On the ratio between point-polyserial and polyserial correlations for non-normal bivariate distributions

Alessandro Barbiero

Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan, Italy

Abstract

It is a well-known fact that for the bivariate normal distribution the ratio between the point-polyserial correlation (the linear correlation after one of the two variables is discretized into k categories with probabilities p_i, $i = 1, \ldots, k$) and the polyserial correlation ρ (the linear correlation between the two normal components) remains constant with ρ, keeping the p_i's fixed. If we move away from the bivariate normal distribution, by considering non-normal margins and/or non-normal dependence structures, then the constancy of this ratio may get lost. In this work, the magnitude of the departure from the constancy condition is assessed for several combinations of margins (normal, uniform, exponential, Weibull) and copulas (Gauss, Frank, Gumbel, Clayton), also varying the distribution of the discretized variable. The results indicate that for many settings we are far from the condition of constancy, especially when highly asymmetrical marginal distributions are combined with copulas that allow for tail-dependence. In such cases, the linear correlation may even increase instead of decreasing, contrary to the usual expectation. This implies that most existing simulation techniques or statistical models for mixed-type data, which assume a linear relationship between point-polyserial and polyserial correlations, should be used very prudently and possibly reappraised.

Keywords: bivariate normal distribution; copula; discretization; latent variable; point-polyserial correlation

On the ratio between point-polyserial and polyserial correlations for non-normal bivariate distributions

Introduction

In behavioral, educational, sociological, and psychological studies, observed variables are often measured using ordinal scales. For example, the Likert scale is widely used in surveys to capture respondents' level of agreement with a statement or their satisfaction with a product or service, typically on a five- or seven-point scale (e.g., 1="Completely disagree" or "Completely unsatisfied," . . . , 5="Completely agree" or "Completely satisfied"). A common approach is to treat these ordinal variables as discretizations of an underlying continuous variable (a latent, i.e., unobserved, variable) representing the degree of agreement with the statement or the level of satisfaction (see, e.g., Bartholomew, 1980; Zhang, Liu, & Pan, 2024). There are also many examples of quantitative variables that are explicitly discretized in social science studies: for instance, income may be recorded as under/over 50K, and alcohol consumption as less/more than two drinks per day. When analyzing this type of data, another common approach is to assign consecutive integer values to each category and then proceed as if the data were measured on an interval scale, assuming the desired distributional properties. (Norman, 2010). Some non-standard techniques, such as non-linear or categorical principal component analysis (De Leeuw, 2006) can be alternatively employed, where the scores corresponding to the ordinal categories are not assigned a priori, but are computed iteratively as the solution to a minimization problem involving an objective loss function (the main algorithm is known as "alternating least squares").

The most common choice for the distribution of the latent variables is the (multivariate) normal, or Gaussian, distribution, as its dependence structure is fully captured by the variance-covariance matrix, and each of its distinct elements can be estimated separately using a bivariate normal distribution (McNeil, Frey, & Embrechts, 2015).

In the literature, analyses of latent variables models with mixed type of continuous and categorical (ordinal) variables have received a great deal of attention (see Song & Lee, 2003, and references therein). Let X_2 be an observed ordinal variable depending on an underlying latent continuous random variable (rv) Z_2 and let Z_1 denote another observed continuous variable. It is often assumed that the joint distribution of Z_1 and Z_2 is bivariate normal. Pearson's correlation coefficient between Z_1 and X_2 is called the point-polyserial correlation, while the correlation between Z_1 and Z_2 is called the polyserial correlation. As a special case, if X_2 is a dichotomous rv, the corresponding correlations are referred to as point-biserial and biserial correlations. The problem of estimating the polyserial correlation based on a bivariate sample has been studied by Cox (1974), who derived its MLE; Olsson, Drasgow, and Dorans (1982) established the relationship between the polyserial and the point-polyserial correlations and compared the MLE of the polyserial correlation with both a two-step estimator and a computationally convenient ad hoc estimator. Bedrick (1995) studied the attenuation of the correlation coefficient (the polyserial correlation) when one of the continuous variables is categorized. The attenuation was shown to depend critically on the distribution of the underlying latent variable and on the scores assigned to the categories. However, attenuation becomes less severe as the number of categories increases, provided that the category scores are carefully chosen. In particular, equally-spaced scores (e.g., $1, 2, \dots, k$) offer reasonable protection against substantial attenuation across a variety of distributions. Cheng and Liu (2016) derived the maximal point-biserial correlation under several non-normal distributions, namely, the uniform, Student's t, exponential, and a mixture of two normal distributions. Barbiero (2025) extended this investigation to the case of more than 2 categories, analysing the behavior of the maximal point-polyserial correlation.

Demirtas and Hedeker (2016) and, later, Demirtas and Vardar-Acar (2017) studied the relationship between the biserial and the point-biserial correlations by devising an algorithm applicable to any underlying distribution other than the (bivariate) normal. The method is claimed to be equally applicable to the ordinal case, in the context of the relationship between the polyserial (before ordinalization) and point–polyserial (after ordinalization) correlations. The authors state that "it works for ordinal-continuous data combinations, and so one can compute the polyserial correlation given the point–polyserial correlation (or vice versa) when the relative proportions of the ordinal categories are specified." The algorithm is based on the generation of a large (of size, say, $N = 100,000$) pseudo-random sample from a bivariate random vector (Z_1, Z_2) constructed through the use of Fleishman polynomials (Fleishman, 1978) applied to an underlying Gaussian random vector (Foldnes & Grønneberg, 2015). Although the numerical experiments carried out in Demirtas and Hedeker (2016) are shown to produce negligible errors (when an analytical solution is also available), the sampling error naturally induced by random simulation can hardly be controlled. The linearity and constancy arguments proposed by Demirtas and Hedeker (2016) are at the basis of the algorithm proposed by Demirtas, Allozi, Hu, Inan, and Ozbek (2017) and designed for concurrently generating binary, ordinal, count, and continuous data. These assumptions had already been employed earlier in the development of a methodology for large-scale simulation of binary and potentially non-normally distributed continuous variables (Demirtas, Hedeker, & Mermelstein, 2012).

Moss and Grønneberg (2023) calculated intervals (referred to as "partial identification sets") for latent correlations based on the distribution of ordinal data, under the assumption that the marginal distributions of the latent variable are known, but only partial information about the copula is available. For typical values of the number of categories of the ordinal variable, such as 5 or 7, it is shown that these sets are rather wide. Having further information about the copula can substantially reduce their width and thus make the partial identification sets more informative.

The aim of this paper is to investigate the relationship (specifically, the ratio) between point-polyserial and polyserial correlations for different bivariate distributions, varying both the type of marginal distributions, focusing on well-known parametric (non-normal) families, and the dependence structure (i.e., the copula linking the two margins into the joint distribution). We will analyze if and to which extent the constancy ratio condition is violated as we move away from the bivariate normal distribution.

The structure of the paper is as follows. In the next section, we recall a known result about the relationship between polyserial and point-polyserial correlations for the bivariate normal distribution. In the third section, we investigate the relationship between the two correlations when the joint continuous distribution is no longer bivariate normal, but is a more general distribution characterized by (possibly) non-normal margins combined through a (possibly) non-Gaussian copula. We will consider, along with the normal, three non-normal marginal distributions (uniform, exponential, Weibull) and four bivariate copulas (Gauss, Frank, Gumbel, and Clayton). We will examine different distributions of the discretized variable, consisting of 2, 3, or 4 categories with varying probability values. The evaluation of the relationship between polyserial and point-polyserial correlations is numerically accomplished by resorting to the `R` packages `copula` and `cubature`, which allow, in order, to construct the desired bivariate distribution combining a pair of margins with a copula, and to evaluate the integral corresponding to the mixed moment between the continuous and the discretized components. The fourth section illustrates an application to a real data set. The last section provides some final remarks and insights. All the results were derived under the `R` statistical environment. The relevant code, which is essential for reproducing the results and further analysis, is available for reference and use.

Point-polyserial correlation under normality

Let (Z_1, Z_2) be a bivariate standard normal rv, and let X_2 be a dichotomization of Z_2 at the threshold ω. Hence, X_2 is a rv which takes the value 1 when $Z_2 \geq \omega$ and the value 0 when $Z_2 < \omega$. Denoting by $\varphi(\cdot)$ the probability density function (pdf) of a standard normal rv and letting $P(X_2 = 1) = \int_{\omega}^{\infty} \varphi(y)dy = p(\omega)$ and $P(X_2 = 0) = q(\omega) = 1 - p(\omega)$, the relationship between $\mathrm{cor}(Z_1, Z_2)$ (the biserial correlation) and $\mathrm{cor}(Z_1, X_2)$ (the point-biserial correlation) is due to Karl Pearson (Pearson, 1909) and is also reported in MacCallum, Zhang, Preacher, and Rucker (2002), where the consequences of dichotomization for measurement and statistical analyses are illustrated and discussed in a more general context:

$$\frac{\operatorname{cor}(Z_1, X_2)}{\operatorname{cor}(Z_1, Z_2)} = \frac{\varphi(\omega)}{\sqrt{p(\omega)q(\omega)}}. \tag{1}$$

It is interesting to examine the plot of this function, displayed in Figure 1, and to note that it is symmetric and attains its unique maximum, equal to $2\varphi(0) = \sqrt{2/\pi} \approx 0.7979$, at $\omega = 0$, corresponding to the "equal-probability" dichotomization ($p(0) = q(0) = 1/2$). It is worth noting that changing the two values of the support of the discrete rv X_2, by default set at 0 and 1, providing that the order is maintained, does not affect the value of the point-biserial correlation (this follows from the well-known invariance of Pearson's correlation under any positive linear transformation).

A generalization of Pearson's point-biserial correlation to the case of discretization into a k point-scale distribution, supported on $\{y_1, y_2, \ldots, y_k\}$, is easily provided, again starting from a bivariate standard normal rv. Let X_2 be the discrete rv obtained by discretizing the continuous component Z_2, i.e., $X_2 = \sum_{i=1}^{k} y_i \mathbb{1}(\theta_{i-1} < Z_2 \leq \theta_i)$, where $\mathbb{1}(\cdot)$ is the indicator function, $y_1 < \cdots < y_k$ are the values that X_2 can take, with corresponding probabilities $p_i = P(X_2 = y_i) = P(\theta_{i-1} < Z_2 \leq \theta_i)$, and the thresholds θ_i applied to the continuous rv Z_2 satisfy $-\infty = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_k = +\infty$.

Since the following relationship holds for the pdf of a standard normal rv:

$$\int x\varphi(x)\mathrm{d}x = -\varphi(x) + \text{constant},$$

it can be proved that the resulting Pearson's correlation coefficient between Z_1 and X_2, i.e., the point-polyserial correlation, is (see Eq.(11) in Olsson et al. (1982)):

$$\rho_{PP} = \operatorname{cor}(Z_1, X_2) = \rho \sum_{i=1}^{k-1} \varphi\left[\Phi^{-1}(P_i)\right](y_{i+1} - y_i) \Bigg/ \sqrt{\sum_{i=1}^{k} y_i^2 p_i - \left(\sum_{i=1}^{k} y_i p_i\right)^2}, \tag{2}$$

where P_i is the cumulative probability of the value y_i and $\rho = \operatorname{cor}(Z_1, Z_2)$. Eq. (2) conveys a powerful result: it points out that under the bivariate normal distribution there is a linear relationship between the polyserial and the point-polyserial correlations, for any possible choice of the number of categories k and for any discrete distribution $\{p_1, \ldots, p_k\}$ sitting on $\{y_1, \ldots, y_k\}$. Obviously, the proportionality constant depends on this latter distribution. By considering the simple case where $y_i = i$ (i.e., consecutive integer scores), and noting that $y_{i+1} - y_i = 1$ for all $i = 1, \ldots, k-1$, the ratio between the point-polyserial correlation and the polyserial correlation of the bivariate normal distribution simplifies to

$$\rho_{PP}/\rho = \sum_{i=1}^{k-1} \varphi\left[\Phi^{-1}(P_i)\right] \Bigg/ \sqrt{\sum_{i=1}^{k} i^2 p_i - \left(\sum_{i=1}^{k} i p_i\right)^2}, \tag{3}$$

which corresponds to the maximum attainable point-polyserial correlation when $\rho = 1$. As an example, for $k = 3$, Figure 2 shows the level curves of the ratio in (3), viewed as a function of p_1 and p_2, with $p_3 = 1 - p_1 - p_2$.

This important result is expected not to hold, or to hold just in an approximate manner, when we move away from the bivariate normal distribution. Demirtas and Hedeker (2016), for example, considered the standardized rvs X and Y obtained by the components Z_1 and Z_2 of a bivariate normal rv with correlation ρ and expressed the relationship between the two rvs via the following linear regression model:

$$Y = \rho X + \epsilon \tag{4}$$

where ϵ is independent of X and Y, and follows $\mathcal{N}(0, 1-\rho^2)$. When one generalizes it to non-normal X and/or Y (both centered and scaled), the same relationship can be assumed to hold with the exception that the distribution of ϵ follows a non-normal distribution. As long as (4) is valid, denoting by X_D a dichotomization of X, it can be proved that

$$\text{cor}(X_D, Y) = \text{cor}(X, Y) \cdot \text{cor}(X_D, X) \tag{5}$$

i.e., the linear association between X_D and Y is assumed to be fully explained by their mutual association with X.

According again to Demirtas and Hedeker (2016), when X is ordinalized to obtain X_O, the fundamental ideas remain unchanged. As long as the assumptions of Eq.(4) are met, the method is equally applicable to the ordinal case in the context of the relationship between the polyserial and point–polyserial correlations, which can be written by generalizing Eq.(5) as

$$\rho_{PP} = \rho \cdot \text{cor}(X_O, X), \tag{6}$$

where $\rho = \mathrm{cor}(X, Y)$ and $\rho_{PP} = \mathrm{cor}(X_O, Y)$. Under ordinalization, the key covariance term, $\mathrm{cov}(X_O, X)$, involves far more complicated integrals than in the dichotomous case. Demirtas and Hedeker (2016) claimed that "even when these integrals can be computed by brute force or numerical integration techniques, the proposed sorting approach makes the problem much less challenging." The validity of (5) has therefore a central role, and taking it for granted can lead to unreliable or contradictory results in cases where it is not satisfied. Demirtas and Hedeker (2016) argued that this formula works approximately quite well even when a non-normal bivariate rv is considered. However, in their study, although considering different non-normal and possibly non-identical margins, they based the construction of the bivariate rv on Fleishman polynomials, which is known to produce continuous bivariate rvs whose copula is rather similar to the Gaussian (Foldnes & Grønneberg, 2015). Eqs. (5) and (6) are also assumed to hold by Demirtas et al. (2017), who proposed a unified framework for concurrently generating data sets that include all four major kinds of variables (i.e., binary, ordinal, count, and normal) when the marginal distributions and a feasible association structure are specified.

It is worth noting that the bivariate normal distribution enjoys another very important property (Lancaster, 1957), usually referred to as an 'extremal' property, which can be stated in this way: if a bivariate rv (Z_1, Z_2) follows the bivariate normal distribution with correlation ρ, then, if we apply any transformations, say T_1 and T_2, to Z_1 and Z_2, the linear correlation between $T_1(Z_1)$ and $T_2(Z_2)$ will be always smaller, in absolute value, than $|\rho|$. In formula:

$$|\mathrm{cor}(T_1(Z_1), T_2(Z_2))| \leq |\mathrm{cor}(Z_1, Z_2)| = |\rho|.$$

From this property, by letting T_1 be the identity function and T_2 an appropriate step function, it follows that the point-biserial (or the point-polyserial) correlation in absolute value is always less than or equal to the biserial or polyserial correlation in absolute value. This property can be stated saying that discretization reduces the strength of linear correlation. However, this holds for the bivariate normal distribution, where normal marginal distributions are combined through a normal copula, but it may not extend to other bivariate distributions. In particular, the correlation may increase in absolute magnitude after one component is discretized. In such cases, it is clear that Eq. (6) cannot hold: since $|\rho| \leq 1$, if $|\rho_{PP}| > |\rho|$, then $|\mathrm{cor}(X_O, X)|$ would have to exceed 1, which is a contradiction.

Point-polyserial correlation under non-normality

If we consider a continuous bivariate rv (X_1, X_2) that is not bivariate normal, then formula (3) generally does not hold and one cannot claim the existence of a linear relationship between the linear correlation coefficient before and after the discretization of X_2. This implies that for fixed k and p_i, defining the distribution of the ordinalized rv, the ratio of the correlation after discretization to the original correlation may not be constant, but varies with the value of the latter. Nevertheless, as discussed in the previous sections, an (approximately) linear relationship between the two correlations is often assumed, as in Demirtas and Vardar-Acar (2017), even when the bivariate rv is known to be non-normal.

In the following subsections, we want to determine the behavior of the ratio between point-polyserial and polyserial correlations, when we consider a bivariate rv (X_1, X_2) with identical continuous, not necessarily normal, marginal distributions. The joint cdf of (X_1, X_2) is $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$, which can be rewritten as $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$, where C is the unique copula of the rv, and $F_1(w) = F_2(w)$ for all $w \in \mathbb{R}$. Discretizing the second component X_2 according to a pre-specified set of probabilities $(p_1, \dots, p_k)$ and category scores $(1, 2, \dots, k)$, we obtain a discrete rv, which we call $X_{2\mathcal{D}}$. Formally, $X_{2\mathcal{D}} = \sum_{i=1}^{k} i \mathbb{1}(\theta_{i-1} < X_2 \leq \theta_i)$ for $i = 1, \dots, k$, with $p_i = P(X_{2\mathcal{D}} = i) = P(\theta_{i-1} < X_2 \leq \theta_i)$, where $\theta_0 \leq \theta_1 \leq \cdots \leq \theta_k$ are a set of thresholds dividing the support of X_2, with θ_0 and θ_k being its lower and upper bounds.

As we have done so far, we will continue to assume full knowledge of the distribution of the bivariate latent variable, that is, both its marginals and its dependence structure, which is captured by its unique copula. This assumption allows us to uniquely determine the relationship between the polyserial and point-polyserial correlations. In this regard, we assume significantly more information than what is available in Moss and Grønneberg (2023).

Copula parameter, Spearman's and Pearson's correlations

For a bivariate rv (X_1, X_2), with continuous margins F_1 and F_2, and copula C, Sperman's correlation ρ_S is defined as the linear correlation between the probability transforms of X_1 and X_2:

$$\rho_S(X_1, X_2) = \mathrm{cor}(F_1(X_1), F_2(X_2)), \tag{7}$$

i.e., as the linear correlation between the components U_1 and U_2 of the copula C.

We recall that for a bivariate distribution, the values of Pearson's ρ and Spearman's ρ_S generally differ: while the latter only depends on the copula C, the former also depends on the margins. For a bivariate rv (X_1, X_2) whose copula is Gaussian with parameter ρ and whose margins are both standard normal, or standard uniform, or exponential with unit rate parameter, Figure 3 displays Pearson's correlation between X_1 and X_2 as a function of ρ (≥ 0). For all these scenarios, $\rho_S(X_1, X_2) = \frac{6}{\pi} \arcsin(\rho/2)$ and this function is very close to the identity function and corresponds to $\mathrm{cor}(X_1, X_2)$ when the margins are standard uniform.

Figure 4, in the left panel, displays Pearson's and Spearman's correlations as functions of the parameter $\theta \geq 1$ of the Gumbel copula linking two exponential rvs with unit rate parameter into a bivariate distribution. We note that Pearson's correlation dominates Spearman's. In the right panel, Pearson's and Spearman's correlations are displayed as a function of the parameter $\theta \geq 0$ of the Clayton copula linking two exponential rvs with unit parameter into a bivariate distribution. We note that now it is Spearman's correlation that dominates Pearson's, and both correlations tend to the upper bound 1 much more slowly than in the previous setting. In both scenarios, we observe that Pearson's and Spearman's correlations are continuous and monotone functions of the parameter θ and span the entire unit interval.

In order to induce a desired value of correlation $\rho > 0$ for the bivariate rv at hand, the following algorithm can be adopted:

1. Set the iteration index t equal to 1; set the error threshold ϵ equal to a reasonably small positive value (e.g., 0.0002)

2. Set the value of Spearman's rank correlation $\rho_S^{(t)}$ equal to ρ

3. Calculate the corresponding value of the copula parameter $\theta^{(t)}$ inducing $\rho_S^{(t)}$, through the function `iRho`. The function `iRho`, provided by the package `copula` (Hofert, Kojadinovic, Maechler, & Jun, 2023) in `R` (R Core Team, 2024), determines ('calibrates') the parameter θ of a selected copula inducing the (desired) value of Spearman's ρ_S. A similar function, `rhoCOP`, is available in the package `copBasic` (Asquith, 2024). We recall that ρ_S, defined though (7), can be expressed as a functional of the unique copula C that can be extracted from the bivariate rv in this way:
$$\rho_S = 12 \int_0^1 \int_0^1 C(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2 - 3$$
or, equivalently,
$$\rho_S = 12 \int_0^1 \int_0^1 u_1 u_2 c(u_1, u_2) \mathrm{d}u_1 \mathrm{d}u_2 - 3,$$
where $c(u_1, u_2; \theta) = \partial^2 C(u_1, u_2)/(\partial u_1 \partial u_2)$ is the copula density (here we consider only copulas that admit a joint density function)

4. Construct the bivariate distribution (margins combined with the copula) and calculate the value of Pearson's $\rho^{(t)}$. It is sufficient to numerically compute the mixed moment using the `cuhre` function of the `cubature` package (Narasimhan, Johnson, Hahn, Bouvier, & Kiêu, 2023); the common marginal expectations and variances can easily obtained from the known marginal distributions.

5. If $|\rho - \rho^{(t)}|/\rho > \epsilon$, set $t \leftarrow t + 1$, appropriately update $\rho_S^{(t)}$ and go back to 3., else stop and use $\rho_S^{(t)}$. This is a reasonable, but not the unique, convergence rule that can be adopted. Alternatively, one can rely on the last two values of Spearman's (or Pearson's) correlation to create a stopping rule.

We found out that an efficient updating rule for Spearman's rho (and then for the copula parameter) that can be used in step (5) works as follows (please note that here we only work with positive correlations):

$$\rho_S^{(t+1)} = \rho^{\log \rho_S^{(t)} / \log \rho^{(t)}}$$

or, equivalently,

$$\log \rho_S^{(t+1)} = \log \rho_S^{(t)} \cdot \log \rho \Big/ \log \rho^{(t)}$$

or

$$\rho_S^{(t+1)} = \left[\rho_S^{(t)}\right]^{\log \rho / \log \rho^{(t)}}. \tag{8}$$

It is clear from Eq. (8) that if $\rho^{(t)}$ is smaller (greater) than the target ρ, then $\rho_S^{(t+1)}$ will be assigned a larger (smaller) value than $\rho_S^{(t)}$, which will induce a value of $\rho^{(t+1)}$ larger (smaller) than $\rho^{(t)}$. This follows from our assumption of an increasing relationship between Spearman's rank correlation and Pearson's linear correlation of the copula-based model F, as observed in the cases depicted in Figure 4 – a point to which we will return shortly.

Using the target ρ as the initial value for ρ_S is reasonable since we know that ρ and ρ^S nearly coincide for a bivariate normal distribution, whereas for other bivariate distributions the two correlations differ, but using ρ as a start value for ρ^S is still suitable. For successive iterations, ρ still plays a benchmark role for updating ρ_S. A similar algorithm for inducing a target Pearson correlation between two discrete rvs was proposed in Barbiero (2021). In Table 1, for illustrative purpose, we reported the sequence of values of $\rho_S^{(t)}$, $\theta^{(t)}$, and $\rho^{(t)}$ $(t = 1, \ldots, 9)$ calculated according the algorithm illustrated above for a pair of exponential rvs with Clayton copula; the last row displays the final values for which the relative error between the target ρ (0.5) and the current $\rho^{(9)}$ is less than $\epsilon = 0.0002$.

We note that this iterative procedure can be skipped if the margins are standard uniform. In this case, in fact, Spearman's and Pearson's correlation coincide, since the former is nothing else than the linear correlation of the copula that can be extracted from the bivariate rv.

It is important to demonstrate why the algorithm here proposed can ensure convergence, i.e., why for any feasible value ρ (in our analysis, lying in the unit interval) it is possible to determine a value of the copula parameter θ (for the copula families here examined) inducing it, given the margins. If the sequence $\rho_S^{(t)}$ converges to some limit ρ_S^{∞}, then, according to Eq. (8), at the limit:

$$\rho_S^{\infty} = (\rho_S^{\infty})^{\frac{\log(\rho)}{\log(\rho^{\infty})}}.$$

Taking the logarithm on both sides, $\log \rho_S^{\infty} = \frac{\log \rho}{\log \rho^{\infty}} \log \rho_S^{\infty}$, and after rearranging: $\log \rho = \log \rho^{\infty}$. Thus, at convergence, $\rho^{\infty} = \rho$, which is the desired solution.

To ensure convergence, the key requirement is that Spearman's correlation is continuous and monotonic in the copula parameter θ, and this holds for all the examined copulas (see Ansari & Rockel, 2024, Section 3.2). Similarly, Pearson's rho exhibits the same behavior once the margins are fixed (see Cherubini, Luciano, & Vecchiato, 2004, p.103). This implies that, for the one-parameter copulas examined, θ quantifies the level of correlation. Consequently, for any given ρ and any pair of margins, there always exists a unique value of θ that induces ρ and the corresponding unique value of Spearman's ρ_S. The updating rule of Eq. (8) for ρ_S is therefore reasonable: the new value of ρ_S (at the iteration step $t+1$) is set equal to the old ρ_S (at t) raised to the power of the ratio between the logarithm of the target ρ to the logarithm of the old ρ. Using the logarithms of positive correlations prevents the updating rule from yielding unfeasible values: $\rho_S^{(t)}$ and $\rho^{(t)}$ remain bounded in $(0, 1)$; using this updating formula also allows to increase ρ_S if the old ρ is under the target, or decrease ρ_S if the old ρ is above the target, so that the new induced value of ρ will be updated in the correct direction with respect to the target, given the fact that both indices are monotonic with respect to the copula parameter.

We can prove that the suggested updating rule of Eq. (8) works perfectly and efficiently when applied to a one-parameter copula that is positively ordered with respect to its parameter $\theta \in (0, \infty)$, for which Spearman's and Pearson's correlation coefficients, given a pair of margins, have the following expressions (unknown to the user): $\rho_S(\theta) = \left(\frac{\theta}{\theta+1}\right)^{\alpha}$ and $\rho(\theta) = \left(\frac{\theta}{\theta+1}\right)^{\beta}$, with $\alpha, \beta > 0$. Since the two correlations are linked by the relationship $\rho(\theta) = [\rho_S(\theta)]^{\beta/\alpha}$, the suggested algorithm would produce the following sequence of correlation (and parameter) values: $\rho_S^{(1)} = \rho$, $\theta^{(1)} = \rho^{1/\alpha}/(1-\rho^{1/\alpha})$, $\rho^{(1)} = \rho^{\beta/\alpha}$; $\rho_S^{(2)} = \rho^{\left(\log\rho/\log\rho^{\beta/\alpha}\right)} = \rho^{\alpha/\beta}$, $\theta^{(2)} = \rho^{1/\beta}/(1-\rho^{1/\beta})$, and finally $\rho^{(2)} = (\rho^{\alpha/\beta})^{\beta/\alpha} = \rho$, and here the algorithm stops. Note that the functions chosen to model the two correlations in terms of the copula parameter are not so arbitrary, as can be suggested by the plot in the right panel of Figure 4, where a Clayton copula is considered.

Indeed, calibrating Pearson's linear correlation reduces to a root-finding problem, the only complication being that the function to evaluate is not available in a simple closed form, but instead requires the numerical computation of a bivariate integral. If the proposed updating rule fails, one can resort to the standard bisection method, which still guarantees convergence to the solution.

The same general calibration idea is adopted, in a multivariate context, by the VITA (VInes To Anything) simulation tool (Grønneberg & Foldnes, 2017; Grønneberg, Foldnes, & Marcoulides, 2022) for generating data with user-specified marginal distributions, covariance matrix, and certain bivariate dependencies. The most challenging objective is to ensure that the resulting d-variate distribution has the user-specified target covariance matrix, which is achieved by solving $d(d-1)/2$ equations in a single variable (the bivariate copula parameter), which is possible thanks to the fact that the functions defining the equations are continuous and increasing in the corresponding copula parameter. These calibrations are performed independently though a two-stage process: an initial high-speed calibration and a final high-precision calibration, both involving the generation of bivariate samples and the application of a standard root-finding algorithm.

Evaluating point-biserial and point-polyserial correlations

Focusing on a bivariate rv (X_1, X_2), with margins F_1 and F_2 and copula C, the formula of the mixed moment after that the second component is discretized into a rv $X_{2\mathcal{D}}$ sitting on $\{1, 2, \ldots, k\}$ with cumulative probabilities P_j is (properly adjusting

Eq.(6) in Olsson et al., 1982):

$$\mathbb{E}(X_1 X_{2\mathcal{D}}) = \sum_{j=1}^{k} j \int_{\mathbb{R}} \mathrm{d}x_1 \int_{F_2^{-1}(P_{j-1})}^{F_2^{-1}(P_j)} x_1 f(x_1, x_2; \theta) \mathrm{d}x_2, \tag{9}$$

where $P_0 = 0$ by convention and $f(x_1, x_2)$ (the joint pdf of the bivariate continuous rv) can be expressed as:

$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2),$$

with f_1 and f_2 denoting the marginal pdfs of X_1 and X_2, c the copula density. The corresponding point-polyserial correlation can then be expressed as

$$\rho_{PP} = \frac{\mathbb{E}(X_1 X_{2\mathcal{D}}) - \mathbb{E}(X_1)\mathbb{E}(X_{2\mathcal{D}})}{\sqrt{\mathrm{var}(X_1)\mathrm{var}(X_{2\mathcal{D}})}},$$

and is readily computed once the mixed moment in (9) is evaluated: to this aim, one can resort to the package `cubature` (Narasimhan et al., 2023) in `R`, which implements adaptive multivariate integration over hypercubes. The calculation is particularly fast when the margins are standard uniform; in this case, conveniently renaming the two rvs as U_1 and U_2, Eq. (9) becomes

$$\mathbb{E}(U_1 U_{2\mathcal{D}}) = \sum_{j=1}^{k} j \int_0^1 \mathrm{d}u_1 \int_{P_{j-1}}^{P_j} u_1 c(u_1, u_2; \theta) \mathrm{d}u_2.$$

In case the margins are normal, exponential, or Weibull, the unbounded support of the bivariate rv may require more computational time and lead to larger approximation error. Some practical computational issues arise when the correlation ρ is close to its natural upper bound. In this case, the copula parameter inducing such a value for a given pair of margins can become very large, requiring the functions `iRho` and `rhoCOP` to be reworked.

Study design

It is obviously impossible to investigate the behavior of the correlations ratio for all the infinitely many possible combinations of copula and margins (and discretized distribution). Here, we will focus on a significant subset. Specifically, we will consider the scenarios arising from all the combinations of the following marginal distributions:

- (standard) normal. For a rv X_1 following this distribution, $\mathbb{E}(X_1) = 0$, $\text{var}(X_1) = 1$, $\text{skewness}(X_1) = 0$, $\text{kurtosis}(X_1) = 3$;
- (standard) uniform. For a rv X_1 following this distribution, $\mathbb{E}(X_1) = 1/2$, $\text{var}(X_1) = 1/12$, $\text{skewness}(X_1) = 0$, $\text{kurtosis}(X_1) = 9/5$;
- exponential with unit rate parameter. For a rv X_1 following this distribution, $\mathbb{E}(X_1) = 1$, $\text{var}(X_1) = 1$, $\text{skewness}(X_1) = 2$, $\text{kurtosis}(X_1) = 9$;
- Weibull with scale parameter $\lambda = 2$ and shape parameter $\kappa = 1.5$. For a rv X_1 following this distribution, $\mathbb{E}(X_1) \approx 1.805491$, $\text{var}(X_1) \approx 1.502761$, $\text{skewness}(X_1) \approx 1.071987$, $\text{kurtosis}(X_1) \approx 4.390404$;

with the following parametric copulas: i) Gauss, ii) Frank, iii) Gumbel, and iv) Clayton. Please note that all the copulas considered here reduce to the comonotonicity copula if their parameter tends to its natural upper bound and reduce to the independence copula if their parameter is or tends to zero (for Gauss, Frank, and Clayton copulas) or is equal to 1 (for Gumbel). We recall that the combination of (standard) normal margins with normal copula corresponds to the bivariate (standard) normal distribution, which represents a sort of benchmark, since in this case we know the ratio of point-polyserial to polyserial correlation is constant, once the final ordinalized distribution is assigned.

For each combination of margins and copula, we considered different correlation values ρ (the biserial/polyserial correlation) for the resulting bivariate continuous distribution, namely, a grid of values spanning the entire unit interval (from 0.01 to 0.99 with constant steps of 0.01). Each value of ρ is properly induced by the copula parameter θ. Then, we computed the point-biserial/polyserial correlation and the corresponding ratio, by considering, for the sake of simplicity, the following final discrete distributions:

- $k = 2$ equal-probability categories; the corresponding expected value is $\mathbb{E}(X_{2\mathcal{D}}) = 3/2$ and the corresponding variance is $\text{var}(X_{2\mathcal{D}}) = 1/4$
- $k = 2$ with $p_1 = 2/3$ and $p_2 = 1/3$; $\mathbb{E}(X_{2\mathcal{D}}) = 4/3$ and $\text{var}(X_{2\mathcal{D}}) = 2/9$

- $k = 3$ with $p_1 = p_2 = p_3 = 1/3$; $\mathbb{E}(X_{2\mathcal{D}}) = 2$ and $\mathrm{var}(X_{2\mathcal{D}}) = 2/3$
- $k = 3$ with $p_1 = p_3 = 1/6$, $p_2 = 2/3$; $\mathbb{E}(X_{2\mathcal{D}}) = 2$ and $\mathrm{var}(X_{2\mathcal{D}}) = 1/3$
- $k = 3$ with $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/6$; $\mathbb{E}(X_{2\mathcal{D}}) = 5/3$ and $\mathrm{var}(X_{2\mathcal{D}}) = 5/9$
- $k = 4$ with $p_1 = p_2 = p_3 = p_4 = 1/4$; $\mathbb{E}(X_{2\mathcal{D}}) = 5/2$ and $\mathrm{var}(X_{2\mathcal{D}}) = 5/4$
- $k = 4$ with $p_1 = p_4 = 1/8$, $p_2 = p_3 = 3/8$; $\mathbb{E}(X_{2\mathcal{D}}) = 5/2$ and $\mathrm{var}(X_{2\mathcal{D}}) = 3/4$
- $k = 4$ with $p_1 = 1/2$, $p_2 = 1/4$, $p_3 = 1/6$, $p_4 = 1/12$; $\mathbb{E}(X_{2\mathcal{D}}) = 11/6$ and $\mathrm{var}(X_{2\mathcal{D}}) = 35/36$

The special and representative cases of the uniform distribution, a symmetric unimodal or reverse-U-shaped distribution, and a decreasing (and thus asymmetric) distribution are easily identifiable. They are graphically displayed in Figure 5.

We considered only positive values of ρ, since whereas the Frank and the Gauss copulas are comprehensive copulas (i.e., they can model the full range of dependence, from countermonotonicity to comonotonicity, including independence), and then they are able to induce all the values of ρ in $[-1, +1]$, the Gumbel and the Clayton copulas can only model positive dependence and thus induce only positive values of linear correlation.

Recall that the linear correlation ρ can attain the value +1 if and only if the two marginal distributions are of the same type (e.g., both normal, or both exponential, or otherwise identical): this is related to the result known as "attainable correlations" (McNeil et al., 2015). Then, the algorithm is always able to find a corresponding value for θ and for ρ_S inducing the assigned ρ. If, instead, the marginal distributions were not of the same type, then ρ_S would achieve the value 1 in case of perfect positive dependence, but ρ would not.

Results

Figure 6 displays, for each combination of copula and marginal distribution examined, the values of the ratio between point-biserial and biserial correlations, ρ_{PP}/ρ, when $p_1 = p_2 = 1/2$. Starting from the case of normal margins, one can note that the values of the ratio cluster around $2\varphi(0) = 0.7979$, which is the constant value of the ratio for the bivariate normal distribution. But whereas for the normal copula the ratio is exactly constant, as theoretically expected, for the other three copulas there is some variability. Specifically, for the Frank copula it exhibits a decreasing trend with respect to ρ, especially for values of ρ close to 1. For the Gumbel copula it shows an increasing trend, and for the Clayton copula a decreasing-increasing-decreasing trend. When ρ tends to 1, the ratio converges to $2\varphi(0)$ for all the copulas.

Moving to the case of standard uniform margins, the situation is similar to the previous case: for the four copulas, now the ratio spans a narrower range and always converges to the value $\sqrt{3}/2 \approx 0.8660$ when ρ tends to 1, which corresponds to the maximum point-biserial correlation for a uniform rv (Barbiero, 2025; Cheng & Liu, 2016). Curiously, now it is the Frank copula that ensures an almost constant trend for the ratio; the Gauss and the Gumbel copulas produce a slightly increasing trend, whereas the Clayton copula a first increasing and then decreasing trend.

If we consider the case of exponential margins, it is immediately apparent that the results are sensibly different from those of the two previous cases: the ranges spanned by the ratio are much larger than before. The Clayton copula is the dependence structure entailing the largest range and also the largest value for the ratio (it can take on values greater than 1.2); it shows a decreasing trend with ρ. The Frank and Gauss copula show a decreasing trend too, but with a narrower range for the ratio. In contrast, the Gumbel copula shows an increasing trend. All the ratios converge to a common value ($\ln 2 \approx 0.6931$, see Cheng and Liu 2016, p.49, and Barbiero 2025) for ρ tending to 1.

Moving to the last case (Weibull margins) the situation is similar to the exponential case, but the range covered by the correlation ratio is now smaller; moreover, for the Clayton copula, the ratio shows an increasing-decreasing trend with ρ. The ratios for ρ tending to 1 converge to a value between 0.77 and 0.78.

Figure 7 displays, for each combination of copula and marginal distribution examined, the values of the ratio between point-polyserial and polyserial correlations, when the discrete distribution consists of $k = 3$ equal-probability categories ($p_1 = p_2 = p_3 = 1/3$). It is quite straightforward to recognize that the results, in terms of behaviour of the correlations ratio, are very similar to those presented in Figure 6. The only substantial difference is that now the ratio shifts towards larger values than before, but the width of the range of values taken by the ratio is almost the same. Now, the ratio can take values greater than 1 even when the margins are Weibull and the dependence structure is modeled by the Clayton copula.

The fact that the linear correlation can inflate after discretization, as foreseen in the previous sections, may be surprising. However, as early as in MacCallum et al. (2002), it was recognized that dichotomization could increase the magnitude of linear correlation. This phenomenon was explained by attributing it to sampling error and small sample size, as stated: "Simply because of sampling error, sample correlations may increase following dichotomization, especially when sample size is small or the sample correlation is small." Our study clarifies how, at the level of rv (or at the population level, as MacCallum et al. (2002) would put it), it is possible for correlation to increase in magnitude after discretization, especially when deviating from the condition of bivariate normality – that is, when considering asymmetric marginal distributions and copulas capable of modeling tail dependence. Furthermore, if, instead of considering a bivariate rv, one were to analyze random samples drawn from it – particularly small-sized samples – then the effect of sampling error would only increase the probability of obtaining a sample point-polyserial correlation greater than the estimated polyserial correlation. It can be observed that the occurrence of such an inflation is also related to the choice of assigning the first consecutive positive integers to the contiguous intervals of the discretized variable. Adopting other scores, if possible – for example, the so-called Pearson scores (Pearson, 1913), corresponding to the conditional mean of the continuous rv over the interval between two consecutive thresholds – may reduce if not prevent the occurrence of correlation inflation.

The other results obtained for this study, related to the remaining six distributions for the discretized variable, are not displayed here but are available as supplementary material. They confirm the primary role of the continuous marginal distribution and the copula in determining the trend and the range of the ratio between point-polyserial and polyserial correlations. The distribution of the discretized variable also affects the ratio: increasing the number of equal-probability categories pushes the ratio up while keeping the margins and the copula fixed. When the number of categories is fixed as well, the effect of making the probabilities unequal is more difficult to interpret and seems to be strictly interconnected with the type of continuous margin and copula.

Trying to summarize what has been presented so far, the results of our numerical study indicate that the ratio between point-polyserial and polyserial correlations is not constant with ρ, although in some specific cases (standard uniform margins and Frank copula) it can be considered as nearly constant, confirming the fact that a constant ratio characterizes the (bivariate) normal distribution only. The range of values that the ratio can span sensibly varies depending on the copula selected, on the marginal, and - to a smaller extent - on the discretization adopted. It is important to emphasize that for some combinations of copula, continuous distribution, and discrete distribution, the ratio between point-polyserial and polyserial correlation can be greater than 1. That is, if we discretize one of the two identical continuous components in a bivariate model, it may occur that instead of decreasing, the correlation increases, and this can seem counter-intuitive at least for those who are used to working with the bivariate normal distribution and its properties. Although this happens for small values of the correlation between the continuous components (see Figures 6 and 7, bottom-left panels), it is nevertheless indicative of how, when moving away from the bivariate normal distribution, some of its peculiar statistical properties are lost.

It is important to notice that given the margins and the final distribution of the discretized variable, the point-polyserial correlation converges to a common value as ρ approaches 1, for all the copulas considered here, which themselves converge to the comonotonicity copula at the same limit. This behavior does not hold as ρ tends to zero: although all the copulas converge to the independence copula, the ratios of the point-polyserial to the polyserial correlations do not seem to converge to a common limit.

Application to real data

The St. Louis Risk Research Project was an observational study to assess the effects of parental psychological disorders on child development. In the preliminary study, 69 families with 2 children were studied. The data set consists of 69 observations on 7 variables (Schafer, 2024): G, Parental risk group; D_1, Symptoms of child 1; D_2 Symptoms of child 2; R_1, Reading score of child 1; V_1, Verbal score of child 1; R_2, Reading score of child 2; and V_2, Verbal score of child 2. The parental risk group was coded 1, 2 or 3, from low or high, and the child symptoms 1 = low or 2 = high. Missing values occur on all variables except G. Here we consider the variable G, whose values are re-coded so that 1 means high risk, 2 medium, and 3 low. We rename it $X_{2\mathcal{D}}$, treating it as a discrete rv with support $\{1, 2, 3\}$. Similarly, we rename the variable R_1 as X_1 for consistency with the notation used in previous sections. After deleting the observations with missing value for X_1, we obtain 48 complete cases. The scatter plot of the two variables is displayed in Figure 8. The sample linear correlation is equal to 0.3919. The marginal frequencies of $X_{2\mathcal{D}}$ are $n_1 = 12$, $n_2 = 19$, $n_3 = 17$. Summary statistics for X_1 are: mean 108.1, median 113, variance 277.8, skewness -0.999, kurtosis 4.082. The distribution of X_1 is thus negatively asymmetric and slightly leptokurtic. Fitting the Weibull distribution to X_1 leads to the following MLEs of the shape and scale parameters $\hat{\lambda} = 114.8$ and $\hat{\kappa} = 8.343$. The Kolmogorov-Smirnov test, neglecting the presence of ties in the sample, accepts the hypothesis that the values of X_1 come from the fitted distribution (p-value=0.4243).

If we accept the idea that $(X_1, X_{2\mathcal{D}})$ derives from a bivariate continuous rv (X_1, X_2) after discretization of the variable X_2, we can assume that (X_1, X_2) is a bivariate rv with identical Weibull margins with shape parameter 8.343 and scale parameter 114.8, and some copula C. This copula can be estimated from the bivariate sample of $(X_1, X_{2\mathcal{D}})$, by considering the corresponding pseudo-observations, i.e, transforming the actually observed values into the corresponding normalized ranks, going from $1/(n+1)$ to $n/(n+1)$. Since the values for both X_1 and $X_{2\mathcal{D}}$ are integers, there may be some ties; in order to compute the normalized ranks, we adopt the so-called 'first' method in the `R` function `rank`, which results in a permutation with increasing values at each index set of ties. Then, on this sample (U_1, U_2) of pseudo-observations, we search for the bivariate copula which fits best, by using the function `BiCopSelect` of the package `VineCopula` (Nagler et al., 2023). It turns out that the best copula for our data is the Frank copula with $\theta = 2.48$. Then, we can construct the estimated joint distribution of (X_1, X_2) by combining the two Weibull margins, with the estimated shape and scale parameters, with this copula. The evaluation of the corresponding correlation coefficient for this bivariate distribution, carried out as done in the previous section by using the package `cubature`, returns the value $\hat{\rho} = 0.3592$. The contour plot of the estimated joint pdf along with the plot of the marginal pdf is displayed in Figure 9.

If we want to compute the theoretical point-polyserial correlation, when one continuous margin is discretized into an ordinal variable having the same distribution of G in the sample ($p_1 = 0.25$, $p_2 = 0.3958$, and $p_3 = 0.3542$), we obtain $\hat{\rho}_{PP} = 0.3471$, and then a ratio between the two correlations around 0.966. Please note that this latter value of correlation is quite smaller than the sample correlation on the original sample, but this discrepancy is obviously due to the fact that we replaced the sample marginal distribution of X_1 with the estimated Weibull distribution and that we estimated the unknown dependence structure of the underlying continuous bivariate distribution as well.

A referee pointed out that since our available data contain a discrete/ordinal variable with ties, using ranks can lead to biased estimates, as the standard theory of bivariate copulas hold when both variables are continuous. To address this issue, we can consider the proposal described in Pan and Joe (2024) and implemented in Nagler and Vatter (2025), according to which for a discrete variable one has to consider both the vector of the empirical cdf and its left-side limit. Following this approach and restricting our attention to one-parameter copulas, it turns out that the best-fitting copula is the Frank copula with parameter $\theta = 2.70$. Therefore, with this correction for ties, the copulas is still Frank, but with a larger parameter value than before. Using this copula as the true dependence structure of the underlying bivariate continuous random variable, and following the same steps as before, we obtain $\hat{\rho} = 0.386$ and $\hat{\rho}_{PP} = 0.373$, with the ratio between the point-polyserial and polyserial correlations equal to 0.966. Therefore, both correlation measures increased, like θ, but the ratio remained (approximately) the same.

Another proposal involves introducing some randomness when computing the pseudo-random observations, denoted as V_i, of the ordinal rv

$$V_i = \frac{\sum_{j=1}^{n} \mathbb{1}(x_{2\mathcal{D},j} < x_{2\mathcal{D},i})}{n} + U_i \frac{\sum_{j=1}^{n} \mathbb{1}(x_{2\mathcal{D},j} = x_{2\mathcal{D},i})}{n}, \quad U_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1),$$

where the first addend is the left-side limit of the empirical cdf of X, and in the second term U_i introduces controlled randomness by weighting the sample frequency of the i-th observed value, thereby preventing ties while preserving marginal probabilities. By adopting this approach, it is evident that the results themselves become random. In fact, when implementing the analysis in `R`, the estimated copula and its parameter(s) vary depending on the chosen simulation seed.

A more natural way to proceed can be to use a full maximum likelihood approach, by assuming that the sample at hand comes from the discretization of one rv of a pair of identically distributed rvs characterized by some unique copula C. This approach is basically the one pursued by de Leon and Wu 2011, Sect.2, who used a Gaussian copula, whereas here we do not limit a priori the choice of the dependence structure of the bivariate latent rv. In this sense, our approach is more similar to that of Chen and Hanson 2017, Eq.18. For the observed rv the log-likelihood function can be constructed as

$$\begin{aligned} \ell(\theta_1, \theta_2, \lambda, \kappa, \theta) &= \sum_{i=1}^{n} \log \left(\int_{\theta_{x_{2\mathcal{D},i}-1}}^{\theta_{x_{2\mathcal{D},i}}} c(F_{X_1}(x_{1i}; \lambda, \kappa), F_{X_2}(x_2; \lambda, \kappa); \theta) f_{X_1}(x_{1i}; \lambda, \kappa) f_{X_2}(x_2; \lambda, \kappa) \mathrm{d}x_2 \right) \\ &= \sum_{i=1}^{n} \log f_{X_1}(x_{1i}; \lambda, \kappa) + \sum_{i=1}^{n} \log \left(\int_{\theta_{x_{2\mathcal{D},i}-1}}^{\theta_{x_{2\mathcal{D},i}}} c(F_{X_1}(x_{1i}; \lambda, \kappa), F_{X_2}(x_2; \lambda, \kappa); \theta) f_{X_2}(x_2; \lambda, \kappa) \mathrm{d}x_2 \right) \end{aligned} \tag{10}$$

where $c(\cdot, \cdot; \theta)$ is the copula density; f_{X_1} and f_{X_2} are the marginal pdfs, λ and κ are the unknown parameters of the Weibull distribution which we assume X_1 and X_2 both follow, θ is the parameter of the Frank copula linking the two margins into the joint distribution, θ_1 and θ_2 are the two thresholds for the rv X_2: $X_{2\mathcal{D}} = \sum_{i=1}^{k} i \mathbb{1}_{\theta_{i-1} < X_2 \le \theta_i}$, $i = 1, 2, 3$, with $\theta_0 = 0$ and $\theta_3 = +\infty$. Maximizing the log-likelihood function (10) with respect to the five unknown parameters $0 < \theta_1 < \theta_2$, $\lambda > 0$, $\kappa > 0$, and θ, one can easily obtain the MLEs numerically, for example using the `solnp` (Ghalanos & Theussl, 2015) function in the `R` package `Rsolnp`. The complication with respect to the analogous procedure of de Leon and Wu (2011) is that while for the Gauss copula the integral in (10) can be computed analytically, for a general copula, it can be computed only numerically. For the problem at hand, the full maximization leads to the following estimates: $\hat{\theta}_1 = 98.18$, $\hat{\theta}_2 = 114.85$, $\hat{\lambda} = 114.55$, $\hat{\kappa} = 8.25$, $\hat{\theta} = 2.62$. Note that the value of the estimate of θ falls between the two values previously calculated by considering the ranks. This approach, although requiring a priori the specification of the parametric families for the margins and for the copula, allows the user to overcome all the theoretical and practical issues related to the direct modeling of the mixed-type rv $(X_1, X_{2\mathcal{D}})$ by modeling the latent rv (X_1, X_2) instead. We remark that the hypothesis of identical continuous margins for the latent bivariate rv is not so restrictive after all and allows for the construction of a pair of rvs whose linear correlation can reach the natural maximum value of $+1$.

Conclusion

This work showed that the alleged constancy of the ratio between point-polyserial and polyserial correlations does not hold if we move away from the bivariate normal distribution, i.e., considering either non-normal marginal distributions or non-normal copulas. This departure is more substantial if we consider asymmetrical marginal distributions (such as exponential) or copulas showing lower or upper tail dependence and being radially non-symmetric (e.g., Clayton, Gumbel). The effect of the discretized distribution on the values of the correlations ratio is a bit less explainable. However, increasing the number of categories k from 2 up to 4, pushes the ratio toward higher values, taking the continuous margins and copula fixed, and the probabilities of the ordinalized rv uniform. These results have clear practical implications: since in many fields the bivariate/multivariate normal distribution does not fit real data well, this suggests that the presumed constancy of the ratio between correlations before and after discretization, which characterizes this model, cannot be maintained. Consequently, the reduced complexity in simulations, model calibration, and likelihood-based estimation, which may be ensured by the constant ratio, is lost.

As a by-product, this work provides an interesting and quite comprehensive picture of how the correlation coefficient for a continuous bivariate distribution is inextricably related to both the margins and the dependence structure identified by its unique copula, and how the linear correlation coefficient is modified when either random component is discretized into 2 or more categories (and the change in correlation still depends on the margins, on the copula, and on the discrete distribution itself). This work also contributes to the implementation of the bijective relationship between the copula parameter and Spearman's rank correlation, which – for a bivariate rv with continuous margins – depends on the copula only. This implementation is crucial, since the function linking the copula parameter to Spearman's rank correlation is rarely expressible in closed form. Therefore, a numerical evaluation is typically required, which can be easily performed by recalling the expression of Spearman's correlation as an integral over the unit square.

Throughout the work, we have always assumed that the discrete distribution is supported over the set of the first k positive integers. Someone may observe that since discrete distributions often arise from the discretization of continuous rvs, it would be more reasonable to assign to the categories numerical scores that are not necessarily integer and/or equally spaced, but are rather related to the support of the parent distribution (if known). Such scores can be obtained by resorting to the optimization of some objective function and yield to so-called optimal scores (Bedrick, 1995; Bedrick & Breslin, 1996) and, as hinted before, may prevent from correlation inflation. Here we preferred to work with consecutive integer scores only, also because the majority of papers about point-polyserial correlation followed the same approach (Demirtas, Ahmadian, Atis, Can, & Ercan, 2016; Demirtas & Hedeker, 2016; Demirtas & Vardar-Acar, 2017; Demirtas & Yavuz, 2015), whether the purpose was to model the mixed-type data or to simulate it.

Although in this study, for the sake of brevity, we necessarily excluded many parametric families of distribution for the margins and parametric copulas for the dependence structure, nonetheless, we developed an `R` implementation that can be easily adapted in order to encompass such marginal or dependence models by simply reworking the relevant code chunks. In order to build a more complete picture, further research can therefore explore the behavior of the correlation ratio for other combinations of copulas and margins. Attention should be paid when calculating the correlation coefficients numerically, since *approximation errors* in the computation of the related double integrals through the package `cubature` may make the final results less reliable. Further to this point, we want to emphasize that the algorithms and the pertaining results presented in this work have been obtained dealing with rvs and their joint distribution directly, and not with some huge sample drawn from it, as other algorithms do (Demirtas et al., 2016, 2017; Demirtas & Hedeker, 2016; Demirtas & Vardar-Acar, 2017; Demirtas & Yavuz, 2015). Hence, there is no *sampling error* related to the finiteness of the sample size. We believe this is a favorable feature of our setup.

We hope that this work can contribute to the discussion on the simulation of mixed continuous-discrete variables, which are common in the fields of psychology, behavioral science, and social sciences. When simulating and inferring from such data, where the framework of the multinormal variable no longer applies, it is crucial to exercise caution when extending its properties and results, or when attempting to characterize the class of non-normal distributions. The work highlights the central role of combining marginal distributions with the copula in defining a latent continuous variable, which should be preferred over the use of transformations of the multinormal. More importantly, it underscores the risk associated with using Pearson's ρ to measure the correlation between variables, due to its often overlooked behavior under monotone but non-linear transformations. This work also sheds light on the conditions under which discretizing a variable can lead to an increase in the linear correlation coefficient—a phenomenon that is not unknown, yet often dismissed as a byproduct of sampling errors or small sample sizes rather than recognized as the result of a more systematic mechanism.

Acknowledgments

The author thanks the editor, the associate editor, and the three anonymous referees for their valuable suggestions and comments, which have improved the quality of the initial draft.

Supplementary material

Foundational results on copulas and copula-based models, which underpin the framework adopted in this paper, are collected in the online appendix for completeness.

A supplementary PDF file contains the graphs of the correlation ratios for six of the scenarios examined that are not displayed in this manuscript.

Relevant `R` code implementing the analyses contained in this work is available on GitHub: `https://github.com/alessandro-barbiero/NNPointPolyCorr`.

Disclosure statement

The author reports there are no competing interests to declare.

Funding

The author acknowledges financial support by the PRIN2022 project 'The effects of climate change in the evaluation of financial instruments' financed by the 'Ministero dell'Università e della Ricerca' with grant number 20225PC98R, CUP Code: G53D23001960006.

References

Ansari, J., & Rockel, M. (2024). Dependence properties of bivariate copula families. *Dependence Modeling*, *12*(1), 20240002.

Asquith, W. H. (2024). copbasic—general bivariate copula theory and many utility functions [Computer software manual]. (R package version 2.2.4)

Barbiero, A. (2021). Inducing a desired value of correlation between two point-scale variables: a two-step procedure using copulas. *AStA Advances in Statistical Analysis*, *105*(2), 307–334.

Barbiero, A. (2025). Maximal point-polyserial correlation for non-normal random distributions. *British Journal of Mathematical and Statistical Psychology*, *78*(1), 341–377.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *42*(3), 293–312.

Bedrick, E. J. (1995). A note on the attenuation of correlation. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 271–280.

Bedrick, E. J., & Breslin, F. C. (1996). Estimating the polyserial correlation coefficient. *Psychometrika*, *61*, 427–443.

Chen, Y., & Hanson, T. (2017). Copula regression models for discrete and mixed bivariate responses. *Journal of Statistical Theory and Practice*, *11*, 515–530.

Cheng, Y., & Liu, H. (2016). A short note on the maximal point-biserial correlation under non-normality. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 344–351.

Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula methods in finance.* Chichester, UK: John Wiley & Sons.

Cox, N. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics*, *30*(1), 171–178.

De Leeuw, J. (2006). Multiple correspondence analysis and related methods. In B. J. Greenacre M. (Ed.), *Nonlinear principal component analysis and related techniques* (pp. 107–133). Boca Raton, FL: Chapman & Hall / CRC.

de Leon, A. R., & Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, *30*(2), 175–185.

Demirtas, H., Ahmadian, R., Atis, S., Can, F. E., & Ercan, I. (2016). A nonnormal look at polychoric correlations: Modeling the change in correlations before and after discretization. *Computational Statistics*, *31*, 1385–1401.

Demirtas, H., Allozi, R., Hu, Y., Inan, G., & Ozbek, L. (2017). Joint generation of binary, ordinal, count, and normal data with specified marginal and association structures in monte-carlo simulations. In D. G. Chen & J. D. Chen (Eds.), *Monte-Carlo simulation-based statistical modeling* (pp. 3–15). Singapore City, Singapore: Springer.

Demirtas, H., & Hedeker, D. (2016). Computing the point-biserial correlation under any underlying continuous distribution. *Communications in Statistics - Simulation and Computation*, *45*(8), 2744–2751.

Demirtas, H., Hedeker, D., & Mermelstein, R. J. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine*, *31*(27), 3337–3346.

Demirtas, H., & Vardar-Acar, C. (2017). Anatomy of correlational magnitude transformations in latency and discretization contexts in Monte-Carlo studies. In D. G. Chen & J. D. Chen (Eds.), *Monte-carlo simulation-based statistical modeling* (pp. 59–84). Singapore City, Singapore: Springer.

Demirtas, H., & Yavuz, Y. (2015). Concurrent generation of ordinal and normal data. *Journal of Biopharmaceutical Statistics*, *25*(4), 635–650.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521–532.

Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika*, *80*, 1066–1083.

Ghalanos, A., & Theussl, S. (2015). Rsolnp: General non-linear optimization using augmented Lagrange multiplier method [Computer software manual]. (R package version 1.16.)

Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines.

Psychometrika, *82*, 1035–1051.

Grønneberg, S., Foldnes, N., & Marcoulides, K. M. (2022). covsim: An R package for simulating non-normal data for structural equation models using copulas. *Journal of Statistical Software*, *102*(3), 1–45.

Hofert, M., Kojadinovic, I., Maechler, M., & Jun, Y. (2023). copula: Multivariate dependence with copulas [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=copula/` (R package version 1.1-2)

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, *44*(1/2), 289–292.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*(1), 19.

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools*. Princeton University Press.

Moss, J., & Grønneberg, S. (2023). Partial identification of latent correlations with ordinal data. *Psychometrika*, *88*(1), 241–252.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., & Erhardt, T. (2023). Vinecopula: Statistical inference of vine copulas [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=VineCopula` (R package version 2.5.0)

Nagler, T., & Vatter, T. (2025). rvinecopulib: High performance algorithms for vine copula modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rvinecopulib` (R package version 0.7.1.1.0)

Narasimhan, B., Johnson, S. G., Hahn, T., Bouvier, A., & Kiêu, K. (2023). cubature: Adaptive multivariate integration over hypercubes [Computer software manual]. Retrieved from `https://bnaras.github.io/cubature/` (R package version 2.1.0)

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education*, *15*, 625–632.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*, 337–347.

Pan, S., & Joe, H. (2024). Assessing copula models for mixed continuous-ordinal variables. *Dependence Modeling*, *12*(1), 20240001.

Pearson, K. (1909). On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika*, *7*(1/2), 96–105.

Pearson, K. (1913). On the measurement of the influence of “broad categories” on correlation. *Biometrika*, *9*(1/2), 116–139.

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Schafer, J. L. (2024). mix: Estimation/multiple imputation for mixed categorical and continuous data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=mix` (R package version 1.0-12)

Song, X.-Y., & Lee, S.-Y. (2003). Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. *Multivariate Behavioral Research*, *38*(1), 57–79.

Zhang, P., Liu, B., & Pan, J. (2024). Iteratively reweighted least squares method for estimating polyserial and polychoric correlation coefficients. *Journal of Computational and Graphical Statistics*, *33*(1), 316–328.

Table 1

Iterative computation of the parameter θ of the Clayton copula ensuring a linear correlation equal to $\rho = 0.5$ between two exponential rvs with unit rate parameter. For each iteration, along with the values of Spearman's and Pearson's correlations and θ, we reported in the last column the values of the relative error, whose maximum tolerated value was set equal to 0.0002.

iteration t	$\rho_S^{(t)}$	$\theta^{(t)}$	$\rho^{(t)}$	$\lvert\rho - \rho^{(t)}\rvert/\rho$
1	0.50000	1.076	0.30488	0.39024
2	0.66733	1.892	0.43218	0.13564
3	0.71591	2.259	0.47473	0.05053
4	0.73276	2.409	0.49025	0.01951
5	0.73908	2.468	0.49618	0.00764
6	0.74153	2.492	0.4985	0.00301
7	0.74249	2.502	0.49941	0.00119
8	0.74287	2.505	0.49977	0.00047
9	0.74303	2.507	0.49992	0.00017

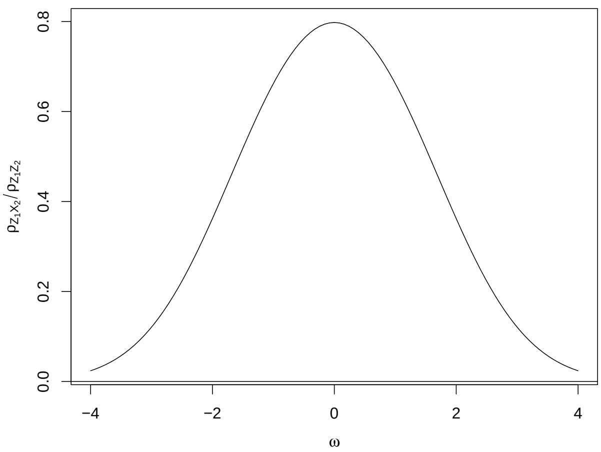

Figure 1. Ratio between point-biserial and biserial correlations as a function of the cut-point ω for a bivariate normal rv - Equation (1); the maximum value, equal to $\sqrt{2/\pi}$, is attained at $\omega = 0$

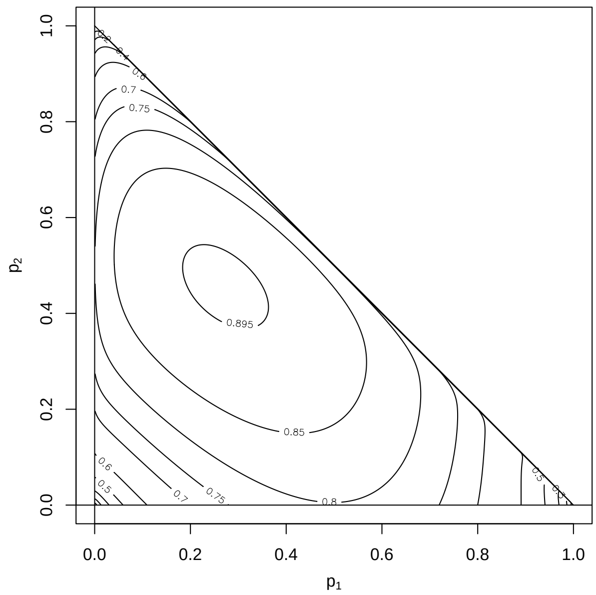

Figure 2. Level curves of the ratio between point-triserial and triserial correlations for a bivariate normal rv as a function of the probabilities p_1 and p_2 (with p_3 being equal to $1 - p_1 - p_2$).

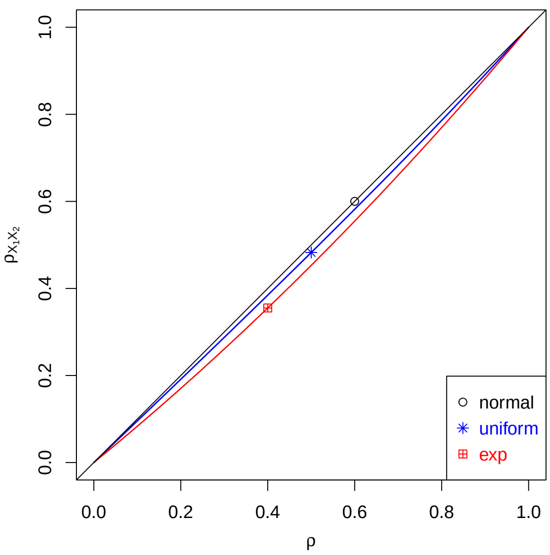

Figure 3. Pearson's correlation for a bivariate rv with Gauss copula with correlation parameter $\rho \geq 0$ and normal, uniform or exponential marginal distributions. The black dotted line represents the first-third orthants bisector (the identity function) and corresponds to Pearson's correlation in case of standard normal margins.

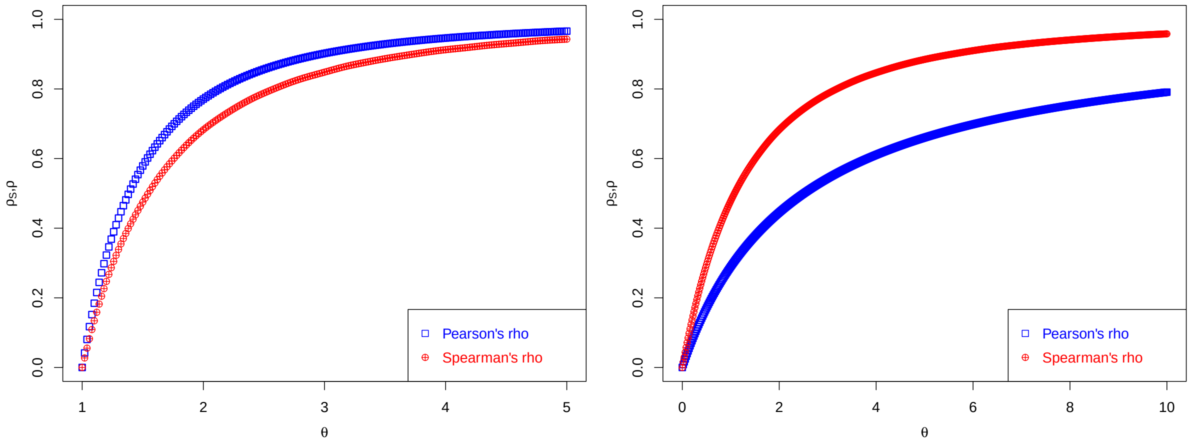

Figure 4. Pearson's and Spearman's correlations for a bivariate distribution with exponential margins with unit rate parameter and Gumbel copula with parameter $\theta \geq 1$ (left) and Clayton copula with parameter $\theta \geq 0$ (right). Please note that the scales for the x axis are different, since we considered a larger interval for the θ parameter of the Clayton copula.

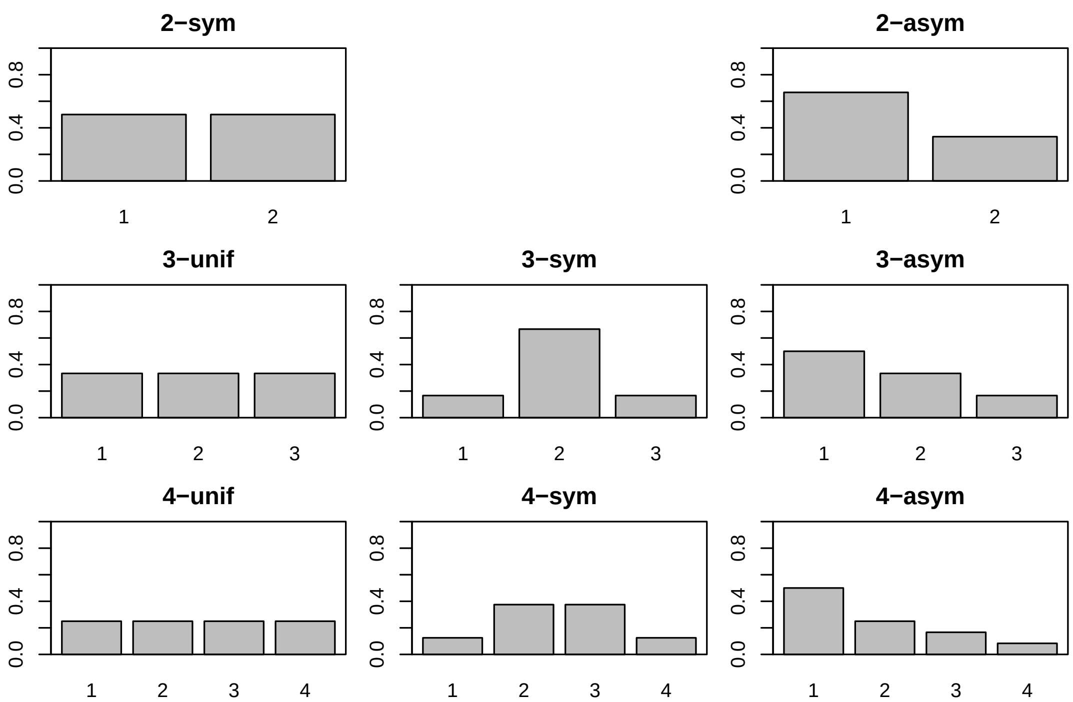

Figure 5. Discretized distributions examined in Section Point-polyserial correlation under non-normality.

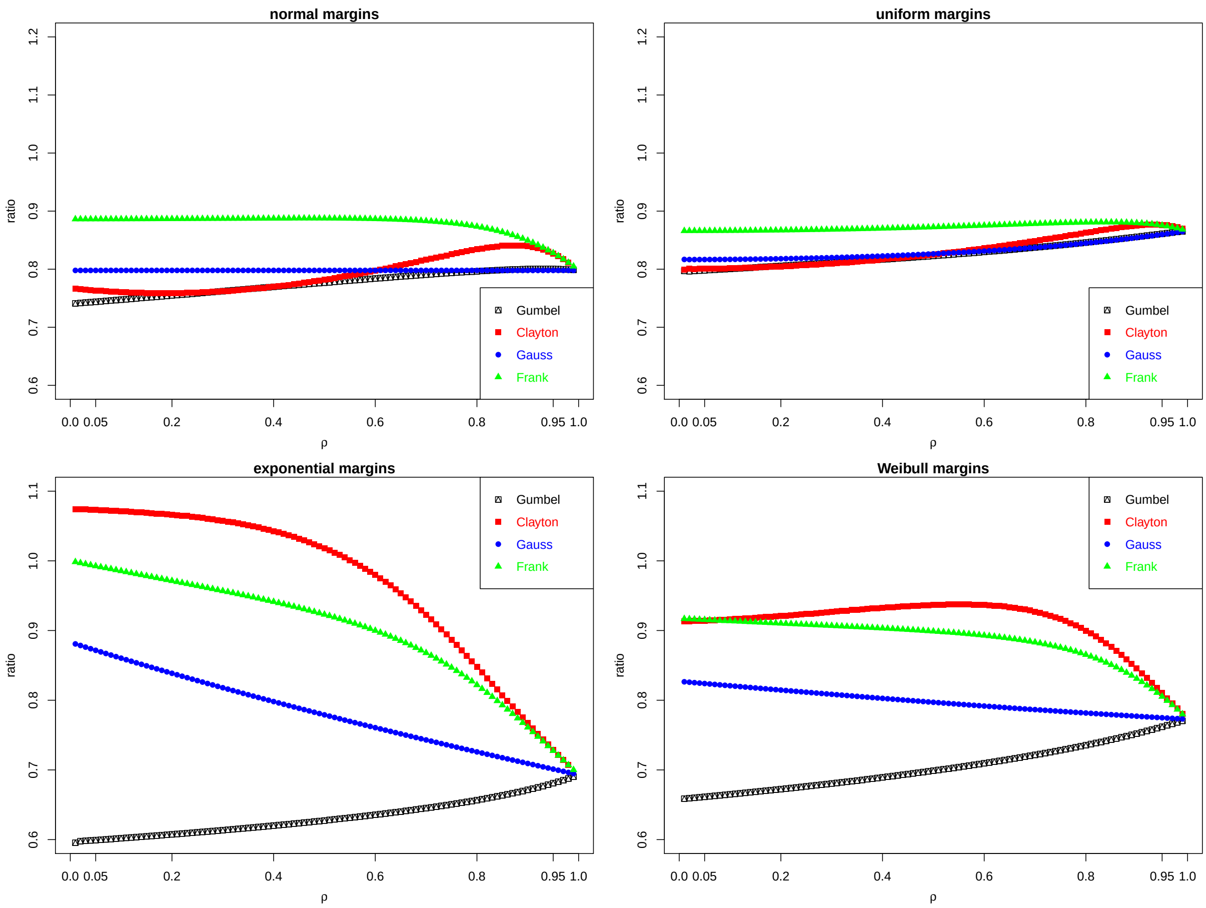

Figure 6. Graphs of the ratio between point-biserial correlation and biserial correlation for several margins, copulas and values of ρ. The final discrete distribution takes on two values with the same probability $p_1 = p_2 = 1/2$.

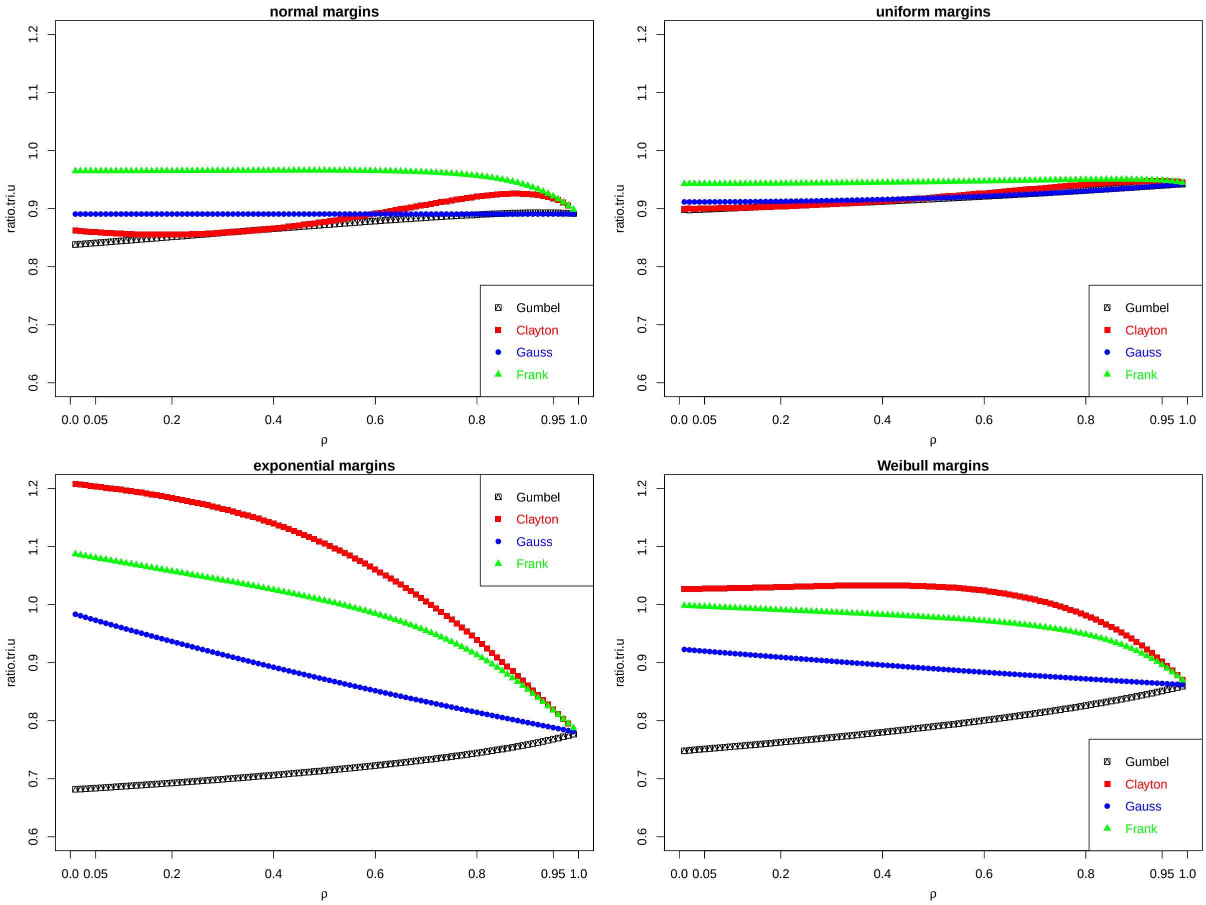

Figure 7. Graphs of the ratio between point-polyserial correlation and polyserial correlation for several margins, copulas and values of ρ for a bivariate rv. Here the number of categories of the ordinalized rv is $k = 3$ and the corresponding probabilities are $p_1 = p_2 = p_3 = 1/3$.

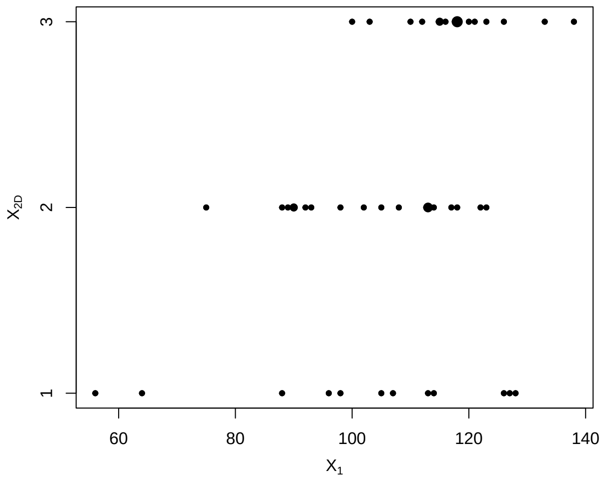

Figure 8. Scatter plot of the variables X_1 and $X_{2\mathcal{D}}$ for the St. Louis dataset. The categories of $X_{2\mathcal{D}}$ have been recoded as 1=high risk, 2=medium risk, 3=low risk. Points $(90, 2)$ and $(115, 3)$ appear twice in the data set; point $(113, 2)$ appears three times, and point $(118, 3)$ four times. The points are displayed as circles, with areas proportional to their corresponding frequencies.

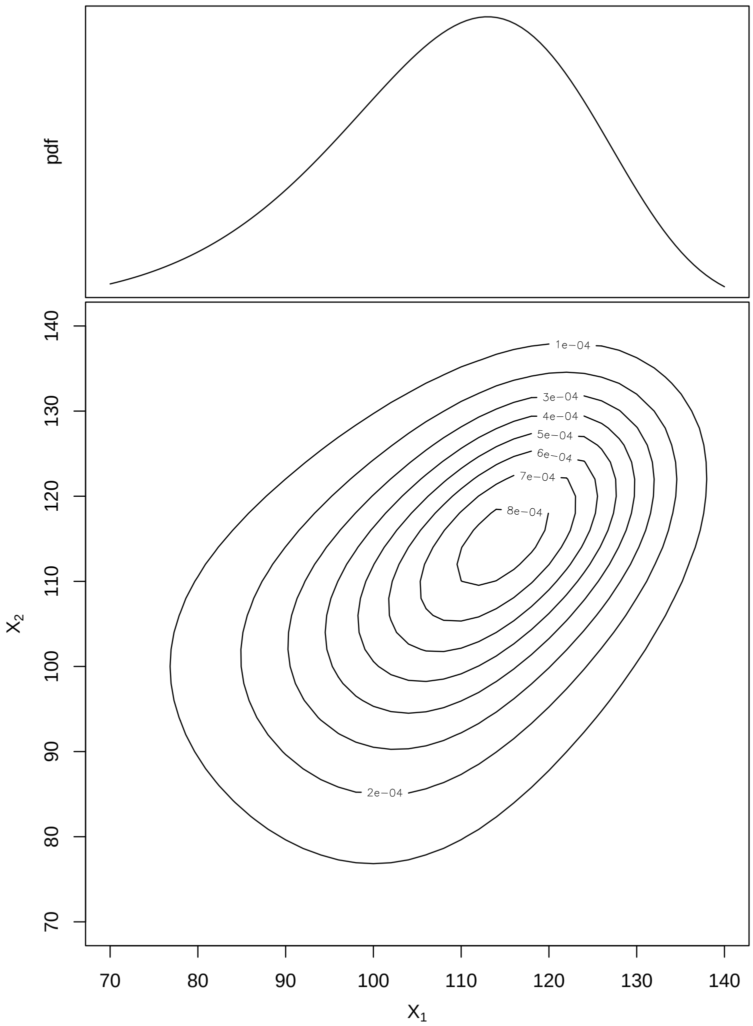

Figure 9. Level curves of the estimated latent joint probability density function underlying the pair of variables $(X_1, X_{2\mathcal{D}})$, with $X_{2\mathcal{D}}$ ordinal and X_1 continuous, from the St Louis dataset (bottom panel). The top panel displays the estimated common marginal distribution of the continuous components, a Weibull distribution with scale parameter $\lambda = 114.8$ and shape parameter $k = 8.343$. The copula linking the two margins into the joint distribution is a Frank copula with parameter $\theta = 2.48$.