Effects of Marginal Specifcations on Copula Estimation

Kazim Azam

No 1053

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

Effects of Marginal Specifications on Copula Estimation

Kazim Azam *

Vrije Universiteit, Amsterdam

August 28, 2014

Abstract

This paper studies the effect of marginal distributions on a copula, in the case of mixed discrete-continuous random variables. The existing literature has proposed various methods to deal with mixed marginals: this paper is the first to quantify their effect in a unified Bayesian setting. Using order statistics based information for the marginals, as proposed by Hoff (2007), we find that in small samples the bias and mean square error are at least half in size as compared to those of empirical or misspecified marginal distributions. The difference in the bias and mean square error enlarges with increasing sample size, especially for low count discrete variables. We employ the order statistics method on firm-level patents data, containing both discrete and continuous random variables, and consistently estimate their correlation.

JEL Classification: C11, C14, C52.

Keywords: Bayesian copula, discrete data, order statistics, semi-parametric,

*The author would like to thank Gianna Boero, Michael Clements, Michael Pitt, Robert Kohn, Alessandro Palandri and Federica Liberini for their invaluable comments. All the suggestions from the participants of the Econometrics workshop at Vrije Univeriteit, Amsterdam are appreciated, in particular special thanks to Andre Lucas and Siem Jan Koopman. Address correspondance: FEWEB/FIN, VU Amsterdam, The Netherlands. E-mail: kazim.azam@vu.nl.

1 Introduction

A Copula approach provides flexibility and ease for multivariate analysis (among) of marginals of different types. Nelsen (2007) and Joe (1997) provide a detailed coverage of the copula theory, and present various copula families available to practitioners. However, their benefits and simplicity rely on the fundamental requirement that the (parametric or non-parametric) marginals are of continuous type. Discrete marginals are permissible, but the uniqueness property of the copula does not hold and can also pose problems when maximizing the copula. Trivedi and Zimmer (2006) proposes to employ a continuation transformation to the discrete variable and then base the likelihood estimation on continuous copula families. With continuous margins, misspecification can be avoided by adopting a pseudo-likelihood approximation for the joint density based on the normalized ranks as proposed by Genest et al. (1995), which allows to attain consistent and asymptotically normal estimates of the copula parameters. For discrete data this method is however inappropriate, due to the ties observed in the ranks.

In the case of a multivariate analysis involving a mixture of both discrete and continuous margins, the practitioner would be unable to use a copula density or probability mass function for estimation purposes. Bayesian methods can in this case provide a possible solution. Pitt et al. (2006) proposes a Bayesian sampling scheme for discrete and continuous margins in a fully parametric Gaussian copula framework. Alternatively for discrete or mixture of discrete-continuous data, Hoff (2007) proposes a method where the marginal distributions are left completely unspecified, while being assumed to be non-decreasing functions. The uniforms obtained through a probability intergral transformation are used to estimate the copula, but are unobtainable, given the missing assumptions on the form of their distribution. The only information available is that the unknown uniforms should obey the same ranking structure as the observed data, so that the inference on the copula parameters is based on a summary statistic which is independent on the nuisance marginal parameters.

In this paper we set out a simulation to study the effects on the copula parameters estimates, when we have mixed discrete-continuous type margins. Our purpose is to evaluate how well the method proposed by Hoff (2007) performs in comparison to the case where all marginals are assumed to be empirically distributed, and to that where all discrete marginals are made continuous through the addition of a random noise. The method based on order statistics is estimated in a Bayesian framework. Also the other two marginal distributions are estimated through Bayesian techniques in order to guarantee results comparability. The novelty of our paper is the computation of the size of the bias and Mean Square Error (MSE) of the copula parameters, under different sample sizes and with various levels of heterogeneity in the discrete random variables. When using the order statistics for the marginals, for small samples, the bias and the MSE are at least half the size compared to the other two methods. With increasing sample the difference in size becomes larger, especially when one of the random variables involved is highly discrete. Misspecified marginals produce the highest bias and MSE, especially in the case of large samples.

After establishing the consistency of Hoff's method, we apply it to an empirical analysis of the joint dependence structure between a firm status of being multinational (binary variable), its expenditure in Research and Development (R&D) (continuous variable) and both its number of patents and trademarks (count variables). Such a multivariate analysis is not feasible, unless assumptions on the direction of causality between the variables of interest are made. We find our results to be in line with existing literature on firms' innovation studies.

In Section 2 we provide details about a Gaussian copula. Section 3 sets out the Bayesian sampling scheme for the marginals and the copula parameters. The simulation details are then given in Section 4 along with the results. In Section 5 we will present an application based on firms level patent data and finally concluding in Section 6.

2 Gaussian Copula Setup

We refer to Sklar's theorem (1959) for the definition of a copula. If H is the multivariate distribution of dimension p, then it can be partitioned into a copula C and the marginal distributions $F_1, \ldots, F_p$, for the random variables $Y_1, \ldots, Y_p$ given by,

$$H(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p)),$$

where $C[0,1]^p \to [0,1]$. The copula distribution can also be stated as,

$$C(u_1, \ldots, u_p) = P(U_1 \leq u_1, \ldots, U_p \leq u_p),$$

where $(u_1, \ldots, u_p)$ are the uniforms obtained through their respective univariate marginal distributions. The Gaussian copula is the most frequently employed copula and it offers to model dependence in a linear correlation manner, but does not require normal marginals (unlike the multivariate normal distribution). It is given as,

$$C(u_1, \ldots, u_p) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)),$$

where Φ is the standard normal Cumulative Distribution Function (CDF), and Φ_p is the CDF of a multivariate normal vector of dimension p. Let us denote a standard normal variable as z_j with zero mean and variance one, which is computed as,

$$z_j = \Phi^{-1}\{F_j(y_j)\}, \text{for } j = 1, \ldots, p. \tag{2.1}$$

Let $z = (z_1, \ldots, z_p)$, then we can define the multivariate normal distribution with zero mean and the covariance matrix equal to the correlation matrix Θ as,

$$z \sim N_p(0, \Theta).$$

Song (2000) states that Gaussian copula density equals

$$|\Theta|^{-1/2}\exp(-\frac{1}{2}z'\Theta^{-1}z)\exp(\frac{1}{2}z'z). \tag{2.2}$$

Equation (2.2) requires the standard normals to be computed through (2.1), where F_j is the marginal distribution for the j^{th} component. We simplify the problem by not having mixture of marginal specifications in a given multivariate analysis. That is, if F_j is specified to be parametric, then $F_{\setminus j}$ (i.e. all other marginal distributions except F_j) will also be parametrically specified as well, and vice versa in the case of non-parametric specifications.

2.1 Full Parametric Copula Specification

Let n be the total number of observations given as $y = y_1, \ldots, y_n$, for $i = 1, \ldots, n$, where each y_i is of dimension p. Then the fully parametric Gaussian copula estimation problem is given as,

$$z_i \sim N_p(0, \Theta),$$
$$y_{ij} = F_j^{-1}\{\Phi(z_{ij})|\beta_j\}, \text{ for all } i \text{ and } j,$$

where F_j is the CDF function for either a continuous or discrete random variable, and β_j is the parameter vector associated with the j^{th} component. If the j^{th} component is continuous F_j^{-1} is a one-to-one function, in case it is discrete then it is a many-to-one function and z_j have to considered as auxiliary variables and sampled along with the copula and the marginal parameters. Our estimation problem here is the same as Pitt et al. (2006), however we do not account for the presence of covariates in the marginal specification.

2.2 Semi-Parametric Copula Specification

We could also specify the marginals non-parametrically, then along with a parametric copula the estimation problem is a semi-parametric based specification. The z_j in this case will have to rely upon some rank-based information on the observed data.

2.2.1 Empirical Distribution $\widetilde{F}_{jn}$

If empirical distributions are employed for all the margins in a multivariate Gaussian copula, then there are no parameters associated to any components, and the modelling problem becomes,

$$\begin{aligned} z_i &\sim N_p(0, \Theta), \\ y_{ij} &= \widetilde{F}_{jn}^{-1}\{\Phi(z_{ij})\}, \\ \widetilde{F}_{jn}(y) &= \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}(y_{ij} \leq y), \text{for all } i \text{ and } j. \end{aligned}$$

$\widetilde{F}_{jn}$ denotes the empirical distribution used for all j instead of a parametric F_j. The division of $n+1$ is to avoid boundary cases. We only need to estimate the correlation matrix Θ and in case any of the random variable is discrete, then the z_j are sampled uniformly from the empirical step size dictated by the observed data to break the ties in the ranks.

2.2.2 Unknown F_j

Hoff (2007) presents a semi-parametric copula estimation technique, which unlike the method explained above, treats all the z_j as auxiliary variables. No assumption is made regarding F_j

and it is treated as completely unknown. The method is applicable to discrete, continuous and mixtures of discrete-continuous data types. Here unlike employing an empirical CDF for all the marginal distributions, we treat all F_j's as completely unknown and hence do not know z. The only information we have regarding F_j is that it is a non-decreasing function. We can also determine the corresponding rank for each observed y_{ij}, let the rank of y_{ij} be k, then the order statistic of y_{ij} is $y_j^{(k)}$. Therefore we know that the unobserved z_{ij} corresponding to y_{ij} has the same rank k, and can be written formally as,

$$y_j^{(k-1)} < (y_{ij} = y_j^{(k)}) < y_j^{(k+1)}, \text{implies,} \tag{2.3}$$

$$z_j^{(k-1)} < (z_{ij} = z_j^{(k)}) < z_j^{(k+1)}. \tag{2.4}$$

From (2.4), we know for certain that z_{ij} has to lie in the interval dictated by the order statistics of the observed data. Based on this information, we set out the Gaussian copula specification as,

$$z_i \sim N_p(0, \Theta),$$
$$y_{ij} = m, \text{ if } \max\Big\{z_{rj};\ F : m-1 \mapsto z_{rj}\Big\} < z_{ij} < \min\Big\{z_{rj}; F : m+1 \mapsto z_{rj}\Big\}, \text{ for all } i \text{ and } j,$$
$$\text{where } m \in M \text{ (discrete outcomes).}$$

In the case of continuous margins and large samples, the interval where z_{ij} lies in becomes smaller, and hence the uncertainty regarding the true value of z_{ij} is reduced.

3 Bayesian Estimation

We can divide the Bayesian sampling scheme into two parts, first β (for parametric margins) and $z = (z_1, \ldots, z_n)$ (if needed) are sampled conditional upon Θ, followed by sampling Θ conditional upon β and z.

3.1 First Stage $p(\beta, z|\Theta)$

3.1.1 Parametric Marginals

In case the marginals are all parametrically specified, then we sample in the following order:

1. Sample from $p(\beta_j|y_{.,j}, z_{.,\setminus j}, \Theta)$, where $y_{.,j}$ denotes all the observations n for the given component j, and $z_{.,\setminus j}$ denotes all the observations from all the other components except j.

2. If j^{th} margin's distribution F_j is continuous, then compute $z_{ij} = \Phi^{-1}\{F_j(y_{ij}|\beta_j)\}$. If F_j is a discrete distribution, we sample z_{ij} from $p(z_{ij}|\beta_j, y_{ij}, z_{i,\setminus j}, \Theta)$ for all i.

The above two steps are repeated for each j in turn. Sampling directly from conditional density of β_j is not always possible, hence Metropolis-Hasting like algorithm are needed. In case a component j has a discrete marginal distribution, we first sample β_j and then conditional upon it z_{ij} are sampled from a truncated univariate Normal distribution. We refer the interested reader to Pitt et al. (2006) (page 542-544) for full details about the sampling scheme for parametric margins.

3.1.2 Non-parametric Marginals

If a semi-parametric copula approach is adopted, where no assumption regarding F_j is made, then there is no β_j to be sampled. In case an empirical distribution is assumed for a discrete random variable, then z_{ij} corresponding to the observed y_{ij} is sampled uniformly through the interval,

$$\Phi(z_{ij}) \sim \text{Unif}\,[\widetilde{F}_j(y_{ij}-1), \widetilde{F}_j(y_{ij})], \text{ for all } i \text{ and } j,$$

where $u_{ij} = \Phi(z_{ij})$. To be clear, z_{ij} is conditionally independent of Θ and uniform sampling is only to break the rank ties. To employ the approach set out by Hoff (2007), we need z_{ij} to be sampled from,

$$z_{ij} \sim p(z_{ij}|z_{i,\setminus j}, y_j^{(k)}, \Theta), \text{ for all } i \text{ and } j,$$

where the conditional density of z_{ij} is conditioned on the correlation matrix Θ and all the standard normals corresponding to the other random variables. The conditioning of $y_j^{(k)}$ implies z_{ij} has to lie in the interval $[z_j^{(k-1)}, z_j^{(k+1)}]$, that is it has to obey the order statistics. Hoff (2007) specifies a full conditional distribution for z_{ij}, which is a truncated univariate Normal distribution with mean and variance accounting for correlation between other other components

$z_{i,\setminus j}$. The major difference in sampling the z here is that the truncation is dictated by the order statistics, whereas in the parametric case (also for the empirical distribution), the truncation is given by the CDF of the discrete parametric distribution, evaluated at y_{ij} and $y_{ij} - 1$ (see Pitt et al. (2006)). This scheme is invariant to either discrete or continuous margins. The full details of the sampling of z here are provided in Hoff (2007) page 273.

3.2 Second Stage

This stage is invariant to what approach was adopted in the previous stage, all we require are the z. We can write the posterior of Θ as,

$$p(\Theta|z) \propto p(\Theta) \times p(z|\Theta).$$

Similar to Hoff (2007), we assume a semi-conjugate prior for the Gaussian copula. The prior $p(\Theta)$ is defined through V, and it has prior given as an inverse-Wishart distribution $(\nu_0, \nu_0 V_0)$, parametrized such that $E[V^{-1}] = V_0^{-1}$, where ν_0 is the degrees-of-freedom and $\nu_0 V_0$ the scale matrix. Θ is computed as,

$$\Theta_{[i,j]} = \frac{V_{[i,j]}}{\sqrt{V_{[i,i]} V_{[j,j]}}}.$$

The posterior of V can then be shown to be proportional to,

$$V|z \sim \text{inverse-Wishart}(\nu_0 + n, \nu_0 V_0 + z'z),$$

from which a sample of V can be obtained, and then Θ computed from the above transformation.

We follow the estimation approach as give by Hoff (2007) rather than Pitt et al. (2006) to compute the posterior of Θ, as our focus is not on an efficient sampling scheme, but rather on studying the effects of the marginal specifications on copula estimation.

4 Simulation

4.1 Data Generating Process

In this section we explain how to simulate data from a multivariate Gaussian copula and provide details about the Data Generating Process (DGP). Through the simulated data we test various marginal specifications and their effect on a Gaussian copula estimation. For some correlation matrix Θ_{DGP} and β, a set of generated y can be sampled as follow:

1. Sample z from $N_p(z|0, \Theta)$.
2. Compute $y_{ij} = F_j^{-1}\{\Phi(z_{ij})|\beta_j\}$, for all i and j.

Where $z = (z_1, \ldots, z_n)$, each component j has n observations given as $z_{.,j} = (z_{1j}, \ldots, z_{nj})'$. Step 2 above implies that we need to be able to compute the inverse CDF of all the chosen parametric marginal distributions. We choose $p = 3$ and alter n such that it ranges from small sample ($n = 10$) to large sample ($n = 500$). The DGP is,

$$z \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & 0.4 \\ 0.8 & 1 & 0.6 \\ 0.4 & 0.6 & 1 \end{pmatrix}\right],$$

$$\begin{aligned}
y_{.,1} &= F_1^{-1}\{\Phi(z_{.,1})|1.5\} && \Rightarrow && F_1(y_{.,1}|1.5) = \text{Exponential}\,(y_{.,1}|\lambda_1), \\
y_{.,2} &= F_2^{-1}\{\Phi(z_{.,2})|6\} && \Rightarrow && F_2(y_{.,1}|6) = \text{Poisson}\,(y_{.,2}|\lambda_2), \\
y_{.,3} &= F_3^{-1}\{\Phi(z_{.,3})|0.6\} && \Rightarrow && F_2(y_{.,1}|0.6) = \text{Bernoulli}\,(y_{.,2}|\lambda_3).
\end{aligned}$$

So the true DGP is a mixture of discrete and continuous marginals, and it stays fixed throughout the simulation.

4.2 Marginal Specifications

Using the DGP, we aim to estimate the correlation matrix Θ by three different type of marginal specifications (MS), for ease of reference we label as MS1, MS2 and MS3.

- **MS1** All three marginal distributions (F_1, F_2 and F_3) are assumed to be completely unknown. Using the order statistics of the observed data, first z, and then the correlation matrix is sampled. This is as described previously and is the method proposed by Hoff (2007).

- **MS2** Assume all the margins are empirically distributed, and compute $z_{ij} = \Phi^{-1}(u_{ij})$, where u_{ij} is uniformly sampled from the interval $[\widetilde{F}_{nj}(y_{ij} - 1), \widetilde{F}_{nj}(y_{ij})]$,

- **MS3** Perform a continuation transformation for the discrete margins, then let $ln\, z_{ij} \sim \mathcal{N}(y_{ij}|\mu_j, \sigma_j)$, for all i and j. Hence all margins are taken to be log normally distributed[1].

Hence we specify three different marginal specifications, the first two correspond to semi-parametric copula estimation, and the third to a fully parametric copula estimation. MS3 takes the discrete marginals and adds a uniform $[0, 1]$ random noise to the observed values, to make them continuous. This is an approach stated in Trivedi and Zimmer (2006), to avoid computational problems generally encountered in likelihood estimation. This transformation along with assuming log normal distribution induces a misspecification. The first margin (originally exponential in the DGP) is also misspecified by assuming a log normal distribution.

4.3 Monte Carlo

The sampling scheme described in the last section is a kind of Gibbs type sampler over the two defined stages. To obtain the posterior density of Θ, we perform 6000 iterations from which every 5th iteration is saved and of the thinned sample we drop the first 200 for burn in. The autocorrelation within the final posterior sample of 1000 is below 0.05 after the 3rd lag. Our quantity of interest is the posterior mean $\mathrm{E}(\Theta|y)$ through all the marginal specifications. To analyse the properties of the various marginal specifications and their effect on the estimation of

[1] We also tried using Normal distribution for each random variable, but encountered numerical instability.

Θ, we have to obtain a distribution for the posterior mean itself, hence we employ Monte Carlo over the DGP. The size of the Monte Carlo simulation is 250, which is sufficient as convergence for the quantities computed is quick. At each Monte Carlo iteration ($s = 1, \ldots, 250$), we obtain a new sample of y through the same DGP, which can be denoted as $y^{(s)}$. We can define the Monte carlo structure as,

$$
\begin{aligned}
&\text{for } s = 1, \ldots, 250,\\
&\text{sample } y^{(s)} \text{ from the DGP},\\
&\text{obtain } \mathrm{E}\left[\Theta|y^{(s)}\right], \text{ for all MS1, MS2 \& MS3.}
\end{aligned}
$$

The above scheme is repeated for various sample sizes ($n = 10, 25, 50, 100, 250, 500$), to get an understanding of the various marginal specifications effect under different sample sizes.

After obtaining the distribution of the posterior mean of Θ for all the three marginal specifications, we compare them in terms of their bias and variance towards the true correlation matrix Θ_{DGP}, defined in Section 4. First, we compute the bias through the difference of $\mathrm{E}\left[\Theta|y^{(s)}\right]$ from Θ_{DGP}, followed by the MSE. We compute the two quantities of interest for all the marginal specifications. The values are computed as,

$$\text{Bias} = \frac{1}{S}\sum_{s=1}^{S}\mathrm{E}\left[\Theta|y^{(s)}\right] - \Theta_{\mathrm{DGP}},$$

$$\text{Mean Squared Error (MSE)} = \frac{1}{S}\sum_{s=1}^{S}\left[\mathrm{E}[\Theta|y^{(s)}] - \Theta_{\mathrm{DGP}}\right]^2.$$

Our interest is particularly in determining the performance of MS1 compared to the other specifications, so we compute the MSE ratio of MS1 with respect to MS2 and MS3,

$$\omega_{12} = \frac{\mathrm{MSE}_{\mathrm{M1}}}{\mathrm{MSE}_{\mathrm{M2}}}, \qquad \omega_{13} = \frac{\mathrm{MSE}_{\mathrm{M1}}}{\mathrm{MSE}_{\mathrm{M3}}}.$$

These quantities are computed for all the entries of the correlation matrix Θ.

4.4 Result: Bias

In Table 1, We present the results from computing the bias through using all the marginal specifications. We use a subscript on Θ to represent bivariate dependence among a particular pair, for example $\Theta_{[1,2]}$ is the correlation between the first and the second margin, namely the Exponential and Poisson distributed margins. We can see that the bias from Hoff's method $\mathrm{B}_{\mathrm{MS1}}$ is lower in all sample sizes and for all the parameters, as compared to $\mathrm{B}_{\mathrm{MS2}}$ (bias from empirically computed margins) and $\mathrm{B}_{\mathrm{MS1}}$ (bias from misspecified margins). For a sample size of $n = 10$, $\mathrm{B}_{\mathrm{MS1}}$ for all the parameters is almost half of $\mathrm{B}_{\mathrm{MS2}}$ and $\mathrm{B}_{\mathrm{MS3}}$. The smaller bias is particularly noticeable for $\Theta_{[1,3]}$ (correlation between exponential and a binary variable) of -0.1137, which is one-third of the bias from the other estimators. The size of the bias from the misspecified model MS3 is similar to MS2 for $n = 10$, which suggests that even a misspecified model can be used instead of MS2 for very small sample sizes. As n increases to 25, the bias from MS1 drops by half for all the parameters. The bias of $\Theta_{[1,2]}$ in the case of MS2 also drops by half, but for $\Theta_{[1,3]}$ and $\Theta_{[2,3]}$ (correlation between poisson and binary variables) the bias reduces slightly. This is also true for the bias from MS3. Through increasing n, we see the bias in MS1 reduces further, and the rate of reduction is faster as compared to MS2 for correlation involving the binary variable ($\Theta_{[1,3]}$ and $\Theta_{[2,3]}$).

For $\Theta_{[1,2]}$ which denotes the correlation between a continuous and high count data, the bias is almost equal for MS1 and MS2 in large samples, which points to the appropriateness of empirically computed margins for continuous and high count data. The bias from MS3 decreases at a slower rate as compared to MS2 as n increases, which shows that using misspecified margins (transforming discrete data) is not appropriate, and will produce wrong results. Overall, we see MS1 produces smaller bias as compared to there estimators, and it especially performs well for dependence analysis involving discrete data. The information contained within the order statistics and building a likelihood conditional upon this ensures an unbiased estimate for the copula parameters of interest.

Table 1: Bias for all marginal specifications

	n=10	n=25	n=50	n=100	n=250	n=500
B_{MS1}						
$\Theta_{[1,2]}$	-0.2074	-0.0794	-0.0509	-0.0312	-0.0183	-0.0108
$\Theta_{[1,3]}$	-0.1137	-0.0334	-0.0319	-0.0120	-0.0127	-0.0053
$\Theta_{[2,3]}$	-0.2177	-0.0840	-0.0639	-0.0349	-0.0222	-0.0119
B_{MS2}						
$\Theta_{[1,2]}$	-0.4153	-0.2041	-0.1184	-0.0622	-0.0352	-0.0245
$\Theta_{[1,3]}$	-0.2891	-0.2123	-0.2021	-0.1642	-0.1602	-0.1539
$\Theta_{[2,3]}$	-0.3797	-0.2658	-0.2278	-0.1910	-0.1725	-0.1639
B_{MS3}						
$\Theta_{[1,2]}$	-0.3985	-0.2389	-0.1907	-0.1525	-0.1358	-0.1262
$\Theta_{[1,3]}$	-0.2731	-0.2263	-0.2136	-0.2086	-0.1985	-0.1962
$\Theta_{[2,3]}$	-0.3995	-0.3084	-0.2990	-0.2919	-0.2790	-0.2775

4.5 Result: MSE

The MSE ratio results for MS1 against the other two specifications are reported in Table 2 for various n. For $n = 10$, the ratio ω_{12} (for Hoff's method against the empirical specification) for all the parameters is less than one, indicating that Hoff's method MS1 produces smaller variance as compared to MS2. This is also true in comparison to MS3 in ω_{13}, and the ratio is quite similar to ω_{12}. For small n we could misspecify for ease of computation, and still get reasonable results. In small n, we see the MSE ratio for the continuous and poisson variable (ω_{12}) is lower compared to other bivariate random variables estimates. This indicates that even though the sample is small, Hoff's method's estimates are close to the true ones. But when one of the variables involved is a binary variable, the uncertainty is large due to only two ranks available, and therefore the ratios ω_{13} and ω_{23} are large in small samples, implying the gain in efficiency from Hoff's method is not that substantial. As n increases, we see the MSE ratio for MS1 against the other two becomes much smaller, which implies even though we are dealing with highly discrete data and large n, we can on average estimate the parameters

Table 2: MSE ratio

	n=10	n=25	n=50	n=100	n=250	n=500
ω_{12}						
$\Theta_{[1,2]}$	0.3633	0.3203	0.3885	0.4681	0.5031	0.4275
$\Theta_{[1,3]}$	0.7640	0.6181	0.4513	0.3107	0.1854	0.0708
$\Theta_{[2,3]}$	0.6550	0.4334	0.3295	0.2098	0.1225	0.0531
ω_{13}						
$\Theta_{[1,2]}$	0.3838	0.2374	0.1580	0.0950	0.0461	0.0233
$\Theta_{[1,3]}$	0.8354	0.5694	0.4184	0.2036	0.1249	0.0442
$\Theta_{[2,3]}$	0.5725	0.3179	0.2033	0.0958	0.0483	0.0188

more efficiently through Hoff's method. More information is present in large n about the true dependence, which MS1 captures. The ratio ω_{12} for $\Theta_{[1,2]}$ increases with n, indicating that computing margins through an empirical distribution for continuous and high count data in large n will become quite similar to Hoff's method in terms of efficiency. This is similar to the use of an empirical distribution for continuous type variables. Although when one of the variable is a binary type, the variance ratio ω_{12} starts dropping as n increases and gets close to zero for $n = 500$, which implies MS1 is more efficient compared to other specifications. When using misspecified margins MS3, which clearly implies that effect of misspecifying the margins and estimating the copula is problematic and not efficient. There are two major aspects of why Hoff's method performs so well, first the interval where each z_{ij} lies can change due to the changing bounds dictated through the ranks, and second, the z are conditioned on Θ in the full conditional probability. An interesting point to note is how for large n, MS2 and MS3 have similar MSE ratios in case of low count data, which supports the case of a transformation as suggested in the previous literature.

In overall terms, using Hoff's method becomes more and more appropriate, when the random variable involved is of low count. For almost continuous like random variables, we can employ an empirical distribution for the marginals with a large n.

5 Application on Firm Level Patents Data

We now present a real data application using the semi-parametric copula methodology, where the copula likelihood is based on the ranks of the observed data. Practitioners often encounter data of varying types, one example is the empirical analysis of innovation economics, where one models a firm's status as multinational headquarter (binary variable), its volume of expenditure in Research and Development (R&D) (continuous variable) and its stock of patents and trademarks (count variable). The literature on innovation typically refers to the process that transforms research (proxied by R&D expenditure) into new technology (proxied by number of patents) as a knowledge production function. This literature has well established the fact that multinational firms are major producers of new technologies (see Gilroy (1993) and Javorcik et al. (2010) among others). While at the same time R&D investments and registered patents positively affect a firm's value (see Cockburn and Griliches (1988) and Blundell et al. (2002) among others). Similar considerations apply to the number of trademarks (Sandner and Block (2011)), particularly when accounting for the investors' valuation of firms that, like most multinationals, are able to protect their trademark portfolio.

The standard empirical method used to approach these variables would be to assume a standard parametric distribution (multinomial, poisson, etc.) and to choose a response variable which can provide closed-form conditional mean equations. For instance a practitioner interested in identifying whether intensive innovation accelerates the growth of the firm, and provides it with sufficient market leadership to become a multinational, might formulate a Probit regression with the binary status of multinational headquarter as a dependent variable. At the same time, we could use the status of multinational headquarter as an explanatory variable in a standard Gaussian type regression, attempting an identification of the R&D expenditure determinants. Finally, following the knowledge production function literature, we could estimate a count model where the R&D expenditure is the input of an innovation generating process. All these approaches utilize the same variables, but the conditional expectation results on these two variables (keeping other variables fixed) through a Probit or a Gaussian regression would not correspond to the same joint distribution. A second problem one might encounter is the endogeneity issue among such variables, which can be difficult to correct, particularly if the specified model has a discrete outcome. Joint dependence estimation provides an alternative for the statistical analysis of the economic relationship between these variables, without needing instruments to correct endogeneity or from the outset choosing a response variable. Through this identified joint distribution one can then proceed with causality analysis, where the marginals are recovered from the copula.

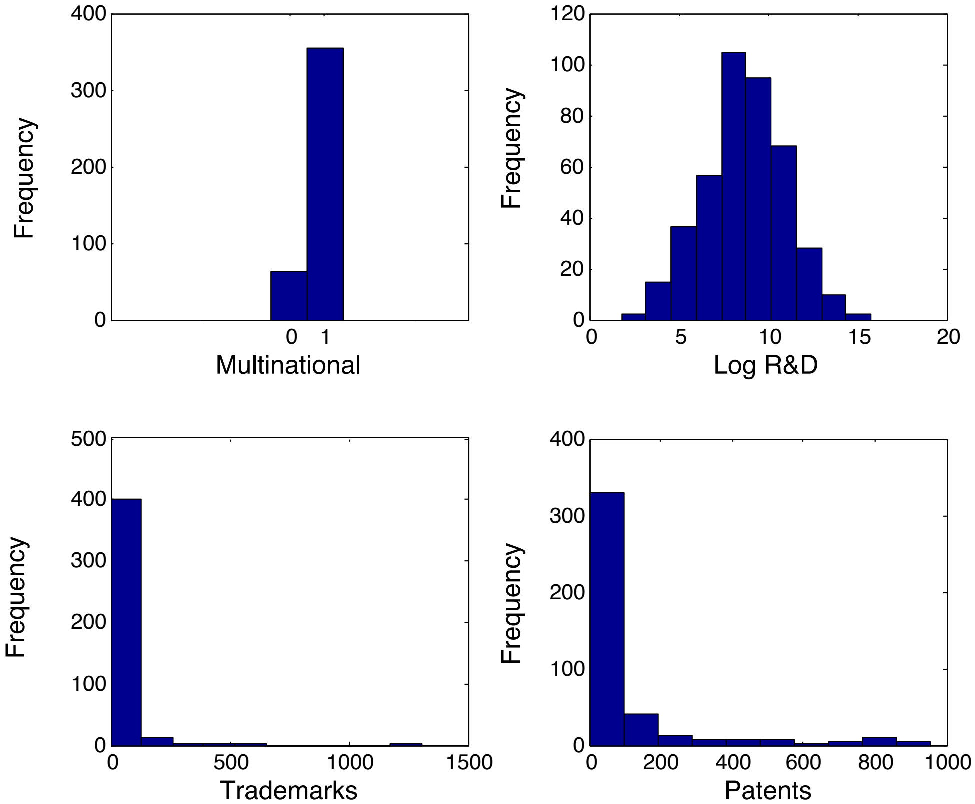
We use the method proposed by Hoff (2007) to understand the multivariate dependence structure among the firm's characteristics, without making any assumption on the direction of the effects. The Pearson correlation coefficients are analogous to the multivariate Gaussian copula under the assumption of normality, but would produce biased estimates. Methods based on rank correlations would work well for non-normal data (such as Kendall's tau and Spearman's rho), but in the presence of ties in the ranks (discrete data) these methods would not provide accurate correlation measure.

We employ the order statistics based multivariate Gaussian copula to compute the correlation between being a multinational (binary), number of patents (count), number of trademarks (count) and R&D expenditure (continuous). We collected firm's level data from Bureau van Dijk's Amadeus, and integrate them with European Patent Office's PATSTATA data. We selected the sample of all European firms defined as Ultimate Owners: these are either standalone firms or parents of corporate groups, some of which are multinational [2]. For each firm, we collected information about the number of patents and trademarks directly owned[3] in 2012, and the R&D expenditure reported in the five years that preceded 2012.

[2]To define an Ultimate Owner, BvD analyses the shareholding structure of a company. It looks for the shareholder with the highest direct or total % of ownership. If this shareholder is independent, it is defined as the Ultimate Owner of the subject company, if the highest shareholder is not independent, the same process is repeated to him until BvD finds an Ultimate Owner.

[3]We count exclusively the number of patents whose ownership is registered at the observed firm, and do not account for the number of patents registered at any of the firm's subsidiaries

Figure 1: Data Histogram



Excluding the firms in the top 5 percentiles of both size and patents distribution, as well as the firms that did not disclose R&D information, reduces the sample to 417 observations. The histogram, presented in Figure 1, gives graphic support to the inappropriateness of standard parametric distributions for trademarks and patents and then proceed with a multivariate normal density analysis.

Table 3: Firm's Innovation (Gaussian copula, MS1)

	Multi	R&D	Trademarks	Patents
Multi	1	-	-	-
R&D	0.57	1	-	-
Trademarks	0.47	0.40	1	-
Patents	0.41	0.48	0.47	1

Table 3 presents the correlation results (posterior mean) for the multivariate Gaussian copula based on MS1 marginals. The Bayesian method performs well and similarly to the simulation procedure, appropriate thinning and burn-in was done to ensure independent posterior sample. We see the correlation is quite high, indicating a strong relationship among the considered firms' characteristics. The correlation between being a multinational and R&D is the strongest, which is quite expected as multinationals sustain large R&D spending as a result of being international, benefiting from comparative advantage and holding the position of market leader and innovator. Similar to Sandner and Block (2011), we find support for the claim that being multinational is more strongly related to number of trademarks (0.47) as compared to number of patents (0.40). It is also expected that R&D is more strongly related to a high number of patents, as proposed in a knowledge production function, and that this relationship is stronger compared to R&D and number of trademarks, which represents direct product protection.

To conclude, the estimation of joint dependence allows us to obtain results qualitatively equivalent to those proposed by the existing literature. Without making strong assumptions about direction of causality, we conclude that firms' innovation mechanisms are strongly related to their status (multinational) and R&D spending.

6 Conclusion

Multivariate analysis among random variables of diverse type can be problematic. In this paper, we evaluated the effect on a Gaussian copula estimation due to various marginal distribution specifications, when we have mixed discrete-continuous variables. In particular, we study the approach of Hoff (2007), where the marginals are left completely unspecified. Along with Hoff's method, two more specifications are employed, one where the marginals were empirically computed, and completely misspecifiying the marginal distributions. The results show that Hoff's method outperforms the other two specifications in all sample sizes. The bias is half as compared to the other methods, and it quickly goes to zero with increasing sample size. Using empirical distribution is quite reasonable, but for low count data the bias persist even with increasing sample. For misspecified margins, regardless of discrete or continuous variable, the bias is large and persistent. In terms of MSE, again Hoff's method has the smallest variance compared to the other two specifications. In small sample, the MSE is similar for correlation estimates of discrete data, but the ratio approaches zero as the sample size increases. For the case of continuous and high count data, the MSE ratio between Hoff's method against empirically computed margins increases as the sample size increases. In case one of the random variable is of a binary type, the MSE through misspecified margins is similar to the empirically computed ones. We also apply the order statistics based method on firm level patent data, and consistently estimate the correlation among vital firms' characteristics, and the results coincide with the existing literature.

References

Blundell, Richard; Griffith, Rachel, and Windmeijer, Frank. Individual effects and dynamics in count data models. *Journal of Econometrics*, 108(1):113 – 131, 2002.

Cockburn, Iain and Griliches, Zvi. Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents. *American Economic Review*, 78(2):419–23, May 1988.

Genest, C.; Ghoudi, K., and Rivest, L.-P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

Gilroy, Bernard Michael. Book review: John h. dunning multinational enterprises and the global economy. Mpra paper, University Library of Munich, Germany, 1993.

Hoff, Peter D. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.*, 1(1): 265–283, 2007.

Javorcik, Beata; Neary, Peter; Eckel, Carsten, and Iacovone, Leonardo. Multi-product firms at home and away: Cost- versus quality-based competence. Economics Series Working Papers 522, University of Oxford, Department of Economics, December 2010.

Joe, H. *Multivariate Models and Dependence Concepts.* Chapman & Hall/CRC, 1997.

Nelsen, R. B. *An Introduction to Copulas.* Springer, 2007.

Pitt, Michael; Chan, David, and Kohn, Robert. Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3):537–554, September 2006.

Sandner, Philipp G. and Block, Joern. The market value of R&D, patents, and trademarks. *Research Policy*, 40(7):969–985, September 2011.

Song, Peter Xue-Kun. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

Trivedi, Pravin K. and Zimmer, David M. Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111, 2006. ISSN 1551-3076.