Modeling correlated counts in reliability engineering

Alessandro Barbiero

Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, via Conservatorio 7 - 20122 Milan (Italy)

Abstract

In this chapter, we review the problem of modelling correlated count data. Among the several methods that can be used for this scope, we focus on the copula approach, illustrating its advantages, but also possible limitations and issues arising in the discrete context if compared to the continuous case. After introducing the basic notions about copulas, the construction of a multivariate joint distribution is discussed and pseudo-random simulation and point estimation of copula-based models for count data are then outlined. Results related to minimum and maximum correlation between two assigned discrete marginal distributions are also described and put in connection with the choice of the copula to be used for modelling correlated counts. A numerical example and an application to a real dataset are provided.

Keywords: copula, cumulative distribution function, dependence structure, discrete variables, joint distribution, marginal distributions, Pearson's correlation

2010 MSC: 60E05, 62F10, 62H20

1. Introduction

Over the last decades, a large amount of literature on discrete bivariate and multivariate distributions has been accumulated. An extensive account of these distributions can be found in [1, 2]. Focusing on reliability engineering and lifetime analysis, discrete distributions are naturally used for modelling count variables, such as the number of rounds fired by a weapon till the first failure;

the number of deaths at a given place over a given time period; the number of cycles prior to the first failure when devices work in cycles; the number of periods successfully completed without failure; the numbers of accidents a worker experiences in different time periods; but they can be usefully employed also when the lifetime (of an item, a system, ...) is measured in days, weeks, months, etc., i.e., when lifetime is measured on a discrete scale. When the researcher has to deal with several phenomena of this type, he needs to take into account and model their statistical association by considering appropriate multivariate distributions.

For a long time, the multivariate normal distribution has been used even for modelling non-continuous correlated data. However, it cannot turn out to be a satisfactory choice, especially when one needs to model data with just a few distinct observed values or presenting an excess of zeros, something which is quite common in several engineering fields. Thus the need for using a multivariate discrete model. [3] suggested that "A multivariate discrete distribution should satisfy some theoretical properties in order to model practical situations. We highlight two of these theoretical properties: a rich enough correlation structure and marginal overdispersion." There are several methods of constructing discrete bivariate (and multivariate) distributions. An exhaustive account is given in [4]. Most of multivariate models for count data take origin from the multivariate Poisson model. This model allows for positive correlations only and for equi-dispersed margins (remember that for a Poisson r.v. expected value and variance have the same value, coinciding with the parameter λ). Extensions of this model are based on mixtures, which are able to (partially) overcome both problems. Apart from these models, other proposals have been suggested, often limited to the bivariate case since a generalization ot higher dimension is not immediate. These models usually have identical marginal distributions or marginal distributions of the same type and the dependence structure is somewhat limited.

Here we will focus on a method which is particularly popular and flexible when used for modelling multivariate continuous distributions, the copula

approach, whose use has been recently adapted to the discrete case. We will highlight the still great convenience of this approach as well as differences and limitations with respect to the original context. Copulas offer a great flexibility both in the dependence structure they can induce and in the choice of marginal distributions. [5] provides a list of desiderata properties of multivariate copula families for modeling multivariate discrete data.

The chapter is structured as follows. In the next section we introduce the concept of copula and explain how it can be exploited for constructing multivariate distributions with arbitrary margins and flexible dependence structures. Section 3 discusses some issues related to Pearson's correlation coefficient, often used as a measure of dependence, which is incorrect, especially when moving far from the multivariate normal distribution, and then in particular when modelling correlated counts. Alternative dependence measures are recalled. Section 4 briefly outlines estimation for copula models, highlighting potential pitfalls one has to take into consideration when handling dependent discrete distributions. Section 5 briefly mentions how copula-based models can include covariates. Section 6 presents two simple numerical examples of modelling of bivariate discrete counts, pointing out from a practical perspective what has been theoretically illustrated in the previous sections. Section 7 illustrates an example of fitting several bivariate discrete models to a real dataset. The last section concludes the chapter with a summary and some brief remarks.

2. Copulas for modelling multivariate distributions

The theory of copulas, whose origins date back to the first half of the last century, has received great and renewed interest in the last two decades especially due to the number of applications in the fields of quantitative finance and actuarial mathematics. The copula approach allows to separate the study of the marginal models from the study of the dependence models and then combine more marginal distributions with a variety of possible dependence structures. About the choice of univariate discrete distributions for modelling count data,

we address the reader to [6].

Let us start with the mathematical definition of copula and with a theorem that formally establishes its relationship with multivariate distributions.

2.1. Definition of copula and Sklar's theorem

A d-dimensional copula is a joint cumulative distribution function (c.d.f.) in $[0,1]^d$ with standard uniform c.d.f.s U_j, $j = 1 \ldots, d$:

$$C(u_1, \ldots, u_d) := P(U_1 \leq u_1, \ldots, U_d \leq u_d). \tag{1}$$

The importance of copulas in the study of multivariate distribution functions is summarized by the theorem of Sklar [7, 8], which here we briefly reprise.

Let F be a joint distribution function with margins $F_1, \ldots, F_d$. Then there exists a copula $C : [0,1]^d \to [0,1]$ such that, for all $x_1, \ldots, x_d$ in $\bar{\mathbb{R}} = [-\infty, +\infty]$,

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{2}$$

If the margins are continuous, then C is unique, otherwise C is uniquely determined on $\mathrm{Ran}(F_1) \times \cdots \times \mathrm{Ran}(F_d)$, with $\mathrm{Ran}(F_j)$ denoting the range of F_j. Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate c.d.f.s, then the function F defined in (2) is a joint distribution function with margins $F_1, \ldots, F_d$. If the margins $F_1, \ldots, F_d$ are continuous, the unique copula C is given by

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)). \tag{3}$$

where F_j^{-1} denotes the generalized inverse of the marginal c.d.f. F_j, i.e., $F_j^{-1}(t) = \inf\{x \in \mathbb{R} : F_j(x) \geq t\}$. Formulas (2) and (3) are fundamental in dealing with copulas. The former shows how joint distributions F are formed by connecting together marginal distributions with copulas C; the latter shows how copulas are extracted from a multivariate c.d.f. with continuous margins. Sklar's Theorem also suggests that, in the case of continuous margins, it is natural to define the notion of the copula of a c.d.f. F with continuous marginal c.d.f.s $F_1, \ldots, F_d$ as the c.d.f. C of $(F_1(X_1), \ldots, F_d(X_d))$.

2.2. Copula bounds and fundamental copulas

We recall that for any copula C the following constraint holds for any $(u_1, \ldots, u_d) \in [0,1]^d$:

$$\max(0, \sum_{j=1}^{d} u_j + 1 - d) \leq C(u_1, u_2, \ldots, u_d) \leq \min(u_1, u_2, \ldots, u_d); \qquad (4)$$

the left and right members of the inequality are called Fréchet lower bound and Fréchet upper bound, respectively [9]. We note that $M(u_1, u_2, \ldots, u_d) = \min(u_1, u_2, \ldots, u_d)$ is itself a copula, named "comonotonicity copula"; $W(u_1, \ldots, u_d) = \max(0, \sum_{j=1}^{d} u_j + 1 - d)$ is a copula only if $d = 2$, the bidimensional "countermonotonicity copula". In the bivariate case, the comonotonicity copula is the copula associated to two perfectly positively dependent r.v.s X_1 and X_2, satisfying the relationship $X_2 = f(X_1)$ almost surely, with f being a strictly increasing function. The countermonotonicity copula is the copula associated to two perfectly negatively dependent r.v.s X_1 and X_2, satisfying the relationship $X_2 = g(X_1)$ almost surely, with g being a strictly decreasing function. Although Fréchet bounds have been given for a copula, they may be given for any multivariate c.d.f. F with margins $F_1, \ldots, F_d$:

$$\max\left\{\sum_{j=1}^{d} F_j(x_j) + 1 - d, 0\right\} \leq F(x_1 \ldots, x_d) \leq \min\left\{F_1(x_1), \ldots, F_d(x_d)\right\};$$

so we have now bounds for F in terms of its own marginal distributions.

The comonotonicity and bivariate countermonotonicity copula are two examples of "fundamental" copulas. Another fundamental copula is the "independence copula", $\Pi(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j$: recalling (2), we can easily see that for a continuous joint c.d.f. F, the univariate margins are independent if and only if the copula of F is Π (in fact, the joint c.d.f. factorizes into the product of its marginal c.d.f.s).

2.3. Examples of implicit and explicit copulas

We provide now two example of parametric copula families, which we will use in the following sections.

2.3.1. Gaussian copula

If $\boldsymbol{Y} \sim N_d(\boldsymbol{\mu}, \Sigma)$ is a d-dimensional Gaussian r.v., with mean vector $\boldsymbol{\mu}$ and correlation matrix Σ, then its copula is a so-called Gaussian copula. Note that the copula of $\boldsymbol{Y}$ is exactly the same as the copula of $\boldsymbol{X} \sim N_d(\boldsymbol{0}, P)$, where P is the correlation matrix of $\boldsymbol{Y}$. By definition

$$C_P^{Ga}(u_1, \ldots, u_d) = P(\Phi(X_1) \leq u_1, \ldots, \Phi(X_d) \leq u_d) = \Phi_P(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)),$$

where Φ is the standard univariate normal c.d.f., $\Phi(z) = \int_{-\infty}^{z} 1/\sqrt{2\pi} e^{-t^2/2} \mathrm{d}t$, and Φ_P denotes the joint c.d.f. of $\boldsymbol{X}$. The Gaussian copula does not have a simple closed form (it belongs to the so-called class of "implicit copulas"), but can be expressed as an integral over the density of $\boldsymbol{X}$; in two dimensions for $\rho \neq 1$ we have that:

$$C_P^{Ga}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{2(1-\rho^2)}} \mathrm{d}s_1 \mathrm{d}s_2. \quad (5)$$

Both independence and comonotonicity copulas are special cases of the Gaussian copula. If $P = I_d$, we obtain the independence copula; if $P = J_d$, the $d \times d$ matrix consisting entirely of ones, then we obtain the comonotonicity copula. Also, for $d = 2$ and $\rho = -1$ the Gaussian copula is equal to the countermonotonicity copula. Thus in two dimensions the Gaussian copula can be thought of as a dependence structure that interpolates between perfect positive ($\rho = 1$) and negative ($\rho = -1$) dependence, where the parameter ρ represents the strength and direction of dependence.

2.3.2. Frank copula

The one-parameter bivariate Frank copula is defined as

$$C(u_1, u_2) = -\frac{1}{\kappa} \ln \left[1 + \frac{(e^{-\kappa u_1} - 1)(e^{-\kappa u_2} - 1)}{e^{-\kappa} - 1} \right],$$

with $\kappa \neq 0$. The Frank copula, which belongs to the wider class of Archimedean copulas, through the choice of its parameter κ, allows both negative ($\kappa < 0$) and positive ($\kappa > 0$) dependence. For $\kappa \to 0$, we have that the Frank copula reduces to the independence copula; for $\kappa \to \infty$, the Frank copula tends to

the upper Fréchet bound; for $\kappa \to -\infty$, the Frank copula tends to the lower Fréchet bound. This means that the Frank copula, like the Gaussian copula, interpolates between perfect positive and perfect negative dependence.

2.4. Derivation of the p.m.f. for a copula-based model with discrete margins

While in the continuous case the derivation of the joint p.d.f. of $(X_1, \ldots, X_d)$ is obtained easily through the use of partial derivatives of the joint c.d.f, for the discrete case, the expression of the joint p.m.f. for the multivariate r.v. with c.d.f. (2) can be derived by using finite differences [10]. Let $\boldsymbol{c} = (c_1 \ldots, c_d)$ be vertices where each c_j is equal to either x_j or $x_j - 1$, $j = 1, \ldots, d$. Then the joint p.m.f. of the d-variate discrete r.v. $(X_1, \ldots, X_d)$ is given by

$$p(x_1, \ldots, x_d) = \sum \operatorname{sgn}(\boldsymbol{c}) C(F_1(c_1), \ldots, F_d(c_d)), \tag{6}$$

where the sum is taken over all vertices $\boldsymbol{c}$ and $\operatorname{sgn}(\boldsymbol{c})$ is given by:

$$\operatorname{sgn}(\boldsymbol{c}) = \begin{cases} 1 & \text{if } c_j = x_j \text{ for an even number of j}'s \\ -1 & \text{if } c_j = x_j \text{ for an odd number of j}'s. \end{cases} \tag{7}$$

For the simplest case ($d = 2$), for a given pair (x_1, x_2) of non-negative integers, we have $2^2 = 4$ possible vertices $\boldsymbol{c}$ and (6) becomes

$$\begin{aligned} p(x_1, x_2) = {} & C(F_1(x_1), F_2(x_2)) + C(F_1(x_1 - 1), F_2(x_2 - 1)) \\ & - C(F_1(x_1), F_2(x_2 - 1)) - C(F_1(x_1 - 1), F_2(x_2)). \end{aligned}$$

From (6) it is clear that one has to evaluate the copula repeatedly for calculating the joint p.m.f. This means that an analytical closed form for the joint c.d.f. would make the computation of the p.m.f. easier and the copula model for count data more usable. Multivariate elliptical copulas (comprising the Gaussian copula) though providing a flexible structure, allowing for both positive and negative linear correlations, do not have a closed form and then for $d > 2$ the computation of p.m.f. requires repeated multivariate numerical integration; for this reason, their use is not straightforward.

2.5. Simulating copulas

Based on Sklar's theorem, if $(X_1, \ldots, X_d)$ has copula C and marginal c.d.f.s $F_1, \ldots, F_d$, then one may simulate from $(X_1, \ldots, X_d)$ in the following way:

1. Simulate a random vector $(U_1, \ldots, U_d)$ with c.d.f. C;
2. Return the random vector $(X_1, \ldots, X_d) := (F^{-1}(U_1), \ldots, F_d^{-1}(U_d))$.

Now, the problem arises on how to simulate from C. Let us first focus on the case $d = 2$. If a bivariate copula $C(u_1, u_2)$ has a sufficiently simple algebraic expression (this occurs for typical bivariate copulas, especially absolutely continuous copulas), then there exists a very efficient, analytical simulation algorithm which can often be applied. It goes under the name of "conditional sampling method" [see, for example, 11, p.78-79]. This method is based on the quantity

$$C_{u_2|u_1}(U_2|u_1) := P(U_2 \leq u_2|U_1 = u_1) = \frac{\partial C(u_1, u_2)}{\partial u_1}, u_2 \in [0, 1],$$

which exists for almost every $u_1 \in (0, 1)$. $C_{u_2|u_1}(U_2|u_1)$ is the c.d.f. of U_2 conditioned on the event that $U_1 = u_1$.

If we want to simulate r.v.s with copula C, we can implement the following steps:

- Simulate independently $U_1 \sim U(0, 1)$ and $V \sim U(0, 1)$
- Compute $U_2 = C^{-1}_{u_2|u_1}(V|u_1)$
- (U_1, U_2) is a random sample from (U_1, U_2) with copula C

If we consider the Frank copula presented in Section 2.3.2, one can easily derive the conditional c.d.f.

$$C_{U_2|u_1}(u_2|u_1) = \frac{\exp(-\kappa(u_1 + u_2)) - \exp(-\kappa u_1)}{\exp(-\kappa(u_1 + u_2)) - \exp(-\kappa u_1) - \exp(-\kappa u_2) + \exp(-\kappa)},$$

which can be inverted analytically:

$$C^{-1}_{U_2|u_1}(v|u_1) = -\frac{1}{\kappa} \log \left[1 - \frac{1 - \exp(-\kappa)}{1 + (v^{-1} - 1) \exp(-\kappa u_1)} \right].$$

The major shortcoming of the conditional sampling method is that it is very difficult to generalize to larger dimensions $d \geq 3$.

In practice, the simulation of r.v. for large d is usually only possible along a sufficiently easy stochastic model, which can be exploited in order to derive a simulation algorithm, e.g., Gaussian copula or more generally elliptical copulas; or Archimedean copulas [11, p.81].

Simulating a Gaussian copula C_P^{Ga} with correlation matrix P can be carried out following these steps:

1. Simulate $\boldsymbol{Z} = (Z_1, \ldots, Z_d) \sim N(\mathbf{0}, P)$, a multivariate normal distribution with mean vector $\mathbf{0}$ and correlation matrix P;
2. Return $\boldsymbol{U} = (\Phi(Z_1), \ldots \Phi(Z_d))$. The r.v. $\boldsymbol{U}$ has copula C_P^{Ga}.

The implementation is very straightforward, due to the large availability of software carrying out the simulation of the multivariate normal (step 1 of the above algorithm). Actually, one of the first attempts to simulate correlated discrete data is due to [12], who employed the Gaussian copula (without explicitly referring to copula theory) to link discrete variables together with assigned marginal distributions and correlation matrix. The simulation algorithm is named NORTA (NORmal To Anything), which gives the idea that starting from a multivariate normal distribution we can construct any other multivariate distribution with arbitrary margins (sharing the Gaussian copula). Indeed, there the focus was on how to set the correlation matrix P in order to ensure the desired correlation matrix $P_{\boldsymbol{X}}$ on the non-normal multivariate distribution. This attempt was reprised and developed later by other authors [see e.g. 13, 14, 15, 16, 17].

3. Copulas and measures of dependence

Often the researcher is interested in synthesizing the statistical dependence between two r.v.s through a scalar "measure of dependence". Here we focus on two kinds of dependence measures: the usual Pearson linear correlation and rank correlations.

3.1. Linear correlation

Pearson's linear correlation coefficient is by far the most popular measure of dependence between r.v.s. For a pair of r.v.s (X_1, X_2) it is defined as

$$\rho_{X_1X_2} = \frac{\mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right]}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}}$$

as far as X_1 and X_2 have both finite variance; it takes values in $[-1, 1]$ and actually measures the strength and direction of the linear relationship between X_1 and X_2. It is the canonical dependence measure in the world of multivariate normal distributions, and more generally for spherical and elliptical distributions; more precisely, the linear correlation (or the correlation matrix, for $d \geq 3$) contains all the information about the dependence structure of the multivariate model. However, empirical research in many applied sciences shows that the distributions of the real world are seldom in this class, and thus Pearson's ρ is not suitable to capture dependence [18]. This is one of its main drawbacks, along with others that [8] denotes as "fallacies", which are often underrated.

A first fallacy is that given two marginal distributions F_1 and F_2 and a correlation value $\rho \in [-1, +1]$, it is not always possible to construct a joint distribution F with margins F_1 and F_2, whose correlation is equal to the assigned ρ. To better explain it, let us introduce the concept of equality in type for random variables. Two r.v.s V and W (or their distributions) are said to be of the same type if there exist constants $a > 0$ and $b \in \mathbb{R}$ such that $V \stackrel{d}{=} aW + b$, with "$\stackrel{d}{=}$" denoting equality in distribution. In other words, distributions of the same type are obtained from one another by location and scale transformations. Then, we can state the following result, concerning "attainable correlations". Let (X_1, X_2) be a random vector with finite-variance marginal cdfs F_1 and F_2 and an unspecified joint cdf; assume also that $\mathrm{var}(X_1) > 0$ and $\mathrm{var}(X_2) > 0$. The following statements hold [8, pp.204-205].

1. The attainable correlations form a closed interval $[\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} < 0 < \rho_{\max}$.
2. The minimum correlation $\rho = \rho_{\min}$ is attained if and only if X_1 and X_2

are countermonotonic. The maximum correlation $\rho = \rho_{\max}$ is attained if and only if X_1 and X_2 are comonotonic.

3. $\rho_{\min} = -1$ if and only if X_1 and $-X_2$ are of the same type, and $\rho_{\max} = 1$ if and only if X_1 and X_2 are of the same type.

Let focus for a while on continuous r.v.s. Clearly, the Gaussian family of distributions with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$ is an example of class distributions of the same type, since we know that if $X \sim N(\mu_x, \sigma_x)$, then any linear transformation $Y = a + bX$, with $a \neq 0, b \in \mathbb{R}$, is still normal. Then, thanks to the above statements, it is possible to join together two any normal distributions into a joint distribution with an assigned correlation $-1 \leq \rho \leq +1$. This result was quite predictable, as we know the bivariate normal distribution satisfies these conditions. Another less trivial example: let X_1 and X_2 be exponential r.v.s with parameters λ_1 and λ_2, respectively. Are the minimum and maximum attainable correlation values equal to -1 and $+1$? We have that $\rho_{\max} = +1$ since X_2 is equal in distribution to $\lambda_1 X_1/\lambda_2$. In fact, if the d.f. of X_1 is $f_1(x) = \lambda_1 \exp(-\lambda_1 x)$, the d.f. of the transformation $\lambda_1 X_1/\lambda_2$ is $f_2(x) = \lambda_2 \exp(-\lambda_2 x)$. In other terms, the exponential r.v.s "scale". However, $-1 < \rho_{\min} < 0$, since it cannot be $X_1 \stackrel{d}{=} -aX_2 + b$ for any $a > 0$ and b (both X_1 and X_2 are defined on $\mathbb{R}^+$). It can be shown that $\rho_{\min} = 1 - \pi^2/6 \approx= -0.645$, for any choice of λ_1, λ_2.

For non-negative count r.v.s, it is then clear that the maximum correlation is $+1$ if and only if X_1 and X_2 are identically distributed; whereas the minimum correlation can never be -1. Determining analytically $\rho_{\min}$ (and $\rho_{\max}$, in case of non-identical margins) for a pair of discrete r.v.s is a challenging task; for the geometric distribution, one can refer to [19], where it is shown that for two identical geometric margins with parameter θ, $\rho_{\min}$ is equal to $\theta - 1$ if $\theta \geq 1/2$, whereas, if $\theta < 1/2$, a numerical procedure is sketched in order to recover $\rho_{\min}$.

If dealing with two discrete r.v.s with finite support, the values $\rho_{\min}$ and $\rho_{\max}$ can be computed by building the cograduation and countergraduation tables [see 16, for an example of calculation]. An empirical and straightforward method

(X_1, X_2)	Geo(1/4)	Pois(4)	NegBin(10,2/3)
Geo(1/2)	$[-0.549, 0.978]$	$[-0.772, 0.891]$	$[-0.743, 0.919]$
Pois(2)		$[-0.912, 0.965]$	$[-0.907, 0.968]$
NegBin(5,1/2)			$[-0.909, 0.989]$

Table 1: Correlation range $[\rho_{\min}, \rho_{\max}]$ for six possible combinations of discrete margins.

that can be employed in order to numerically derive the extremal correlations $\rho_{\min}$ and $\rho_{\max}$ between two (discrete, continuous or mixed-type) r.v.s is the following [20]:

1. Generate two random samples from the two univariate distributions independently, using a large number of observations (e.g., $n = 100,000$).
2. Sort the two samples in the same direction, and compute the sample correlation, which corresponds to the upper bound $\rho_{\max}$.
3. Sort the two samples in opposite directions (i.e., in ascending order for one of the variables, and in descending order for the other). Then, compute the sample correlation, which corresponds to the lower bound $\rho_{\min}$.

The rationale of the above algorithm clearly relies on the result about attainable correlations, point 2). For illustrative purpose, we used the algorithm above, by setting $n = 10^7$, in order to compute the range $[\rho_{\min}, \rho_{\max}]$ for six possible combinations of discrete margins. We considered three common families of count distributions, namely, geometric, Poisson and negative binomial (a generalization of the geometric). The results are reported in Table 3.1. Note that the extremal correlations can be quite far from the corresponding limits -1 and $+1$ of the correlation coefficient. Actually, as can be easily observed considering the combination of two negative binomials, as the discrete distributions tend to resemble a continuous one (i.e., as most of the probability mass tends to spread over a huge number of integers and not to concentrate on a few values), then the extremal correlations tend to their corresponding limits ± 1.

Another fallacy of the linear correlation can be outlined as follows. Given two margins F_1 and F_2 and a feasible linear correlation ρ (where by feasible we mean

comprised between the $\rho_{\min}$ and $\rho_{\max}$ discussed above), the joint distribution F having margins F_1 and F_2 and correlation ρ is not unique. In other terms, the marginal distributions and pairwise correlations of a r.v. do not univocally determine its joint distribution. [8] presented an example, where two normal distributions can be linked together forming two different joint distributions with the same correlation, by simply selecting two different copulas (a Gaussian copula and a linear combination of Fréchet-bound copulas). We will provide an analogous example related to the discrete case in Section 7.

3.2. Rank correlations

The first fallacy of Pearson's correlation can be overcome, when handling continuous distributions only, by employing two other dependence measures, namely, Spearman's rho [21] and Kendall's tau [22], also known as "rank correlations".

Spearman's rho between two r.v.s X_1 and X_2 with marginal c.d.f.s F_1 and F_2, respectively, is defined as

$$\rho^S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2));$$

whereas the definition of Kendall's tau is

$$\rho_\tau(X_1, X_2) = \mathbb{E}(\text{sgn}(X_1 - X_2)(\tilde{X}_1 - \tilde{X}_2)),$$

where sgn is the usual sign function, taking value $+1$ or -1 according whether its argument is positive or negative, and $(\tilde{X}_1, \tilde{X}_2)$ is an independent copy of (X_1, X_2). In higher dimensions, the Spearman's rho matrix for the general d-variate random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is given by $\rho^S(\boldsymbol{X}) = \rho(F_1(X_1), \ldots, F_d(X_d))$; the Kendall's tau matrix of $\boldsymbol{X}$ may be written as $\rho_\tau(\boldsymbol{X}) = \text{cov}(\text{sgn}(\boldsymbol{X} - \tilde{\boldsymbol{X}}))$, where $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$.

For a bivariate sample $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, with $\boldsymbol{x}_j = (x_{ij})^T$, $i = 1, \ldots, n$, $j = 1, 2$, the sample version of Spearman's rho is $\hat{\rho}_S(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho(\text{rank}(\boldsymbol{x}_1), \text{rank}(\boldsymbol{x}_2))$, where the function rank assigns a number from 1 to n corresponding to the position

of x_{1j} (or x_{2j}) in an ascending order. The sample version of Kendall's tau can be defined as

$$\hat{\rho}_\tau(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\binom{n}{2}},$$

where two points in $\mathbb{R}^2$, denoted by (x_{1t}, x_{2t}) and (x_{1u}, x_{2u}), are said to be concordant if $(x_{1t} - x_{1u})(x_{2t} - x_{2u}) > 0$ and to be discordant if $(x_{1t} - x_{1u})(x_{2t} - x_{2u}) < 0$.

Differently from Pearson's ρ, these two measures i) are both able to capture monotone dependence between two random variables; ii) in the continuous case, their value depend only on the copula C and not on the margins (i.e. they are "margin-free"); iii) they take the value 1 when the margins are comonotonic and the value -1 when they are countermonotonic. These three properties do not keep holding when we move to the discrete case [23]. Indeed, with discrete variables, the definition itself of these two measures leads to ambiguity, due to the stepwise nature of c.d.f. and then the presence in the sample data of the so called "ties" (observations having the same value). Depending on the choice of margins, the two rank correlations may or may not span the entire interval $[-1, 1]$ [see e.g. 24, chapter 5]. Several rescaled versions of Kendall's τ and Spearman's ρ were introduced in an attempt to correct it, but none of them is margin-free, however, and some of them do not reach the bounds ± 1 [23].

4. Inference for copula models

Let consider a multivariate copula-based discrete model whose joint c.d.f. has the following representation:

$$F(x_1, \ldots, x_d; \theta_1, \ldots, \theta_d, \theta) = C(F_1(x_1; \theta_1), \ldots, F_d(x_d; \theta_d); \theta), \tag{8}$$

where F_j is the marginal c.d.f. corresponding to the j-th margin, characterized by marginal parameter θ_j, $j = 1, \ldots, d$, and θ is the copula parameter (θ and the θ_j can be scalar or vectors). Henceforth, we suppose that the functions F_j and C are known, except for the values of their parameters: this means that we

move within a parametric framework; for non-parametric and semi-parametric cases, we address the reader to [25, pp.247-251]. The p.m.f. p of $(X_1, \ldots, X_d)$ can be derived recalling (6). Now, if a d-variate random sample of size n, $[x_{ij}], i = 1, \ldots, n, j = 1, \ldots, d$, is available, the log-likelihood functions for the univariate margins are

$$\ell_j(\theta_j) = \sum_{i=1}^{n} \log p_j(x_{ij}; \theta_j), \quad j = 1, \ldots, d \tag{9}$$

where p_j is the marginal p.m.f. of X_j; and the total log-likelihood function can be written as

$$\ell(\theta, \theta_1, \ldots, \theta_d) = \sum_{i=1}^{n} \log p(x_{i1}, \ldots, x_{id}; \theta_1, \ldots, \theta_d, \theta). \tag{10}$$

Parameter estimates can be simultaneously recovered maximizing the log-likelihood function (10), thus implementing a full maximum likelihood estimation (MLE). Such a maximization can be usually solved only numerically. Alternatively, efficient estimation of the model parameters is succeeded by the inference function of margins (IFM) method, which consists of a two-step approach. At the first step, the univariate log-likelihoods (9) are maximized independently of the copula parameter; at the second step, the joint log-likelihood (10) is maximized over θ with the values of univariate parameters θ_j fixed as estimated at the first step. The parameter estimation is thus decomposed into two smaller problems: fitting the marginal distributions (as if they were independent) and then fitting the existing dependence structure. From a computational cost perspective, estimation by IFM method becomes more advantageous than full MLE as the dimension d increases. Asymptotic efficiency of the IFM estimator has been studied by Joe (2005) for a number of multivariate models and is overall shown to be highly efficient compared to standard maximum likelihood, except for extreme cases near the Fréchet bounds.

Another method uses empirical estimates of either Spearman's or Kendall's rank correlation to infer an estimate for the copula parameter. One needs a theoretical relationship between one of the rank correlations and the copula

parameter and substitute empirical values of the rank correlation into this relationship to get estimates of the copula parameter. If, for example, for a certain (bivariate) parametric copula family we have that $\rho_\tau = T(\theta)$, for some function T, and T is invertible, so that $\theta = T^{-1}(\rho_\tau)$, then an estimate for θ is $T^{-1}(\hat{\rho}_\tau)$, with $\hat{\rho}_\tau$ being the sample version of ρ_τ. This method is usually referred to as method of moments.

5. Issues with copulas for discrete data

The copula concept is slightly less natural for multivariate discrete distributions. As seen when discussing Sklar's theorem, this is because there is more than one copula that can be used to join the margins to form the joint c.d.f., as the following example shows [8].

Example (Copulas of bivariate Bernoulli)**.** *Let* (X_1, X_2) *have a bivariate Bernoulli distribution satisfying*

$$P(X_1 = 0, X_2 = 0) = 1/8, \qquad P(X_1 = 1, X_2 = 1) = 3/8$$
$$P(X_1 = 0, X_2 = 1) = 2/8, \qquad P(X_1 = 1, X_2 = 0) = 2/8$$

Clearly, $P(X_1 = 0) = P(X_2 = 0) = 3/8$ *and the marginal distributions* F_1 *and* F_2 *of* X_1 *and* X_2 *are the same. From Sklar's theorem – see Eq.* (2) *– we know that* $P(X_1 \leq x_1, X_2 \leq x_2) = C(P(X_1 \leq x_1), P(X_2 \leq x_2))$ *for all* x_1, x_2 *and some copula* C*. Since* $Ran(F_1) = Ran(F_2) = \{0, 3/8, 1\}$*, clearly the only constraint on* C *is that* $C(3/8, 3/8) = 1/8$*. Any copula fulfilling this constraint is a copula of* (X_1, X_2)*, and there are infinitely many such copulas.*

This fact does not cause any issues in the modelling/simulation step, since Formula (2) always returns a valid joint c.d.f. even if some of the F_j are discrete; some problems arise at the estimation stage, when one has to make inference about the parameter of the copula that is assumed to link the margins. The fact that from two or more correlated discrete r.v.s the copula that can be extracted is not unique (indeed, there are infinite copulas that can be extracted)

naturally poses an unidentifiability problem when one needs to estimate the dependence structure. This issue leads to some important consequences; from a a practical point of view, the most important is that inference for the dependence parameter θ under a parametric copula model should not resort to rank-based approach, but rather to maximum likelihood estimation. Resuming and citing [23], "copula models provide a viable approach to the construction of multivariate distributions with given margins, even in the discrete case", and "When dealing with count data, however, modeling and interpreting dependence through copulas is subject to caution. Furthermore, inference (and particularly rank-based inference) for copula parameters from discrete data is fraught with difficulties". For a detailed discussion of this topic, which is still debated among the statistical community, we address the reader to the thorough work of Genest [23] and to the recent paper by Faugeras [26], where the author seems much more reluctant than [23] to extend copulas from the continuous to the discrete case.

6. Regression models

Copula-based models are susceptible to the introduction of explanatory variables. Model parameters can be regressed towards different sets of covariates, increasing the goodness-of-fit of the model to the data and at the same time its complexity [27, 28]. Covariates are commonly used for the marginal parameters; however, since the interest of the researcher often lies on studying the dependence structure rather than the marginal properties and copula measures the association between marginal distributions, the use of covariates in its parameters, allowing for direct modeling of association, has been recently increasing. Following the approach of [10], if we consider a bivariate copula-based parametric model for count r.v.s Y_1 and Y_2, their joint c.d.f. F, by slightly adapting (8), can be written as:

$$F(y_1, y_2; \theta_1, \theta_2, \theta) = C(F_1(y_1; \theta_1), F_2(y_2; \theta_2), \theta)$$

Suppose the data are (y_{ij}, x_{ij}), $i = 1, \ldots, n$, $j = 1, 2$, with x_{ij} a vector of covariates for the i-th observation associated to the j-th random component. One can easily introduce covariates x_{ij} on the copula-based parametric model by assuming that the j-th margin is $y_{ij} \sim F_j(;\theta_{ij})$, with $\theta_{ij} = (\mu_{ij} = g(\beta_j^T x_{ij}), \gamma_j)$ where μ_{ij} denotes the mean parametrized by a suitable link function g to accommodate the covariates, β_j the vector of the regression coefficients, and γ_j the vector of marginal parameters not depending on covariates. Furthermore, one can introduce a regression part for the copula parameter θ, which can be specified through an appropriate covariate function on θ, $s(\theta_i) = b^T x_i$, with parameter vector b. Possible covariate functions for the copula parameter of several copula families are reported in [10, Table 1]. For parameter estimation, one can use the standard MLE method, maximizing the log-likelihood function, which can be written as

$$L(\beta_1, \gamma_1, \beta_2, \gamma_2, b) = \sum_{i=1}^{n} \log p(y_{i1}, y_{i2}; \beta_1, \gamma_1, \beta_2, \gamma_2, b),$$

where p is the joint p.m.f. (6).

7. Numerical example

In this example, focusing on the simple bivariate case, we show how to construct different bivariate models with the same margins X_1 and X_2 but different dependence structure, i.e. how can connect the same margins through different (families of) copulas. Let consider two univariate r.v.s X_1, which we assume to be binomial, and X_2, which we assume to follow the Poisson law. The p.m.f. of X_1 is

$$p_1(x_1; n, p) = \binom{n}{x_1} \theta^{x_1} (1-\theta)^{n-x_1}, x_1 = 0, 1, \ldots, n; n \in \mathbb{N}, 0 < \theta < 1;$$

the p.m.f. of X_2 is

$$p_2(x_2; \lambda) = \frac{\lambda^{x_2} e^{-\lambda}}{x_2!}, x_2 = 0, 1, 2, \ldots, \lambda > 0.$$

We set the marginal parameters as follows: $n = 4$, $\theta = 0.5$, $\lambda = 2$. As copulas linking the two margins we consider the Gaussian and the Frank copula.

(x_1, x_2)	0	1	2	3	≥ 4	tot
0	0.05595	0.00649	0.00007	0.00000	0.00000	0.0625
1	0.07427	0.14545	0.02916	0.00111	0.00001	0.25
2	0.00512	0.11478	0.18662	0.06237	0.00611	0.375
3	0.00001	0.00394	0.05453	0.11146	0.08006	0.25
4	0.00000	0.00000	0.00029	0.00551	0.05670	0.0625
tot	0.13534	0.27067	0.2707	0.18045	0.14288	1

Table 2: Joint p.m.f. of the r.v. (X_1, X_2) with binomial and Poisson margins connected by Gaussian copula with $\rho = 0.9$.

We describe how to recover the p.m.f. and compute Pearson's correlation and compare the results under both dependence structures, varying the dependence parameter (ρ for the former copula, κ for the latter).

If X_1 and X_2 are connected through the Gaussian copula with parameter ρ, Eq.(5), then the probability values of the joint p.m.f. can be computed numerically, as double integrals over rectangles of the joint normal p.d.f. For example, computing the probability $p(0,0)$ according to Eq. (6) is equivalent to compute the following:

$$p(0,0) = F(0,0) = C(F_1(0), F_2(0); \rho) = \int_{-\infty}^{\Phi^{-1}(F_1(0))} \int_{-\infty}^{\Phi^{-1}(F_2(0))} \phi_2(s,t;\rho)\mathrm{d}s\mathrm{d}t \tag{11}$$

For our example, $F_1(0) = 0.0625$ and then $\Phi^{-1}(0.0625) = -1.534121$; $F_2(0) = 0.1353353$ and then $\Phi^{-1}(0.1353) = -1.10152$. If we set the dependence parameter ρ equal to 0.9, the integral on the right side of (11) is equal to 0.05594503. The other probability values can be calculated the same way. Computation of the double integral can be performed through the function `pmvnorm` comprised in the `R` package `mvtnorm`. Table 2 displays (part of) the joint p.m.f. when $\rho = 0.9$ (we truncated the values of the Poisson margin at $x_2 = 4$).

For the bivariate discrete r.v. (X_1, X_2) it is possible to compute Pearson's correlation $\rho_{X_1X_2}$, which we expect to be different from the correlation of the Gaussian copula ρ. In order to do that, we need to compute the mixed moment

$\mathbb{E}(X_1X_2)$, preferably employing a joint p.m.f. computed over a larger support grid than that of Table 2. In order to tackle the infinite support of the Poisson variate and numerically compute the value of the mixed moment $\mathbb{E}(X_1X_2)$ with a small error margin, we can follow the expedient suggested in [29, 30] and (temporarily) truncate it to a very high quantile; in this case taking a threshold equal to $x_{2\,\max} = 13$ is reasonable, since $F_2(x_{2\,\max}) \approx 1$. Then,

$$\mathbb{E}(X_1X_2) = \sum_{i=0}^{4}\sum_{j=0}^{\infty} i \cdot j \cdot p(i,j) \approx \sum_{i=0}^{4}\sum_{j=0}^{\infty} i \cdot j \cdot p^*(i,j)$$

where

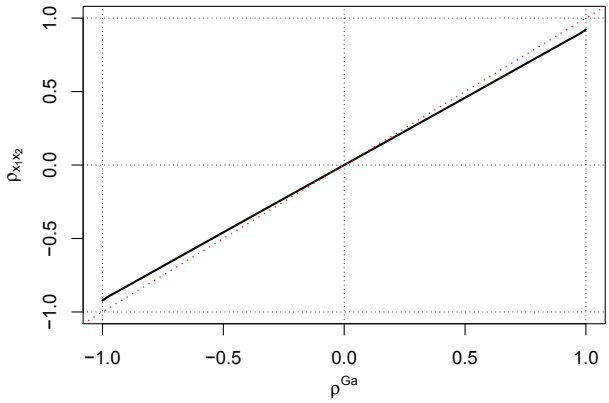
$$p^*(i,j) = \begin{cases} p(i,j) & \text{if } j < x_{2\,\max} \\ \sum_{j=x_{2\,\max}}^{\infty} p(i,j) & \text{if } j \geq x_{2\,\max} \end{cases}$$

is the joint p.m.f. of the truncated version of the bivariate discrete r.v.

In Figure 1, we graphically display the relationship between the correlation parameter ρ^{Ga} of the bivariate Gaussian copula, which takes all the values in $[-1, +1]$, and the correlation coefficient $\rho_{X_1X_2}$ between the two discrete margins $X_1 \sim \text{Binom}(n = 4, \theta = 0.5)$ and $X_2 \sim \text{Pois}(\lambda = 2)$. The function $\rho_{X_1X_2} = G(\rho^{Ga})$ is a strictly increasing function passing through the origin. From Figure 1, it is also evident how the minimum and maximum correlation between the two discrete margins connected by a Gaussian copula are not -1 and $+1$, but -0.92171 and $+0.92171$. This empirically confirms the result given in Section 3, since the binomial and Poisson are clearly not distributions of the same type. It is also evident that for $\rho^{Ga} = 0$, also $\rho_{x_1x_2} = 0$ (for $\rho = 0$, the Gaussian copula reduces to the independence copula) and that the relationship between $\rho_{x_1x_2}$ and ρ^{Ga} is almost linear over the entire interval $[-1, +1]$. More importantly, we have that $|\rho_{X_1X_2}| \leq |\rho|$, which was empirically noticed by [12] and is a more general result due to [31], reprised by [32, p.155]. For example, if we set $\rho^{Ga} = 0.9$, we have that $\rho_{X_1X_2} = 0.8259$.

We can develop an analogous exercise by using the Frank copula, parametrized by $\kappa \in \mathbb{R}$, which spans the entire dependence spectrum, like the Gaussian copula. As we did for the case of the Gaussian copula, we can reconstruct the joint

Figure 1: Relationship (represented by the solid line) between the correlation parameter ρ^{Ga} of the Gaussian copula and the correlation of the Binomial (with $n = 4, \theta = 0.5$) and the Poisson ($\lambda = 2$) margins connected by the same copula. The red dashed line indicates the I-III quadrant bisector.



p.m.f. of the r.v. (X_1, X_2) recalling again (6). For example, letting $\kappa = 14.21$, in order to compute $p(0,0)$ we calculate

$$p(0,0) = F(0,0) = C(F_1(0), F_2(0)) =$$
$$= -\frac{1}{14.21} \log\left[1 + \frac{(e^{-14.21 \cdot 0.00625} - 1)(e^{-14.21 \cdot 0.13534} - 1)}{e^{-14.21} - 1}\right] = 0.04914.$$

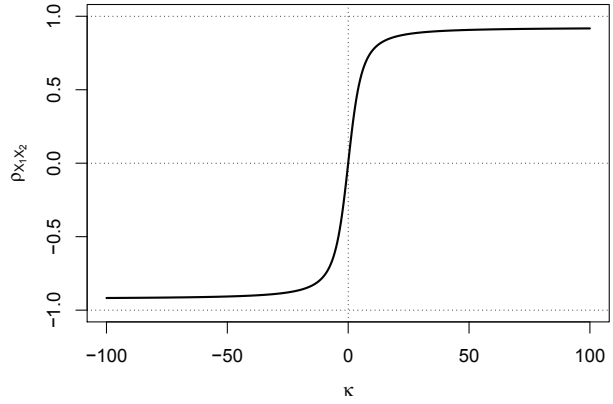
Table 3 displays the joint pmf for $\kappa = 14.21$. As done before, by building the p.m.f. for several values of κ, we can reconstruct the relationship between

(x_1, x_2)	0	1	2	3	≥ 4	tot
0	0.04914	0.01305	0.00031	0.00001	0.00000	0.0625
1	0.08151	0.15245	0.01566	0.00036	0.00003	0.125
2	0.00466	0.10394	0.21277	0.04837	0.00526	0.375
3	0.00002	0.00121	0.04094	0.12062	0.08720	0.125
4	0.00000	0.00002	0.00099	0.01110	0.05039	0.0625
tot	0.13534	0.27067	0.2707	0.18045	0.14288	1

Table 3: Joint p.m.f. of the r.v. (X_1, X_2) with binomial and Poisson margins connected by Frank copula with $\kappa = 14.21$.

the correlation coefficient $\rho_{X_1X_2}$ and the dependence parameter κ, which is displayed in Figure 2. Note that $\rho_{X_1X_2}$ is an increasing function of κ, passing through the origin (when the Frank copula coincides with the independence copula, the two margins are independent and then uncorrelated); when κ goes to $+\infty$ ($-\infty$), which corresponds to the comonotonicity (countermonotonicity) case, then ρ tends asymptotically but quite slowly to the limit $\rho_{\max}$ ($\rho_{\min}$). The value $\kappa = 14.21$ we employed before in the example of computation of $p(x_1, x_2)$ yields a correlation equal to 0.8259, the same induced by the Gaussian copula with $\rho^{Ga} = 0.9$.

Figure 2: Relationship (represented by the solid line) between the correlation parameter κ of the Frank copula and the correlation between the Binomial (with $(n = 4, \theta = 0.5)$) and the Poisson ($\lambda = 2$) margins connected by the same copula.



The comparison between Tables 2 and 3 shows that though sharing the same value of linear correlation, the two joint distributions are quite different (one can just consider the $(0, 0)$ probabilities). Through this example we showed that it is possible to construct two bivariate discrete distributions with the same choice of margins and the same value of (linear) correlation, but with a different joint distribution. This is a practical counterexample that indicates how Pearson's ρ is in general unable to characterize the dependence structure of a bivariate distribution (its second fallacy discussed in Section 3.1).

Table 4: Bivariate distribution of the data taken from [33]: number of flight aborts by 109 aircrafts in the first and second consecutive six months of a one-year period.

$x_1 \setminus x_2$	0	1	2	3	4	tot
0	34	20	4	6	4	68
1	17	7	0	0	0	24
2	6	4	1	0	0	11
3	0	4	0	0	0	4
4	0	0	0	0	0	0
5	2	0	0	0	0	2
tot	59	35	5	6	4	109

8. An application to real data

In this section, we fit several bivariate copula-based discrete models to a dataset taken from the literature. The data, considered in [33] (see Table 4), consist of the number of aborts by 109 aircrafts in two (first = x_1, second = x_2) consecutive 6 months of 1-year period. Summary statistics for the dataset are $\bar{x}_1 = 0.624$, $\bar{x}_2 = 0.725$, $\mathrm{var}(x_1) = 1.024$, $\mathrm{var}(x_2) = 1.062$. The sample correlation coefficient between x_1 and x_2 is $\hat{\rho}_{x_1 x_2} = -0.1609$, which denotes a negative dependence. In order to fit these data, we adopt the copula approach described in this chapter: we separate the modeling of the margins from the modeling of the dependence structure. As to the marginal distributions, it was shown that the geometric distribution could be a plausible model for both x_1 and x_2; then this can be a first choice. Alternatively, we can fall back on one-parameter discrete Lindley distribution [34], whose p.m.f. is $p(x;\theta) = \sum_{i=0}^{1}(-1)^i(1+\theta(x+1)/(1+\theta)e^{-\theta(x+i)})$, for $x = 0, 1, 2, \ldots$, with $\theta > 0$. We can add a further degree of complexity by considering the discrete Weibull distribution [35], which can be regarded as a generalization of the geometric distribution. The expression of the p.m.f. of the discrete Weibull distribution with parameters q and β is $p(x;q,\beta) = q^{x^\beta} - q^{(x+1)^\beta}$, for $x = 0, 1, 2, \ldots$, with $0 < q < 1$ and $\beta > 0$. If $\beta = 1$, the discrete Weibull distribution reduces to

a geometric distribution with parameter $\theta = 1 - q$. As for the modelling of the dependence structure, here we limit to consider the Gaussian and the Frank copula, which are able to handle negative dependence. We considered the bivariate models displayed in Table 5 (just a selection of all possible combinations), for which we computed the MLEs for all the parameters, the maximum value of the log-likelihood function ℓ, and the value of the AIC (Akaike Information Criterion) index, $2r - 2\ell$, where r is the number of model parameters. In terms of AIC, the best model among the eight considered here is the one linking two geometric margins through a Gaussian copula (first line of Table 5). The corresponding theoretical joint frequencies are displayed in Table 6. The second best model is the one linking two discrete Lindley distributions through the Gaussian copula (third row). The results indicate that complicating the model by fitting the two-parameter discrete Weibull distribution to the margins is not relatively convenient: as expected, the value of the log-likelihood function increases, but also the AIC increases; note that all the estimated β values, for both x_1 and x_2, are very close to 1, the value for which the discrete Weibull degenerates into a geometric distribution. We reaffirm that the models of Table 5 represent just a small part of the infinite copula-based models that can be devised by exploiting the construction principle expressed by Eq.(2).

Table 5: Synthetic results about the bivariate models fitted to the data taken from [33]. For each model (consisting of a parametric distribution for x_1, a parametric distribution for x_2 and a parametric copula for their dependence structure) the values of the parameter estimates, the log-likelihood function and AIC are computed. Legend: Geo=Geometric, DW=Discrete Weibull, DL=Discrete Lindley, Ga=Gaussian, F=Frank

x_1	p_1	θ_1	q_1	β_1	x_2	p_2	θ_2	q_2	β_2	copula	ρ	τ	ℓ	AIC
Geo	0.6132	-	-	-	Geo	0.5806	-	-	-	Ga	-0.2261	-	-244.4287	494.8574
Geo	0.6157	-	-	-	Geo	0.5769	-	-	-	F	-	-1.1158	-244.7584	495.5168
DL	-	1.3456	-	-	DL	-	-1.2453	-	-	Ga	-0.2140	-	-244.5413	495.0826
DL	-	1.3533	-	-	DL	-	1.2359	-	-	F	-	-1.0744	-244.9184	495.8368
DW	-	-	0.3763	0.9660	DW	-	-	0.4534	1.1370	Ga	-0.2353	-	-243.7156	497.4312
DW	-	-	0.3746	0.9695	DW	-	-	0.4551	1.1290	F	-	-1.1536	-244.1309	498.2618
Geo	0.6156	-	-	-	DW	-	-	0.4528	1.1365	Ga	-0.2302	-	-243.7573	495.5145
Geo	0.6182	-	-	-	DW	-	-	0.4548	1.1286	F	-	-1.1343	-244.1652	496.3303

Table 6: Theoretical joint frequencies for Mitchell & Paulson data [33] under the copula-based model with geometric margins and Gaussian copula. Cell borders highlight the cells' groupings.

(x_1, x_2)	0	1	2	3	≥ 4	tot
0	35.12	17.60	7.95	3.50	2.67	66.84
1	16.51	5.84	2.17	0.82	0.51	25.85
2	6.93	2.01	0.68	0.24	0.14	10.00
3	2.83	0.70	0.22	0.07	0.04	3.87
4	1.14	0.25	0.07	0.02	0.01	1.50
≥ 5	0.75	0.14	0.04	0.01	0.01	0.94
tot	63.28	26.54	11.13	4.67	3.37	109

The estimates of Table 5 were derived according to the full MLE; alternatively, one can implement the IFM method (see Section 4), which is computationally more convenient. For the "best" model, it works as follows: one first computes the MLEs of the parameters θ_1 and θ_2, of the two geometrically distributed margins, as if they were independent, which are equal to $\hat{\theta}_{1,IFM} = 1/(1+\bar{x}_1) = 0.6158$ and $\hat{\theta}_{2,IFM} = 1/(1+\bar{x}_2) = 0.5798$, and then one can maximize the log-likelihood with respect to the dependence parameter only, plugging in the two estimates above: $\hat{\rho}_{IFM} = -0.2251$. The three estimate derived through the IFM method are very close to the corresponding full MLEs.

We can compute the customary chi-square statistic as an absolute measure of goodness-of-fit on the "best" model. In order to do it, we have first to group cells and sum up the values of the theoretical contingency table of Table 6, in order to ensure a minimum frequency of 5 for each grouping. A possible grouping is there displayed. Then we calculate $\chi^2 = \sum_{g=1}^{G}(e_g - o_g)^2/e_g = 5.408$, where e_g and o_g are the expected and observed frequency of the g-th grouping, respectively, $g = 1, \ldots, G = 8$. The corresponding p-value of the chi-square test statistic, under the null hypothesis that the observed bivariate sample actually comes from the selected bivariate model, is 0.248, which attests the model fits the data adequately.

9. Conclusions

In this chapter we reviewed the statistical modelling of correlated count data via copulas. After presenting the basic notions of copulas, we discussed how copula-based models can be constructed for modelling multivariate correlated count data, how to recover the joint probability mass function, how to simulate pseudo-random samples, and how to make inference, cautioning about possible issues arising from the discrete nature of variables. A numerical example is presented that practically shows how to derive the joint probability function and the linear correlation for two bivariate models sharing the same margins but having different dependence structures. Finally, a dataset taken from the literature has been analysed and fitted using several copula-based distributions, whose goodness-of-fit has been compared through a customary index.

Rather than discussing the whole literature on the modelling of correlated count data, which would have been a very arduous task, since it is quite scattered in many and recent works, we preferred to highlight and linger on several connotative points, illustrating them through some numerical examples.

Although a very intuitive and flexible tool for building up multivariate discrete distributions, copulas are however problematic when the dimension increases: the joint probability function may be cumbersome to compute and pseudo-random simulation can be not straightforward. Moreover, the discrete nature of the data adds some questions related to estimation. Statistical research is facing these issues in order to spread and facilitate the use of copulas for building appropriate stochastic models for real data.

References

[1] S. Kocherlakota, K. Kocherlakota, Bivariate discrete distributions, Wiley Online Library, 1992.

[2] N. Johnson, S. Kotz, N. Balakrishnan, Discrete multivariate distributions, Wiley Online Library, 1997.

[3] J. M. Sarabia, E. Gómez-Déniz, Multivariate Poisson-Beta distributions with applications, Communications in Statistics - Theory and Methods 40(6) (2011) 1093–1108.

[4] C.-D. Lai, Constructions of discrete bivariate distributions, in: N. Balakrishnan, E. Castillo, J. A. Sarabia (Eds.), Advances in Distribution Theory, Order Statistics, and Inference, Springer, New York, 2006.

[5] A. K. Nikoloulopoulos, Copula-based models for multivariate discrete response data, in: Copulae in Mathematical and Quantitative Finance, Springer, 2013, pp. 231–249.

[6] A. K. Nikoloulopoulos, D. Karlis, On modeling count data: a comparison of some well-known discrete distributions, Journal of Statistical Computation and Simulation 78 (3) (2008) 437–457.

[7] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, Publications de l'Institut de Statistique de l'Université de Paris 8 (1959) 229–231.

[8] A. J. McNeil, R. Frey, P. Embrechts, Quantitative risk management: Concepts, techniques and tools, Princeton University Press, 2005.

[9] M. Fréchet, Sur les tableaux de corrélation dont les marges sont données, Annales de l'Université de Lyon. Section A: Sciences mathématiques et astronomie 3(14) (1951) 53–77.

[10] A. K. Nikoloulopoulos, D. Karlis, Modeling multivariate count data using copulas, Communications in Statistics - Simulation and Computation 39 (1) (2010) 172–187.

[11] J. Mai, M. Scherer, Financial engineering with copulas explained, Palgrave Macmillan, New York, 2014.

[12] M. Cario, B. Nelson, Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, Tech. rep., Department

of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois (1997).

[13] L.-F. Lee, On the range of correlation coefficients of bivariate ordered discrete random variables, Econometric Theory 17 (1) (2001) 247–256.

[14] H. V. Ophem, A general method to estimate correlated discrete random variables, Econometric Theory 15 (2) (1999) 228–237.

[15] L. Madsen, D. Dalthorp, Simulating correlated count data, Env Ecol Stat 14 (2) (2007) 129–148.

[16] P. A. Ferrari, A. Barbiero, Simulating ordinal data, Multivariate Behavioral Research 47 (4) (2012) 566–589.

[17] L. Madsen, D. Birkes, Simulating dependent discrete data, Journal of Statistical Computation and Simulation 83 (4) (2013) 677–691.

[18] C. Genest, J. G. Nešlehová, Modeling dependence beyond correlation, in: J. L. Lawless (Ed.), Statistics in Action: A Canadian Outlook, CRC Press, London, 2014, pp. 59–78.

[19] M. Huber, N. Maric, Minimum correlation for any bivariate geometric distribution, Alea 11 (1) (2014) 459–470.

[20] H. Demirtas, D. Hedeker, A practical way for computing approximate lower and upper correlation bounds, American Statistician 65(2) (2011) 104–109.

[21] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 15 (1) (1904) 72–101.

[22] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.

[23] C. Genest, J. Nešlehová, A primer on copulas for count data, ASTIN Bulletin: The Journal of the IAA 37 (2) (2007) 475–515.

[24] M. Denuit, J. Dhaene, M. Goovaerts, R. Kaas, Actuarial theory for dependent risks: measures, orders and models, John Wiley & Sons, 2006.

[25] H. Joe, Dependence Modelling with Copulas, CRC Press, Boca Raton, 2015.

[26] O. P. Faugeras, Inference for copula modeling of discrete data: a cautionary tale and some facts, Dependence Modeling 5 (1) (2017) 121–132.

[27] R. Winkelmann, Econometric Analysis of Count Data, 5th Edition, Springer-Verlag, Berlin, 2008.

[28] A. C. Cameron, P. K. Trivedi, Regression analysis of count data, Vol. 53, Cambridge University Press, 2013.

[29] A. Barbiero, P. A. Ferrari, Simulation of correlated poisson variables, Applied Stochastic Models in Business and Industry 31 (5) (2015) 669–680.

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

[31] C. A. Klaassen, J. A. Wellner, et al., Efficient estimation in the bivariate normal copula model: normal margins are least favourable, Bernoulli 3 (1) (1997) 55–77.

[32] S. Kotz, D. D. Mari, Correlationa nd Dependence, Imperial College Press, London, 2001.

[33] C. Mitchell, A. Paulson, A new bivariate negative binomial distribution, Naval Research Logistics Quarterly 28 (1981) 359–374.

[34] E. Gómez-Déniz, E. Calderín-Ojeda, The discrete lindley distribution: properties and applications, Journal of Statistical Computation and Simulation 81 (11) (2011) 1405–1416.

[35] T. Nakagawa, S. Osaki, The discrete weibull distribution, IEEE Transactions on Reliability 24 (5) (1975) 300–301.