econometrics

MDPI

Article

A Note on Identification of Bivariate Copulas for Discrete Count Data

Pravin Trivedi [1] and David Zimmer [2,*]

[1] Department of Economics, Indiana University Bloomington, 100 South Woodlawn Avenue, Bloomington, IN 47405-7104, USA; trivedi@indiana.edu

[2] Department of Economics, Western Kentucky University, 1906 College Heights Blvd., Bowling Green, KY 42101, USA

* Correspondence: david.zimmer@wku.edu; Tel.: +1-270-745-2880

Academic Editor: Marc S. Paolella

Received: 5 August 2016; Accepted: 7 February 2017; Published: 15 February 2017

Abstract: Copulas have enjoyed increased usage in many areas of econometrics, including applications with discrete outcomes. However, Genest and Nešlehová (2007) present evidence that copulas for discrete outcomes are not identified, particularly when those discrete outcomes follow count distributions. This paper confirms the Genest and Nešlehová result using a series of simulation exercises. The paper then proceeds to show that those identification concerns diminish if the model has a regression structure such that the exogenous variable(s) generates additional variation in the outcomes and thus more completely covers the outcome domain.

Keywords: ties; Monte Carlo; Gaussian; Clayton; Gumbel

JEL Classification: C35; C52

1. Introduction

The copula approach for constructing joint distributions has gained popularity in recent years in applied econometric studies, including models with discrete outcomes (Van Ophen (1999) [1]; Cameron et al. (2004) [2]; Zimmer and Trivedi (2006) [3]; Bien et al. (2011) [4]; Winkelmann (2012) [5]). While copula researchers have long understood that a multivariate discrete distribution does not possess a unique copula representation (Marshall (1996) [6]), recent research also indicates that *any* copula applied to discrete data is not identified. The lack of identification of the copula in a model for discrete data, as explained by Genest and Nešlehová (2007) [7], arises when one of the marginal distributions is discontinuous. Although Genest and Nešlehová present findings for other discontinuous settings, this paper focuses on their main emphasis: count outcomes.

We derive motivation from research in the areas of health economics and demography, where, due to count outcomes having small means, the *empirical* support present in the data is far smaller than the theoretically *infinite* support of count outcomes. For example, the widely-used Medical Expenditure Panel Survey, published by a unit of the U.S. Department of Health and Human Services, asks respondents their number hospital discharges in a calendar year. Not surprisingly, because most respondents report zero hospital discharges, the mean number of annual discharges is small (e.g., 0.085 discharges in the 2014 wave of the survey). Reflecting our health economic motivation, the remainder of this paper emphasizes low-mean settings.

This paper shows that the identification problem appears to shrink when the count outcomes more completely cover the outcome domain. We present two ways in which this might occur. First, coverage of the domain improves as the means of the outcome variables become larger. Second, coverage of the domain also improves if the marginal distributions have regression structures, as the addition of covariates changes marginal distributions to conditional distributions.

2. Background on Bivariate Copulas

A bivariate copula is a two-dimensional cumulative distribution function (cdf) with uniform margins $[0,1]$ and support contained in $[0,1]^2$. For detailed treatments of copulas, see Joe (1997) [8]; McNeil et al. (2005) [9]; Nelsen (2006) [10]; Trivedi and Zimmer (2007) [11]. The practical usefulness of copulas follows from Sklar's (1959) theorem [12], which holds that the copula parameterizes a multivariate distribution in terms of its marginals. Thus, for random variables y_1 and y_2 with respective marginal distributions $F_1(y_1)$ and $F_2(y_2)$, the bivariate distribution $F(y_1, y_2)$ can be expressed as

$$F(y_1,y_2) = C(F_1(y_1), F_2(y_2); \theta), \tag{1}$$

where, throughout this paper, the copula function C is assumed to be indexed by a scalar-valued dependence parameter θ.

Equation (1) provides a fairly general approach to modeling complex joint distributions. By plugging the known marginal distributions (F_1, F_2) into a copula function, the right hand side of Equation (1) provides a parametric representation of the unknown, or difficult to work with, joint distribution on the left hand side. Results in this paper rely on the following three commonly-employed copulas.

		θ domain	Kendall's τ
Gaussian	$\Phi_G\left(\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2)); \theta\right)$	$(-1\,, 1)$	$\frac{2}{\pi}\arcsin(\theta)$
Clayton	$\left(F_1(y_1)^{-\theta} + F_2(y_2)^{-\theta} - 1\right)^{-1/\theta}$	$(0, \infty)$	$\frac{\theta}{\theta+2}$
Gumbel	$\exp\left(-(\tilde{u}_1^{\theta} + \tilde{u}_2^{\theta})\right)^{1/\theta}$	$[1, \infty)$	$1-\frac{1}{\theta}$

In this notation, the symbol Φ represents the cdf of the standard normal distribution, $\Phi_G(\cdot,\cdot)$ is the standard bivariate normal distribution with Pearson correlation θ, and $\tilde{u}_j = -\ln F_j(y_j)$. The Gaussian copula has a symmetric shape, owing to its reliance on the normal distribution. The Clayton and Gumbel copulas, by contrast, are symmetric in their arguments, but asymmetric in their tail dependence patterns, with Clayton dependence stronger in the lower tail, and Gumbel dependence concentrated in the upper tail. Because magnitudes of dependence parameters are not comparable across copulas, it is standard to convert those to measures of concordance, such as Kendall's τ.

With the focus of this paper being count outcomes, the marginals (F_1, F_2) both follow Poisson distributions, a common distributional choice in applied econometric work. (Another common choice is the closely-related negative binomial distribution, which is a Poisson with exchangeable iid heterogeneity. Due to the exchangeable iid nature of that heterogeneity, the main message of this paper also applies to negative binomial marginals).

A number of approaches to estimating copulas appear in the literature. In fully parametric settings, such as those considered in this paper, one may maximize the full likelihood function, or first maximize the marginals and then treat them as given while maximizing the likelihood for θ (Joe (2005) [13]). Genest et al. (1995) [14], Shih and Louis (1995) [15], and Kim et al. (2007) [16] advocate a two-step approach in which the marginals are estimated nonparametrically using empirical distributions. McNeil et al. (2005) [9] (Chapter 5) discuss an approach that involves first calculating Kendall's τ and then converting it to θ.

This paper opts for the aforementioned full maximum likelihood approach based on the probability mass function (pmf) version of the copula, which can be computed so long as the researcher knows (or assumes) specific forms for the marginal distributions and copula. The pmf is calculated as

$$\begin{aligned} c(F_1(y_1), F_2(y_2); \theta) &= C(F_1(y_1), F_2(y_2); \theta) - C(F_1(y_1 - 1), F_2(y_2); \theta) \\ &- C(F_1(y_1), F_2(y_2 - 1); \theta) + C(F_1(y_1 - 1), F_2(y_2 - 1); \theta). \end{aligned} \tag{2}$$

Then taking the natural logarithm of expression (2) and summing over all observations gives the log likelihood function.

3. Drawbacks of Copulas for Discrete Outcomes

If the margins (F_1, F_2) are continuous, then the corresponding copula in Equation (1) is unique. If (F_1, F_2) are not both continuous, the joint distribution function can always be expressed as (1), although in such a case the copula lacks uniqueness (see Schweizer and Sklar (1983) [17] (Chapter 6)). This usually does not pose a problem in applied settings, as researchers use copulas because the joint distribution $F(y_1, y_2)$ is either not known or is difficult to work with. Genest and Nešlehová (2007) [7] state "The fact that there exist (infinitely many) copulas for the same discrete joint distribution does not invalidate models of this sort."

A much more serious problem is that estimates of the dependence parameter θ are biased when either F_1 or F_2 is noncontinuous. Consider two variables (y_1, y_2) that arise from copula $C(\cdot, \cdot; \theta)$. Each observation (y_{1i}, y_{2i}), where i indexes observations, can be viewed as arising from a latent pair (u_{1i}, u_{2i}) where $y_{1i} = F_1^{-1}(u_{1i})$ and $y_{2i} = F_2^{-1}(u_{2i})$, and (u_1, u_2) is a random sample from the copula. When F_1 or F_2 are continuous, Genest and Nešlehová (2007) [7] show that estimates of dependence are identical for both (y_1, y_2) and (u_1, u_2). Thus, an unbiased estimate of the dependence parameter $\widehat{\theta}$ can be obtained.

However, when F_1 or F_2 is discontinuous, then the marginal distributions have jumps that cause the inverses F_1^{-1} and F_2^{-1} to have plateaus. Genest and Nešlehová (2007) [7] show that those plateaus potentially lead to biased estimates of θ. To illustrate, we borrow from their Definition 1 and Example 1 (pp. 477–479). First, Sklar's Theorem asserts that, when F_1 and F_2 are continuous, the functions $F(F_1^{-1}(u_1), F_2^{-1}(u_2))$ and $F(F_1^{-1}(u_{1\leftarrow}), F_2^{-1}(u_{2\leftarrow}))$ are the same, which is one of the important foundations of copula inference (Genest and Favre (2007) [18]). The notation $u_{j\leftarrow}$ indicates the limit of u_j as it approaches from above. But if F_1 or F_2 is discontinuous, then transformations that lead to a unique copula in the continuous case now lead to different objects, some of which are copulas, and some of which are not.

As a simple example, let y_1 and y_2 be binary variables with $\Pr(y_1 = 0) = p$, $\Pr(y_2 = 0) = q$, and $\Pr(y_1 = 0, y_2 = 0) = r < \min(p, q)$. Then,

$$F(F_1^{-1}(u_1), F_2^{-1}(u_2)) = \begin{cases} 0 & if\ u = 0\ or\ v = 0 \\ r & if\ (u,v) \in (0,p] \times (0,q] \\ q & if\ (u,v) \in (p,1] \times (0,q] \\ p & if\ (u,v) \in (0,p] \times (q,1] \\ 1 & if\ (u,v) \in (p,1] \times (q,1] \end{cases}$$

while

$$F(F_1^{-1}(u_{1\leftarrow}), F_2^{-1}(u_{2\leftarrow})) = \begin{cases} r & if\ (u,v) \in [0,p) \times [0,q) \\ q & if\ (u,v) \in [p,1) \times [0,q) \\ p & if\ (u,v) \in [0,p) \times [q,1] \\ 1 & if\ (u,v) \in [p,1) \times [q,1] \end{cases}$$

such that the two no longer coincide (see Proposition 1 in Genest and Nešlehová (2007) [7] (p. 479) for an elaboration on this idea).

Various methods have been proposed to accommodate discrete margins, including Bayesian data augmentation (Smith and Khaled (2012) [19]) and continuous extensions of discrete variables (Denuit and Lambert (2005) [20]). The remainder of this paper illustrates that, in count data settings, the identification problem diminishes if the count outcomes more completely cover the outcome domain, such as when means increase or the model has a regression structure.

4. "Ties" in Count Variables

For count variables, one way to think about the identification problem is in terms of "ties", where multiple observations of an outcome measure assume the same value (Li et al. (2016) [21]; Pappadà et al. (2016) [22]). Naturally, a count outcomes with many ties also tends to have poor coverage of the outcome domain. Denuit and Lambert (2005) [20] provide the formula for the probability of a tie for arbitrary discrete marginals. In the following notation $y_{j,k}$ denotes an observation other than $y_{j,i}$. Re-expressing the formula for count outcomes, the probability that any two independent observations are tied is

$$\begin{aligned} \Pr(\text{tie}) &= \Pr(y_{1,i} = y_{1,k}) + \Pr(y_{2,i} = y_{2,k}) - \Pr(y_{1,i} = y_{1,k}\,,\ y_{2,i} = y_{2,k}) \\ &= \textstyle\sum_{y_1=0}^{\infty}[f_1(y_1)]^2 + \sum_{y_2=0}^{\infty}[f_2(y_2)]^2 \\ &- \textstyle\sum_{y_1=0}^{\infty}\sum_{y_2=0}^{\infty}\left[\begin{array}{c} C(F_1(y_1),F_2(y_2);\theta) + C(F_1(y_1-1),F_2(y_2-1);\theta) \\ -C(F_1(y_1),F_2(y_2-1);\theta) - C(F_1(y_1-1),F_2(y_2);\theta) \end{array}\right] \end{aligned} \tag{3}$$

For simplicity, assume that y_1 and y_2 share the same mean μ. Table 1 calculates this formula for the three aforementioned copulas, each with dependence set to $\tau = 0.25$, 0.50, or 0.75, and each with Poisson marginals. (Applying the formula requires replacing the infinities with large finite numbers.) Keeping an eye toward our health economics motivation, the table intentionally focuses on small values for μ. As highlighted by Denuit and Lambert (2005) [20], the probabilities of ties appear to diminish as the means of y_1 and y_2 increase. And because the partition of the unit interval induced by the quantile functions becomes finer as μ increases, the lack of identification of θ likewise should diminish as μ increases.

Table 1. Probabilities that any two independent observations are tied, based on Equation (3).

μ	$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
	Gaussian	**Clayton**	**Gumbel**	**Gaussian**	**Clayton**	**Gumbel**	**Gaussian**	**Clayton**	**Gumbel**
0.5	0.92	0.92	0.92	0.91	0.91	0.90	0.88	0.89	0.87
0.6	0.82	0.82	0.82	0.81	0.81	0.80	0.78	0.79	0.77
0.7	0.74	0.74	0.74	0.73	0.73	0.72	0.70	0.70	0.69
0.8	0.68	0.68	0.68	0.66	0.66	0.66	0.63	0.64	0.62
0.9	0.63	0.62	0.62	0.61	0.61	0.61	0.58	0.58	0.57
1.0	0.58	0.58	0.58	0.57	0.57	0.56	0.53	0.53	0.52
1.1	0.55	0.54	0.54	0.53	0.53	0.53	0.50	0.49	0.49
1.2	0.51	0.51	0.51	0.50	0.50	0.50	0.46	0.46	0.46
1.3	0.49	0.49	0.49	0.47	0.47	0.47	0.44	0.43	0.42
1.4	0.46	0.46	0.46	0.45	0.45	0.45	0.42	0.41	0.41
1.5	0.45	0.44	0.44	0.43	0.43	0.43	0.40	0.39	0.39

Monte Carlo Evidence

This concept is illustrated by several Monte Carlo experiments. Experiments 1−4 are as follows:

- Step 1: Randomly draw simulated Poisson variates y_1 and y_2 with means μ_1 and μ_2 from the three aforementioned copulas, each with dependence set to $\tau = 0.25, 0.50$, or 0.75. The experiments consider sample sizes of $N = 100$ and $N = 2500$.
- Step 2: Estimate the copulas using the log likelihood function generated from Equation (2).
- Step 3: Replicate steps 1 and 2 1000 times, and report the mean and standard deviation of $\widehat{\theta}$.

The experiments are then repeated several times after increasing the means, all the while focusing on small-mean settings, in keeping with our health economics motivation.

Results for this set of experiments appear in the top panels of Tables 2–10. Those results show that copulas for discrete count outcomes fail to capture the true dependence magnitudes at extremely small means, which suggests lack of identification of the dependence parameter in such settings. Only in Experiment 4, where the means are larger than 1, do the estimates of $\widehat{\theta}$ fall closer to their true values. But even in Experiment 4, the Clayton and Gumbel copulas still appear to miss their true values.

Experiments 1−4 confirm the Genest and Nešlehová word of caution regarding copulas applied to discrete outcomes. The experiments also suggest that identification problems diminish as probabilities of ties decrease. However, what recourse do practitioners have who apply copulas to count data in small-mean settings? The following section provides evidence that the introduction of covariates facilitates identification.

Table 2. Gaussian with true $\theta = 0.38$ (such that $\tau = 0.25$).

			N = 100		**N = 2500**	
	μ_1	μ_2	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	0.844	0.048	0.844	0.009
Experiment 2	0.45	0.50	0.604	0.079	0.604	0.016
Experiment 3	0.75	0.80	0.436	0.099	0.434	0.021
Experiment 4	1.05	1.10	0.370	0.095	0.372	0.023
			Discrete covariate			
Experiment 5	0.15	0.20	0.382	0.186	0.378	0.037
Experiment 6	0.45	0.50	0.378	0.131	0.379	0.026
Experiment 7	0.75	0.80	0.385	0.111	0.380	0.027
Experiment 8	1.05	1.10	0.376	0.097	0.380	0.020
			Continuous covariate			
Experiment 9	0.15	0.20	0.390	0.227	0.380	0.039
Experiment 10	0.45	0.50	0.379	0.131	0.380	0.025
Experiment 11	0.75	0.80	0.384	0.110	0.380	0.025
Experiment 12	1.05	1.10	0.376	0.097	0.380	0.029

Table 3. Gaussian with true $\theta = 0.71$ (such that $\tau = 0.50$).

	μ_1	μ_2	**N = 100**		**N = 2500**	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	0.918	0.036	0.914	0.008
Experiment 2	0.45	0.50	0.805	0.048	0.802	0.010
Experiment 3	0.75	0.80	0.734	0.058	0.735	0.012
Experiment 4	1.05	1.10	0.704	0.057	0.705	0.011
			Discrete covariate			
Experiment 5	0.15	0.20	0.705	0.123	0.711	0.023
Experiment 6	0.45	0.50	0.711	0.079	0.711	0.015
Experiment 7	0.75	0.80	0.713	0.064	0.711	0.012
Experiment 8	1.05	1.10	0.713	0.058	0.712	0.011
			Continuous covariate			
Experiment 9	0.15	0.20	0.711	0.128	0.710	0.024
Experiment 10	0.45	0.50	0.715	0.076	0.711	0.015
Experiment 11	0.75	0.80	0.715	0.063	0.710	0.013
Experiment 12	1.05	1.10	0.712	0.057	0.711	0.011

Table 4. Gaussian with true $\theta = 0.92$ (such that $\tau = 0.75$).

	μ_1	μ_2	**N = 100**		**N = 2500**	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	0.975	0.014	0.977	0.003
Experiment 2	0.45	0.50	0.942	0.022	0.944	0.005
Experiment 3	0.75	0.80	0.925	0.023	0.926	0.005
Experiment 4	1.05	1.10	0.918	0.024	0.917	0.005
			Discrete covariate			
Experiment 5	0.15	0.20	0.911	0.053	0.921	0.010
Experiment 6	0.45	0.50	0.921	0.032	0.921	0.006
Experiment 7	0.75	0.80	0.921	0.027	0.920	0.005
Experiment 8	1.05	1.10	0.924	0.023	0.920	0.004
			Continuous covariate			
Experiment 9	0.15	0.20	0.913	0.057	0.921	0.010
Experiment 10	0.45	0.50	0.921	0.031	0.921	0.006
Experiment 11	0.75	0.80	0.922	0.026	0.920	0.005
Experiment 12	1.05	1.10	0.922	0.023	0.920	0.004

Table 5. Clayton with true $\theta = 0.67$ (such that $\tau = 0.25$).

	μ_1	μ_2	**N = 100**		**N = 2500**	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
				No covariate		
Experiment 1	0.15	0.20	2.76	0.596	2.67	0.110
Experiment 2	0.45	0.50	1.05	0.331	1.00	0.060
Experiment 3	0.75	0.80	0.676	0.282	0.650	0.054
Experiment 4	1.05	1.10	0.737	0.299	0.709	0.054
				Discrete covariate		
Experiment 5	0.15	0.20	1.04	0.981	0.675	0.252
Experiment 6	0.45	0.50	0.758	0.421	0.668	0.083
Experiment 7	0.75	0.80	0.713	0.322	0.677	0.063
Experiment 8	1.05	1.10	0.716	0.279	0.671	0.053
				Continuous covariate		
Experiment 9	0.15	0.20	1.24	1.18	0.675	0.290
Experiment 10	0.45	0.50	0.752	0.435	0.677	0.089
Experiment 11	0.75	0.80	0.719	0.318	0.670	0.060
Experiment 12	1.05	1.10	0.714	0.296	0.672	0.053

Table 6. Clayton with true $\theta = 2$ (such that $\tau = 0.50$).

	μ_1	μ_2	**N = 100**		**N = 2500**	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
				No covariate		
Experiment 1	0.15	0.20	3.35	0.834	3.18	0.136
Experiment 2	0.45	0.50	1.92	0.492	1.88	0.092
Experiment 3	0.75	0.80	1.86	0.501	1.78	0.089
Experiment 4	1.05	1.10	2.15	0.514	2.10	0.095
				Discrete covariate		
Experiment 5	0.15	0.20	2.34	1.55	2.00	0.269
Experiment 6	0.45	0.50	2.11	0.763	2.00	0.150
Experiment 7	0.75	0.80	2.14	0.608	2.01	0.103
Experiment 8	1.05	1.10	2.11	0.516	2.01	0.093
				Continuous covariate		
Experiment 9	0.15	0.20	2.48	1.86	1.97	0.398
Experiment 10	0.45	0.50	2.11	0.726	2.01	0.136
Experiment 11	0.75	0.80	2.08	0.593	2.01	0.114
Experiment 12	1.05	1.10	2.10	0.509	2.01	0.095

Table 7. Clayton with true $\theta = 6$ (such that $\tau = 0.75$).

	μ_1	μ_2	N = 100		N = 2500	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	5.15	1.71	4.60	0.221
Experiment 2	0.45	0.50	4.52	1.11	4.27	0.197
Experiment 3	0.75	0.80	5.20	1.26	5.35	0.217
Experiment 4	1.05	1.10	6.63	1.52	6.35	0.254
			Discrete covariate			
Experiment 5	0.15	0.20	7.38	4.96	6.00	0.689
Experiment 6	0.45	0.50	6.59	2.10	6.03	0.332
Experiment 7	0.75	0.80	6.58	1.88	6.05	0.277
Experiment 8	1.05	1.10	6.48	1.61	6.07	0.258
			Continuous covariate			
Experiment 9	0.15	0.20	7.53	7.53	5.97	0.810
Experiment 10	0.45	0.50	6.58	1.98	6.03	0.338
Experiment 11	0.75	0.80	6.46	1.73	6.03	0.276
Experiment 12	1.05	1.10	6.34	1.56	6.05	0.279

Table 8. Gumbel with true $\theta = 1.33$ (such that $\tau = 0.25$).

	μ_1	μ_2	N = 100		N = 2500	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	3.58	0.789	3.39	0.120
Experiment 2	0.45	0.50	1.88	0.222	1.85	0.041
Experiment 3	0.75	0.80	1.47	0.150	1.46	0.028
Experiment 4	1.05	1.10	1.32	0.113	1.31	0.021
			Discrete covariate			
Experiment 5	0.15	0.20	1.40	0.238	1.33	0.038
Experiment 6	0.45	0.50	1.35	0.146	1.33	0.028
Experiment 7	0.75	0.80	1.35	0.128	1.33	0.025
Experiment 8	1.05	1.10	1.36	0.120	1.33	0.024
			Continuous covariate			
Experiment 9	0.15	0.20	1.43	0.230	1.33	0.040
Experiment 10	0.45	0.50	1.35	0.149	1.33	0.027
Experiment 11	0.75	0.80	1.36	0.135	1.33	0.025
Experiment 12	1.05	1.10	1.35	0.123	1.33	0.024

Table 9. Gumbel with true $\theta = 2$ (such that $\tau = 0.50$).

	μ_1	μ_2	**N = 100**		**N = 2500**	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	5.98	1.86	5.46	0.379
Experiment 2	0.45	0.50	2.92	0.445	2.85	0.086
Experiment 3	0.75	0.80	2.28	0.300	2.22	0.053
Experiment 4	1.05	1.10	1.97	0.204	1.95	0.040
			Discrete covariate			
Experiment 5	0.15	0.20	2.21	0.620	2.01	0.088
Experiment 6	0.45	0.50	2.06	0.327	2.01	0.057
Experiment 7	0.75	0.80	2.04	0.267	2.00	0.048
Experiment 8	1.05	1.10	2.05	0.236	2.01	0.044
			Continuous covariate			
Experiment 9	0.15	0.20	2.30	1.51	2.02	0.092
Experiment 10	0.45	0.50	2.07	0.310	2.01	0.059
Experiment 11	0.75	0.80	2.05	0.261	2.00	0.050
Experiment 12	1.05	1.10	2.05	0.230	2.01	0.044

Table 10. Gumbel with true $\theta = 4$ (such that $\tau = 0.75$).

	μ_1	μ_2	**N = 100**		**N = 2500**	
			Mean of $\widehat{\theta}$	**St. dev. of $\widehat{\theta}$**	**Mean of $\widehat{\theta}$**	**St. dev. of $\widehat{\theta}$**
			No covariate			
Experiment 1	0.15	0.20	11.7	5.30	10.2	0.772
Experiment 2	0.45	0.50	6.10	1.58	5.79	0.274
Experiment 3	0.75	0.80	4.66	0.988	4.45	0.167
Experiment 4	1.05	1.10	3.98	0.687	3.83	0.120
			Discrete covariate			
Experiment 5	0.15	0.20	52.6	79.7	4.05	0.340
Experiment 6	0.45	0.50	6.34	17.9	4.03	0.200
Experiment 7	0.75	0.80	4.32	1.01	4.02	0.159
Experiment 8	1.05	1.10	4.19	0.794	4.02	0.133
			Continuous covariate			
Experiment 9	0.15	0.20	73.3	90.3	4.05	0.349
Experiment 10	0.45	0.50	5.62	14.9	4.03	0.189
Experiment 11	0.75	0.80	4.24	0.923	4.02	0.154
Experiment 12	1.05	1.10	4.19	0.780	4.02	0.133

5. Identification Through Covariates

This section presents evidence that, even with many ties, copulas applied to count data for which the marginals are conditioned nontrivially upon covariates encounter fewer identification problems. The reason is that, with covariates, the arguments to the copula functions are expected means, rather than the outcome variables themselves, and those expected means are continuous.

To illustrate, the Monte Carlo experiments in the previous section are modified: the Poisson marginals include a single explanatory variable, denoted x, common to each marginal. We consider separately experiments in which x is a discrete dummy variable and where it is continuous. The experiments proceed as follows:

- Step 1: Randomly generate the explanatory variable x. In the discrete case, it assumes values -2 and -1 with equal probability, so that the mean is -1.5. For purposes of comparison, in the continuous case x is uniform $(-2, -1)$, so that the mean is also -1.5. The values x are generated once and held fixed for each replication of the Monte Carlo experiment.
- Step 2: Randomly draw simulated Poisson variates y_1 and y_2 from the aforementioned copulas. Rather that setting the means of y_1 and y_2 directly as in the previous section, the means are $\mu_1 = \exp(b_1 x)$ and $\mu_2 = \exp(b_2 x)$, with coefficients b_1 and b_2 specified so that μ_1 and μ_2 are the same as in Table 2. (Note: in this setup, it is not possible to generate a Poisson variable with mean smaller than 0.50 when x is a $0/1$, which is why x is rescaled to be a $-2/-1$ variable, rather than the traditional $0/1$.)
- Step 3: Estimate the copula using the log likelihood function generated from Equation (2).
- Replicate steps 2 and 3 1000 times, and report the mean and standard deviation of $\widehat{\theta}$.

These experiments appear in the middle and bottom panels of Tables 2–10. Even in Experiments 5 and 9, which have the smallest means, the addition of an explanatory variable returns an accurate estimate of dependence for the Gaussian copula. By contrast, findings are somewhat more mixed for the Clayton and Gumbel copulas. For the large-sample experiments ($N = 2500$), the Clayton and Gumbel copulas appear to accurately estimate dependence, even in low-mean settings. But in the small-sample experiments ($N = 100$), both the Clayton and Gumbel copulas appear to struggle to find their true values, although their performances do appear to improve as the means increase.

6. Discussion

Owing to their flexibility and ease of estimation, copulas have enjoyed increased usage in many areas of econometrics, but questions remain regarding identifiability of the dependence parameter when modeling discrete outcomes. Genest and Nešlehová (2007) [7] present evidence that copulas are not identified in discrete settings, particularly when those discrete outcomes follow count distributions. This paper argues that those concerns diminish if the model has a regression structure and sufficient variation is induced in $E[y \mid x]$. The same could be true in the event that the count outcomes are influenced by unobserved heterogeneity, which is tantamount to having unobserved regressors. However, asymmetric copulas, such as Clayton and Gumbel, appear to require larger datasets before the benefits of large means and/or covariates manifest themselves.

Author Contributions: The authors contributed equally to research and writing.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Ophem, H. A general method to estimate correlated discrete random variables. *Econom. Theory* **1999**, *15*, 228–237.
2. Cameron, A.C.; Li, T.; Trivedi, P.; Zimmer, D. Modeling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *Econom. J.* **2004**, *7*, 566–584.
3. Zimmer, D.; Trivedi, P. Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand. *J. Bus. Econ. Stat.* **2006**, *24*, 63–76.
4. Bien, K.; Nolte, I.; Pohlmeier, W. An inflated multivariate integer count hurdle model: An application to bid and ask quote dynamics. *J. Appl. Econom.* **2011**, *26*, 669–707.
5. Winkelmann, R. Copula bivariate probit models: With an application to medical expenditures. *Health Econ.* **2012**, *21*, 1444–1455.
6. Marshall, A. Copulas, marginals, and joint distributions. *Lect. Notes Monogr. Ser.* **1996**, *28*, 213–222.
7. Genest, C.; Nešlehová, J. A primer of copulas for count data. *Astin Bull.* **2007**, *37*, 475–515.
8. Joe, H. *Multivariate Models and Dependence Concepts*; Chapman & Hall: New York, NY, USA, 1997.

9. McNeil, A.; Frey, R.; Embrechts, P. *Quantitative Risk Management: Concepts, Techniques, and Tools*; Princeton University Press: Princeton, NJ, USA, 2005.
10. Nelsen, R.B. *An Introduction to Copulas*, 2nd ed.; Springer Verlag: Berlin, Germany, 2006.
11. Trivedi, P.; Zimmer, D. *Copula Modeling: An Introduction for Practitioners*; Now Publishers Inc.: Delft, The Netherlands, 2007.
12. Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **1959**, *8*, 229–231.
13. Joe, H. Aymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivar. Anal.* **2005**, *94*, 401–419.
14. Genest, C.; Ghoudi, K.; Rivest, L.-P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **1995**, *82*, 543–552.
15. Shih, J.; Louis, T. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **1995**, *51*, 1384–1399.
16. Kim, G.; Silvapulle, M.; Silvapulle, P. Comparison of semiparametric and parametric methods for estimating copulas. *Comput. Stat. Data Anal.* **2007**, *51*, 2836–2850.
17. Schweizer, B.; Sklar, A. *Probability Metric Spaces*; North-Holland: New York, NY, USA, 1983.
18. Genest, C.; Favre, A.-C. Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **2007**, *12*, 347–368.
19. Smith, M.; Khaled, M. Estimation of copula models with discrete margins via Bayesian data augmentation. *J. Am. Stat. Assoc.* **2012**, *107*, 290–303.
20. Denuit, M.; Lambert, P. Constraints on concordance measures in bivariate discrete data. *J. Multivar. Anal.* **2005**, *93*, 40–57.
21. Li, Y.; Li, Y.; Qin, Y.; Yan, J. Copula modeling for data with ties. *arXiv* **2016**, arXiv:1612.06968
22. Pappadà, R.; Durante, F.; Salvadori, G. Quantification of the environmental structural risk with spoiling ties: Is randomization worthwhile? *Stoch. Environ. Res. Risk Assess.* **2016**, doi:10.1007/s00477-016-1357-9.