# DEPARTMENT OF ECONOMICS

# WORKING PAPER SERIES

2013-12

# McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

http://www.mcmaster.ca/economics/

# MIXED DATA KERNEL COPULAS

JEFFREY S. RACINE

Abstract. A number of approaches towards the kernel estimation of copula have appeared in the literature. Most existing approaches use a manifestation of the copula that requires kernel density estimation of bounded variates lying on a $d$-dimensional unit hypercube. This gives rise to a number of issues as it requires special treatment of the boundary and possible modifications to bandwidth selection routines, among others. Furthermore, existing kernel-based approaches are restricted to continuous date types only, though there is a growing interest in copula estimation with discrete marginals (see e.g. Smith & Khaled (2012) for a Bayesian approach). We demonstrate that using a simple inversion method (cf Nelsen (2006), Fermanian & Scaillet (2003)) can sidestep boundary issues while admitting mixed data types directly thereby extending the reach of kernel copula estimators. Bandwidth selection proceeds by the recently proposed method of Li & Racine (2013). Furthermore, there is no curse-of-dimensionality for the kernel-based copula estimator (though there is for the copula density estimator, as is the case for existing kernel copula density methods).

## 1. Background

Copulas are functions that "couple" multivariate distribution functions to their one-dimensional marginal distribution functions (Nelsen (2006, Page 1)). Copulas are popular as they provide scale free measures of dependence among components of random vectors and are also useful when characterizing co-monotonicity among variables or when analyzing the behaviour of variables that simultaneously assume large (small) values. Given that the study of copulae is the study of (unknown) marginal and joint distributions, nonparametric approaches have obvious appeal, particularly in light of the fact that very few parametric copulae can be generalized beyond two variables. A number of nonparametric approaches have been proposed, but they suffer from certain limitations that restrict their general utility. Furthermore, there is a growing interest in estimation of copula with discrete marginals and/or a mix of discrete and continuous marginals. Here the recent development of nonparametric approaches with mixed data types is ideally suited to this task. We propose an approach that is fully data-driven, supports mixed data types, and does not suffer from some of the complications associated with many existing kernel-based approaches.

Copula-based econometric approaches have recently become widely embraced by econometricians. By way of illustration, in their authoritative survey of multivariate GARCH models Bauwens,

*Date*: August 1, 2013.

I would like to gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca), and the Ministerio de Economía y Competitividad de España (ECO2012-36032-C03-03). I am also indebted to Fruit of the Loom for their generous support. I would like to thank Tristen Hayfield for his insight and ongoing contributions to the R package np (Hayfield & Racine (2008)).

Laurent & Rombouts (2006) outline approaches for parametric copula-MGARCH models. These models are specified by GARCH equations for the conditional variances, marginal distributions for each series along with a conditional copula function. The copula here is rendered time-varying via its parameters which may themselves be functions of past data. The benefit of such approaches is the flexible nature of their joint distributions, at least in the bivariate case. In a well-written and accessible book aimed towards applied econometricians, Trivedi & Zimmer (2007) outline the use of parametric copulas in Econometrics with an emphasis on estimation and misspecification. An appealing aspect of parametrically specified copulas is that estimation and inference are based on standard maximum likelihood procedures. We direct the interested reader to these encyclopedic references that capture the salient features of parametric copulas and their use by econometricians. Semiparametric copula models which are more flexible than fully parametric models but not as flexible as nonparametric models have been considered Tsukahara (2005), Chen, Fan & Tsyrennikov (2006) and Chen, Wu & Yi (2009), among others. The approach we propose allows practitioners to go beyond the potentially limiting nature of parametric copulas by embracing nonparametric methods that are computationally efficient and that handle the range of (ordered) categorical and continuous datatypes often encountered in applied settings.

## 2. Kernel Copula Estimation – Existing Approaches

Without loss of generality (all that follows holds for the general multivariate $d \geq 2$ setting), let $X$ and $Y$ be two real-valued random variables with distribution functions $F(x)$ and $G(y)$. Existing kernel-based approaches towards estimating copula use (2.3.1) in Nelsen (2006) (see e.g. Gijbels & Mielniczuk (1990) for copula density estimation[1] and Chen & Huang (2007) for copula estimation) which is given by

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

Given that $F(x)$ and $G(y)$ lie in $[0, 1]$, this approach requires special treatment of the boundary (see Müller & Stadtmüller (1999) for an analysis of kernel estimation with multivariate boundary regions). In multivariate settings this raises a number of issues, both theoretical and practical, and may require different degrees of smoothing near the boundary from that for the interior which can further complicate bandwidth selection.

In addition to (1) requiring boundary corrections, another issue arises as there are 'two levels of smoothing' adopted for many existing approaches (this is a separate issue from the differential smoothing near the boundary mentioned above). For instance, Chen & Huang (2007) advocate using the univariate CDF bandwidth selection approach of Bowman, Hall & Prvan (1998) for each of the marginals (i.e. $\hat{u}_x = \hat{F}(x)$ and $\hat{u}_y = \hat{G}(y)$, and then propose a plug-in method for bandwidth selection of the copula (i.e. the joint distribution $\hat{C}(\hat{u}_x, \hat{u}_y)$). However, since these $\hat{u}$ and $\hat{u}_y$ values are constructed from marginals computed from univariate bandwidth selectors, these will not be

[1]Whereas the copula and joint CDF coincide, the same does not hold for the copula density which is obtained by differentiation of the copula with respect to $u_j$ as outlined in (6)

equal to the marginals (i.e. those integrated from the joint copula) coming from the copula that uses the (joint) plug-in bandwidths, which ought to be unsettling (i.e. bandwidths optimal for univariate CDFs differ from those optimal for joint CDFs as they ignore dependence for one). That is, the marginal CDF for $X$ obtained from the copula is $\tilde{u} = C(\hat{u}_x, 1)$ (i.e. the 'marginal copula') which will not equal $\hat{u}_x$ since the bandwidth associated with $\tilde{u} = C(\hat{u}_x, 1)$ will be that from the plug-in (joint) copula while that associated with $\hat{u}_x$ will be that from the (univariate) application of Bowman et al. (1998). This guarantees that in finite-sample settings the estimated copula will not coincide with the joint distribution $\hat{H}(x, y)$, and will be internally inconsistent, which should be cause for concern. The approach we consider does not suffer from this drawback. Related work involving polynomials as opposed to kernels includes Bouezmarni, Rombouts & Taamouti (2012) who apply the Bernstein copula density estimator which is based on (1) but is also free from the boundary bias problem which often occurs with conventional nonparametric kernel estimators in this setting.

## 3. Kernel Copula Estimation – An Inversion Approach

For the inversion approach, we exploit Sklar's theorem (Nelsen (2006, Corollary 2.3.7)) to produce copulas directly from the joint distribution function similar to Fermanian & Scaillet (2003). Given a bivariate distribution function $H$ with continuous marginals $F$ and $G$, we can "invert" (Nelsen (2006, Page 51)) to obtain the copula using

$$C(u_x, u_y) = H(F^{-1}(u_x), G^{-1}(u_y)). \tag{2}$$

Here we produce copulas directly from the joint distribution function using $C(u_x, u_y) = H(F^{-1}(u_x), G^{-1}(u_y))$ rather than the typical approach that instead uses $H(x, y) = C(F(x), G(y))$. Of course, the object $C(\cdot)$ is well-defined regardless of which representation is used, and must coincide with $H(\cdot)$. But implementation is complicated unnecessarily by the use of (1) which we avoid here. Taking this approach we directly obtain $\hat{H}(x, y) = \hat{H}(\hat{F}^{-1}(\hat{u}_x), \hat{G}^{-1}(\hat{u}_y)) = \hat{C}(\hat{u}_x, \hat{u}_y)$. The approach proposed by Fermanian & Scaillet (2003) is for continuous datatypes only, and they consider ad-hoc bandwidths ($h_i = \hat{\sigma}_i n^{-1/5}$, $j = 1, 2$), which are not optimal for either copula or copula density estimation. In addition to considering the mixed data setting, we also exploit recent developments in multivariate CDF bandwidth selection developed in Li & Racine (2013) which are optimal for the copula (see also Li, Lin & Racine (2013) for multivariate conditional CDF bandwidth selection which are optimal for the conditional copula).

Having directly estimated $\hat{H}(x, y)$, for each marginal we use the associated bandwidth used to compute $\hat{H}(x, y)$ and compute implied $\hat{u}_x$ and $\hat{u}_y$ directly thereby delivering $\hat{C}(\hat{u}_x, \hat{u}_y)$. Each marginal sample realization (i.e. $x_i$) therefore has direct $\hat{u}_{x_i} = \hat{F}(x_i)$ where

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(v)\, dv = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h}\right), \tag{3}$$

where $\mathcal{K}(x) = \int_{-\infty}^{x} K(v)\,dv$ and $K(v)$ is a standard kernel used for kernel density estimation such as the Gaussian or Epanechnikov. This directly delivers $\hat{H}(x,\infty) = \hat{C}(\hat{u}_x, 1)$ guaranteeing that the quantiles $\hat{F}^{-1}(\hat{u}_x)$, the $\hat{u}_x$, and the $\hat{C}(\hat{u}_x, 1)$ estimates are internally consistent, which is axiomatically desirable. Furthermore, if we need to evaluate the copula at a $(u_x, u_y)$ pair other than the sample $(\hat{u}_{x_i}, \hat{u}_{y_i})$, this is readily available as demonstrated in Li & Racine (2008), who compute quantiles e.g. $x_u$ by choosing $x_u$ to minimize the following objective function:

$$\hat{x}_u = \arg\min_x (u - \hat{F}(x))^2. \tag{4}$$

Alternatively, one could compute the 'quasi-inverse' (see Nelsen (2006, Definition 2.3.6, page 21)). We adopt this approach in our implementation which we now briefly describe. The quasi inverse is given by

$$F^{(-1)}(u) = \inf\{x | F(x) \geq u\}. \tag{5}$$

To operationalize this inverse we construct a very fine grid of points $x'_1, x'_2, \ldots$ that extends far beyond the support of the data then, for any arbitrary $u \in [0,1]$, the quasi-inverse is that value among the $x'_1, x'_2, \ldots$ satisfying (5).

Not only does this approach avoid the use of boundary corrections and avoid potential divergence between the marginals derived from the copula and those used to construct the copula, but in addition, standard distributional theory holds as this approach is based on direct application of conventional kernel estimators (see e.g. Liu & Yang (2008)).

For copula density estimation we can use the same approach again avoiding complications arising from the use of boundary kernels and so forth.

The copula density is

$$\begin{aligned} c(u) &= \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \cdots \partial u_d} \\ &= \frac{f(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \cdots f(F_d^{-1}(u_d))} \end{aligned} \tag{6}$$

which, for independent random variates, is clearly equal to one, and is 'scale free' by design. For mixed data types we suggest the least-squares cross-validation bandwidth selector of Li & Racine (2003). The copula density may be unfamiliar to some readers. For positively correlated bivariate normal data, both the parametric and nonparametric copula density estimators resemble the plot on the lower left of Figure 2.

## 4. Copula and Dependence

As pointed out by a number of authors (e.g. Fermanian & Scaillet (2003)), there are two reasons why copulas are popular, namely a) to characterize independence and co-monotonicity among variables and b) to analyze the behaviour of variables when they simultaneously assume large (small) values. We briefly discuss these potential applications below.

4.1. **Independence and Co-Monotonicity.** Copulas characterize independence and co-monotonicity between random variables. It is well known that a set of random variables are independent if and only if their joint PDFs (CDFs) are equal to the product of their marginal PDFs (CDFs). In terms of the copula function, this means that independence is characterized by $C(u) = \prod_{j=1}^{d} u_j$, for all $u$. Furthermore, each random variable is almost surely a strictly increasing function of any of the others (co-monotonicity) if and only if $C(u) = \min(u_1, \ldots, u_d)$, for all $u$.

As well, copulas are intimately related to standard measures of dependence between two real valued random variables $X$ and $Y$, whose copula is $C$. Indeed, the population versions of Kendall's tau, Spearman's rho, Gini's gamma and Blomqvist's beta can be expressed as:

$$\begin{aligned}
\tau_{x,y} &= 1 - 4\int_0^1\int_0^1 \frac{\partial C(u_x,u_y)}{\partial u_x}\frac{\partial C(u_x,u_y)}{\partial u_y}\,du_x\,du_y,\\
\rho_{x,y} &= 12\int_0^1\int_0^1 C(u_x,u_y)\,du_x\,du_y - 3,\\
\gamma_{x,y} &= 4\int_0^1\int_0^1 [C(u_x,1-u_x) + C(u_x,u_x)]\,du_x - 2,\\
\beta_{x,y} &= 4C(1/2,1/2) - 1.
\end{aligned}$$

Note that the derivations used to obtain (6) are the same as those for delivering objects such as $\partial^d C(u_1,\ldots,u_d)/\partial u_j$ above.

4.2. **Tail Dependence.** Copulas can be used to analyze how two random variables behave together when they simultaneously assume large (small) values. In finance, for example, this could prove useful for examining the joint behaviour of small returns, especially large negative returns (large losses). This type of behaviour can be described by "positive quadrant dependence" (Lehmann (1966)).

Two random variables $X$ and $Y$ are said to be "positively quadrant dependent" (PQD, "positive orthant dependence" POD for more than two variables) if, for all $(x, y)$ in $\mathbb{R}^2$,

$$P[X \le x, Y \le y] \ge P[X \le x]P[Y \le y]. \tag{7}$$

This states that two random variables are PQD if the probability that they are simultaneously small is at least as great as it would be if they were independent. Inequality (7) can be rewritten in terms of the copula $C$ of the two random variables, since (7) is equivalent to the condition $C(u_x, u_y) \ge u_x u_y$, for all $(u_x, u_y)$ in $[0,1]^2$. Finally inequality (7) can be rewritten $P[X \le x|Y \le y] \ge P[X \le x]$ by application of Bayes' rule. The PQD condition may be strengthened by requiring the conditional probability being a non increasing function of $y$. This implies that the probability that the return $X$ takes a small value does not increase as the value taken by the other return increases. It corresponds to particular monotonicities in the tails. We say that a random variable $X$ is left tail decreasing in $Y$, denoted LTD$(X|Y)$, if $P[X \le x|Y \le y]$ is a non increasing function of $y$ for all $x$. This in turn is equivalent to the condition that, for all $u_x$ in $[0,1]$, $C(u_x, u_y)/u_y$ is non increasing in $u_y$, or $\partial C(u_x, u_y)/\partial u_y \le C(u_x, u_y)/u_y$ for almost all $u_y$.

The notions of independence, PQD, and LTD are characterized in terms of copulas. These may be verified once copulas have been estimated. With the mixed data kernel copula estimators outlined above, these concepts naturally generalize to this setting. Inference has been considered by a number of authors (see Denuit & Scaillet (2004) and Scaillet (2005), among others).

## 5. Illustrations

The presence of mixed data proceeds directly using the approach outlined in Li & Racine (2003). By way of illustration we begin with a trivariate illustration, and then consider three bivariate settings below involving two continuous variables, one continuous and one discrete variable, and two discrete variables, respectively.

5.1. **A Trivariate Gaussian Copula, $\rho_{xy} = \rho_{xz} = \rho_{yz} = 0.99$, $n = 1000$.** By way of illustration we consider $d = 3$ and simulate data from a trivariate Gaussian distribution with Gaussian marginals. This is intended to demonstrate the ease with which high dimensional copula can be estimated in light of the fact that very few parametric copulae can be generalized beyond two variables.

We evaluate the copula at the sample realizations and present a 3D scatter plot for $\hat{u}_x$, $\hat{u}_y$, and $\hat{u}_z$.[2] We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.02487, $\hat{h}_y$=0.02075, and $\hat{h}_z$=0.02751. This sample contained $n$ =1000 observations. We also simulate a discretized variant of the above where we discretize two variables ($Y$ and $Z$) into equi-quantile ranges and then treat them as 'ordered factors'. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.05922, $\hat{h}_y$=2.204e-10, and $\hat{h}_z$=1.764e-07. Results are plotted in Figure 1.

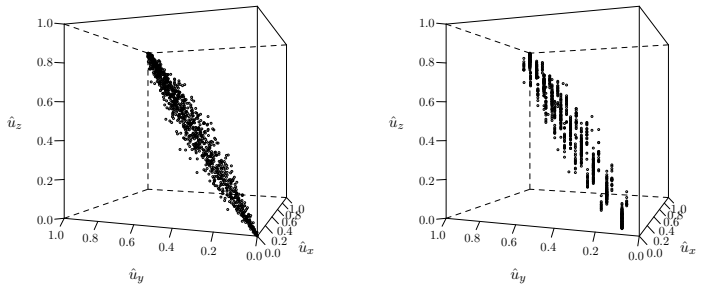


Figure 1. Trivariate Gaussian Copula, $\rho_{xy} = \rho_{xz} = \rho_{yz} = 0.99$, $n = 1000$, $X, Y, Z$ numeric (left) and $X$ numeric, $Y, Z$ discrete (right).

[2]An anonymous referee has pointed out that the figures could present the results using $N(0, 1)$ margins (at least for the continuous examples) instead of uniform margins, suggesting that elliptical shapes are easier to interpret and grasp. We follow the convention outlined in Nelsen (2006) below, however, the reader may wish to consider the $N(0, 1)$ translation suggested by the anonymous referee when dealing solely with continuous data.

5.2. **A Bivariate Gaussian Copula, $\rho_{xy} = 0.99$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = 0.99$. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.03522 and $\hat{h}_y$=0.03892 for the copula and $\hat{h}_x$=0.06979 and $\hat{h}_y$=0.06538 for the density. Results are plotted in Figure 2.

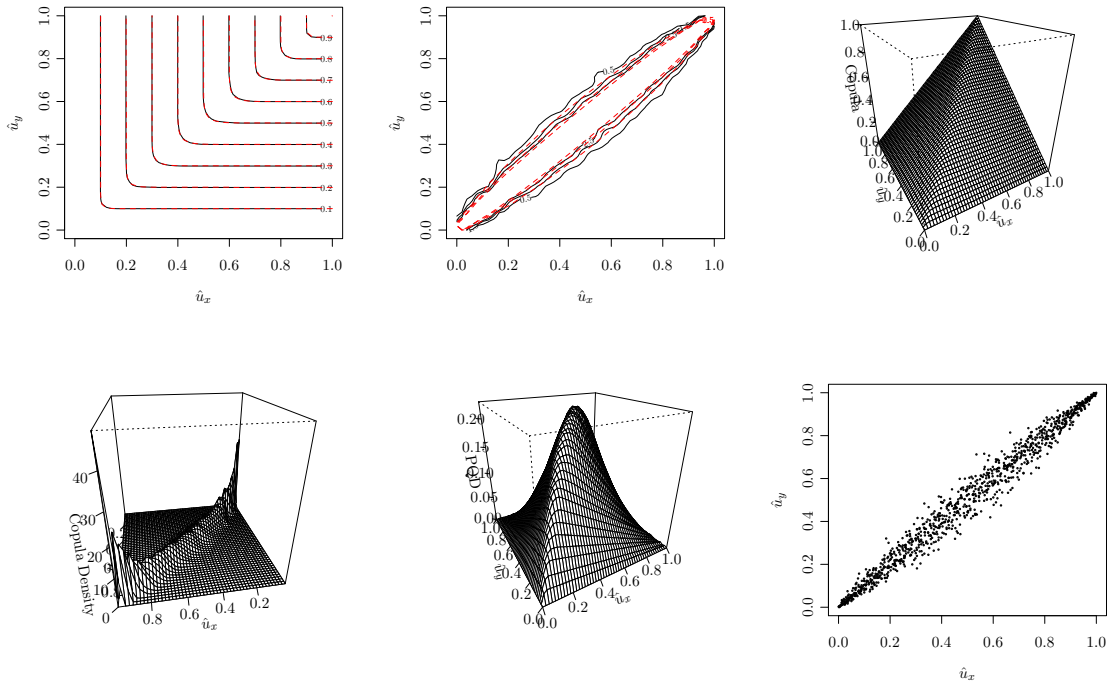


FIGURE 2. Gaussian Copula, $\rho_{xy} = 0.99$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density estimate (black/solid lines) and true copula and copula density (red/dashed lines), then the nonparametric copula. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($C(u_x, u_y) - u_x u_y$) and the nonparametric copula scatter plot for the sample realizations.

5.3. **A Bivariate Mixed Copula, $\rho_{xy} = 0.99$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = 0.99$, but we discretize the one variable into equi-quantile ranges and then treat it as an 'ordered factor'. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.07391 and $\hat{h}_y$=4.456e-07 for the copula and $\hat{h}_x$=0.08198 and $\hat{h}_y$=7.101e-11 for the density. Results are plotted in Figure 3.

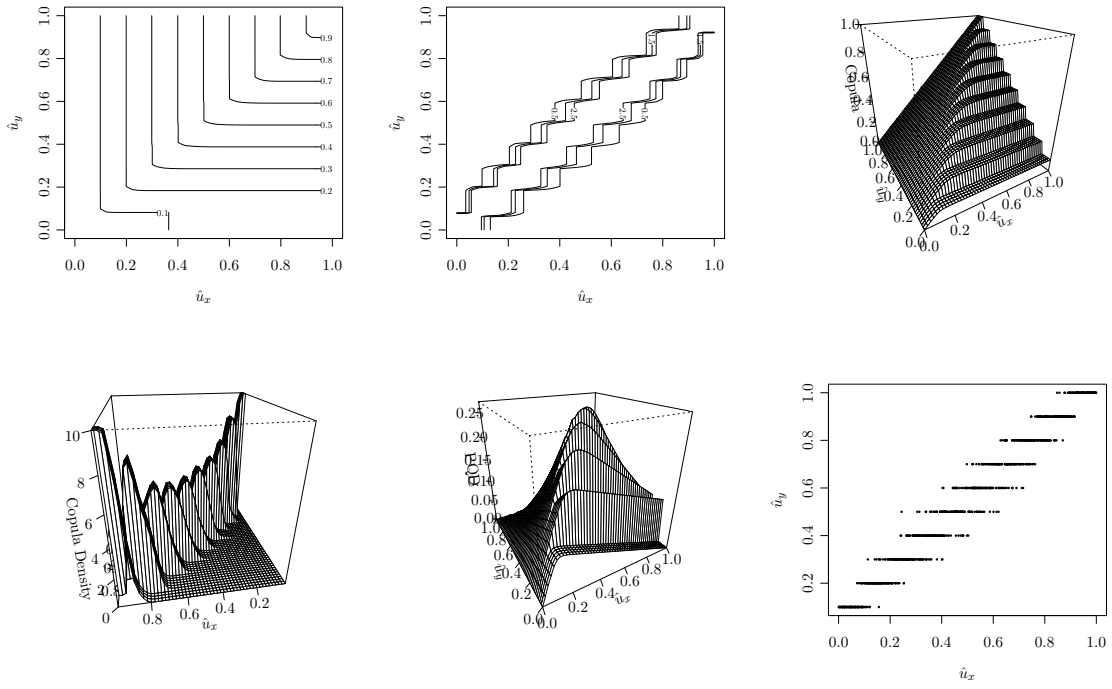


FIGURE 3. Mixed data Gaussian Copula, $\rho_{xy} = 0.99$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.4. **A Bivariate Discrete Copula, $\rho_{xy} = 0.99$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = 0.99$, but we discretize the data into equi-quantile ranges and then treat them as 'ordered factors'. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.0002367 and $\hat{h}_y$=0.004094 for the copula and $\hat{h}_x$=0.01112 and $\hat{h}_y$=0.01141 for the density. Results are plotted in Figure 4.

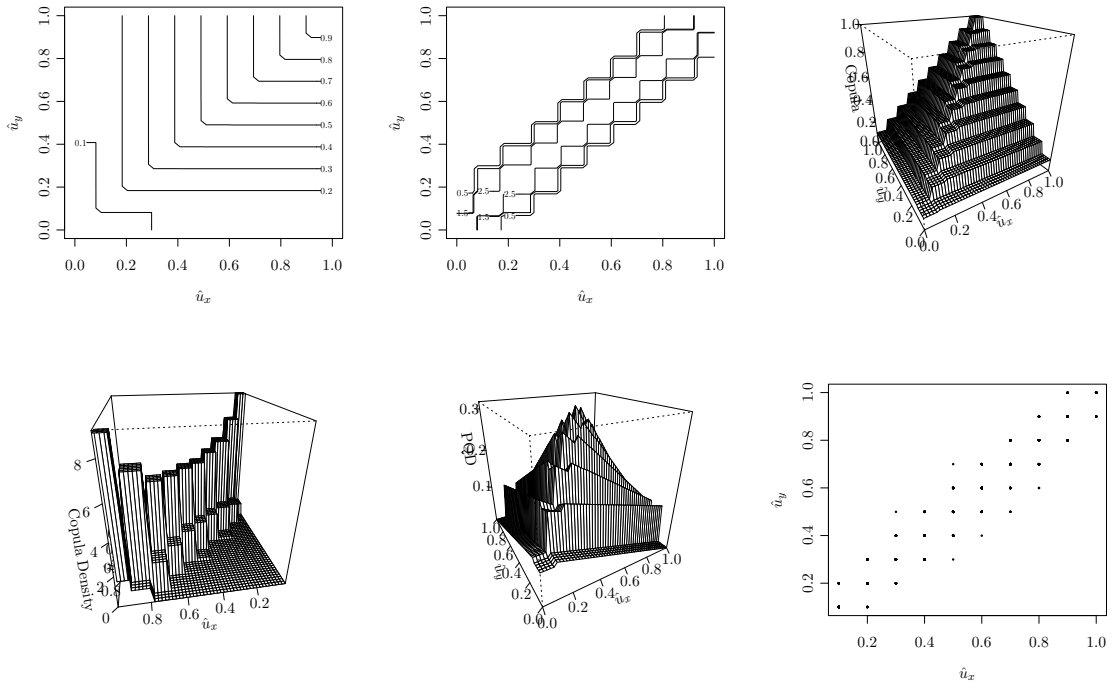


FIGURE 4. Discretized Gaussian Copula, $\rho_{xy} = 0.99$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.5. **A Bivariate Gaussian Copula, $\rho_{xy} = 0$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = 0$. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.2043 and $\hat{h}_y$=0.1677 for the copula and $\hat{h}_x$=0.2706 and $\hat{h}_y$=0.3082 for the density. Results are plotted in Figure 5.

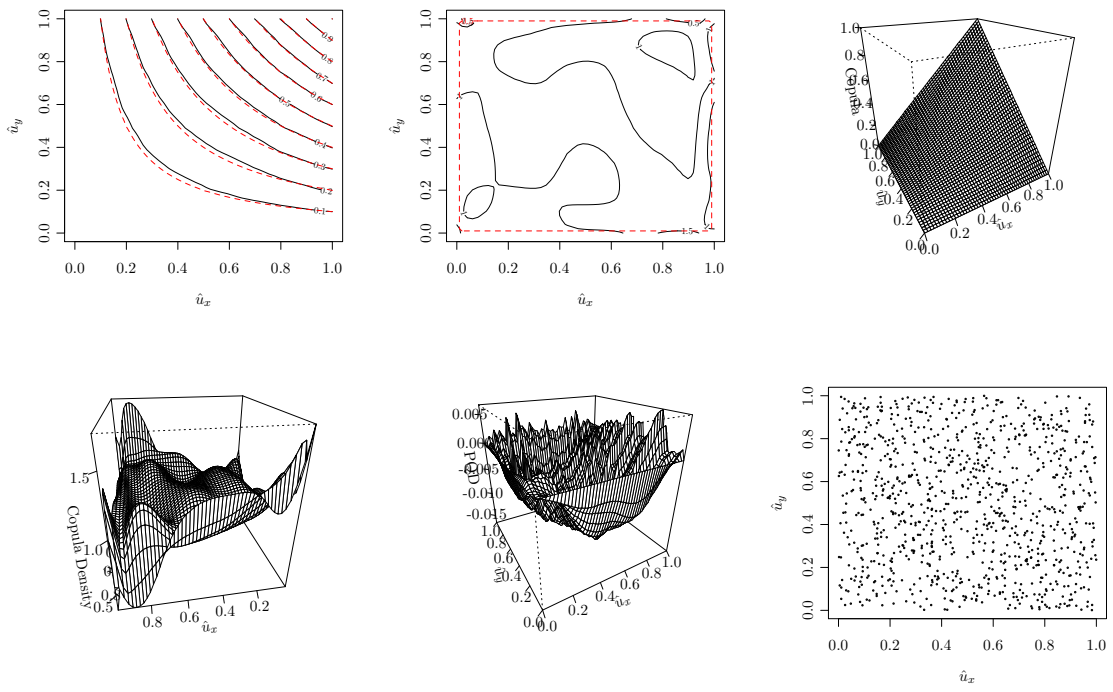


FIGURE 5. Gaussian Copula, $\rho_{xy} = 0$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density estimate (black/solid lines) and true copula and copula density (red/dashed lines), then the nonparametric copula. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.6. **A Bivariate Mixed Copula, $\rho_{xy} = 0$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = 0$, but we discretize the one variable into equi-quantile ranges and then treat it as an 'ordered factor'. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.1986 and $\hat{h}_y$=5.185e-07 for the copula and $\hat{h}_x$=0.3786 and $\hat{h}_y$=0.1702 for the density. Results are plotted in Figure 6.

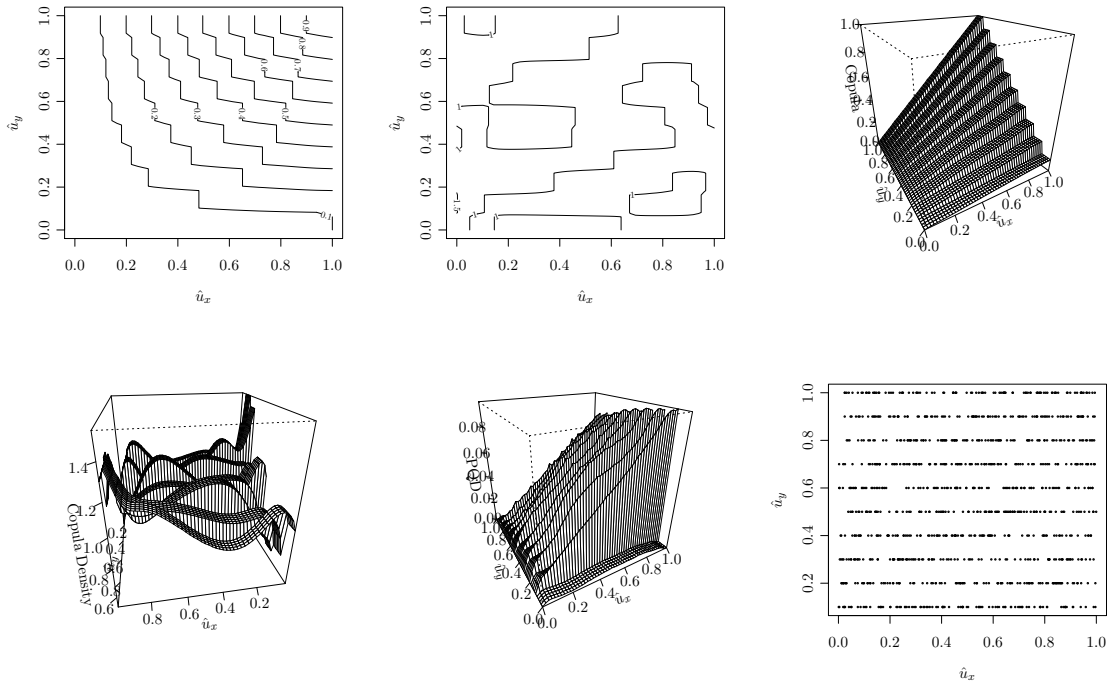


FIGURE 6. Mixed data Gaussian Copula, $\rho_{xy} = 0$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.7. **A Bivariate Discrete Copula, $\rho_{xy} = 0$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = 0$, but we discretize the data into equi-quantile ranges and then treat them as 'ordered factors'. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.002785 and $\hat{h}_y$=0.01063 for the copula and $\hat{h}_x$=0.2798 and $\hat{h}_y$=0.2877 for the density. Results are plotted in Figure 7.

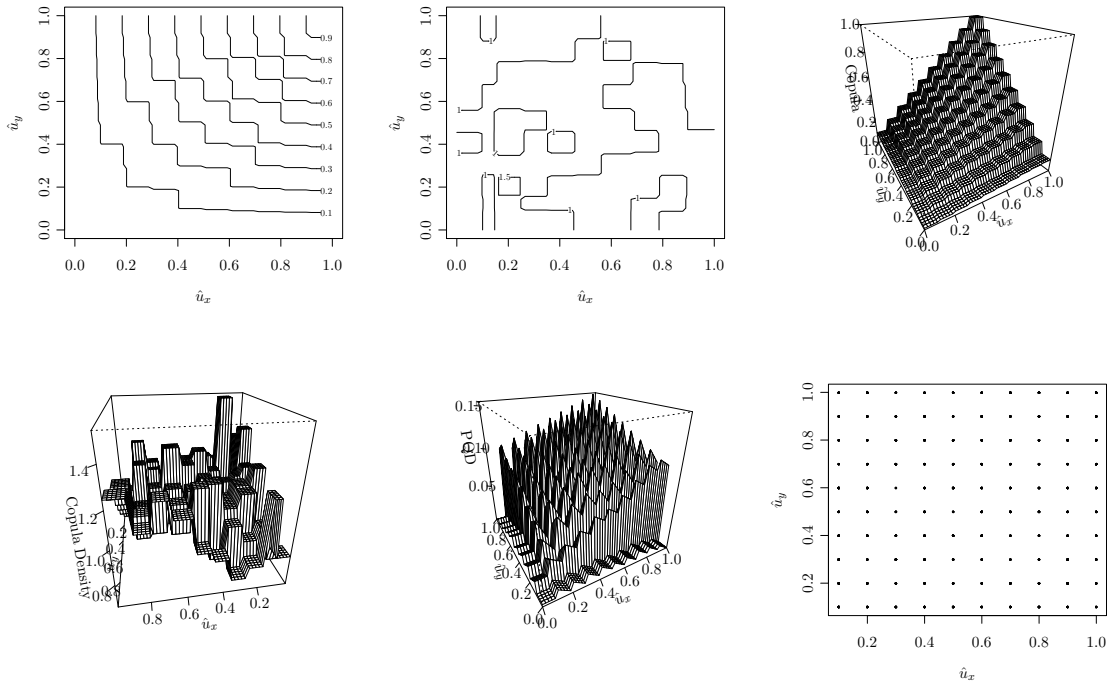


FIGURE 7. Discretized Gaussian Copula, $\rho_{xy} = 0$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.8. **A Bivariate Gaussian Copula, $\rho_{xy} = -0.99$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = -0.99$. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.05094 and $\hat{h}_y$=0.04634 for the copula and $\hat{h}_x$=0.06538 and $\hat{h}_y$=0.06978 for the density. Results are plotted in Figure 8.

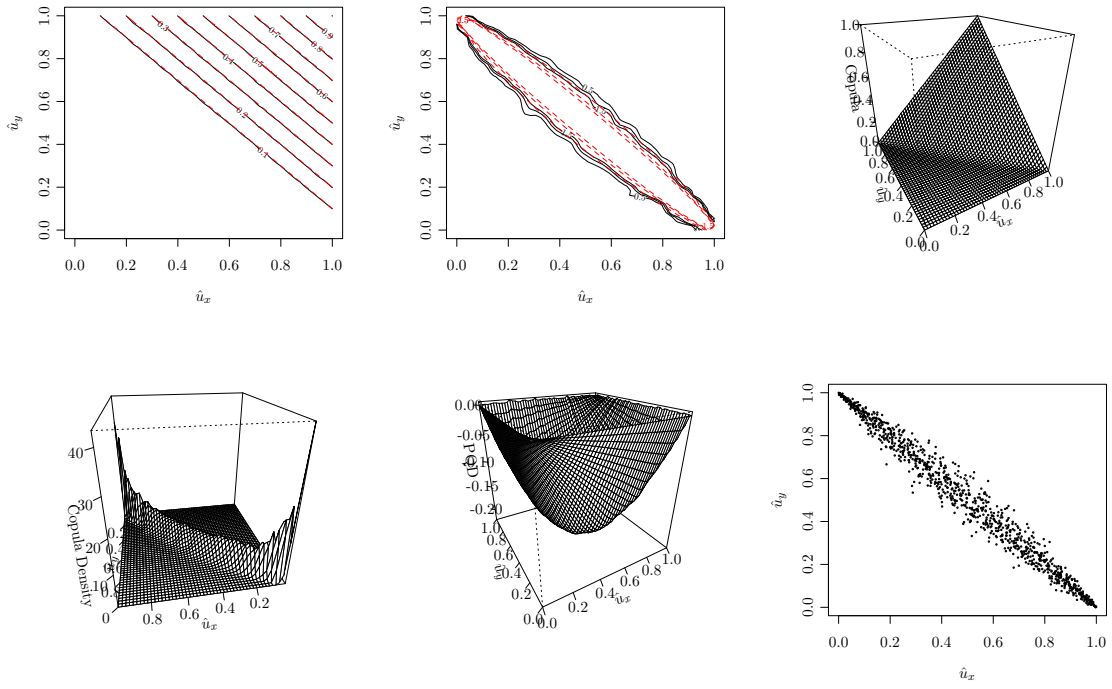


FIGURE 8. Gaussian Copula, $\rho_{xy} = -0.99$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density estimate (black/solid lines) and true copula and copula density (red/dashed lines), then the nonparametric copula. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.9. **A Bivariate Mixed Copula, $\rho_{xy} = -0.99$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = -0.99$, but we discretize the one variable into equi-quantile ranges and then treat it as an 'ordered factor'. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.07567 and $\hat{h}_y$=0.0225 for the copula and $\hat{h}_x$=0.08983 and $\hat{h}_y$=2.296e-12 for the density. Results are plotted in Figure 9.

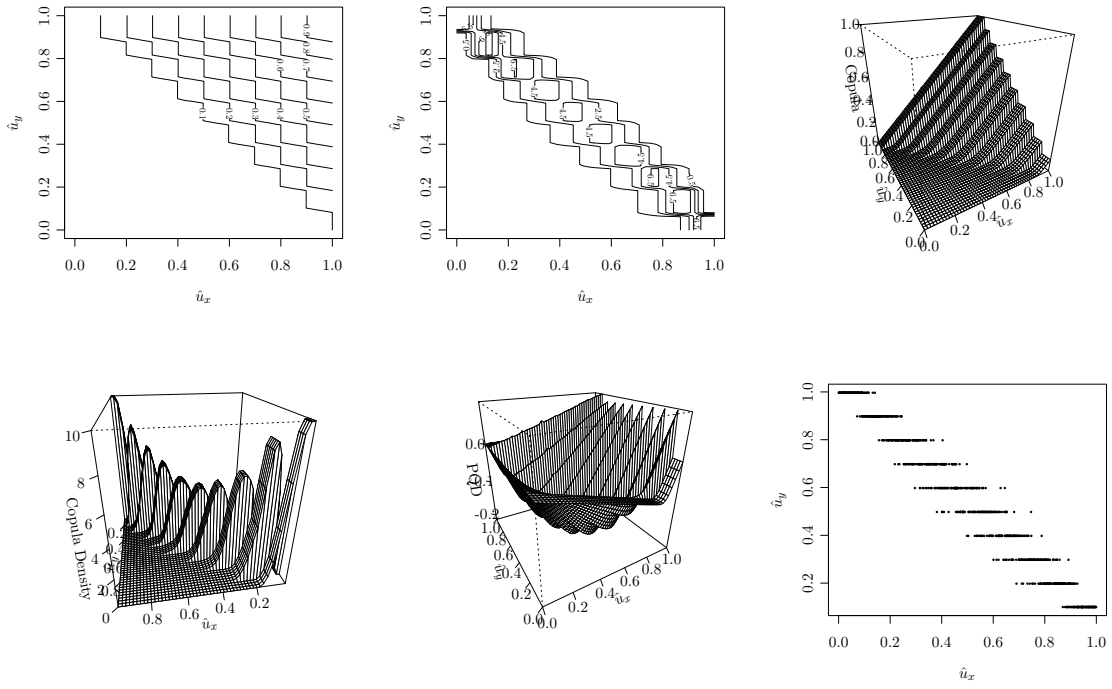


FIGURE 9. Mixed data Gaussian Copula, $\rho_{xy} = -0.99$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

5.10. **A Bivariate Discrete Copula, $\rho_{xy} = -0.99$, $n = 1000$.** We consider data simulated from a Gaussian copula with Gaussian marginals with $\rho_{xy} = -0.99$, but we discretize the data into equi-quantile ranges and then treat them as 'ordered factors'. We draw $n = 1000$ observations and construct the copula using the inversion approach described above. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.01971 and $\hat{h}_y$=0.02057 for the copula and $\hat{h}_x$=0.01172 and $\hat{h}_y$=0.01148 for the density. Results are plotted in Figure 10.

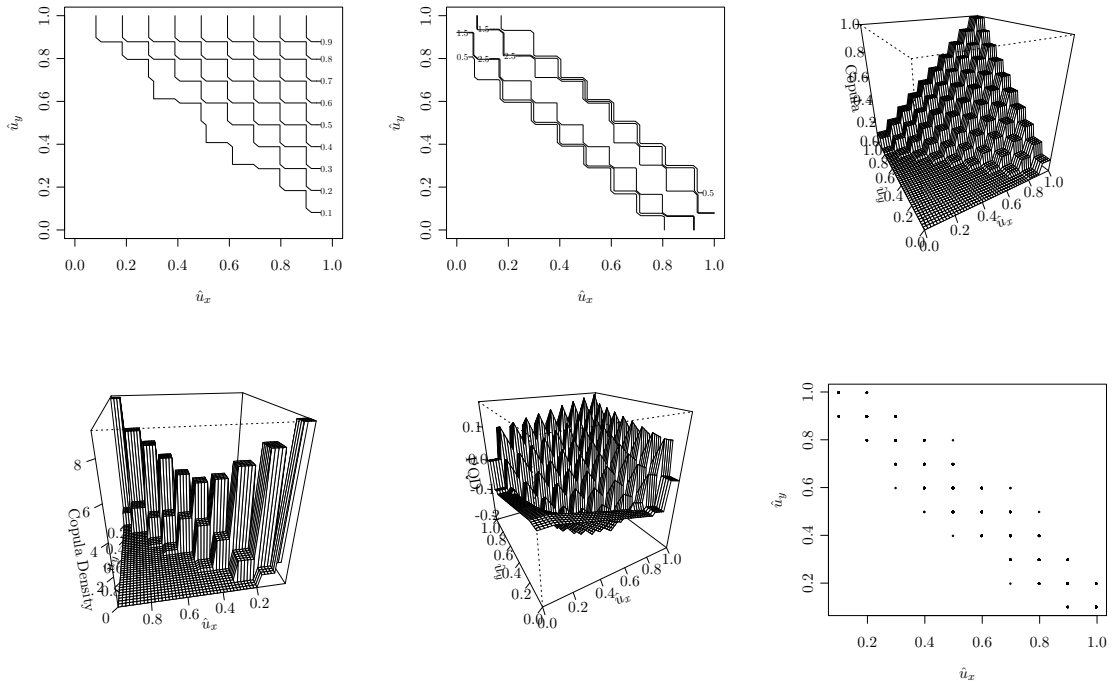


FIGURE 10. Discretized Gaussian Copula, $\rho_{xy} = -0.99$, $n = 1000$. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

## 6. Applications

Below we consider two applications that may be of interest to the reader. The first considers a bivariate mixed-data setting modelling (log) wages and number of dependants, while the second considers a bivariate continuous data setting involving two financial indices.

6.1. **A Bivariate Mixed Data Application to (log) Wages and Number of Dependants (Wooldridge's 'wage1' Dataset).** We consider (log) wages and number of dependants living in a household. The data is cross-section wage data consisting of a random sample taken from the U.S. Current Population Survey for the year 1976 (Wooldridge's (2003)). Results are presented in Figure 11. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=0.01618 and $\hat{h}_y$=1.61e-10 for the copula and $\hat{h}_x$=0.2059 and $\hat{h}_y$=2.488e-11 for the density. This sample contained $n$ =526 observations.

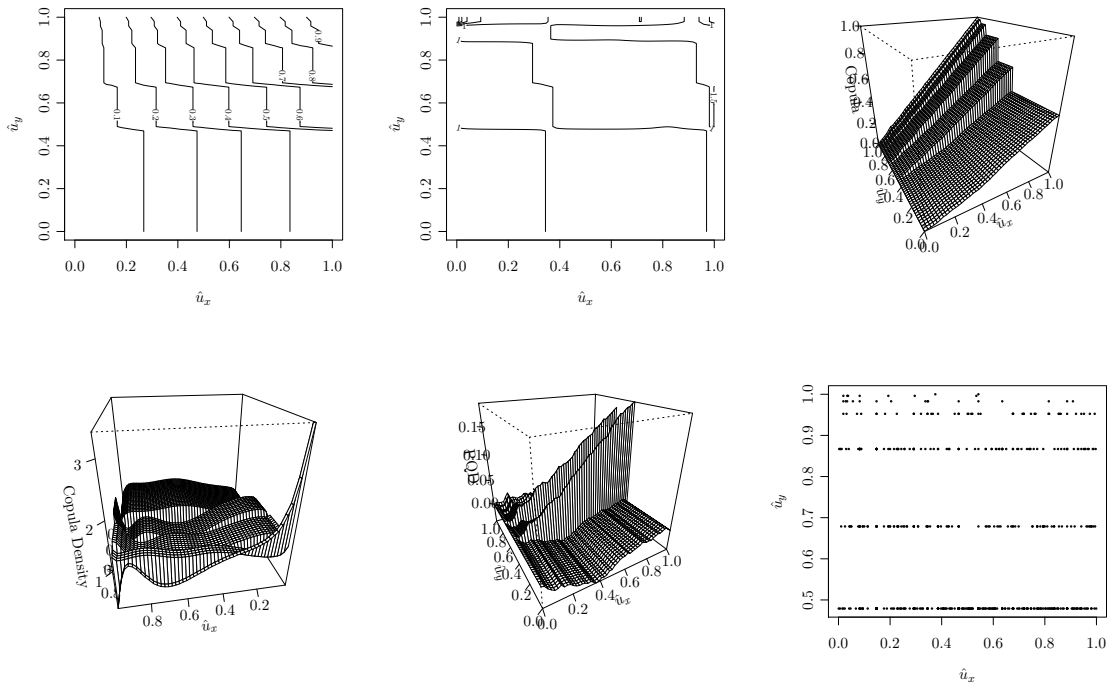


FIGURE 11. Bivariate Copula for (log) wages and number of dependants. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

This illustration demonstrates that copula and various measures can be readily computed in mixed-data settings.

6.2. **A Bivariate Continuous Data Application to the Merval and Hang Seng Stock Indices.** In Finance, issues of diversification and co-movement play a key role in portfolio analysis. We consider daily closing values of two indices, Merval (Buenos Aires) and Hang Seng (Hong Kong), for the dates 1996-10-08 through 2012-07-30. Index values are paired for trade days common to both to ensure temporal pairing is correct. Results are presented in Figure 12. We use multivariate least-squares cross-validation and obtain bandwidths $\hat{h}_x$=11.06 and $\hat{h}_y$=147.1 for the copula and $\hat{h}_x$=14.33 and $\hat{h}_y$=194.3 for the density. This sample contained $n$ =3766 observations.

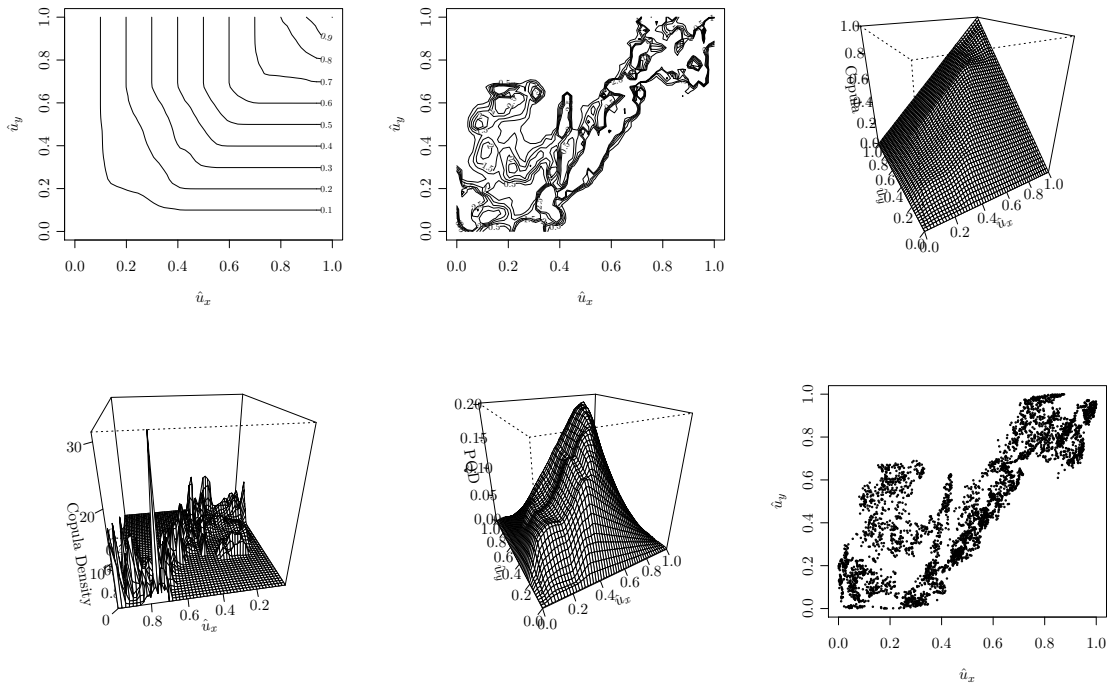


FIGURE 12. Bivariate Copula for the Merval and Hang Seng indices. The first row of figures present contour plots for the nonparametric copula and copula density, then the nonparametric copula itself. The second row of figures presents the nonparametric copula density, the nonparametric dependence measure PQD ($\hat{C}(\hat{u}_x, \hat{u}_y) - \hat{u}_x\hat{u}_y$) and the nonparametric copula scatter plot for the sample realizations.

We observe that the dependence measure PQD plotted in Figure 12 is uniformly non-negative indicating that these two indices display this feature (a formal test could be predicated on Scaillet (2005)).

By way of comparison, we fit a parametric Gaussian copula and compare and contrast the contour plots for the parametric and nonparametric copulas. Results are presented in Figure 13. It is evident from this dataset that the parametric copula imposes structure on the data that may distort subsequent analysis.

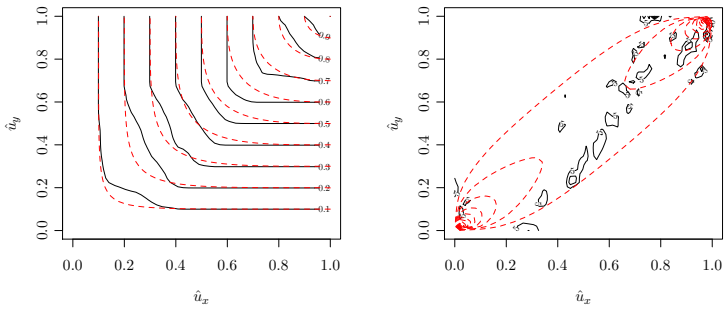


FIGURE 13. Comparison of parametric versus nonparametric bivariate Copula for the Merval and Hang Seng indices. The figures present contour plots for the parametric and nonparametric copula and copula density, respectively (the nonparametric contours are black/solid lines, parametric red/dashed).

## 7. SUMMARY

We apply recently developed methods for optimal bandwidth selection and kernel estimation of (unconditional) PDFs and CDFs to the problem of copula modeling. Nonparametric methods are particularly well-suited to this problem domain. Furthermore, by taking an approach based on inversion of marginal CDFs and PDFs we avoid the need for boundary kernels which can be challenging in multidimensional settings, particularly as the dimension $d$ increases beyond 2. In addition, no new theory is required. Our approach is fully general, delivers $\sqrt{n}$-consistent estimates of the copula that are dimension free hence circumvents the curse of dimensionality that plagues nonparametric approaches (the copula density, however, does suffer from this limitation). Measures of dependence can then be computed directly from the estimated copula. An implementation in the R package 'np' (Hayfield & Racine (2008, Version 0.50-1, function 'npcopula')) is available for the interested reader.

## REFERENCES

Bauwens, L., Laurent, S. & Rombouts, J. (2006), 'Multivariate GARCH models: a survey', *Journal of Applied Econometrics* **21**(1), 79–109.

Bouezmarni, T., Rombouts, J. & Taamouti, A. (2012), 'A nonparametric copula based test for conditional independence with applications to granger causality', *Journal of Business and Economic Statistics* **30**(2), 275–287.

Bowman, A., Hall, P. & Prvan, T. (1998), 'Bandwidth selection for the smoothing of distribution functions', *Biometrika* **85**, 799–808.

Chen, S. X. & Huang, T. (2007), 'Nonparametric estimation of copula functions', *Canadian Journal of Statistics* **35**, 265–282.

Chen, X., Fan, Y. & Tsyrennikov, V. (2006), 'Efficient estimation of semiparametric multivariate copula models', *Journal of the American Statistical Association* **101**, 1228–1240.

Chen, X., Wu, W. & Yi, Y. (2009), 'Efficient estimation of copula-based semiparametric markov models', *Annals of Statistics* **37**, 4214–4253.

Denuit, M. & Scaillet, O. (2004), 'Nonparametric tests for positive quadrant dependence', *Journal of Financial Econometrics* **2**(3), 422–450.

Fermanian, J.-D. & Scaillet, O. (2003), 'Nonparametric estimation of copulas for time series', *Journal of Risk* **5**, 847–860.

Gijbels, I. & Mielniczuk, J. (1990), 'Estimating the density of a copula function', *Comm. Statist. Theory Methods* **19**, 445–464.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
**URL:** *http://www.jstatsoft.org/v27/i05/*

Lehmann, E. (1966), 'Some concepts of dependence', *Annals of Mathematical Statistics* **37**, 1137–1153.

Li, H. & Racine, J. S. (2013), Cross-validated estimation of cumulative distribution functions with categorical and continuous data, Manuscript, McMaster University.

Li, Q., Lin, J. & Racine, J. S. (2013), 'Optimal bandwidth selection for nonparametric conditional distribution and quantile functions', *Journal of Business and Economic Statistics* **31**(1), 57–65.

Li, Q. & Racine, J. (2003), 'Nonparametric estimation of distributions with categorical and continuous data', *Journal of Multivariate Analysis* **86**(2), 266–292.

Li, Q. & Racine, J. S. (2008), 'Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data', *Journal of Business and Economic Statistics* **26**(4), 423–434.

Liu, R. & Yang, L. (2008), 'Kernel estimation of multivariate cumulative distribution function', *Journal of Nonparametric Statistics* **20**(8), 661–677.

Müller, H. G. & Stadtmüller, U. (1999), 'Multivariate boundary kernels and a continuous least squares principle', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(2), 439–458.

Nelsen, R. B. (2006), *An Introduction to Copulas*, second edn, Springer-Verlag.

Scaillet, O. (2005), 'A kolmogorov- smirnov type test for positive quadrant dependence', *Canadian Journal of Statistics* **33**, 415–427.

Smith, M. S. & Khaled, M. A. (2012), 'Estimation of copula models with discrete margins via bayesian data augmentation', *Journal of the American Statistical Association* **107**(497), 290–303.

Trivedi, P. & Zimmer, D. (2007), *Copula Modeling: An Introduction for Practitioners*, NOW.

Tsukahara, H. (2005), 'Semiparametric estimation in copula models', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **33**(3), pp. 357–375.

Wooldridge, J. M. (2003), *Introductory Econometrics*, Thompson South-Western.

## R Code for the 'wage1' example

The following R code is based on the R package 'np' (Hayfield & Racine (2008, Version 0.50-1, function 'npcopula')) and generates the example based on Wooldridge's (2003) 'wage1' data. Note that the plots exploit the 'tikzDevice' package that allows the use of TEX symbols directly in figures.

```
data(wage1)
mydat <- with(wage1, data.frame(lwage = lwage, numdep = ordered(numdep)))
bw.copula <- npudistbw(~lwage + numdep, ckertype = ckertype, bwmethod = "cv.cdf",
    data = mydat)

q.min <- 0
q.max <- 1
grid.seq <- seq(q.min, q.max, length = n.eval)
u <- cbind(grid.seq, grid.seq)
q.density.min <- 0.025
q.density.max <- 0.975
grid.density.seq <- seq(q.density.min, q.density.max, length = n.eval)
u.density <- cbind(grid.density.seq, grid.density.seq)

## Full range copula on evaluation grid
mycopula <- npcopula(bws = bw.copula, data = mydat, u = u)

## Full range empirical copula
mycopula.emp <- npcopula(bws = bw.copula, data = mydat)

## Restricted range copula on evaluation grid
bw.density <- npudensbw(~lwage + numdep, ckertype = ckertype, bwmethod = "cv.ml",
    data = mydat)
mycopula.density <- npcopula(bws = bw.density, data = mydat, u = u.density)
C.xy <- mycopula$copula
c.xy <- mycopula.density$copula

## Copula contour plot
contour(x = grid.seq, y = grid.seq, z = matrix(C.xy, n.eval, n.eval), xlab = "$\\hat u_x$",
    ylab = "$\\hat u_y$")
## Copula density contour plot
contour(x = grid.seq, y = grid.seq, z = matrix(c.xy, n.eval, n.eval), xlab = "$\\hat u_x$",
    ylab = "$\\hat u_y$", levels = seq(0.5, 2.5, by = 1))

## Copula perspective plot
persp(x = grid.seq, y = grid.seq, z = matrix(C.xy, n.eval, n.eval), xlab = "$\\hat u_x$",
    ylab = "$\\hat u_y$", zlab = "Copula", zlim = c(0, 1), ticktype = "detailed",
    theta = 330, phi = 25)

## Copula density perspective plot
persp(x = grid.density.seq, y = grid.density.seq, z = matrix(c.xy, n.eval, n.eval),
    xlab = "$\\hat u_x$", ylab = "$\\hat u_y$", zlab = "Copula Density",
    ticktype = "detailed", theta = 170, phi = 25)
```

```
## Measure of positive quadrant dependence plotted via persp
PQD <- mycopula$copula - mycopula$u1 * mycopula$u2
persp(x = grid.seq, y = grid.seq, z = matrix(PQD, n.eval, n.eval), xlab = "$\\hat u_x$",
    ylab = "$\\hat u_y$", zlab = "PQD", ticktype = "detailed", theta = 330,
    phi = 25)

## Empirical copula scatter plot
plot(mycopula.emp$u1, mycopula.emp$u2, xlab = "$\\hat u_x$", ylab = "$\\hat u_y$",
    cex = 0.25)
```

McMaster University, Hamilton, Ontario, Canada, racinej@mcmaster.ca