Research Article **Open Access**
Special Issue: Salzburg Workshop on Dependence Models & Copulas

Olivier P. Faugeras*

Inference for copula modeling of discrete data: a cautionary tale and some facts.

DOI 10.1515/demo-2017-0008
Received December 20, 2016; accepted June 5, 2017

Abstract: In this note, we elucidate some of the mathematical, statistical and epistemological issues involved in using copulas to model discrete data. We contrast the possible use of (nonparametric) copula *methods* versus the problematic use of parametric copula *models*. For the latter, we stress, among other issues, the possibility of obtaining impossible models, arising from model misspecification or unidentifiability of the copula parameter.

Keywords: copula; discrete data; parametric model; statistical inference; unidentifiability

MSC2010: 62A01, 62H20, 62H12

1 Introduction

In 2005, T. Mikosch [20] wrote his famous cautionary article whose primary objective was to temper the optimism of practitioners about copulas as a panacea "to solve all problems of stochastic dependence". It stirred some controversy and heated debate within the copula community, see for instance the responses [21] to the article. In the aftermath of the great financial crisis of 2008, the debate was revived and diffused in the mainstream audience, in particular following Salmon [33], where the blame was put on faulty models based on Gaussian copulas of Li [17]: the narrative was that Gaussian copula models failed to take into account tail dependence and risk which manifests itself in case of an extreme, "black swan" event like a systemic crisis.[1]

Our goal in this article has a similar cautionary objective, yet is much less broad in scope: our intent is to warn the practitioner who would like to naively extend the copula paradigm to discrete data, especially by using parametric copula models for making inference. More precisely, we organize the discussion as follows: Section 2 is a theoretical discussion of copulas associated to discrete data, of their indeterminacy and of their topological properties. In particular, we show how probabilistic constructions can be used to extend nonparametric copula methods in the discrete case and obtain consistent empirical copulas, in spite of the indeterminacy of the copulas associated to a discrete distribution. These positive results for *copula methods*

***Corresponding Author: Olivier P. Faugeras:** Toulouse School of Economics - Université Toulouse Capitole, Manufacture des Tabacs, Bureau MF319, 21 Allée de Brienne, 31000 Toulouse, France, E-mail: olivier.faugeras@gmail.com

1 As it is not the topic of this article to discuss the causes of the financial crisis nor our intent to jump-start the debate on mathematical models in finance, we will not delve too much on that matter. Nonetheless, let us simple state, for honesty, our belief that the issue may be more subtle and profound that simply blaming faulty copula models and may reside at a higher ontological level: as the late Keynes put forward after painfully learning from his practice of the Stock Markets [14], we believe that a distinction should be made between randomness (contingency but statistical regularity in repeated experiments) and uncertainty (mere contingency). In inanimate physical systems, possible stability of the underlying mechanisms may translate at the phenomenological level in terms of randomness, see Bunge [2, 3]. To the contrary, in complex, partially self-determining and ever-evolving human systems like the Stock Market, viewing observational data resulting from human decisions through probabilistic categories paves the way for disaster in times of crisis: when everybody wants to sell, all assets correlates and hedges derived from statistical models based on past data eventually break down when they are the most needed... Such view is likely not to fit within the Overton window of the statistical community...

are contrasted in Section 3 with the use of parametric *copula models* for discrete data: we delineate the mathematical, statistical, practical and epistemological issues involved with making inference from such discrete parametric copula models. Our arguments are based on discussing [12]'s proposed inference methodology applied to their example of a bivariate Bernoulli distribution.

2 A theoretical discussion of copulas associated to discrete data: topological properties and nonparametric estimation

2.1 A probabilisitic view on copulas

Let $\mathbf{X}$ be a d-variate real-valued random vector with c.d.f. F, and corresponding vector of marginal c.d.f.s $\mathbf{G} = (G_1, \ldots, G_d)$, namely $G_i(x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)$. We denote vectors by bold letters, and interpret operations between vectors componentwise.

Recall that a d-dimensional copula function $C : [0,1]^d \mapsto [0,1]$ is defined analytically as a grounded, d–increasing function, with uniform marginals whose domain is $[0,1]^d$ (see Nelsen [23]). Alternatively, it can be viewed from the probabilistic standpoint as the restriction to $[0,1]^d$ of the multivariate c.d.f. of a random vector $\mathbf{U}$, called a *copula representer*, whose marginals are uniformly distributed on $[0,1]$ (see Rüschendorf [31, 32]). The now well-known Sklar's Theorem asserts that, for every random vector $\mathbf{X} \sim F$, there exists a copula function $C \in \mathcal{C}$ *connecting, or associated with* (the law of) $\mathbf{X}$, in the sense that:

$$F(\mathbf{x}) = C \circ \mathbf{G}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d. \tag{1}$$

Let us denote by $\mathcal{C}(F)$ the set of copula functions associated with F in Sklar's Theorem (1). When F is continuous, $\mathcal{C}(F)$ reduces to a single copula C: the latter can be defined from F either analytically by $C = C^\star := F \circ \mathbf{G}^{-1}$, where $\mathbf{G}^{-1} = (G_1^{-1}, \ldots, G_d^{-1})$ is the vector of marginal quantile functions, or probabilistically as the c.d.f. of the multivariate Probability Integral Transform, namely $C(\mathbf{u}) = C^{\star\star}(\mathbf{u}) := P(\mathbf{G}(\mathbf{X}) \leq \mathbf{u})$, $\mathbf{u} \in [0,1]^d$. $C^\star$ and $C^{\star\star}$ are called respectively pre-copulas of the first and second types, in the terminology of Deheuvels [6].

Such distinction becomes relevant whenever discontinuity is present, as the mapping C obtained from relation (1) is (a sub-copula) only uniquely determined by F on the range of $\mathbf{G}$: neither $C^\star$ nor $C^{\star\star}$ are genuine copula functions belonging to $\mathcal{C}(F)$, see Genest and Nešlehová [12]. In other words $C \in \mathcal{C}(F)$, as a functional parameter, *is not identifiable from F alone*. The sub-copula derived from relation (1) can be extended on the whole unit cube $[0,1]^d$ to a copula $C \in \mathcal{C}(F)$ in multiple ways. In particular, probabilistic constructions of a copula representer $\mathbf{U}$ associated with $\mathbf{X} \sim F$,

$$\begin{array}{ccc} \mathbf{X} & \sim & F \\ \downarrow & & \\ \mathbf{U} & \sim & C \end{array}$$

can be based on:

i) the d-variate Distributional Transform

$$\mathbf{U} = \mathbf{G}(\mathbf{X}, \mathbf{V}), \tag{2}$$

where $G_j(x_j, \lambda) = P(X_j < x_j) + \lambda P(X_j = x_j)$, $j = 1, \ldots, d$, $\lambda \in [0,1]$, and $\mathbf{V} \sim R$ is a vector of uniform $[0,1]$ marginals, independent of $\mathbf{X}$ (see Moore and Spruill [22], Rüschendorf [30–32], Nešlehová [24], Faugeras [8]);

ii) Probabilistic Continuation, i.e. by taking the limit of

$$\mathbf{U}_h = \hat{\mathbf{G}}_h(\mathbf{X}_h), \tag{3}$$

in distribution along a subsequence, where $\hat{\mathbf{G}}_h$ is the vector of marginal c.d.f. of the continued $\mathbf{X}_h = \mathbf{X} + h\mathbf{Z}$, where $\mathbf{Z}$ is continuous and $h \downarrow 0$, (see Faugeras [7, 8]);

iii) More generally, one can generalize the d-variate Distributional Transform (2) into a *Conditional* Distributional Transform by choosing different randomizers $\mathbf{V_x}$ for each different discontinuity point $\mathbf{x}$ of $\mathbf{G}$. More precisely, let us denote by $\mathcal{D}(F)$ the set of discontinuity points of F, and set, for each $\mathbf{x} \in \mathcal{D}(F)$, $\mathbf{V_x}$ distributed according to a copula distribution $R_{\mathbf{x}}$. Then, by defining, conditionally on $\mathbf{X} = \mathbf{x}$,

$$\mathbf{U} = \mathbf{G}(\mathbf{x}, \mathbf{V_x}), \tag{4}$$

one obtains a copula representer $\mathbf{U}$ associated with $\mathbf{X}$, see [31].

One interest of these explicit probabilistic constructions is to make apparent that, contrary to the continuous case where the copula $C \in \mathcal{C}(F)$, defined as the law of $\mathbf{G}(\mathbf{X})$, depends solely on the law F of $\mathbf{X}$ alone, this is no longer the case in the non-continuous case: the set $\mathcal{C}(F)$ of copulas associated to F not only depends on the choice of the randomization mechanism i), ii), or iii) but also on the choice of the distribution of the randomizers, i.e. on the distributions of $\mathbf{V}$ or $\mathbf{V}_x$ for the Distributional Transform construct i) or iii), or on the distribution of $\mathbf{Z}$ in the Probabilistic Continuation construct ii). (Obviously, some choices are more natural than others, see Subsection 2.3).

We already stress the profound statistical/epistemological consequences this will have when it comes to inference: in the continuous case, F (and consequently the copula $C = F \circ \mathbf{G}^{-1}$) can be given unambiguous factual content w.r.t. the real world as the limit of the empirical frequency F_n, thanks to the Glivenko-Cantelli Theorem. However, in the discontinuous case, $C \in \mathcal{C}(F)$ defined as, say, the law of $\mathbf{G}(\mathbf{X}, \mathbf{V})$, depends also on an arbitrary/subjective choice of randomizers. In other words, in the discrete case, *the copula does not exist* in the ontological/epistemological meaning, in the sense that it can not be inferred from data alone in an objective way (only the sub-copula defined on $\overline{Ran\mathbf{G}}$ can), even though one can say the copula exists in the idealistic meaning, as a perfectly valid mathematical construct.

These subtle Popperian issues of correspondence of concepts to reality and their relevance for the practitioner will be made more clear when we discuss them on an example of parametric copula model in Section 3. (See in particular the discussions on model unidentifiability in Subsections 3.3.3 and 3.3.4). We now complement our theoretical discussion by investigating the topological properties of empirical copula functions in the next subsection.

2.2 Empirical copulas and topological properties

The issue involved in the indeterminacy of "the" copula functions associated to a non-continuous c.d.f. matters in particular for the so-called "empirical copulas" associated with the empirical c.d.f. F_n,

$$F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} 1_{\mathbf{X_i} \leq \mathbf{x}},$$

based on a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots$ of copies of $\mathbf{X}$ distributed according to F. Indeed, as F_n is discrete, C_n in Sklar's Theorem,

$$F_n(\mathbf{x}) = C_n \circ \mathbf{G}_n(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

where $\mathbf{G}_n$ is the corresponding vector of marginal e.c.d.f.s, is no longer unique and can no longer be defined, in parallel with the continuous case, as $C_n^\star := F_n \circ \mathbf{G}_n^{-1}$, or as $C_n^{\star\star}(\mathbf{u}) := P^\star(\mathbf{G}_n(\mathbf{X}_n^\star) \leq \mathbf{u})$, with $\mathbf{X}_n^\star \sim F_n$, conditionally on the sample, and where $P^\star$ is the corresponding conditional probability. Indeed, these empirical "copulas" $C_n^\star$ and $C_n^{\star\star}$, introduced respectively by Rüschendorf [29] under the name of multivariate rank order function and Deheuvels [5, 6] under the name of empirical dependence function, do not have uniform marginals and hence are not genuine copula functions associated with F_n. This is a drawback in practice, as one always has a finite sample, although, for a continuous F, this matter is of little relevance asymptotically.

However, when F has a discrete component, the issue become more severe: Deheuvels [4], Sempi [34] and Lindner and Szimayer [18] among other have shown that if C_n and C are copula functions associated with F_n, F, with respective marginals $\mathbf{G}_n, \mathbf{G}$, then, $F_n \xrightarrow{d} F$ if and only if

i) the margins weakly converge: $F_{n,j} \xrightarrow{d} F_j$, $j = 1, \ldots, d$;

ii) and C_n converges uniformly to C on $\overline{Ran\mathbf{G}}$.

Quoting Lindner and Szimayer, "since in [the discrete] case, the copula of $\mathbf{X}$ does not need to be unique, convergence of the copulas on $[0,1]^d$ cannot be expected" and they present a counter-example. See also Nešlehová [24] or Marshall [19] for a similar discussion. Hence, the possibility of completely inferring the whole unidentified C from some empirical copula C_n seems hopeless at first sight.

2.3 Convergence of empirical copulas: maximal coupling and weak convergence

Nonetheless, the author has shown in [7, 8] that one can apply the probabilistic constructions i) and ii) of Subsection 2.2 to the empirical c.d.f in order to obtain empirical copula representers $\mathbf{U}_n \sim C_n$ and $\mathbf{U} \sim C$ associated with F_n, F respectively, with C_n, C *genuine* copula functions. Moreover, a maximal coupling construction of the empirical measure in the ergodic setting yields a.s. convergence of $\mathbf{U}_n \to \mathbf{U}$. This translates analytically into the a.s. convergence of the corresponding genuine empirical copula functions $C_n \to C$, uniformly on the whole $[0,1]^d$:

$$\sup_{\mathbf{u}\in[0,1]^d} |C_n(\mathbf{u}) - C(\mathbf{u})| \overset{a.s.}{\to} 0.$$

This coupling construction can be regarded as a strengthening of Skorohod's Representation Theorem. It implies an implicit "Discontinuous Mapping Theorem", which yields consistency of functionals of the copula representers as simple corollaries. In particular, it was applied to prove the ergodic consistency of estimates of discrete extension of dependence coefficients, such as versions of Kendall's τ, see Faugeras [8].

The apparent paradox of such a uniform convergence result of copula functions, in the discrete case, on the whole $[0,1]^d$ and not solely on $\overline{Ran\mathbf{G}}$, comes from the fact that the copula functions in Deheuvels' result are left unspecified, which leaves some space for counterexamples. Whereas the indeterminacy on the copula functions involved is resolved in [8], by the specific constructions of the copula representers $\mathbf{U}_n, \mathbf{U}$. In particular, for the Distributional Transform construction, the author used a randomizer $\mathbf{V} \sim \Pi$ with independent marginals, so that the dependence coefficients, computed on the level of the observation $\mathbf{X}$, matches those computed on the level of the copula $\mathbf{U}$, and are not perturbed by some artificial local dependence introduced by the Statistician in the procedure. (See also Remark 2.3 (a) in [31]).

These results complements those obtained by Rüschendorf (see Theorem 4.1 in [31]), who study the weak convergence of the (sequential) empirical copula process $\sqrt{n}(C_n(.)-C(.))$: by using the probabilistic approach based on the Distributional Transform, he was able to extend the weak convergence result he obtained earlier for the continuous case in [29] to the discontinuous case. Similar results where also later obtained by Genest et al. [11], via purely analytical methods. See [8] for a discussion of these results.

All together, these positive results show that *copula methods* can be extended to discrete data, in particular to make nonparametric estimation of dependence coefficients. We will now address in the next Section 3 the question of the use of *copula models* for discrete distributions and of their inference.

3 Parametric inference in discrete copula models

In the remainder, we restrict to the bivariate case and simplify notation: let $(X, Y) \sim H$ be a bivariate vector with c.d.f. H, with marginals $X \sim F$, $Y \sim G$. Let $\mathcal{C}(H)$ be the set of copula functions compatible with H in Sklar's Theorem (1).

3.1 Statement of the problem

Up to a recent time, because of the non-unicity of the copula function whenever at least one marginal has a jump in its c.d.f., parametric copula models had mostly been used in applications for continuous data. However, the tide is turning, see e.g. [16], [13], [25], [26] [36], [1], [15], [38], [27] for some instances where

discrete data is analyzed via a parametric copula model. In particular, Genest and Nešlehová's [12] investigate, in a very informative paper, the use of copulas for discrete data: they present many subtleties and counter-intuitive results dispersed and/or overlooked in the literature and also argue in favor of extending the usage of parametric copula models to non-continuous data. (See also Marshall [19] for an early investigation on these issues). The setting is as follows: for (X, Y) distributed as an unknown H, one gives oneself a parametric copula model $\mathcal{M} = \{C_\theta, \theta \in \Theta\}$, indexed by a copula parameter θ. Given some (say) i.i.d. data $((X_1, Y_1), \ldots, (X_n, Y_n))$, can one make statistical inference of the copula parameter θ from the data?

[12]'s illustrate their defense of parametric copula models for discrete data on an example of a bivariate Bernoulli distribution for which they propose a Maximum Likelihood estimation procedure. Their message appears mixed: on the one hand, they warn the reader, stressing that "When dealing with count data, however, modeling and interpreting dependence through copulas is subject to caution. Furthermore, inference [...] for copula parameters from discrete data is fraught with difficulties." On the other hand, they state that "Insofar as the dependence parameter θ is identifiable, however, its estimation remains possible via fully parametric maximum likelihood estimation techniques, although exact conditions under which identifiability is guaranteed are yet to be delineated. Once obstacles posed by inference are resolved, and in view of their richness and flexibility, copula-based models are likely to become as attractive for discrete variables as they have grown for continuous data."

We intent to challenge such a positive view by elucidating some of the difficulties and traps a practitioner might fall into if he were to use parametric copula models for inference purposes for discrete data. Our arguments will be based on a detailed discussion of [12]'s toy bivariate Bernoulli example which we now present.

3.2 [12]'s inference methodology for a bivariate Bernoulli with FGM copula

Let $(X, Y) \in \{0, 1\}^2$ be a bivariate Bernoulli distribution parametrized, as in [12], by the marginals $p := P(X = 0)$, $q := P(Y = 0)$ and $r := P(X = 0, Y = 0)$. A valid discrete distribution H is obtained whenever the parameters p, q, r satisfy the set of constraints,

$$0 \le p, q \le 1, \quad \max(0, p + q - 1) \le r \le \min(p, q), \tag{5}$$

the latter inequality being derived from Fréchet-Hoeffding bounds.

[12]'s use such a simple discrete distribution as a pedagogical example to illustrate their parametric copula estimation methodology: they describe this bivariate Bernoulli distribution as resulting from a combination of univariate Bernoulli marginals with a FGM copula The latter is defined analytically, for $(u, v) \in [0, 1]^2$, by (see e.g. [23]),

$$C_\theta(u, v) = uv + \theta uv(1 - u)(1 - v), \quad \theta \in [-1, 1], \tag{6}$$

where θ is an unknown parameter for the copula function. In particular, [12]'s assert that "Standard Maximum likelihood estimate works when the joint distribution of (X, Y) is of copula type, for some $C \in \mathcal{C}_\theta$". Maximum likelihood estimates $\hat{p}_n, \hat{q}_n, \hat{r}_n$ of p, q, r can be computed with the corresponding empirical frequencies. As θ is constrained by

$$C_\theta(p, q) = r, \tag{7}$$

"the maximum likelihood estimate is the unique value $\hat{\theta}_n$ such that

$$C_{\hat{\theta}_n}(\hat{p}_n, \hat{q}_n) = \hat{r}_n. \tag{8}$$

Standard theory then implies that this estimation is consistent and asymptotically Normal. In particular, θ is then estimable."

In the present case, an explicit expression of the estimator $\hat{\theta}_n$ of the copula parameter is available: inverting (7) for the copula (6) yields the parameter as a function of r, p, q:

$$\theta = \frac{r - pq}{pq(1 - p)(1 - q)}. \tag{9}$$

Plugging ML estimators in the latter equation (9) yields the desired explicit expression. For a model where $p = 0.3$, $q = 0.4$, and $r = 0.1452$, they infer the parameter value $\theta = 0.5$.

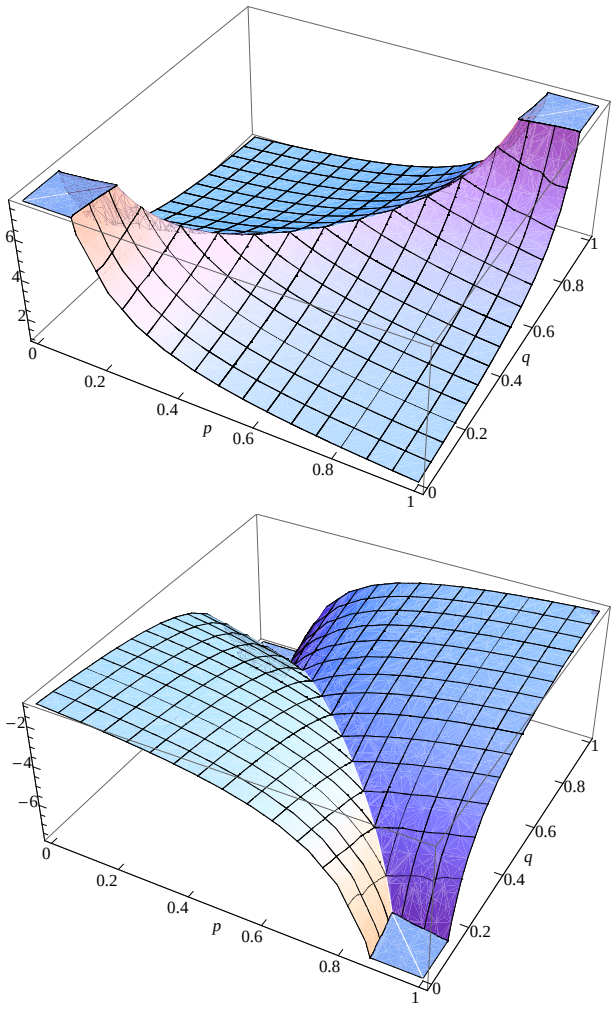

Figure 1: θ^+ (top figure) and θ_- (bottom) as functions of (p, q)

3.3 Some fundamental issues for inference in discrete parametric copula models

3.3.1 The copula model chosen might be valid only for some restricted discrete distributions

In the FGM copula (6), the restriction on the range of the copula parameter θ

$$-1 \leq \theta \leq 1 \tag{10}$$

is to ensure that the copula is a genuine one, in particular that it is a 2-increasing function, see [23]. Using the restrictions on r of (5) in (9), one sees that the copula parameter must lies between $\theta_- \leq \theta \leq \theta^+$, with

$$\begin{aligned} \theta_- &= \frac{\max(0, p+q-1) - pq}{pq(1-p)(1-q)}, \\ \theta^+ &= \frac{\min(p, q) - pq}{pq(1-p)(1-q)}. \end{aligned}$$

Their curves as a function of p, q are plotted in Figure 1.

It is clear that for some valid values of p, q, r satisfying (5), one may obtain a θ value which is larger than 1 or lower than -1, i.e. which is not compatible with (6). In other words, not all bivariate Bernoulli distribution can be described with an FGM copula. In mathematical terms, one may have that

$$\mathcal{C}(H) \cap \mathcal{M} = \emptyset,$$

for the family $\mathcal{M}$ of parametric copulas chosen. As a consequence, the mere existence of the copula model and its θ parameter depends on the (unknown) p, q, r values of the original distribution. It is thus problematic that the validity of the copula model might depend on the value of the parameter which has to be estimated.

Let us illustrate how such an issue can materialize in practice, with the following scenarios:

- **[12]'s favorable scenario:**
  Assume, as in [12]'s example, that the true but unknown values of the parameters are $p = 0.3$, $q = 0.4$, $r = 0.1452$ for the distribution H. These yield a value (here $\theta = 0.5$) for the copula parameter which falls into the domain of validity of the copula model. In such a case, the true value $\theta = 0.5$ of the copula parameter can be consistently estimated by some ML estimator (8) from the data, as [12]'s assert.
- **Unfavorable scenario:**
  Assume our unknown distribution H has the same values $p = 0.3$, $q = 0.4$, but, say, a different r. The constraint (5) on the latter would become $0 \le r \le 0.3$, and $|\theta|$ would be larger than 1 whenever $|r - 0.12| > 0.0504$, viz. $r > 0.1704$ or $r < 0.0696$. Imagine then the following sub-cases:
  - **Case 1: Impossible model**
    Assume the true values are $p = 0.3$, $q = 0.4$, $r = 0.175$. One obtains a valid distribution for H, but an invalid copula model with $\theta = 1.09127 > 1$ for such an H distribution.
  - **Case 2: Inferring a parameter value in an invalid model**
    Worse, add sampling fluctuations to the preceding subcase. Assume the true values of the H parameters remain $p = 0.3$, $q = 0.4$, $r = 0.175$, (so that you have an invalid copula model for the distribution H), but one computes from the data, the estimates, say $\hat{p} \approx 0.31$, $\hat{q} \approx 0.41$, $\hat{r} \approx 0.174$. The estimated value of the copula parameter becomes $\hat{\theta} \approx 0.906$, which is within the bounds (10) of the FGM copula: taking into account sampling fluctuations, you might wrongly empirically infer that your FGM copula model is valid whereas in reality it is not!
  - **Case 3: Inferring an incorrect copula parameter in a correctly specified model**
    Conversely, assume your true H model is $p = 0.3$, $q = 0.4$, $r = 0.17$, so that (9) yields $\theta = 0.992063$: you have a valid FGM copula model. Assume the empirical estimates are now $\hat{p} \approx 0.29$, $\hat{q} \approx 0.39$, $\hat{r} \approx 0.171$, which gives an estimated $\hat{\theta} \approx 1.182 > 1$. One gets the opposite situation of case 2: you might wrongly empirically infer that your FGM copula model is invalid whereas in reality it is valid!

Let us summarize the above argument in the following Table 1:

Table 1: Bernoulli distribution with FGM copula

Bernouilli	FGM Copula Model	Valid Model/Inference?
$\max(0, p+q-1) \le r \le \min(p, q)$	$\theta \in [-1, 1]$	
[12]'s case: $p = 0.3$, $q = 0.4$, $r = 0.1452$	$\theta = 0.5$	Yes
$p = 0.3$, $q = 0.4$, $0 \le r \le 0.3$	$0.0696 \le r \le 0.1705$	
Case 1: $p = 0.3$, $q = 0.4$, $r = 0.175$	$\theta = 1.09127 > 1$	No
Case 2: $p = 0.3$, $q = 0.4$, $r = 0.175$	$\theta = 1.09127 > 1$	No
$\to \hat{p} \approx 0.31$, $\hat{q} \approx 0.41$, $\hat{r} \approx 0.174$	$\hat{\theta} \approx 0.906 < 1$	Yes
Case 3: $p = 0.3$, $q = 0.4$, $r = 0.17$	$\theta = 0.992063 < 1$	Yes
$\to \hat{p} \approx 0.29$, $\hat{q} \approx 0.39$, $\hat{r} \approx 0.171$	$\hat{\theta} \approx 1.182 > 1$	No

Hence, these cases illustrate the fact that, in finite sample, in a copula model for discrete data, the Statistician can make both mistakes of inferring a seemingly correct copula parameter value in an impossible model, or rejecting a correctly specified model from an apparently incorrect copula parameter value. Consequently, *the issues of copula model specification and of inferring from data a copula parameter of interest in a correctly specified model can become intertwined*.

3.3.2 The copula parameter may not be identifiable

[12]'s hint the possible classical issue of parameter unidentifiability, i.e. that the mapping $\theta \mapsto C_\theta$ may not be injective. More precisely, for a d-dimensional discrete distribution, let $k_i = 1, 2, \ldots, \infty$, $i = 1, \ldots, d$, the (possibly infinite) number of points of the ith coordinate with strictly positive mass. The d-dimensional analogue of (7) yields $\prod_{i=1}^{d}(k_i - 1)$ non trivial equations on the parameter θ. When d and/or k_i increase, this either restricts the set of discrete distributions which are identifiable through a given copula family, or one is likely to require from the statistician copulas with a parameter θ of large dimension, for this system to have a unique solution.

Let us stress the practical implications of these constraints:

- On the one hand, most parametric copula families in the literature are one or two dimensional, see [23]. Therefore, the practitioner may have trouble to find a suitable family in his catalogue of copulas to satisfy the identifiability requirement.
- On the other hand, such a copula modeling becomes less interesting. Indeed, one may argue that instead of modeling a discrete distribution with $\prod_{i=1}^{d} k_i - 1$ probability mass values, the copula approach is interesting because it "separates" the modeling of marginals (with $\sum_{i=1}^{d}(k_i - 1)$ probability mass values) from modeling the "dependence" through a parameter θ. Whenever this dependence parameter is low-dimensional, copula modeling gives an economical description of such a distribution. Requiring a copula family with a high dimensional dependence parameter thus mitigates the alleged advantageous copula representation w.r.t. dimension-reduction: one falls back on a sort of curse of dimensionality.
- At last, a practitioner who would overlook this issue would make an implicit systematic modeling error, i.e. he approximates H by a family of distributions which does not contain H. In other words, *he analyzes the data with a copula model which can not generate the distribution of the data*. Hence, all his inferences will be biased by an unquantified systematic modeling error.

3.3.3 The copula model itself is unidentifiable

One may argue that the objection raised in Section 3.3.1 comes from the fact that "'FGM copulas can only model relatively weak dependence" ([23] p.78) and that a way to alleviate such objection is to choose another copula model.

1. **Plackett's copula**: In particular, one may argue that the "natural" copula family to suit such a bivariate Bernoulli distribution is Plackett's one. Indeed, the latter had been invented as an extension for continuous random variables of contingency tables, see Nelsen [23] Section 3.3.1. For Plackett's copula, the parameter θ must lie in $[0, \infty[$. (7) defines θ as,

$$\theta = \frac{r(1 - p - q + r)}{(p - r)(q - r)}, \quad \theta \neq 1.$$

   Plugging the values of [12] yields $\theta = 1.6389$.
2. **AMH copula**: Alternatively, the same analysis can be performed using an Ali-Mikhail-Haq copula, see [23] p. 28. The parameter θ must be constrained to lie in $[-1, 1]$ for the copula to be 2-increasing, and is this time expressed as,

$$\theta = \frac{r - pq}{r(1 - p)(1 - q)}.$$

   A numerical application with the values of [12] yields $\theta = 0.413223$.
3. We let the reader continue the reasoning, with his peculiar favorite copula family. In particular, he can try Gaussian copulas, as advocated e.g. in [13, 16, 36].

The values are summarized in Table 2:

This discussion illustrates the statistical/epistemological issues we already hinted in Section 2.1 on the nonexistence, at the ontological level, of the copula: the data can be said to have been generated from several copula models, whose choice is left to the statistician's personal taste. Although the copula parameter can be

Table 2: Bernoulli distribution with several copula models

Bernouilli $\max(0, p+q-1) \le r \le \min(p, q)$	Copula Model	θ
$p = 0.3, q = 0.4, r = 0.1452$	FGM, $\theta \in [-1, 1]$	$\theta = 0.5$
$p = 0.3, q = 0.4, r = 0.1452$	Plackett, $\theta > 0$	$\theta = 1.6389$
$p = 0.3, q = 0.4, r = 0.1452$	AMH, $\theta \in [-1, 1]$	$\theta = 0.413223$

estimated once an assumption on the copula model has been made, as asserted in [12], the latter assumption can not be tested from the data, and there is no guarantee that the real data generating process originates from the chosen copula model.

In other words, *the copula model itself is unidentifiable*: one can not tell empirically whether the distribution H comes from a given copula model $\mathcal{M}_1 = \{C_\theta, \theta \in \Theta\}$ or another one $\mathcal{M}_2 = \{C'_{\theta'}, \theta' \in \Theta'\}$. Arguing that Plackett's is the most natural copula is aesthetics, or that one should favor Gaussian copulas because algorithms are available is ad-hoc. As [12]'s warn of in their Section 3, given a multivariate distribution H with discrete margins, the set of copula functions $\mathcal{C}(H)$ which are compatible with it, in the sense that Sklar's Theorem (1) is satisfied, can be quite large and one only knows that the set of possible copulas lie between Carley's bounds C_H^+, C_H^-.

3.3.4 The copula parameter looses substance w.r.t. the random phenomenon under study

The previous point has important ontological and semantic implications: when can one say that the copula parameter "exists" w.r.t. the real world and what is its interpretation? Our analysis of the bivariate Bernoulli distribution clearly demonstrated that one can fit as many copula models with the data, provided these models fit the constraints enumerated above. In particular, we have "inferred" for each model, some value for the "dependence parameter" θ, e.g. $\theta = 0.5$ for FGM's model, $\theta = 1.6389$ for Plackett's one, and $\theta = 0.413223$ for AMH's one. How do we reconcile these divergent values of the copula parameter ? We hardly see how, because each value is valid only within its parent copula model. Do these parameter values tell us something more that we could not infer from estimating the (p, q, r) parameters of the distribution H directly? In the bivariate Bernoulli distribution example with parametrisation (p, q, r), the marginals are independent if and only if $r = pq$. In other words, one can directly evaluate from $r - pq$ the departure from independence, whereas the copula parameter, is, in the best of cases, a biased, *implied*, measure of dependence, as it now depends on the margins. See [12] Section 5.3 for an insightful discussion on dependence modeling and coefficients in the discrete case.

In short, the naive statistician may find an estimated value of the parameter but will be unable to assess neither its validity nor meaning for the problem under study, since he can not exactify the underlying copula family. Since the value of the copula parameter only has meaning within the chosen copula model, and that the choice of the latter is arbitrary (with the proviso that the chosen copula family passes the constraints of Sections 3.3.2, 3.3.1), the parameter looses signification w.r.t. the original distribution and thus w.r.t. the concrete problem under study. The inferences drawn and the interpretation of theses parameter values by practitioners could be very misleading.

4 Conclusion

We have argued that modeling a discrete distribution with a parametric copula and its marginals is fraught with difficulties. Apart from possible model misspecification and parameter unidentifiability, our main concern is the non uniqueness of the candidate copula model, viz. model unidentifiability. Although it is (asymp-

totically) possible to exactify the true distribution H of the phenomenon under study via its empirical counterpart H_n, we hope to have clearly demonstrated the issues involved at the copula level: when H is posited to result from a parametric copula+marginals model, it is no longer possible to infer from H_n which copula family $\mathcal{M} = \{C_\theta, \theta \in \Theta\}$ is the "right" one, although it is possible to infer the parameter value once a specific family has been chosen.

One could argue that our reasoning based on [12]'s toy example of a bivariate Bernoulli distribution should be regarded as unconvincing because such a distribution is too simple for copula modeling to be useful. The pedagogical virtue of such a model is precisely its simplicity, which makes transparent the issues involved so that it becomes easier to delineate them. Moreover, if foundational issues arise in such a simple model, *a fortiori* they will not disappear magically in a more complicated one. The risk for the Statistician is that, by burying these issues under the complexity of the model, he might miss them.

Mathematically inclined statisticians may dismiss these matters as mere "speculations": as a formal science, mathematics is not interested in the concrete existence of its objects, or more precisely in their correspondence with real things. From a strictly mathematical point of view, it is perfectly legitimate to choose a particular copula family among many, as long as it passes the difficulties raised in Sections 3.3.1 and 3.3.2. Yet, from a scientific point of view, when one wants to model a particular concrete existing phenomenon, one looks for objectivity and correspondence with reality, see Bunge [3]. Moreover, such epistemological issues have serious practical consequences, which may prove catastrophic, as we reminded in the Introduction.

Subjectivists or Bayesians may argue that it does not matter to find "the" true (copula) model, as long as one finds "a" copula model fitting the data. In addition to possible overfitting (see Vapnik [37]), positing a particular parametric copula family hides a subjective arbitrary choice, which is rendered explicit when one uses the probabilistic view on copulas of Section 2.1. Let us draw a parallel with the use of undefined utility functions in economics: such relativist stance is popular in neoclassical economics, see e.g. [10], where it is claimed that it does not matter whether the assumptions of a model are true as long as their consequences are. In a letter to L. Walras, H. Poincaré [28] pointed out that utility is (to some degree) an arbitrary function, and that while such arbitrary functions may be used in mathematical reasonings, one must try to eliminate them in the end results or consequences of such reasonings: "if the arbitrary functions still occur in these consequences, the latter won't be false, but they will be devoid of interest because they will be subordinated to the arbitrary conventions laid down at the beginning". Hence, one may obtain a model which is not logically nor mathematically false but which may be vacuous, see also Tao [35] p. 358-359.

An appeal to Occam's razor principle could also be invoked: in view of the model uncertainty at the copula level and of the issues we raised in Section 3, wisdom suggests to dispense with this intermediary step of copula modeling and not to multiply *copula models* which have a degree of arbitrariness/subjectivity. Instead, the recommendation for the practitioner is the following: use the *nonparametric copula methods* mentioned in Section 2.3, in particular (empirical) copulas based on the Distributional Transform with an independent randomizer. This allows to obtain e.g. consistent estimates of a dependence measure at the copula level which matches the corresponding dependence measure at the observational level for which asymptotic results are readily available. See also [9] for an illustration on how these nonparametric copula methods can be combined with Mass Transportation techniques in order to define a general notion of multivariate quantile and their related depth areas.

We hope that such a mixed message will find a favorable echo in the copula community. Although somehow critical, our intent was to try to remain friendly and as factual and logical as possible. We believe that both positive and negative views on a topic should be aired and openly discussed in order to make progress in Science. We had the opportunity to test our message on the audience of the conference "Workshop on Dependence Models and Copulas", held in Salzburg, Austria, September 19-22, 2016. To that regard, we are grateful to the organizers and participants of this conference for the friendly discussions we had, in particular with C. Genest and L. Rüschendorf. We are also thankful to I. Gijbels who suggested to us that it might be worthwhile to publicize our concerns to a larger audience by submitting a paper to the special issue of the journal "Dependence Modeling" related to the workshop.

References

[1] Bücher, A. and I. Kojadinovic (2015). An overview of nonparametric tests of extreme-value dependence and of related statistical procedures. In *Extreme Value Modeling and Risk Analysis Methods and Applications*, pp. 377–398. Chapman and Hall/CRC, Boca Raton FL.

[2] Bunge, M. (1988). Two faces and three masks of probability. In *Probability in the Sciences*, pp. 27–50. Kluwer Academic Publishers, Dordrecht.

[3] Bunge, M. A. (2006). *Chasing Reality: Strife Over Realism*. University of Toronto Press, Toronto.

[4] Deheuvels, P. (1978). Caractérisation complète des lois extrêmes mutivariées et de la convergence des types extrêmes. *Pub. Inst. Stat. Univ. Paris 23*(3-4), 1–36.

[5] Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés, un test non paramétrique d'indépendance. *B. Ac. Roy. Belg. 65*(f.6), 274–292.

[6] Deheuvels, P. (2009). A multivariate Bahadur-Kiefer representation for the empirical copula process. *J. Math. Sci. 163*(4), 382–398.

[7] Faugeras, O. P. (2013). Sklar's theorem derived using probabilistic continuation and two consistency results. *J. Multivariate Anal. 122*, 271–277.

[8] Faugeras, O. P. (2015). Maximal coupling of empirical copulas for discrete vectors. *J. Multivariate Anal. 137*, 179–186.

[9] Faugeras, O. P. and L. Rüschendorf (2017). Markov morphisms: a combined copula and mass transportation approach to multivariate quantiles. *Math. Appl.*, to appear. Available at http://dx.doi.org/10.14708/ma.v45i1.2921.

[10] Friedman, M. (1953). The methodology of positive economics. In *Essays in Positive Economics*, pp. 3–16, 30–43. University of Chicago Press, Chicago IL.

[11] Genest, C., J. G. Nešlehová, and B. Rémillard (2014). On the empirical multilinear copula process for count data. *Bernoulli 20*(3), 1344–1371.

[12] Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *Astin Bull. 37*(2), 475–515.

[13] Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat. 1*(1), 265–283.

[14] Keynes, J. M. (1939). Professor Tinbergen's Method. *Econ. J. 49*(195), 558–577.

[15] Kojadinovic, I. (2017). Some copula inference procedures adapted to the presence of ties. *Comput. Statist. Data Anal. 112*, 24–41.

[16] Lee, L.-f. (2001). On the range of correlation coefficients of bivariate ordered discrete random variables. *Economet. Theor. 17*(1), 247–256.

[17] Li, D. X. (2000). On default correlation: a copula function approach. *J. Fix. Income 9*(4), 43–54.

[18] Lindner, A. M. and A. Szimayer (2005). A limit theorem for copulas. *Discussion Paper 433, Sonderforschungsbereich 386, Available at https://epub.ub.uni-muenchen.de/1802/1/paper_433.pdf*.

[19] Marshall, A. W. (1996). Copulas, marginals, and joint distributions. In *Distributions with Fixed Marginals and Related Topics*, pp. 213–222. Institute of Mathematical Statistics, Hayward CA.

[20] Mikosch, T. (2006a). Copulas: Tales and facts. *Extremes 9*(1), 3–20.

[21] Mikosch, T. (2006b). Copulas: Tales and facts–rejoinder. *Extremes 9*(1), 55–62.

[22] Moore, D. S. and M. C. Spruill (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Ann. Stat. 3*(3), 599–616.

[23] Nelsen, R. B. (2006). *An Introduction to Copulas*. Second edition. Springer, New York.

[24] Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *J. Multivariate Anal. 98*(3), 544–567.

[25] Nikoloulopoulos, A. K. and D. Karlis (2009). Finite normal mixture copulas for multivariate discrete data modeling. *J. Stat. Plan. Infer. 139*(11), 3878–3890.

[26] Nikoloulopoulos, A. K. and D. Karlis (2010). Regression in a copula model for bivariate count data. *J. Appl. Stat. 37*(9), 1555–1568.

[27] Pappadà, R., F. Durante, and G. Salvadori (2016). Quantification of the environmental structural risk with spoiling ties: is randomization worthwhile? *Stoch. Environ. Res. Risk Assess*, to appear. Available at http://dx.doi.org/10.1007/s00477-016-1357-9.

[28] Poincaré, H. (1965[1901]). Letter to Léon Walras. In *Correspondence of Léon Walras and Related Papers. Vol. I, 1857-1883. Vol. II, 1884-1897. Vol. III, 1898-1909, and Indexes by W. Jaffé*, pp. 164–165. North Holland Publishing Co., Amsterdam.

[29] Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Stat. 4*(5), 912–923.

[30] Rüschendorf, L. (1981). Stochastically ordered distributions and monotonicity of the OC-function of sequential probability ratio tests. *Math. Operationforsch. Stat. Ser. Statist. 12*(3), 327–338.

[31] Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *J. Stat. Plan. Infer. 139*(11), 3921–3927.

[32] Rüschendorf, L. (2013). *Mathematical Risk Analysis. Dependence, Risk Bounds, Optimal Allocations and Portfolios*. Springer, Berlin.

[33] Salmon, F. (2009). Recipe for disaster: the formula that killed Wall Street. *Wired Magazine 17*(3).

[34] Sempi, C. (2004). Convergence of copulas: critical remarks. *Rad. Mat. 12*(2), 241–249.
[35] Tao, T. (2014). *Analysis. Volume I.* Third edition. Hindustan Book Agency, New Delhi.
[36] van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Economet. Theor. 15*(2), pp. 228–237.
[37] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Second edition. Springer, New York.
[38] Yan, L., L. Yang, Q. Yichen, and Y. Jun (2016). Copula modeling for data with ties. Available at: https://arxiv.org/abs/1612.06968.