the ML estimate $\hat{\sigma}^2$ of σ^2. The term $SE = \hat{\sigma}/\sqrt{n}$ is an estimated standard error for $\hat{\gamma}$ and a Wald confidence interval for γ is $(\hat{\gamma} \pm z_{1-\alpha/2} SE)$.

Rosenthal (1966); Gans and Robertson (1981) showed that $\hat{\gamma}$ has a tendency to converge slowly to normality and to have distributional irregularity, bias, and skewness problems, especially when the true absolute value γ is large. O'Gorman and Woolson (1988); Carr et al. (1989) pointed out that better convergence occurs using the Fisher-type transform

$$\hat{\xi} = \frac{1}{2} \log \frac{1 + \hat{\gamma}}{1 - \hat{\gamma}},$$

whose asymptotic variance equals the asymptotic variance of $\hat{\gamma}$ multiplied by $(1 - \gamma^2)^{-2}$. A confidence interval can be constructed for ξ and then inverted to one for γ, by using the inverse transformation

$$\hat{\gamma} = \frac{e^{2\hat{\xi}} - 1}{e^{2\hat{\xi}} + 1}.$$

3. Analysis of the relationship between ρ and γ for ordinalized bivariate normal distribution

Discretization of continuous variables is commonly encountered in practice. Based on observed nominal age, income, temperature, and depression score, one can derive ordinal variables such as young-middle-old age, low-medium-high income, cold-cool-average-hot temperature, no-mild-moderate-severe depression. Discretization is usually avoided by statisticians for intuitive and valid reasons, the most prominent of which is the power and information loss. Some problems related to identification of statistical models obtained by discretization have been recently raised by Grønneberg and Foldnes (2019). However, simplicity, better interpretability and comprehension of the effects of interest, and superiority of some categorical data measures such as odds ratio have been argued by proponents of discretization (Liu et al., 2002).

The objective of this section is the determination of association magnitude changes when the univariate components of a bivariate normal distribution are ordinalized, i.e., the range of each continuous component is divided into contiguous intervals, which are assigned an ordered category. A similar and extensive analysis has been conducted by Demirtas and Vardar-Acar (2017) for investigating the effects of discretization of continuous variables on the magnitude of linear correlation. If the underlying continuous distribution is bivariate normal, it can be proved that discretization (i.e., assignment of consecutive positive integer values to contiguous intervals) always preserves the sign of linear correlation and more importantly leads to a reduction in magnitude. This result, which was empirically observed in many simulation studies (see, e.g., Bollen and Barb, 1981) and claimed to hold only "in large samples" by Demirtas and Vardar-Acar (2017), is just a consequence of a previous more general theoretical result, named Lancaster's theorem (Lancaster, 1957), which states that the correlation of a bivariate normal cannot increase whatever transformations are applied to its univariate components (see also Mari and Kotz, 2001, p.155, where it is reported as an "extremal property" of the bivariate normal distribution).

For a bivariate normal distribution with correlation coefficient ρ the following relationship holds between ρ and Kendall's rank correlation τ:

$$\tau = \frac{2}{\pi} \arcsin \rho. \tag{2}$$

Equation (2) holds also for most elliptical distributions, e.g., for bivariate Student's t, and for most bivariate distributions whose dependence structure is described by an elliptical copula (McNeil and Frey and Embrechts, 2005). If both univariate margins are discretized/ordinalized through pre-specified thresholds, one can compute some measure of ordinal association on the resulting bivariate ordinal variable, such as Goodman and Kruskal's γ: we know (see the previous section) that for continuous bivariate rvs, such as the bivariate normal, the definitions of γ and τ coincide. Thus, it makes sense to analyze the relationship between τ and γ as a measure of the change in association before and after ordinalization.

Let us start from a very simple case, i.e., dichotomization of the margins of a bivariate normal random variable $(Z_1, Z_2)^\intercal$ with mean vector $(\mu_1, \mu_2)^\intercal$ and linear correlation ρ. Consider the bivariate ordinal variable (X, Y) obtained as follows:

$$X = \begin{cases} x_1 & Z_1 \le \mu_1 \\ x_2 & Z_1 > \mu_1 \end{cases} \quad Y = \begin{cases} y_1 & Z_2 \le \mu_2 \\ y_2 & Z_2 > \mu_2 \end{cases}, \tag{3}$$

with $x_1 < x_2$ and $y_1 < y_2$ being two ordered categories. We have therefore that the marginal probabilities for X and Y are $P(X = x_1) = P(X = x_2) = 1/2$ and analogously $P(Y = y_1) = P(Y = y_2) = 1/2$. Since we know that for a bivariate normal distribution $P(Z_1 \le \mu_1, Z_2 \le \mu_2) = P(Z_1 > \mu_1, Z_2 > \mu_2) = \frac{1}{4} + \frac{1}{2\pi} \arcsin \rho$, irrespective of the values of marginal variances (see e.g. McNeil and Frey and Embrechts, 2005), then the joint probability of (X, Y) is that displayed in the following table:

X, Y	y_1	y_2	tot
x_1	$1/4 + 1/(2\pi) \arcsin \rho$	$1/4 - 1/(2\pi) \arcsin \rho$	$1/2$
x_2	$1/4 - 1/(2\pi) \arcsin \rho$	$1/4 + 1/(2\pi) \arcsin \rho$	$1/2$
tot	$1/2$	$1/2$	1

Based on it, by using Equation (1), one can compute the gamma coefficient as a function of ρ:

$$\gamma = \frac{4\pi \arcsin \rho}{\pi^2 + 4(\arcsin \rho)^2}, \tag{4}$$

or, recalling (2), in terms of τ:

$$\gamma = \frac{2\tau}{1 + \tau^2}, \tag{5}$$

whose graph is plotted in Figure 1. This graph clearly shows how Goodman-Kruskal's gamma for the dichotomized rv is a strictly increasing and odd function of its analogue Kendall's tau for the bivariate normal distribution and, in absolute value, is always larger than or equal to τ, the equality holding when $\tau = 0$ (i.e., when the two normal components are independent), when τ equals 1 (perfectly positively correlated normal rvs) or when τ equals -1 (perfectly negatively correlated normal rvs). It is worth noticing that in this context γ is equal to $+1$ (-1) only in case of perfect monotonic (countermonotonic) relationship between X and Y.

========================
Insert Figure 1 about here
========================

By choosing thresholds different from the corresponding marginal means of the bivariate normal rv, which we now assume without loss of generality to have standard components, we can numerically obtain the value of γ corresponding to any value of τ. Similarly, one can discretize either or both continuous random components into more than two categories, applying different sets of thresholds (or, equivalently, assigning different marginal distributions to the final ordinal variables). In general, let us consider a bivariate standard normal distribution (Z_1, Z_2) whose two continuous components are discretized into two ordinal variables X and Y according to the following scheme:

$$X = \begin{cases} x_1 & Z_1 < \theta_1 \\ \cdots & \\ x_H & \theta_{H-1} \leq Z_1 < \theta_H = +\infty \end{cases}$$

and

$$Y = \begin{cases} y_1 & Z_2 < \eta_1 \\ \cdots & \\ y_K & \eta_{K-1} \leq Z_2 < \eta_K = +\infty \end{cases}$$

where $\theta_1 < \theta_2 < \cdots < \theta_H = \infty$ and $\eta_1 < \eta_2 < \cdots < \eta_K = \infty$ constitute two sets of thresholds. If we denote with $F(i,j)$ the bivariate joint cumulative probability $P(X \leq x_i, Y \leq y_j)$ of the bivariate ordinal variable (X,Y), $i = 1,\ldots,H$, $j = 1,\ldots,K$, and let $F_1(i) := P(X \leq x_i)$ and $F_2(j) := P(Y \leq y_j)$ be the two marginal cumulative distributions, we have the following relationship between F_1 (and, similarly, F_2), and the thresholds on Z_1 (on Z_2):

$$F_1(i) = \Phi(\theta_i) \text{ and } F_2(j) = \Phi(\eta_h),$$

and the joint cdf of (X,Y) is

$$F(i,j) = \Phi_\tau(\theta_i, \eta_j) = \Phi_\tau(F_1^{-1}(i), F_2^{-1}(j)), \tag{6}$$

where Φ_τ is the joint cdf of a bivariate normal with standard components and rank correlation τ, and F_1^{-1} (F_2^{-1}) is the generalized inverse of F_1 (F_2). The joint probabilities $p_{ij} = P(X = x_i, Y = y_j)$ are then obtained as

$$p_{ij} = F(i,j) - F(i-1,j) - F(i,j-1) + F(i-1,j-1) = \sum_{t=0}^{1}\sum_{v=0}^{1} F(i-t, j-v) \cdot (-1)^{t+v} \tag{7}$$

and on these p_{ij} the gamma coefficient can be computed by using Equation (1).

In the following study, we will consider three different possible types of ordinal marginal distributions, namely, i) uniform (i.e., equiprobable categories) ii) symmetrical (non-uniform), based on normal scores, and iii) asymmetrical (triangular). We will examine which is the relationship between τ and γ under these three macro-settings, by varying the number of categories, that for the sake of simplicity we will assume to be the same for the two margins.

Figure 2 displays the graphs of the (τ, γ) curve when the two marginal distributions have the same number of equiprobable categories $H = K = 3; 4; 5; 10; 20$. Table 3 reports the values of γ for τ ranging from -1 to $+1$, with step length of 0.1, and several values of the common number of categories of the two identical margins. Note that the (τ, γ) curve, for a given H, is strictly increasing and always passes through the points $(-1,-1)$,$(0,0)$ and $(1,1)$, and, as one can expect, by increasing H (i.e., when the ordinal distributions “resemble” a continuous one), it tends to get close to the 45 degrees line passing through the origin. Note that γ is an odd function of τ, i.e., $\gamma(\tau) = -\gamma(-\tau)$. For a given value of $\tau > 0$, γ is a decreasing function of H; for a given value of

$\tau < 0$, γ is an increasing function of K; as K tends to ∞, γ seems to converge, though slowly, to τ. It is worth observing, however, that even for a moderately large number of categories, there is a substantial difference between the values of τ and γ before and after ordinalization. When $H = 100$, there is still a difference of the order of 2% between the values of τ and γ.

==========================
Insert Table 3 about here
==========================

==========================
Insert Figure 2 about here
==========================

We now move to non-uniform symmetrical ordinal distributions; in particular, we consider ordinal distributions whose probability mass function resembles somehow the probability density function of the normal variable. For this aim, the $(-4, +4)$ interval, whose probability for a standard normal variable is almost equal to 1, is divided into H equal-width adjacent intervals, $\left(-4 + \frac{8}{H}(i-1), -4 + \frac{8}{H}i\right)$, $i = 1, \ldots, H$. A similar construction was employed by Becker (1989) (where it is named "uniform cut-points (UCP) method" and is employed for studying similarities between uniform association models and ordinalized bivariate normal distribution) and Ferrari and Barbiero (2012) (for studying the effects of discretization of a bivariate normal distribution on the correlation coefficient). To the i-th interval, the i-th ordered category is associated whose probability is thus equal to $\Phi(-4 + 8i/H) - \Phi(-4 + 8(i-1)/H)$, $i = 1, \ldots, H$; the residual probabilities of $(-\infty, -4)$ and $(4, +\infty)$ are assigned to the first and last category, respectively. Thus the probability distribution is always symmetrical about the central category (if H is odd) or categories (if H is even). The probabilities obtained according to this scheme are graphically displayed, for $k = 5$ and $k = 10$, in Figure 3.

Table 4 reports the values of Goodman and Kruskal's γ for τ ranging from -1 to $+1$, with step length of 0.1, and several values of the common number of categories of the two identical symmetrical margins. Note that as in the case of equiprobable categories, γ is an odd function of τ. For a given value of $\tau > 0$ and for $K > 3$, γ is a decreasing function of K; for a given value of $\tau < 0$ and for $K > 3$, γ is an increasing function of K (for $K = 2$, we obtain again a uniform distribution with two equally probable categories). As K tends to ∞, γ converges to τ, but much more slowly than in the case of uniform margins. When $K = 100$, the difference between τ and γ in absolute values is non-negligible at all: for the examined settings, it ranges between 3.5 and 4.5%. This fact can be easily explained just looking at the barplot on the right of Figure 3: among the K categories, only the central ones have appreciable probabilities, so the "actual" number of categories is (much) smaller than K and then a larger value of K is required for obtaining the same difference between τ and γ that is achieved in case of uniform distributions.

==========================
Insert Figure 3 about here
==========================

========================
Insert Table 4 about here
========================

In the end, we considered strongly asymmetrical distributions, namely triangular distributions with H categories, where the probability of the i-th category is proportional to i (through some positive constant α). Since under this assumption $\sum_{i=1}^{H} p_i = \sum_{i=1}^{H} \alpha i = \alpha \frac{H(H+1)}{2}$, we have that $p_i = \frac{2i}{H(H+1)}$. Choosing this distribution as the marginal distribution of both X and Y, Table 5 reports the corresponding values of γ for τ ranging from -1 to $+1$, with step length of 0.1, and for several values of the common number of categories of the two identical triangular margins. One can notice that differently from the two cases analyzed before, due to the asymmetrical nature of the margins, γ is no longer an odd function of τ, being $\gamma(\tau)$ generally different from $-\gamma(-\tau)$.

For a given value of $\tau > 0$, γ is a decreasing function of K; for a given value of $-0.9 < \tau < 0$, γ is an increasing function of K. When $\tau = -0.9$, we have a curious behavior of γ as a function of K, which is displayed in Figure 4.

========================
Insert Figure 4 about here
========================

It is worth underlining how the values of $|\gamma|$ are significantly larger than the corresponding $|\tau|$, especially when the number of categories is small. This feature is more apparent for negative values of τ. For example, when $\tau = -0.5$ and $H = 5$, the value of γ is -0.69721. Even when discretizing the two continuous components through an extremely large number of ordered categories, such as 100, the difference between the values of τ and γ is still non-negligible (not smaller than 2.5% in absolute value).

========================
Insert Table 5 about here
========================

4. Building a bivariate ordinal variable with prescribed margins and gamma coefficient

Researchers can be interested in building and simulating samples from a bivariate (multivariate) ordinal distribution with prescribed marginal distributions and (pairwise) levels of association, expressed in terms of gamma coefficient. Such a concern may arise when one has to show appropriateness and study performance or robustness of a novel statistical technique; since it is not always possible to do it by using analytic arguments, unless in elementary cases, on the contrary, generating data replicates that mimic the real data characteristics of interest allows one to study the performance of the statistical method in any given setting. For example, researchers may be interested in assigning and preserving in their simulation study the marginal structures as well as ordinal associations (Demirtas, 2006). In other circumstances, their concern is matching marginal means, variances, skewnesses and kurtoses, and intercorrelations (Vale and Maurelli, 1983). van der Ark and van Aert (2015), with the aim of investigating the performance of different types of confidence intervals for γ, constructed several bivariate ordinal distributions by assigning each a fixed value of γ and imposing constraints on the margins, by considering uniform or skewed distributions.

However, when assigning the margins and association (correlation) value for two random variables, a uniqueness issue arises for the corresponding bivariate distribution. In fact, one of the fallacies of Pearson's correlation is that if we consider a bivariate random vector (X, Y) with assigned marginal distributions F_1 and F_2 and a feasible value ρ, then F_1, F_2, and ρ do not in general determine the joint distribution F univocally; on the contrary, there may exist several (even infinite) joint distributions whose margins are F_1 and F_2 and whose linear correlation is ρ. This issue is not overcome even by other dependence measures such as Spearman's ρ and Kendall's τ, and possibly by Kruskal and Goodman's γ. Nevertheless, if we restrict our focus on the family of joint distributions obtained by discretizing a specific class of continuous random vectors, we can find out that there is unique joint distribution satisfying the match on margins and association value.

Let us start with the simplest case. For a bivariate ordinal distribution with dichotomous margins, i.e., for a joint probability p_{ij}, with $i = 1, 2$, $j = 1, 2$, the problem consists of finding the solution to the following equation system:

$$\begin{cases} p_{11}p_{22}(1-\gamma) & = p_{12}p_{21}(1+\gamma) \\ p_{12} & = 1 - p_{\cdot 1} - p_{22} \\ p_{21} & = 1 - p_{1\cdot} - p_{22} \\ p_{11} & = p_{1\cdot} - p_{12} \end{cases}$$

which becomes

$$\begin{cases} (p_{1\cdot} + p_{\cdot 1} + p_{22} - 1)p_{22}(1-\gamma) & = (1 - p_{\cdot 1} - p_{22})(1 - p_{1\cdot} - p_{22})(1+\gamma) \\ p_{12} & = 1 - p_{\cdot 1} - p_{22} \\ p_{21} & = 1 - p_{1\cdot} - p_{22} \\ p_{11} & = p_{1\cdot} - p_{12} \end{cases}$$

and then

$$\begin{cases} 2\gamma p_{22}^2 - p_{22}(1 + 3\gamma - 2\gamma(p_{1\cdot} + p_{\cdot 1})) + (1+\gamma)\cdot(1 - p_{1\cdot})(1 - p_{\cdot 1}) & = 0 \\ p_{12} & = 1 - p_{\cdot 1} - p_{22} \\ p_{21} & = 1 - p_{1\cdot} - p_{22} \\ p_{11} & = p_{1\cdot} - p_{12} \end{cases}$$

From the first equation of the system above we derive that p_{22} is the unique real root (the one with sign minus) of the second-order equation $ax^2 + bx + c = 0$ with

$$\begin{cases} a & = 2\gamma \\ b & = -(1 + 3\gamma - 2\gamma \cdot p_{1\cdot} - 2\gamma \cdot p_{\cdot 1}) \\ c & = (1+\gamma)\cdot(1 - p_{1\cdot})\cdot(1 - p_{\cdot 1}); \end{cases}$$

the other joint probabilities can be then easily computed in cascade from the remaining equations. Thus, the problem has a unique feasible solution. For example, if we assign the margins $p_{1\cdot} = 0.5$, $p_{\cdot 1} = 0.5$, and set $\gamma = 3/5$, we obtain the following second-degree equation: $3x^2 - 4x + 1$, whose roots are $p_{22,1} = 1$ and $p_{22,2} = 1/3$; taking the latter leads to the solution:

$$\begin{cases} p_{22} &= 1/3 \\ p_{12} &= 1/6 \\ p_{11} &= 1/3 \\ p_{12} &= 1/6 \end{cases}$$

Note that this set of probabilities corresponds to the joint distribution of the bivariate ordinal variable obtained by discretizing a bivariate normal (3) with $\rho = 1/2$ (or, equivalently, $\tau = 1/3$, see Equation (4)).

If we assign the margins $p_{1\cdot} = 0.6$, $p_{\cdot 1} = 0.3$, and set $\gamma = .25$, we obtain the following solution:

$$\begin{cases} p_{22} &= 0.305 \\ p_{12} &= 0.095 \\ p_{11} &= 0.205 \\ p_{12} &= 0.395 \end{cases}$$

For larger numbers of categories, there will be infinite solutions, which are not easy to derive, due to the nonlinear nature of the problem: the gamma coefficient is not linear in the p_{ij}'s, and then the equation system is non-linear, as we have already experienced with a 2×2 contingency table. This means that there are several bivariate ordinal distributions p_{ij}, whose margins are the assigned $p_{i\cdot}$ and $p_{\cdot j}$, and whose gamma coefficient is equal to a prespecified value $\gamma \in [-1, +1]$. If one is interested in building just *one* bivariate ordinal variable with assigned margins and gamma, he/she can restrict the attention to the class of bivariate ordinal variables with the assigned margins derived by ordinalization of a standard bivariate normal distribution; then he/she can exploit the results of the previous section, by accommodating the value of the parameter τ in order to match the assigned value γ between the assigned ordinal margins. Due to the monotonic relationship between γ and τ, whose form depends on the selected margins, and to the fact that γ can span its entire natural range $[-1, +1]$, which we proved empirically, we are sure there will be a unique bivariate ordinalized distribution satisfying the requested requisites. This means that among all the bivariate distributions with assigned margins and γ, we select the unique one whose underlying continuous latent model is the bivariate normal.

Writing the relationship between the two association measures as $\gamma = g(\tau; F_1, F_2)$, we just need to find the (unique) root τ of the equation $\gamma - g(\tau; F_1, F_2) = 0$, being γ an assigned number in $[-1, +1]$ and F_1, F_2 assigned ordinal distributions. Recovering the correct τ is a task that can be carried out by using an iterative procedure, which requires setting an initial value. Since we have empirically found that γ after discretization is - in absolute value - larger than the rank correlation τ of the bivariate normal rv, one can use $\tau^{(0)} = \gamma$ as a starting value for the unknown τ. One can then determine the corresponding value of the bivariate ordinal rv $\gamma^{(0)}$, and then iteratively adjust the value of τ according to some updating rule, till the γ converges to the assigned value. The updating process shall take into account that when τ is zero, γ is zero, too; a proposal is suggested in the following algorithm:

1. Set $\tau^{(0)} \leftarrow 0, \gamma^{(0)} \leftarrow 0$; let $\epsilon > 0$ be an arbitrarily small number
2. Set $t \leftarrow 1$ and $\tau^{(t)} \leftarrow \gamma$
3. Compute $F(i, j; \tau^{(t)})$ by using (6)
4. Compute $p(i, j; \tau^{(t)})$ by using (7)
5. Compute $\gamma^{(t)}$ for $p(i, j; \tau^{(t)})$ by using (1)
6. If $|\gamma^{(t)} - \gamma| < \epsilon$ stop; else
   set $t \leftarrow t + 1$,
   $\tau^{(t)} \leftarrow \tau^{(t-1)} + m^{(t)}(\gamma - \gamma^{(t-1)})$, with $m^{(t)} = \dfrac{\tau^{(t-1)} - \tau^{(t-2)}}{\gamma^{(t-1)} - \gamma^{(t-2)}}$; go back to 3.

Given the monotonic relationship between τ and γ for the bivariate normal model and its ordinalized counterpart, the above algorithm should be able to recover the value of τ inducing the target γ in a few steps, for any choice of γ and of F_1 and F_2.

Example 1. Consider the following margins for X: $p_{1\cdot} = p_{2\cdot} = p_{3\cdot} = p_{4\cdot} = 0.25$, and for Y: $p_{\cdot 1} = 0.1$, $p_{\cdot 2} = 0.2$ $p_{\cdot 3} = 0.3$, $p_{\cdot 4} = 0.4$, and assign to γ the value 0.5. After only 5 iterations, the algorithm illustrated above recovers the joint distribution ensuring the target γ and the assigned margins, by ordinalization of a bivariate standard normal variable; it is displayed in Table 6.

==========================
Insert Table 6 about here
==========================

This joint distribution is not the unique satisfying the requirements on margins and gamma. A different one can be obtained as a convex combination of the cograduation and countergraduation tables (Table 7):

==========================
Insert Table 7 about here
==========================

The joint probabilities $p^*_{ij} = \lambda p^M_{ij} + (1 - \lambda) p^W_{ij}$, with $\lambda = 0.7293$, preserve the margins and ensure the target γ. It is worth underlining that the non-linearity of γ does not allow its direct derivation for a convex combination of two joint probability distributions. In fact, the value of γ for a convex combination of two joint distributions is not equal to the same convex combination of the corresponding γ's; in the case of combination of cograduation and countergraduation tables, it is not equal to $\lambda \cdot 1 + (1 - \lambda) \cdot (-1) = 2\lambda - 1$.

5. Application to inference

If a bivariate ordinal sample of size n is available, one can assume it is an *i.i.d.* sample from an ordinalized bivariate normal distribution, or, more generally, from a bivariate continuous distribution whose unique copula is the Gaussian one (see, in this sense, Grønneberg and Foldnes, 2019, for a more detailed account on identifiability issues). One has then to estimate the value of the unknown τ (or ρ) and those of the unknown thresholds which define the marginal distributions of X and Y; this can be done by numerically maximizing the joint log-likelihood of the observed sample with respect to all the unknown parameters simultaneously (see Olsson, 1979, for details):

$$\max \sum_{h=1}^{H}\sum_{k=1}^{K} \log p_{hk}(\tau, \theta_1, \ldots, \theta_{H-1}, \eta_1, \ldots, \eta_{K-1}) n_{hk}, \tag{8}$$

where p_{hk} is obtained from (7) being F derived according to (6), $\theta_H = \eta_K = +\infty$; n_{hk} is the frequency of (x_h, y_k) in the sample.

A starting value $\tau^{(0)}$ for τ can be the sample τ, $\hat{\tau}$; for the thresholds, $\theta_h^{(0)} = \Phi^{-1}(\hat{F}_1(h))$, $\eta_k^{(0)} = \Phi^{-1}(\hat{F}_2(k))$, where $\hat{F}$ denotes the empirical cdf:

$$\hat{F}_1(x) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{x_t \le x}, \hat{F}_2(y) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{y_t \le y},$$

being (x_t, y_t) the t-th bivariate observation in the sample.

Instead of applying the full maximum likelihood approach, one can resort to the two-stage maximum likelihood method. At the first stage, the two marginal log-likelihood functions for X and Y are maximized separately as if they were independent, thus obtaining the MLEs of both sets of thresholds; for our case, $\hat{\theta}_h = \Phi^{-1}(\hat{F}_1(h))$, $h = 1, \ldots, H-1$, $\hat{\eta}_k = \Phi^{-1}(\hat{F}_2(k))$, $k = 1, \ldots, K-1$. Then, the joint log-likelihood function is maximized with respect to the only remaining parameter τ, setting the threshold parameters equal to the corresponding MLEs obtained at the first stage. The implementation of this method is much less time-consuming than the first one and often the resulting estimator is nearly as efficient as the full MLE (Joe and Xu, 1996).

As an alternative to both the previous approaches, based on the log-likelihood function, one can resort to the following method of moments for estimating τ and the thresholds:

1. Compute the empirical cumulative distribution functions $\hat{F}_1$ and $\hat{F}_2$ on the two margins of the bivariate sample.

   Compute the sample value of γ, $\hat{\gamma}_M$, on the bivariate sample.
2. Compute the value of τ, $\hat{\tau}_M$, of the underlying bivariate normal distribution inducing $\hat{\gamma}_M$ given $\hat{F}_1$ and $\hat{F}_2$, by resorting to the iterative procedure of Section 4. That is, since we can write $\gamma = g(\tau; F_1, F_2)$ and $\tau = g^{-1}(\gamma; F_1, F_2)$, $\hat{\tau}_M = g^{-1}(\hat{\gamma}_M; \hat{F}_1, \hat{F}_2)$

Given the way the estimates are derived in both stages, this method can be classified as a two-stage method of moments, too. The unknown parameters are in fact estimated in two sequential steps: first, the thresholds for Z_1 and Z_2 are estimated independently based on the empirical cdf of X and Y, respectively; then, the dependence parameter τ is estimated based on the estimates computed at the first stage and on the sample value of γ.

Example 2. Let us consider the bivariate ordinal sample whose joint distribution is displayed in Table 8.

=========================
Insert Table 8 about here
=========================

The value of $\hat{\gamma}$ here is 0.4632. As starting values for the parameter estimates we can consider:

$$\begin{cases} \tau^{(0)} & = 0.4632 \\ \theta_1^{(0)} & = \Phi^{-1}(10/30) = -0.4307 \\ \theta_2^{(0)} & = \Phi^{-1}(20/30) = 0.4307 \\ \eta_1^{(0)} & = \Phi^{-1}(5/30) = -0.9674 \\ \eta_2^{(0)} & = \Phi^{-1}(15/30) = 0 \end{cases}$$

By maximizing the joint log-likelihood function in (8), we derive the MLEs:

$$\begin{cases} \hat{\tau} = 0.2803 \\ \hat{\theta}_1 = -0.4280 \\ \hat{\theta}_2 = 0.4312 \\ \hat{\eta}_1 = -0.9657 \\ \hat{\eta}_2 = -0.0024 \end{cases}$$

The asymptotic standard error of $\hat{\tau}$ is 0.1418. The joint ordinalized distribution obtained by using these MLEs is reported in Table 9.

=========================
Insert Table 9 about here
=========================

If we apply the two-stage ML method to the same data, at the first stage, by maximizing the two marginal log-likelihoods of X ad Y, we get the estimates $\theta_1^{(0)}$, $\theta_1^{(1)}$, $\eta_1^{(0)}$, and $\eta_1^{(0)}$ for the thresholds. At the second stage, we obtain an estimate of τ equal 0.2793, just slightly smaller than the full MLE. Its asymptotic standard error is 0.1382.

By employing the two-stage method of moments, first, we find the estimates of γ, F_1 and F_2, which are the same as the estimates of the two-stage ML method; then, based on them and on the algorithm of Section 4, we find the value $\hat{\tau} = 0.2816$, which is slightly different from the estimates computed via the full and two-stage ML methods. In contrast to these latter, for the two-stage moment estimate of τ, it is not possible to provide a value for its (asymptotic) standard error.

6. Real data

In a now classic study of mental health in Manhattan, New York, Srole and Fischer (1978) explore the relationship, among others, between mental impairment (X) and parents' socioeconomic status (Y). Table 10, from that study, has been used extensively to illustrate the utility and application of models for ordered categorical data.

==========================
Insert Table 10 about here
==========================

The sample value of Goodman-Kruskal's γ is 0.15429. By assuming a bivariate standard normal distribution underlying the bivariate ordinal data, we can derive the MLE for Kendall's tau and the thresholds for the two margins, see Table 11. Note that the MLE of τ is 0.10762, quite smaller than the corresponding estimate of γ, confirming the empirical results of the study of Section 3. However, the estimate is significantly greater than zero, with an associated standard error of 0.01718, which indicates strong evidence that mental health status and parents' socioeconomic status are positively associated (i.e., higher socioeconomic status is associated with better mental health), even if the strength of that association is quite small. Note also that all the MLEs of the thresholds are highly significant, except for θ_3, which is the central threshold for the variable Parents' Socioeconomic Status.

==========================
Insert Table 11 about here
==========================

If one applies the two-stage maximum likelihood method, then the estimates of the thresholds differ even if only by a little from those obtained via the full maximum likelihood method, as well as the estimate of τ, which is 0.10760 (with standard error 0.01706).

The method of moment estimate of τ can be recovered applying the algorithm in Section 4 and is equal to 0.10665, resulting slightly different from the other two estimates.

==========================
Insert Table 12 about here
==========================

Table 12 displays the expected joint frequencies $n_{ij}^{(o)}$ under the bivariate ordinalized distribution whose parameters are set equal to the corresponding MLEs. Note that under this model the marginal frequency distributions are not equal to the analogue observed ones of Table 10 (even if they are very close to each other). This is because the full MLEs of the thresholds are not equal to the corresponding marginal MLEs (which are used as starting values for the optimization routine). Comparing the observed an theoretical contingency tables, one can notice some slight discrepancies between homologous frequencies. In order to (approximately) evaluate the goodness-of-fit of the suggested ordinalized bivariate normal model on the data at issue, we computed the usual chi-squared statistic based on the theoretical and observed frequencies (all theoretical joint frequencies are greater than 5, so there is no need for pooling cells); its value is 8.84. Under the null hypothesis that the data come from the ordinalized bivariate normal distribution with parameters equal to their MLEs, this statistic approximately follows a chi square distribution with a number of degrees of freedom equal to $24 - 9 - 1 = 14$, where 24 is the number of pooled frequencies and 9 is the number of estimated parameters. The corresponding p-value is 0.841, leading us to comfortably accept the null hypothesis. The value of the log-likelihood ratio statistic, given by $-2\sum_{i=1}^{H}\sum_{j=1}^{K} n_{ij}\log(n_{ij}^{(o)}/n_{ij})$ is equal to 8.959 (p-value 0.834) and thus leads to the same conclusion.

Other more sophisticated models may fit the sample data better. For some examples of alternatives (namely, association models) fitting the Midtown Manhattan Mental study data, see for example Becker (2014). In particular, a uniform association model can be used: the expected cell frequencies, calculated by using the `vcdExtra` `R` package, Friendly (2017), are displayed in Table 13. Note that the expected frequencies under this uniform association model are quite close to those under the ordinalized bivariate normal model, confirming the findings of Becker (1989) also stressed in Kateri (2014). To prove this, one can compute the Kullback-Leibler distances between the two theoretical joint distributions ($p_{ij}^{(o)}$ and $p_{ij}^{(u)}$, for the ordinalized normal and uniform association models, respectively): $K_1 = \sum\sum p_{ij}^{(o)}\log(p_{ij}^{(o)}/p_{ij}^{(u)})$ and $K_2 = \sum\sum p_{ij}^{(u)}\log(p_{ij}^{(u)})/p_{ij}^{(o)}$. The values are computed as $K_1 = 0.0001721$ and $K_2 = 0.0001724$, which are very small, proving that the two distributions are very close to each other.

==========================
Insert Table 13 about here
==========================

7. Conclusions and further research

We focused on the bivariate standard normal variable and studied the change in association before and after ordinalization, measured by Kendall's τ and Goodman and Kruskal's γ. This analysis has been facilitated i) by the analytic relationship between Pearson's correlation ρ and Kendall's τ for the normal distribution, and ii) by the wide availability of software implementations of the bivariate normal cdf (despite of its non-closed expression) which is required when computing the joint probabilities of the bivariate ordinal distribution resulting from discretization. We empirically investigated the relationship between τ and γ, by considering several specific configurations for the two final marginal distributions (uniform, unimodal or bimodal symmetric, triangular). The study confirmed a somewhat expected result, i.e., the association measure tends to inflate after discretization, and to a larger extent when the number of ordered categories is small. For a same number of categories, we also highlighted which is the effect of the marginal probabilities in the change of association. Based on these results, we also elaborated a scheme for building a bivariate ordinal distribution with assigned margins and value of association. This procedure can be useful when one has to simulate a large number of samples from correlated ordinal random variables with known marginal probabilities and specified association or correlation. Moreover, the procedure implicitly suggests an estimator for τ to be used as an alternative to that obtained through maximum likelihood estimation.

Other bivariate continuous distributions or bivariate copulas can be explored when the interest is in analyzing the effects of ordinalization on Goodman and Kruskal's γ or in constructing a bivariate ordinal variable with assigned marginal distributions and association; one should prefer the parametric copula families satisfying conditions i) and ii) above mentioned. Moreover, one should focus on so-called "comprehensive" bivariate copulas (Nelsen, 2006, p.15), i.e., copulas able to model continuously the whole range of dependence from the lower to the upper Fréchet bounds passing through the product copula. As for i), there is a wide variety of parametric copulas with a closed-form expression of their joint cdf. As for ii), an analytic expression linking the copula parameter to its Kendall's correlation is not always available. For example, for the Frank family of copulas,

$$C(u_1, u_2; \theta) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right], \theta \in \mathbb{R} \setminus \{0\}$$

we have the following relationship between Kendall's τ and θ:

$$\tau = 1 - \frac{4}{\theta}[1 - D_1(\theta)],$$

with $D_1(\theta) = \frac{1}{x} \int_0^x \frac{t}{e^t - 1} \mathrm{d}t$, which can be computed numerically. Note that letting the parameter θ go to 0, the Frank copula boils down to the product copula, with $\tau(\theta = 0) = 0$; when $\theta \to -\infty$, C reduces to the countermonotonicity copula; when $\theta \to +\infty$, C reduces to the comonotonicity copula. The Plackett family of copulas has the following form:

$$C(u_1, u_2; \theta) = \frac{1 + (\theta - 1)(u_1 + u_2) - \sqrt{[1 + (\theta - 1)(u_1 + u_2)]^2 - 4\theta(\theta - 1)u_1 u_2}}{2(\theta - 1)},$$

with $\theta \in (0, +\infty) \setminus \{1\}$. When $\theta \to 1$, it reduces to the product copula, whereas for $\theta \to 0$ it tends to the countermonotonicity copula and for $\theta \to \infty$ to the comonotonicity copula. For this family, there does not appear to be a closed form expression for Kendall's τ (Nelsen, 2006, p.171).

Another point to inspect is whether Kendall's tau for a bivariate continuous distribution is always smaller - in absolute value - than Goodman and Kruskal's gamma computed on any ordinalized version thereof. We have empirically shown that this happens for the bivariate normal distribution, in a certain sense reversing what happens with respect to the correlation coefficient, whose absolute value is always diminished by discretization of both components (Lancaster's theorem). Actually, it happens that if we discretize the two components of a bivariate Student's t distribution when $\rho = \tau = 0$, the corresponding value of γ between the two ordinalized variables is generally different from zero, and then γ does not generally inherit the sign of τ and the inequality $|\gamma| \geq |\tau|$ is no longer true. For example, if we consider two identical marginal distributions with probabilities 1/3 and 2/3 for the two ordered categories, then the bivariate Student's t with uncorrelated components is discretized into the following bivariate ordinal variable

X, Y	y_1	y_2
x_1	0.1153	0.2180
x_2	0.2180	0.4487

whose value of γ is about 0.0424. In the graphs of Figure 5, the relationship between τ and γ for an ordinalized bivariate Student's t. The reader can also refer to Jin and Yang-Wallentin (2017) for some potential alternatives for the bivariate normal distribution assumption as un underlying stochastic model for ordinal data, where the authors study robustness against misspecification of the underlying distribution with respect to the polychoric correlation estimation.

==========================
Insert Figure 5 about here
==========================

As an aside, the paper illustrated how bivariate ordinal models can be built for simulation studies, alternatively to existing procedures employing categorical marginal models. Extension of the proposed procedures to the multivariate case is not straightforward. Analogous issues in determining the existence of a multivariate binary/ordinal variable with assigned margins and Pearson's correlation matrix have been examined by Chaganty and Joe (2006); Barbiero and Ferrari (2017). As we noticed, for a multivariate ordinal variable, the gamma association matrix is not necessarily positive semidefinite, so once all the margins are assigned, it is not straightforward to establish whether an assigned correlation matrix can be a feasible association matrix or not. However, if one selects a valid correlation matrix, this is a feasible Kendall's tau correlation matrix for any choice of the (continuous) margins, and thus, by the arguments related to the change in magnitude after discretization, it should also be a feasible gamma association matrix for any choice of the ordinal margins.

The points raised in this conclusive Section can be addressed by future research.

References

Agresti, A. (2010). Analysis of ordinal categorical data (Vol. 656). John Wiley & Sons.

Barbiero, A., Ferrari, P. A. (2017). An R package for the simulation of correlated discrete variables. Communications in Statistics-Simulation and Computation, 46(7), 5123-5140.

Beaujean, A. A. (2014). Latent variable modeling using R: A step-by-step guide. Routledge.

Bollen, K. A., Barb, K. H. (1981). Pearson's r and coarsely categorized measures. American Sociological Review, 232-239.

Bryson, K., & Phillips, D. (1975). Method for Classifying Interval-Scale and Ordinal-Scale Data. Sociological Methodology, 6, 171-190

Becker, M. P. (1989). On the bivariate normal distribution and association models for ordinal categorical data. Statistics & Probability Letters, 8(5), 435-440.

Becker, M. P. (2014). Ordered categorical data. Wiley StatsRef: Statistics Reference Online.

Berry, K.J., Johnston, J.E., Mielke, P.W. Jr (2018) The Measurement of Association - A Permutation Statistical Approach, Springer

Carr, G. J., Hafner, K. B., Koch, G. G. (1989). Analysis of rank measures of association for ordinal data from longitudinal studies. Journal of the American Statistical Association, 84(407), 797-804.

Chaganty, N. R., Joe, H. (2006). Range of correlation matrices for dependent Bernoulli random variables. Biometrika, 93(1), 197-206.

Demirtas, H. (2006). A method for multivariate ordinal data generation given marginal distributions and correlations. Journal of Statistical Computation and Simulation, 76(11), 1017-1025.

Demirtas, H. Vardar-Acar, C. (2017) Anatomy of Correlational Magnitude Transformations in Latency and Discretization Contexts in Monte-Carlo Studies, in: Ding-Geng (Din) Chen, John Dean Chen Eds, Monte-Carlo Simulation-Based Statistical Modeling, Springer, Singapore, 2017

Ferrari, P. A., Barbiero, A. (2012). Simulating ordinal data. Multivariate Behavioral Research, 47(4), 566-589.

Friendly, M. (2017). vcdExtra: 'vcd' Extensions and Additions. R package version 0.7-1. https://CRAN.R-project.org/package=vcdExtra

Gans, L. P., Robertson, C. A. (1981). The behavior of estimated measures of association in small and moderate sample sizes for 2×3 tables. Communications in Statistics-Theory and Methods, 10(16), 1673-1686.

Gilula, Z., Krieger, A. M., Ritov, Y. (1988). Ordinal association in contingency tables: some interpretive aspects. Journal of the American Statistical Association, 402: 540-544.

Goodman, L.A., Kruskal, W.H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49, 732-764.

Grønneberg, S., Foldnes, N. (2019). A Problem with Discretizing Vale-Maurelli Simulation Studies. Psychometrika, 84(2), 554-561.

Kendall, M.G. (1945). The treatment of ties in rank problems. Biometrika, 33, 239-251.

Kendall, M.G., Gibbons, J.D. (1990). Rank Correlation Methods, fifth ed. Edward Arnold, London.

Kateri, M. (2014). Contingency table analysis. Methods and Implementation Using R. Birkhauser, New York.

Kampen, J., Swyngedouw, M. (2000). The ordinal controversy revisited. Quality & Quantity, 34(1), 87-102.

Kvålseth, T. O. (2017). An alternative measure of ordinal association as a value-validity correction of the Goodman-Kruskal gamma. Communications in Statistics-Theory and Methods, 46(21), 10582-10593.

Kvålseth, T. O. (2018). Measuring association between nominal categorical variables: an alternative to the Goodman-Kruskal lambda. Journal of Applied Statistics, 45(6), 1118-1132.

Jin, S., Yang-Wallentin, F. (2017). Asymptotic robustness study of the polychoric correlation estimation. Psychometrika, 82(1), 67-85.

Joe, H., Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. UBC, Department of Statistics, Technical Report, 166.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika, 44: 289-292.

Laird, N. M. (1979). A note on classifying ordinal scale data. In: K. F. Schuessler (ed.), Social Methodology. San Francisco: Jossey Bass, pp. 303-310.

Lazarsfeld, P. F., Henry, N. W. (1968). Latent structure analysis. Houghton Mill, Boston, MA.

Liu, H., Hussain, F., Tan, C. L., Dash, M. (2002). Discretization: An enabling technique. Data mining and knowledge discovery, 6(4), 393-423.

Mari, D. D., Kotz, S. (2001). Correlation and dependence. World Scientific.

McNeil, A. J., Frey, R., Embrechts, P. (2005). Quantitative risk management: Concepts, techniques and tools (Vol. 3). Princeton: Princeton University Press.

Muthen, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22(1-2), 43-65.

Nelsen, R.B. (2006). An introduction to copulas, 2nd edition, Springer, New York.

O'Gorman, T. W., Woolson, R. F. (1988). Analysis of ordered categorical data using the SAS system. In Proceeding of the 13th Annual SAS Users Group Conference, 957-963

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44(4), 443-460.

Rousson, V. (2007) The gamma coefficient revisited. Statistics & Probability Letters, 77: 1696-1704.

Rosenthal, I. (1966). Distribution of the sample version of the measure of association, gamma. Journal of the American Statistical Association, 61(314), 440-453.

Srole, L. E., Fischer, A. K. (1978). Mental health in the metropolis: The Midtown Manhattan study, Rev. New York U Press.

Vale, C. D., Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. Psychometrika, 48(3), 465-471.

van der Ark, L. A., van Aert, R. C. (2015). Comparing confidence intervals for Goodman and Kruskal's gamma coefficient. Journal of Statistical Computation and Simulation, 85(12), 2491-2505.

Woods, C. M. (2009). Consistent small-sample variances for six gamma-family measures of ordinal association. Multivariate behavioral research, 44(4), 525-551.

Table 1: Graphic for a simulated weakly monotonic relationship

	1	2	3	4	5	6
1	×	×				
2		×	×			
3			×	×		
4				×	×	
5					×	×
6						×

Table 2: A trivariate probability distribution and its corresponding bivariate marginal distributions

x	y	z	$p(x,y,z)$
1	1	1	1/8
1	1	2	3/8
1	2	1	2/8
2	1	2	1/8
2	2	2	1/8

x,y	1	2	x,z	1	2	y,z	1	2
1	4/8	2/8	1	3/8	3/8	1	1/8	4/8
2	1/8	1/8	2	0	2/8	2	2/8	1/8

Table 3: Values of Goodman and Kruskal's gamma for several combinations of τ and number of categories of the two identical marginal triangular distributions

τ, K	2	3	5	7	10	20	50	100
−1	−1.00000	−1.00000	−1.00000	−1.00000	−1.00000	−1.00000	−1.00000	−1.00000
−0.9	−0.99448	−0.99321	−0.99036	−0.98698	−0.98016	−0.95799	−0.92968	−0.91622
−0.8	−0.97561	−0.97031	−0.95286	−0.93261	−0.90805	−0.86492	−0.82918	−0.81522
−0.7	−0.93960	−0.92266	−0.87790	−0.84384	−0.81053	−0.76170	−0.72647	−0.71358
−0.6	−0.88235	−0.84482	−0.77885	−0.73856	−0.70299	−0.65525	−0.62312	−0.61176
−0.5	−0.80000	−0.74150	−0.66519	−0.62433	−0.59043	−0.54731	−0.51950	−0.50986
−0.4	−0.68966	−0.61803	−0.54185	−0.50458	−0.47493	−0.43854	−0.41572	−0.40792
−0.3	−0.55046	−0.47841	−0.41176	−0.38119	−0.35756	−0.32926	−0.31185	−0.30596
−0.2	−0.38462	−0.32625	−0.27699	−0.25537	−0.23898	−0.21966	−0.20792	−0.20398
−0.1	−0.19802	−0.16537	−0.13923	−0.12805	−0.11966	−0.10987	−0.10397	−0.10199
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.1	0.19802	0.16537	0.13923	0.12805	0.11966	0.10987	0.10397	0.10199
0.2	0.38462	0.32625	0.27699	0.25537	0.23898	0.21966	0.20792	0.20398
0.3	0.55046	0.47841	0.41176	0.38119	0.35756	0.32926	0.31185	0.30596
0.4	0.68966	0.61803	0.54185	0.50458	0.47493	0.43854	0.41572	0.40792
0.5	0.80000	0.74150	0.66519	0.62433	0.59043	0.54731	0.51950	0.50986
0.6	0.88235	0.84482	0.77885	0.73856	0.70299	0.65525	0.62312	0.61176
0.7	0.93960	0.92266	0.87790	0.84384	0.81053	0.76170	0.72647	0.71358
0.8	0.97561	0.97031	0.95286	0.93261	0.90805	0.86492	0.82918	0.81522
0.9	0.99448	0.99321	0.99036	0.98698	0.98016	0.95799	0.92968	0.91622
1	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Table 4: Values of Goodman and Kruskal's gamma for several combinations of τ and number of categories of the two identical marginal symmetrical distributions (uniform cut-points)

$\tau.K$	2	3	5	7	10	20	50	100
−1	−1	−1	−1	−1	−1	−1	−1	−1
−0.9	−0.99448	−0.99698	−0.99501	−0.99391	−0.99233	−0.98624	−0.9589	−0.935
−0.8	−0.97561	−0.9864	−0.97816	−0.97373	−0.96625	−0.92727	−0.86415	−0.83419
−0.7	−0.9396	−0.96561	−0.94663	−0.9343	−0.90815	−0.83339	−0.75986	−0.73083
−0.6	−0.88235	−0.93147	−0.89509	−0.86429	−0.81635	−0.72489	−0.65291	−0.62681
−0.5	−0.8	−0.8799	−0.81381	−0.76197	−0.70225	−0.60945	−0.54487	−0.52252
−0.4	−0.68966	−0.80262	−0.69807	−0.63458	−0.57406	−0.49032	−0.43628	−0.41811
−0.3	−0.55046	−0.6843	−0.55126	−0.4895	−0.43685	−0.36909	−0.3274	−0.31363
−0.2	−0.38462	−0.50907	−0.38064	−0.33245	−0.29397	−0.24663	−0.21834	−0.2091
−0.1	−0.19802	−0.27417	−0.19424	−0.16799	−0.14776	−0.12347	−0.10919	−0.10456
0	0	0	0	0	0	0	0	0
0.1	0.19802	0.27417	0.19424	0.16799	0.14776	0.12347	0.10919	0.10456
0.2	0.38462	0.50907	0.38064	0.33245	0.29397	0.24663	0.21834	0.2091
0.3	0.55046	0.6843	0.55126	0.4895	0.43685	0.36909	0.3274	0.31363
0.4	0.68966	0.80262	0.69807	0.63458	0.57406	0.49032	0.43628	0.41811
0.5	0.8	0.8799	0.81381	0.76197	0.70225	0.60945	0.54487	0.52252
0.6	0.88235	0.93147	0.89509	0.86429	0.81635	0.72489	0.65291	0.62681
0.7	0.9396	0.96561	0.94663	0.9343	0.90815	0.83339	0.75986	0.73083
0.8	0.97561	0.9864	0.97816	0.97373	0.96625	0.92727	0.86415	0.83419
0.9	0.99448	0.99698	0.99501	0.99391	0.99233	0.98624	0.9589	0.935
1	1	1	1	1	1	1	1	1

Table 5: Values of Goodman and Kruskal's gamma for several combinations of τ and number of categories of the two identical marginal triangular distributions

$\tau.K$	2	3	5	7	10	20	50	100
−1	−1	−1	−1	−1	−1	−1	−1	−1
−0.9	−1	−0.99467	−0.99648	−0.99377	−0.98691	−0.96792	−0.93751	−0.92107
−0.8	−0.99933	−0.97729	−0.96629	−0.94876	−0.92616	−0.88070	−0.83783	−0.82001
−0.7	−0.98525	−0.94097	−0.90166	−0.86891	−0.83421	−0.77835	−0.73459	−0.71792
−0.6	−0.93982	−0.87473	−0.80945	−0.76715	−0.72757	−0.67084	−0.63032	−0.61554
−0.5	−0.85853	−0.77767	−0.69721	−0.65222	−0.61317	−0.56094	−0.52560	−0.51304
−0.4	−0.74150	−0.65396	−0.57108	−0.52900	−0.49425	−0.44974	−0.42065	−0.41048
−0.3	−0.59061	−0.50867	−0.43537	−0.40046	−0.37255	−0.33779	−0.31557	−0.30788
−0.2	−0.41063	−0.34734	−0.29328	−0.26853	−0.24914	−0.22540	−0.21041	−0.20526
−0.1	−0.20993	−0.17580	−0.14740	−0.13465	−0.12476	−0.11275	−0.10522	−0.10263
0	0	0	0	0	0	0	0	0
0.1	0.20632	0.17431	0.14685	0.13436	0.12462	0.11271	0.10521	0.10263
0.2	0.39722	0.34185	0.29120	0.26742	0.24857	0.22525	0.21039	0.20525
0.3	0.56388	0.49802	0.43112	0.39810	0.37130	0.33745	0.31551	0.30786
0.4	0.70149	0.63884	0.56453	0.52516	0.49212	0.44911	0.42053	0.41044
0.5	0.80905	0.76064	0.68892	0.64700	0.61008	0.55992	0.52540	0.51298
0.6	0.88837	0.85940	0.80082	0.76119	0.72370	0.66940	0.62999	0.61544
0.7	0.94297	0.93060	0.89462	0.86338	0.83011	0.77651	0.73411	0.71777
0.8	0.97707	0.97283	0.96064	0.94475	0.92278	0.87863	0.83717	0.81978
0.9	0.99482	0.99377	0.99154	0.98915	0.98460	0.96620	0.93662	0.92070
1	1	1	1	1	1	1	1	1

Table 6: Joint distribution ensuring the target $\gamma = 0.5$ and the assigned margins, by ordinalization of a bivariate standard normal variable

X, Y	1	2	3	4	tot
1	0.0593	0.0782	0.0726	0.0399	0.25
2	0.0248	0.0600	0.0863	0.0789	0.25
3	0.0120	0.0415	0.0817	0.1147	0.25
4	0.0039	0.0203	0.0594	0.1664	0.25
tot	0.1	0.2	0.3	0.4	1

Table 7: Cograduation (M, left) and countergraduation (W, right) tables based on the margins of the joint distribution displayed in Table 6. Note that being the distribution of X symmetrical, we have that $p^M_{i,j} = p^W_{I-i+1,j}$, $\forall i = 1, \ldots, 4$, with $I = 4$

X, Y	1	2	3	4	tot
1	0.1	0.15	0	0	0.25
2	0	0.05	0.2	0	0.25
3	0	0	0.1	0.15	0.25
4	0	0	0	0.25	0.25
tot	0.1	0.2	0.3	0.4	1

X, Y	1	2	3	4	tot
1	0	0	0	0.25	0.25
2	0	0	0.1	0.15	0.25
3	0	0.05	0.2	0	0.25
4	0.1	0.15	0	0	0.25
tot	0.1	0.2	0.3	0.4	1

Table 8: Empirical joint distribution of a bivariate ordinal sample considered in the Example 2 of Section 5

X, Y	1	2	3	tot
1	3	4	3	10
2	1	4	5	10
3	1	2	7	10
tot	5	10	15	30

Table 9: Expected joint probabilities for the ordinalized bivariate normal distribution obtained by maximum likelihood estimation of the sample of Table 8

	1	2	3	tot
1	0.0985	0.1319	0.1039	0.3343
2	0.0480	0.1178	0.1667	0.3325
3	0.0206	0.0822	0.2303	0.3331
tot	0.1671	0.3319	0.5010	1

Table 10: The Midtown Manhattan Study: Mental Health and Parents' Socioeconomic Status

Parents' Socioeconomic Status	Mental Health				
	Well	Mild symptom formation	Moderate symptom formation	Impaired	total
A (high)	64	94	58	46	262
B	57	94	54	40	245
C	57	105	65	60	287
D	72	141	77	94	384
E	36	97	54	78	265
F (low)	21	71	54	71	217
total	307	602	362	389	1660

Table 11: MLEs for the Midtown Manhattan Mental study data, assuming an ordinalized bivariate normal model. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (maximized value of the log-likelihood function: −5171.345). The results are obtained by using the package `maxLik`

parameter	estimate	st.error
τ	0.10762	0.01718 ***
θ_1	−1.00344	0.03702 ***
θ_2	−0.51024	0.03219 ***
θ_3	−0.05579	0.03074 .
θ_4	0.55185	0.03253 ***
θ_5	1.12411	0.03900 ***
η_1	−0.89627	0.03573 ***
η_2	0.11931	0.03080 ***
η_3	0.72393	0.03386 ***

Table 12: Expected joint frequencies under the ordinalized bivariate normal model whose parameters are obtained by applying the maximum likelihood method to the data of Table 10

x, y	Well	Mild	Moderate	Impaired	total
A	67.8	102.0	50.1	42.1	262.0
B	53.0	93.0	50.7	47.5	244.2
C	55.8	106.8	61.9	62.3	286.9
D	65.8	138.8	86.2	93.9	384.7
E	39.1	91.8	61.5	73.3	265.7
F	25.6	69.3	51.5	70.2	216.6
total	307.2	601.6	361.8	389.4	1660

Table 13: Expected joint frequencies under the ordinalized bivariate normal model whose parameters are obtained by applying the maximum likelihood method to the data of Table 10

x, y	Well	Mild	Moderate	Impaired	total
A	65.3	104.4	50.1	42.1	262
B	54.2	94.9	49.9	45.9	245
C	55.9	107.2	61.7	62.2	287
D	65.3	137.0	86.4	95.3	384
E	39.0	89.6	61.8	74.7	265
F	27.3	68.8	52.0	68.8	217
total	307	602	362	389	1660

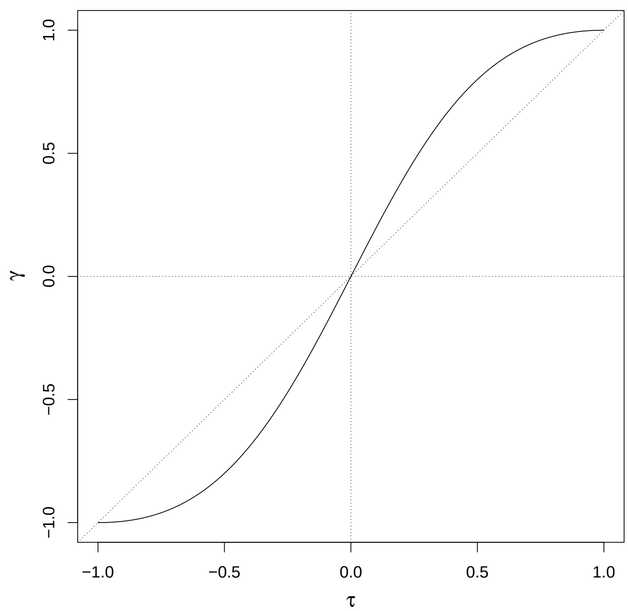

Figure 1: Graph of the gamma coefficient for a dichotomized bivariate normal rv with Kendall's rank correlation τ, see Equation (3). The dashed line is the 45 degrees line passing through the origin

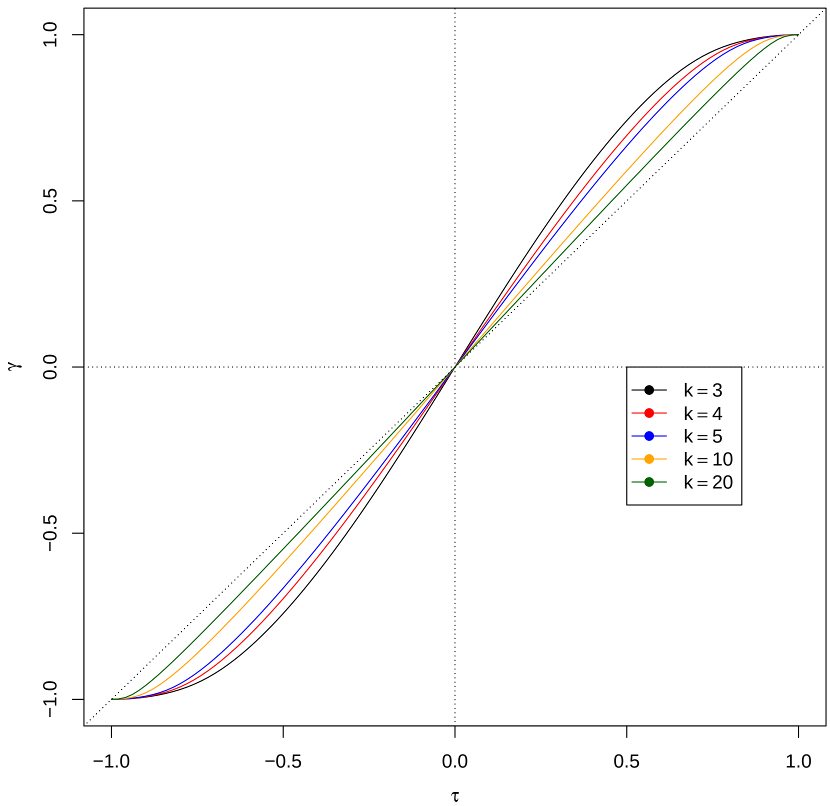

Figure 2: Kruskal's γ for an ordinalized bivariate normal rv with Kendall's rank correlation τ; the categories of the two ordinal variables are set both equal to an integer K and have uniform probabilities $1/K$. The dashed line is the 45 degrees line passing through the origin

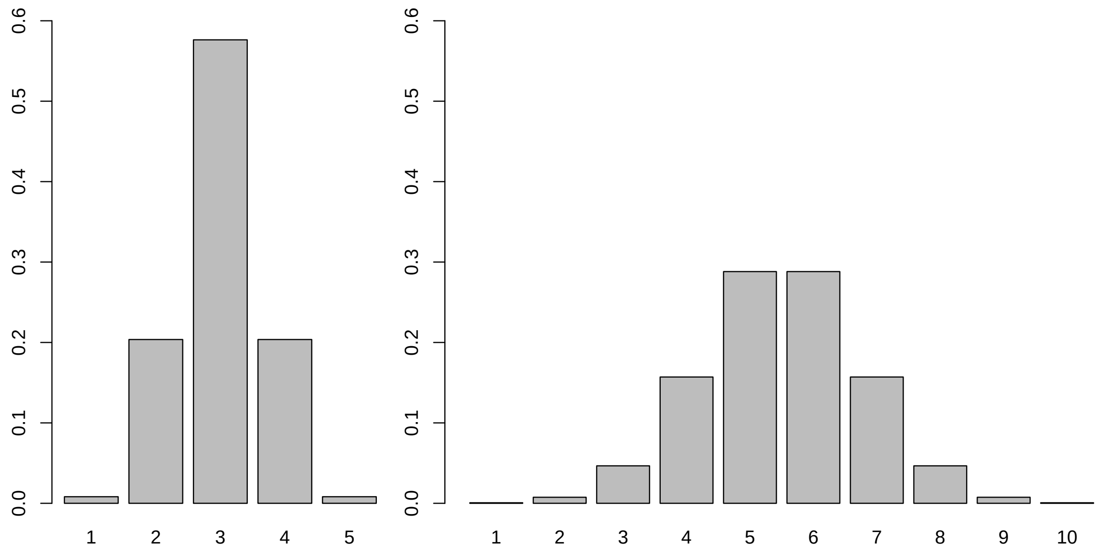

Figure 3: Barplot of two symmetrical non-uniform distributions, with 5 (left) and 10 (right) categories, obtained by mimicking the probability density function of a standard normal variable

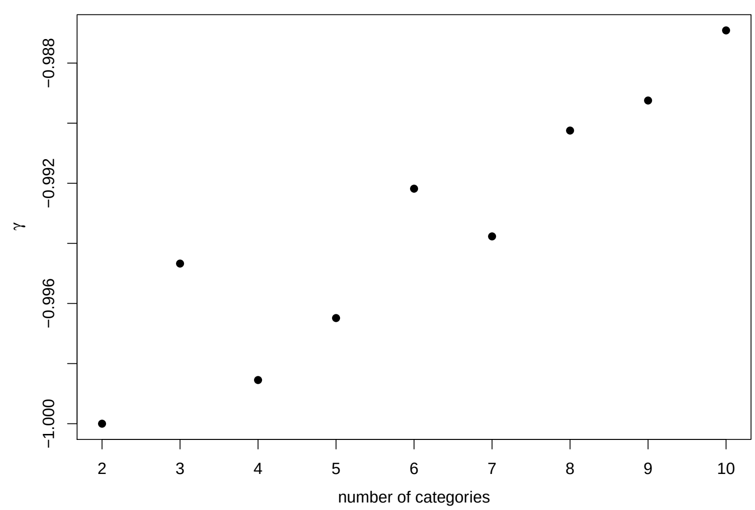

Figure 4: Kruskal's γ for an ordinalized bivariate normal rv with Kendall's rank correlation $\tau = -0.9$ as a function of the common number of categories (from 2 to 10) of the two ordinal triangular variables

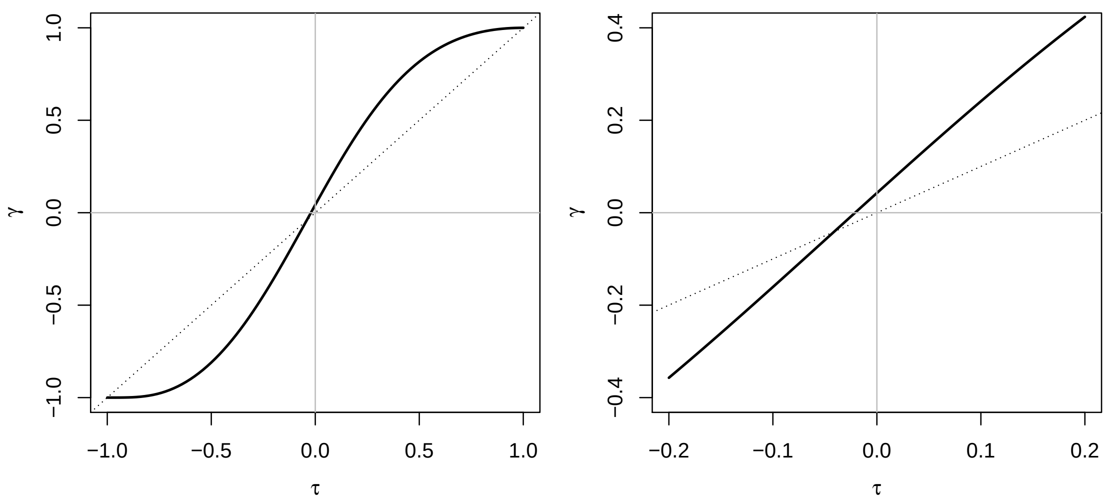

Figure 5: Relationship between τ and γ for a bivariate Student's t rv before and after ordinalization (marginal probabilities for the two identical ordinalized variables are 1/3 and 2/3). In both graphs, the dotted line is the bisector of the first and third orthants. In the right graph, it can be better appreciated the behavior of γ when τ is closer to zero.