AIMS Mathematics, 6(12): 13488–13502.
DOI:10.3934/math.2021782
Received: 13 May 2021
Accepted: 15 September 2021
Published: 18 September 2021

http://www.aimspress.com/journal/Math

Research article

Nonparametric estimation of the measure of functional dependence

Qingsong Shan and Qianning Liu*

Department of Statistics, Jiangxi University of Finance and Economics, Nanchang, China

* **Correspondence:** Email: qianningliu@outlook.com.

Abstract: In this paper, we propose a beta kernel estimator to measure functional dependence (MFD). The MFD not only can measure the strength of linear or monotonic relationships, but it is also suitable for more complicated functional dependence. We derive the asymptotic distribution of the proposed estimator and then use several simulated examples to compare our estimator with the traditional measures. Our simulation results demonstrate that beta kernel provides high accuracy in estimation. A real data example is also given to illustrate one possible application of the new estimator.

Keywords: mutual complete dependence; beta kernel; functional dependence; monparametric estimation; copula
Mathematics Subject Classification: 62G05, 62G07

1. Introduction and motivation

The study of association or dependence plays an important role in statistics. One of the important aspects of this is how to measure the strength of various associations among random variables. Among the measures of associations between random variables, Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ are the most prominent ones. But they only measure linear or monotonic relationships, not suitable for a general nonlinear relationship. For example, when the relationship between two random variables is parabolic-shaped, none of the measures above will be applicable. Thus, a measure addressing this issue is desirable.

Two random variables are said to be mutually complete dependent (MCD) when they have mutual functional relationship. This concept was first introduced by Lancaster [1]. This is also known as the strongest dependence. In this relation, one variable is completely predictable of the other. Siburg and Stoimenov [2] constructed a measure of MCD for continuous random variables. Tasena and Dhompongsa [3] extended this measure to the multivariate case. In their papers, they proposed to measure the distance between two copulas by a modified Sobolev norm. For details about inner product and Sobolev norm on copula space, we refer to Darsow and Olsen [4].

Dette et al. [5] found that a simple modification of the measure of MCD can be used to measure the strength of functional dependence (MFD). Note that this functional dependence includes a wider range of dependence since it could be nonlinear or even nonmonotonic. The discrete form of MFD was given in Shan et al. [6]. Given the theoretical definition of MFD, the next question will be how to estimate it. Similar to the measure of MCD, MFD is also constructed based on copula. So a straightforward way of estimating MFD contains two steps. First, estimating the copula or its density. Second, using the estimated copula or its density to estimate the MFD. Generally in step one, there are two approaches to estimate the copula or its density, parametric way or nonparametric way. In the parametric way, one assumes a parametric model for copula, and estimates its parameters by the method of maximum likelihood (MLE). However, unlike distribution functions, the dependence structure, i.e., the copula, is usually hidden behind the data set, which makes the claim of having prior knowledge of copula family quite questionable. So in this paper, we will mainly consider the nonparametric way.

In this article, we propose estimating the MFD using the kernel method with the beta kernel. We introduce the new estimators and study their asymptotic properties in Section 3. Using Monte Carlo simulations, we investigate the finite sample performance of the proposed estimators relative to traditional measures of dependence. Our simulation results, reported in Section 4, show that the new estimators are accurate and stable with the choice of different parameters in a given model. In Section 5, a real data example is used to illustrate the new estimators. The paper is concluded by some remarks in Section 6.

2. Preliminaries

This section introduces some notations and concepts that will be used in the remainder of the article. We will focus on bivariate continuous distributions. Consider an independent and identically distributed sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ of a bivariate random vector $(\mathbf{X}, \mathbf{Y})$ with joint distribution function H, and marginal distribution functions F and G, respectively. Then, by Sklar's Theorem [7], there exists a unique copula $C : I^2 = [0, 1]^2 \to I = [0, 1]$ such that

$$H(x, y) = C(F(x), G(y)). \tag{2.1}$$

Therefore, the copula density is given by

$$c(F(x), G(y)) = \frac{\partial^2}{\partial x \partial y} C(F(x), G(y)). \tag{2.2}$$

The measure of MCD is based on the norm of functions in the copula space $\mathbb{C}$. The Sobolev norm for copula takes the following form

$$\|C\|^2 = \int_0^1 \int_0^1 \left[\left(\frac{\partial C}{\partial u} \right)^2 + \left(\frac{\partial C}{\partial v} \right)^2 \right] dudv. \tag{2.3}$$

The above norm has the following properties.

Proposition 1. *The Sobolev norm for copulas satisfies* $\|C\|^2 \in [2/3, 1]$ *for all* $C \in \mathbb{C}$. *Moreover, the following properties hold:*

i. $\|C\|^2 = 2/3$ *if and only if X and Y are independent.*

ii. $\|C\|^2 = 1$ *if and only if X and Y are MCD.*

Inspired by this proposition, Siburg and Stoimenov [2] proposed the measure of MCD, which can be defined as follows.

Definition 1. *Given two continuous random variables X, Y with copula C, we define*

$$\rho(X, Y) = (3\|C\|^2 - 2)^{1/2}. \tag{2.4}$$

$\rho(X, Y)$ can be interpreted as a normalized Sobolev distance of C from the independent copula denoted by Π:

$$\rho(X, Y) = \sqrt{3}\|C - P\| = \frac{\|C - \Pi\|}{\|C_m - \Pi\|},$$

where C_m is a MCD copula. A close look at the MCD shows that it can be decomposed into two opposite functional dependencies, i.e., Y is a function of X and X is a function of Y. Therefore, the measure of functional dependence (MFD) can be derived by modifying the measure for MCD, which is the main idea of Dette et al. [5]. Its discrete form was discussed in Shan et al. [6]. The construction of MFD is based on the following propositions.

Proposition 2. *Let X and Y be two random variables with copula C. Then,*

i. X and Y are independent if and only if $\partial_1 C_{X,Y}(u, v) = v$ for Lebesgue almost all $(u, v) \in I^2$.

ii. Y is almost surely (a.s.) a Borel function of X if and only if $\partial_1 C_{X,Y}(u, v) \in \{0, 1\}$ for Lebesgue almost all $(u, v) \in I^2$.

Proposition 3. *For any $C_{X,Y} \in \mathbb{C}$, we have $\|\partial_1 C_{X,Y}\|_2^2 \in [1/3, 1/2]$, Moreover,*

i. $\|\partial_1 C_{X,Y}\|_2^2 = 1/3$ if and only if X and Y are independent.

ii. $\|\partial_1 C_{X,Y}\|_2^2 = 1/2$ if and only if Y is a.s. a Borel function of X.

Notice that $\|\partial_1 C_{X,Y}\|$ reaches its boundaries at two extreme cases. $\|\partial_2 C_{X,Y}\|$ has similar propositions which we will not reproduce here. Using these definitions and propositions, we can define

$$\rho_1^2(Y|X) = 6\int_0^1 \int_0^1 \left(\frac{\partial C}{\partial u}\right)^2 dudv - 2, \tag{2.5}$$

$$\rho_2^2(X|Y) = 6\int_0^1 \int_0^1 \left(\frac{\partial C}{\partial v}\right)^2 dudv - 2. \tag{2.6}$$

This is a standardized form of $\|\partial_1 C_{X,Y}\|$. Hence ρ_1 inherits similar properties:

i. $\rho_1 = 0$ if and only if X and Y are independent.

ii. $\rho_1 = 1$ if and only if Y is a.s. a Borel function of X.

Similarly, $\rho_2 = 1$ indicates X is a.s. a Borel function of Y. Those properties suggest that $\rho_i (i = 1, 2)$ can be used to measure functional dependence. We can assess the strength of dependence from the magnitude of ρ_i.

3. Nonparametric estimation

The measures ρ_1, ρ_2 and ρ are all constructed based on copulas. They both can be stated in terms of copulas or copula densities. In other words, (2.5) can be written as

$$\rho_1^2 = 6 \int_0^1 \int_0^1 \left(\int_0^v c(u, y)dy \right)^2 dudv - 2, \tag{3.1}$$

and (2.6) can be written as

$$\rho_2^2 = 6 \int_0^1 \int_0^1 \left(\int_0^u c(x, v)dx \right)^2 dudv - 2. \tag{3.2}$$

Accordingly, there are two approaches to estimate the measures, through copula or its density. We will focus on the latter in this paper.

The estimation of the copula density has been discussed in many papers. For example, Kauermann et al. [8] estimated copula density with B-spline. Genest et al. [9] estimated copula density through wavelets. It is recognized that the estimation of copula density involves more technical difficulties than usual density estimation. One of the big issues in this respect is boundary bias. Several methods have been proposed to address this issue. Omelka et al. [10] suggested an improved version of mirror-reflection estimator. Charpentier et al. [11] suggested to use transformation estimator. Chen [12] suggested to use beta kernels whose support matches the support of copulas. Geenens et al. [13] used probit transformation to reduce boundary bias effect in kernel estimation of copula density. Majdara and Nooshabadi [14] provides a novel method in estimating copula density in high dimension space.

Dette et al. [5] discussed the asymptotic behavior of the estimation ρ_1 based on symmetric kernels and Eq (2.5). In this paper, we will use the beta kernel in estimating ρ_1 and ρ_2. Beta kernel smoothing was considered by Harrell and Davis [15]. Chen [12], Chen [16] applied beta kernel smoothing in density estimation, and found that the beta estimator can reduce boundary bias and variance compared with local linear estimators for densities with finite support. Following the same idea, Charpentier et al. [11] proposed the beta kernel based estimator for copula density.

3.1. Nonparametric estimation of the copula density

Let $(X_1, Y_1), (X_2, Y_2), \cdots (X_n, Y_n)$ be a sample from $H_{X,Y}(x, y)$ with unknown marginals. Denote the copula corresponding to $H_{X,Y}(x, y)$ by $C(u, v)$. We assume that both H and C are completely unknown. We start by the most convenient situation in which we assume that the copula C is twice differentiable. Let c denote the density of the copula. Usually copulas are estimated via pseudo-observations $(\hat{F}(X_i), \hat{G}(Y_i))$, where $\hat{F}$ and $\hat{G}$ are the empirical distribution functions, i.e.,

$$U = \hat{F}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}(X_i \le x), \qquad \text{and} \qquad V = \hat{G}(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}(Y_i \le y), \tag{3.3}$$

with $\mathbb{1}(A)$ being the usual indicator function. The beta kernel based estimator of the copula density at point $(u, v) \in [0, 1]^2$ is

$$\hat{c}_h(u, v) = \frac{1}{n} \sum_{i=1}^{n} K(U_i, \frac{u}{h} + 1, \frac{1-u}{h} + 1) K(V_i, \frac{v}{h} + 1, \frac{1-v}{h} + 1),$$

where $K(\cdot, \alpha, \beta)$ is the density of the beta distribution with parameters α and β, i.e.,

$$K(x, \alpha, \beta) = \frac{x^{\alpha}(1-x)^{\beta}}{B(\alpha, \beta)}, \quad x \in [0, 1],$$

with $B(\alpha, \beta) = \Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)$. For convenience, the same bandwidth is used in both kernels in $\hat{c}_h(u, v)$. Charpentier et al. [11] claimed the asymptotic normality of $\hat{c}_h(u, v)$by showing that, for all $(u, v) \in [0, 1]^2$,

$$\sqrt{nh}[\hat{c}_h(u, v) - c(u, v)] \xrightarrow{L} N(0, \sigma(u, v)^2),$$

as $nh \to \infty$ and $h \to 0$, where "$\xrightarrow{L}$" means convergence in distribution. Nagler [17] provided detailed proof and gave the bias and variance of $\hat{c}_h(u, v)$ in Proposition 4. He also discussed bandwidth selection for $\hat{c}_h(u, v)$.

Proposition 4. *Let $c(u, v)$ be twice continuously differentiable on $(0, 1)^2$, and $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Then, for all $(u, v) \in (0, 1)^2$,*

$$\begin{aligned} \text{Bias}[\hat{c}_h(u, v)] &= h_n[(1-2u)c_u(u, v) + (1-2v)c_v(u, v) \\ &\quad + \frac{1}{2}u(1-u)c_{uu}(u, v) + \frac{1}{2}v(1-v)c_{vv}(u, v) + o(h_n), \end{aligned}$$

$$\text{Var}[\hat{c}_h(u, v)] = \frac{1}{4nh_h\pi} \frac{c(u, v)}{\sqrt{u(1-u)v(1-v)}} + o\left(\frac{1}{nh_n}\right).$$

Note that there is a little difference between Proposition 4, which is on the interior of $(0, 1)^2$, and Charpentier's claim [11] , which is on the whole $[0, 1]^2$. Since the definition of functional dependence is based on the integration of copula densities, which will not be affected by the boundaries. We will consider the whole $[0, 1]^2$ in the remainder.

3.2. Nonparametric estimation of MFD via the copula density

Since all of the three measures, ρ_1, ρ_2 and ρ, are constructed in similar manner, we only take ρ_1 as an example and show how to estimate it through estimating the copula density using the beta kernel. Let $\ell^{\infty}([0, 1]^2)$ be the space of the collection of all uniformly bounded real-valued functions defined on $[0, 1]^2$, equipped with the uniform metric m defined as

$$m(f_1, f_2) = \sup_{x \in [0,1]^2} |f_1(x) - f_2(x)|, \quad f_1, f_2 \in \ell^{\infty}([0, 1]^2). \tag{3.4}$$

Define $\phi_i : \ell^{\infty}([0, 1]^2) \to \mathbb{R}$, $i = 1, 2$, by

$$\phi_1 : c(u, v) \to \int_0^1 \int_0^1 \left(\int_0^v c(u, y)dy \right)^2 dudv,$$

$$\phi_2 : c(u, v) \to \int_0^1 \int_0^1 \left(\int_0^u c(x, v)dx \right)^2 dudv.$$

Then, the three measures, ρ_1^2, ρ_1^2 and ρ^2, are functionals of $c(\cdot, \cdot)$. So, it suffices to show that ϕ_1 and ϕ_2 are Hadamard differentiable.

Theorem 1. *Let $c(u,v)$ be twice continuously differentiable on $[0,1]^2$, and $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Then,*

$$\sqrt{nh}(\hat{\rho}_1^2 - \rho_1^2) \xrightarrow{L} \phi_1'(N(0, \sigma^2(u,v))),$$

where $\phi_1'(l(u,y)) = \int_0^1 \int_0^1 \left(2 \int_0^v c(u,y)dy \int_0^v l(u,y)dy\right) dudv$.

Proof. ρ_1, ρ_2 can be represented as a map ϕ_1, ϕ_2: $\ell^\infty([0,1]^2) \to \mathbb{R}([0,1])$ via $\rho_1 = \phi_1(c)$ and $\rho_2 = \phi_2(c)$, respectively. The function space $\ell^\infty([0,1]^2)$ is equipped with the uniform metric m. For all converging sequences $t_n \to 0$ and $l_n \to l$ such that $c + t_n l_n \in \ell^\infty([0,1]^2)$ for every n, we have

$$\begin{aligned}
\frac{\phi_1(c + t_n l_n) - \phi_1(c)}{t_n} &= \frac{1}{t_n} \int_0^1 \int_0^1 \left[\left(\int_0^v (c(u,y) + t_n l_n(u,y))dy \right)^2 - \left(\int_0^v (c(u,y))dy \right)^2 \right] dudv \\
&= \int_0^1 \int_0^1 \frac{1}{t_n} \left(\int_0^v (c(u,y) + t_n l_n(u,y))dy + \int_0^v c(u,y)dy \right) \\
&\qquad \cdot \left(\int_0^v (c(u,y) + t_n l_n(u,y))dy - \int_0^v c(u,y)dy \right) dudv \\
&= \int_0^1 \int_0^1 \left(\int_0^v (c(u,y) + t_n l_n(u,y))dy + \int_0^v c(u,y)dy \right) \\
&\qquad \cdot \left(\int_0^v (l_n(u,y))dy \right) dudv.
\end{aligned}$$

So, the Hadamard derivative of ϕ_1 at c is

$$\phi_1'(h) = \int_0^1 \int_0^1 \left(2 \int_0^v c(u,y)dy \int_0^v l(u,y)dy \right) dudv.$$

Therefore, according to the Delta method [18],

$$\phi_1'(N(0, \sigma(u,v)^2)) = \int_0^1 \int_0^1 \left(2 \int_0^v c(u,y)dy \int_0^v N(0, \sigma(u,y)^2)dy \right) dudv.$$

This completes the proof. □

The asymptotic distributions of $\hat{\rho}_1^2$ and $\hat{\rho}^2$ can be derived in exactly the same manner, so we omit their details.

3.3. Nonparametric estimation of MFD via copula functions

In the following, we show that estimators of MFD through copula have the same asymptotic distributions as those established through copula density. As an example, let's check the asymptotic distribution of $\hat{\rho}^2$.

Mapping a copula density to an MFD can be decomposed into two steps as follows

$$c \xrightarrow{\varphi} C \xrightarrow{\psi} \text{MFD}\,.$$

The first map is a double-integration that is linear and continuous, and thus, it is Hadamard-differentiable. We only need to check the second map.

Let $\mathbb{D}_2^1([0,1]^2)$ be the Sobolev space, $D_1, D_2 \in \mathbb{D}_2^1([0,1]^2)$, and define the inner product

$$\langle D_1, D_2 \rangle = \int_{[0,1]^2} \nabla D_1 \cdot \nabla D_2 d\lambda,$$

where ∇ is gradient. The Sobolev norm induced by the inner product in $\mathbb{D}_2^1([0,1]^2)$ is

$$|D|^2 = \langle D, D \rangle = \int_{[0,1]^2} \left[\left(\frac{\partial D}{\partial u} \right)^2 + \left(\frac{\partial D}{\partial v} \right)^2 \right] du dv,$$

for $D \in \mathbb{D}_2^1([0,1]^2)$.

Let $\mathfrak{C} \subset \mathbb{D}_2^1([0,1]^2)$ be the copula space and $C \in \mathfrak{C}$ is a copula. Define $\psi : \mathfrak{C} \mapsto \mathbb{R}$ by $\psi(C) = |C|^2$. Then, the derivative of ψ at C along D is:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{t_n}(\psi(C + t_n H_n) - \psi(C)) &= \lim_{n\to\infty} \frac{1}{t_n}(\langle C + t_n H_n, C + t_n H_n \rangle - \langle C, C \rangle) \\
&= \lim_{n\to\infty} \frac{1}{t_n}(\langle C, C \rangle + 2\langle t_n H_n, C \rangle + \langle t_n H_n, t_n H_n \rangle - \langle C, C \rangle) \\
&= \lim_{n\to\infty} 2\langle H_n, C \rangle \\
&= 2\langle H, C \rangle.
\end{aligned}$$

The last step follows from Theorem 2.3 in [19]. This result shows the convergence of $\hat{\rho}^2$, and straightforward calculations will show that it is consistent with the asymptotic distribution in Theorem 1.

4. Implementation and simulations

4.1. Choosing the evaluation grid

In copula density estimation, Nagler [20] suggested using a grid that is equally spaced after a transformation by the inverse Gaussian cdf, which is shown in Figure 1. Our simulation results below show that evaluating copula density at a set of grid points in a similar pattern will improve the accuracy of estimators of MCD. To compare the impact of the choice of grid, we considered two copula families, the Gaussian copula with parameters $0, 0.1, 0.2, 0.5, 0.8, 0.9$ and the Gumbel copula with parameters $1, 10/9, 10/7, 10/3, 5, 10$. Two samples of sizes 200 and 1000 were taken from each copula, respectively. First, copula densities are estimated from each sample based on the KDEcopula package. Then, the estimated copula density was evaluated on two sets of grid points: the usual grid with equally spaced points and a normalized grid. From the discretized copula density, we calculate the estimate of MFD. Figures 2 and 3 show the mean absolute error (MAE) of estimators with sample size 200 and 1000 under 500 replication. In each case, we find the MAE of estimators based on an equally spaced grid, labeled "equal", are significantly higher than the MAE of the same estimators based on the transformed grid, labeled "norm".

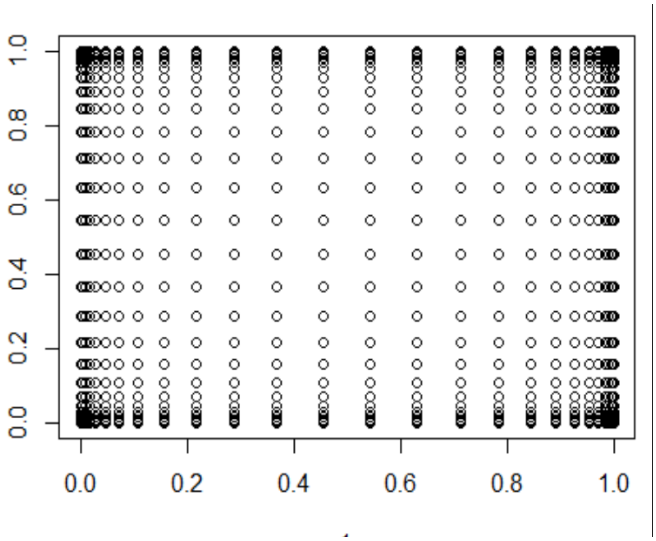


Figure 1. A grid which is equally spaced after inverse Gaussian cdf transformation.

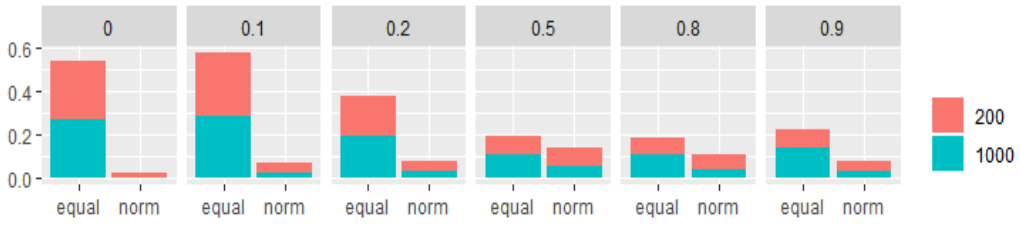


Figure 2. MAE of estimators of MFD for samples drawn from normal copula.

Figure 3. MAE of estimators of MFD for samples drawn from Gumbel copula.

4.2. Simulation

In this section, we explore the finite sample performance of the proposed estimators using the mean squared error (MSE). To put all estimators on the same scale, we standardize MFD. In other words, we use ρ_1, ρ_2 and ρ. The corresponding estimators will be denoted by $\hat{\rho}_1(Y \mid X)$, $\hat{\rho}_2(X \mid Y)$ and $\hat{\rho}(X, Y)$. For two-dimensional density estimation, using cross-validation to choose the bandwidth is computationally expensive. Therefore, in all simulations, a rule-of-thumb bandwidth is used. More precisely, the bandwidth is selected based on the asymptotic mean integrated squared error (AMISE)-optimality with respect to the Frank copula. For further details on bandwidth selection in this context,

we refer to Nagler [20]. In all simulations reported here, the integration is calculated over a grid of 30×30 points. For the choice of the grid, we adopt the method in Nagler [20]. That is, we apply the Gaussian cumulative distribution function to equally spaced 30 knots on a line segment $[-3, 3]$. The final two-dimensional grid is shown in Figure 1. This choice takes into account the fact that copula densities usually have high fluctuation on the boundary and corners. Putting more evaluating points on those regions will reduce approximation errors.

In Tables 1 and 2, we present the simulated MSE of the estimators of ρ_1 and ρ_2 for samples size 50, 100 and 200 for Gaussian copula. These results are based on 1000 replications. Both copulas are generated by R package "copula". We find that our estimators have reasonable precision in all cases. As the sample size increases, MSE is getting smaller. And there is no significant difference in MSE for different θ values, which indicates our estimator is stable for the choice of θ's.

Table 1. Simulated MSE of the estimates when the underlying copula is Gaussian copula with correlation θ.

		$n = 50$	$n = 100$	$n = 200$
$\theta = 0$	$\hat{\rho}_1(Y \mid X)$	2.3×10^{-3}	1.5×10^{-3}	8.0×10^{-4}
	$\hat{\rho}_2(X \mid Y)$	2.3×10^{-3}	1.5×10^{-3}	7.3×10^{-3}
$\theta = 0.3$	$\hat{\rho}_1(Y \mid X)$	6.1×10^{-3}	3.8×10^{-3}	2.2×10^{-3}
	$\hat{\rho}_2(X \mid Y)$	6.2×10^{-3}	3.8×10^{-3}	2.1×10^{-3}
$\theta = 0.6$	$\hat{\rho}_1(Y \mid X)$	8.8×10^{-3}	4.2×10^{-3}	2.1×10^{-3}
	$\hat{\rho}_2(X \mid Y)$	8.9×10^{-3}	4.1×10^{-3}	2.1×10^{-3}
$\theta = 0.9$	$\hat{\rho}_1(Y \mid X)$	2.5×10^{-3}	1.1×10^{-3}	4.9×10^{-4}
	$\hat{\rho}_2(X \mid Y)$	2.3×10^{-3}	1.1×10^{-3}	5.0×10^{-4}

Table 2. Simulated MSE of the estimates when the underlying copula is Clayton copula with parameter θ.

		$n = 50$	$n = 100$	$n = 200$
$\theta = 0.2$	$\hat{\rho}_1(Y \mid X)$	3.7×10^{-3}	2.3×10^{-3}	1.8×10^{-3}
	$\hat{\rho}_2(X \mid Y)$	3.7×10^{-3}	2.3×10^{-3}	1.8×10^{-3}
$\theta = 0.5$	$\hat{\rho}_1(Y \mid X)$	7.2×10^{-3}	4.0×10^{-3}	2.3×10^{-3}
	$\hat{\rho}_2(X \mid Y)$	7.2×10^{-3}	4.1×10^{-3}	2.3×10^{-3}
$\theta = 1$	$\hat{\rho}_1(Y \mid X)$	9.3×10^{-3}	4.9×10^{-3}	2.6×10^{-3}
	$\hat{\rho}_2(X \mid Y)$	9.3×10^{-3}	4.9×10^{-3}	2.5×10^{-3}
$\theta = 2$	$\hat{\rho}_1(Y \mid X)$	7.5×10^{-3}	3.5×10^{-3}	1.8×10^{-3}
	$\hat{\rho}_2(X \mid Y)$	7.4×10^{-3}	3.5×10^{-3}	1.9×10^{-3}
$\theta = 5$	$\hat{\rho}_1(Y \mid X)$	3.1×10^{-3}	1.2×10^{-3}	6.1×10^{-4}
	$\hat{\rho}_2(X \mid Y)$	3.1×10^{-3}	1.2×10^{-3}	6.0×10^{-4}

4.3. Comparison of measures

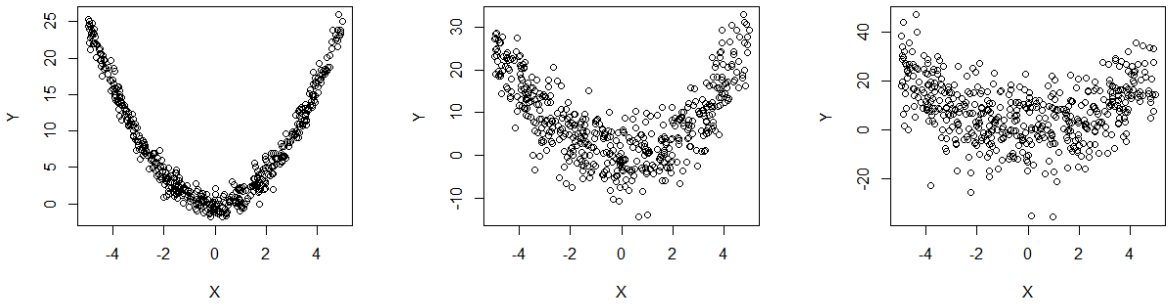
In the second part of the simulation, we will compare the performance of MFDs with other measures of dependence, e.g., linear correlation coefficient r, Spearman's ρ and Kendall's τ under several different types of relationships. We choose three different dependence structures: elliptical

distributions, monotonic dependence and regressional dependence, represented by normal copula, cubic function and quadratic function, respectively.

The first example is a quadratic function. 500 data are generated from the following model,

$$Y = X^2 + \varepsilon, \tag{4.1}$$

where $\varepsilon \sim N(0, \sigma)$ and $\sigma = 1$, 5, and 10, respectively (see Figure 4). To obtain the copula data, we apply the empirical marginal distributions to the data, i.e., apply a rank transformation as shown in (3.3) to the data generated by the model (4.1). Then beta kernel estimation is applied to get the estimations of ρ_1 and ρ_2.



(a) $\sigma = 1$ (b) $\sigma = 5$ (c) $\sigma = 10$

Figure 4. Scatter plot of model $Y = X^2 + \varepsilon$, sample size $N = 500$.

Obviously, neither Spearman's ρ nor Kendall's τ is suitable for this situation. The simulation results in Table 3 also showed that they are almost 0 in all cases. $\hat{\rho}(X, Y)$, on the other hand, is much higher than both Spearman's ρ and Kendall's τ, especially for small σ. This indicates that the type of dependence is functional, not monotonic. And the magnitude of $\hat{\rho}(X, Y)$ tells that the strength of dependence is getting weaker as σ increases. In a comparison of $\hat{\rho}_1(Y \mid X)$ and $\hat{\rho}_2(X \mid Y)$, we find that functional dependence is stronger in the Y to X direction than the other direction since $\hat{\rho}_1(Y \mid X)$ is higher. Again, as σ increases, the strength of dependence in this direction is also getting weaker.

Table 3. Estimators based on a sample of size 500 when the underlying relationship is a parabola.

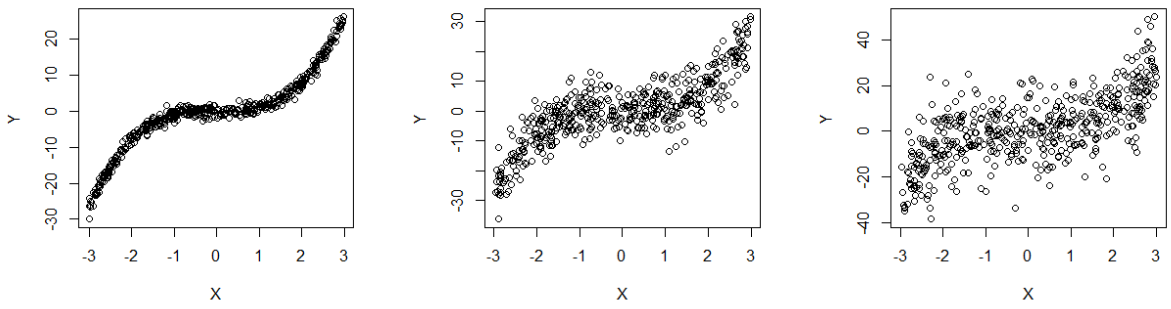
	$\hat{\rho}_1(Y \mid X)$	$\hat{\rho}_2(X \mid Y)$	$\hat{\rho}(X, Y)$	Spearman's ρ	Kendall's τ
σ=1	0.51	0.25	0.40	-0.01	-0.02
σ=5	0.41	0.20	0.32	-0.02	-0.02
σ=10	0.27	0.13	0.21	-0.01	-0.01

To compare the performance of MFD with Kendall's τ and Spearman's ρ in monotonic dependence. 500 data are generated from the following model,

$$Y = X^3 + \varepsilon, \tag{4.2}$$

where $\varepsilon \sim N(0, \sigma)$ and $\sigma = 1$, 5 and 10. The scatter plot of model (4.2) is in Figure 5 and the simulation results are in Table 4. As shown in Table 4, the values of MFD has a similar decreasing pattern as the

other two measures when σ increases. Indeed, a cubic function is one type of functional dependence, so MFDs are capable of measuring the strength of monotonic dependence. The values of $\hat{\rho}_1(Y \mid X)$ and $\hat{\rho}_2(X \mid Y)$, which measure the strength of functional dependence in two directions (Y to X and X to Y) separately, are close to each other, obviously this is because model (4.2) is symmetric.



(a) $\sigma = 1$ (b) $\sigma = 5$ (c) $\sigma = 10$

Figure 5. Scatter plot of model $Y = X^3 + \varepsilon$, sample size $N = 500$.

Table 4. Estimators based on a sample of size 500 when the underlying relationship is cubic.

	$\hat{\rho}_1(Y \mid X)$	$\hat{\rho}_2(X \mid Y)$	$\hat{\rho}(X, Y)$	Spearman's ρ	Kendall's τ
σ=1	0.89	0.89	0.89	0.82	0.95
σ=5	0.63	0.62	0.63	0.55	0.74
σ=10	0.42	0.41	0.41	0.35	0.50

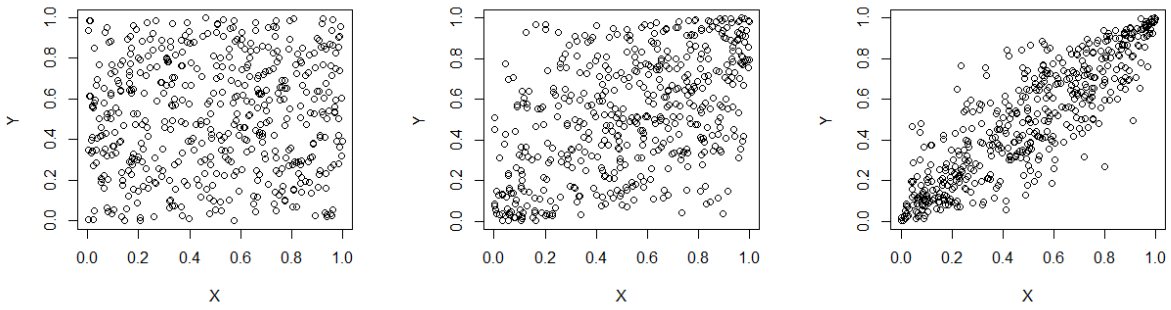
Next, we take into account the Pearson's correlation coefficient. We take normal copulas as an example, which is

$$C_r(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-r^2}} \exp\left\{-\frac{t^2 + s^2 - 2rts}{2(1-r^2)}\right\} dtds, \tag{4.3}$$

with $r = 0.1, 0.5$ and 0.9. The scatter plots of Gaussian copulas are in Figure 6. Table 5 shows the simulation results. As expected, those measures show no significant difference in measuring the dependence of elliptical distributions.

Table 5. Estimators based on a sample of size 500 when the underlying copula is Gaussian copula.

	$\hat{\rho}_1(Y \mid X)$	$\hat{\rho}_2(X \mid Y)$	$\hat{\rho}(X, Y)$	Spearman's ρ	Kendall's τ
$r = 0.1$	0.09	0.09	0.09	0.08	0.11
$r = 0.5$	0.33	0.33	0.33	0.29	0.42
$r = 0.9$	0.73	0.73	0.73	0.68	0.87



(a) $r = 0.1$ (b) $r = 0.5$ (c) $r = 0.9$

Figure 6. Normal copulas with parameter r.

The comparison of MFDs with other measures in models 4.1–4.3 shows that MFDs have good adaptability for different types of relationships.

5. Application

The measurement of functional relationships has many possible applications ([21, 22]). The Communities and Crime Data Set [23] contains community crime rate of 1994 communities with 123 possibly related variables. We will use functional dependence measure as a criteria for variable selection to choose the variables which have the most impact on community crime rate. We calculate the measures ρ, ρ_1, and ρ_2 given in Eqs (2.4)–(2.6), respectively by using beta kernel estimation for community crime rate and each of other variables. Variables with higher values of the measures have a greater impact on community crime rate. The Table 6 shows 9 variables with highest functional dependence measures and Table 7 gives the explanation of the abbreviations. Notice that the measures can detect strong non-linear relationships. As shown in Figure 7, two of the selected variables (PctIlleg and racePctWhite), showed clear nonlinear relations with the crime rate.

Table 6. 15 Variables with highest scores in functional dependence measure.

	Variable	$\hat{\rho}_1(Y \mid X)$	$\hat{\rho}_2(X \mid Y)$	$\hat{\rho}(X, Y)$
1	PctKids2Par	0.56	0.57	0.57
2	PctIlleg	0.53	0.54	0.54
3	PctFam2Par	0.53	0.54	0.53
4	NumIlleg	0.52	0.53	0.53
5	PctTeen2Par	0.49	0.49	0.49
6	racePctWhite	0.49	0.49	0.49
7	FemalePctDiv	0.49	0.49	0.49
8	NumUnderPov	0.48	0.48	0.48
9	TotalPctDiv	0.48	0.48	0.48

Table 7. Selected variables for the communities crime rate data.

Variable	Attribute
PctKids2Par	Percentage of kids in family housing with two parents
PctIlleg	Percentage of kids born to never married
PctFam2Par	Percentage of families (with kids) that are headed by two parents
NumIlleg	Number of kids born to never married
PctTeen2Par	Percent of kids age 12–17 in two parent households
racePctWhite	Percentage of population that is caucasian
FemalePctDiv	Percentage of females who are divorced
NumUnderPov	Number of people under the poverty level
TotalPctDiv	Percentage of population who are divorced

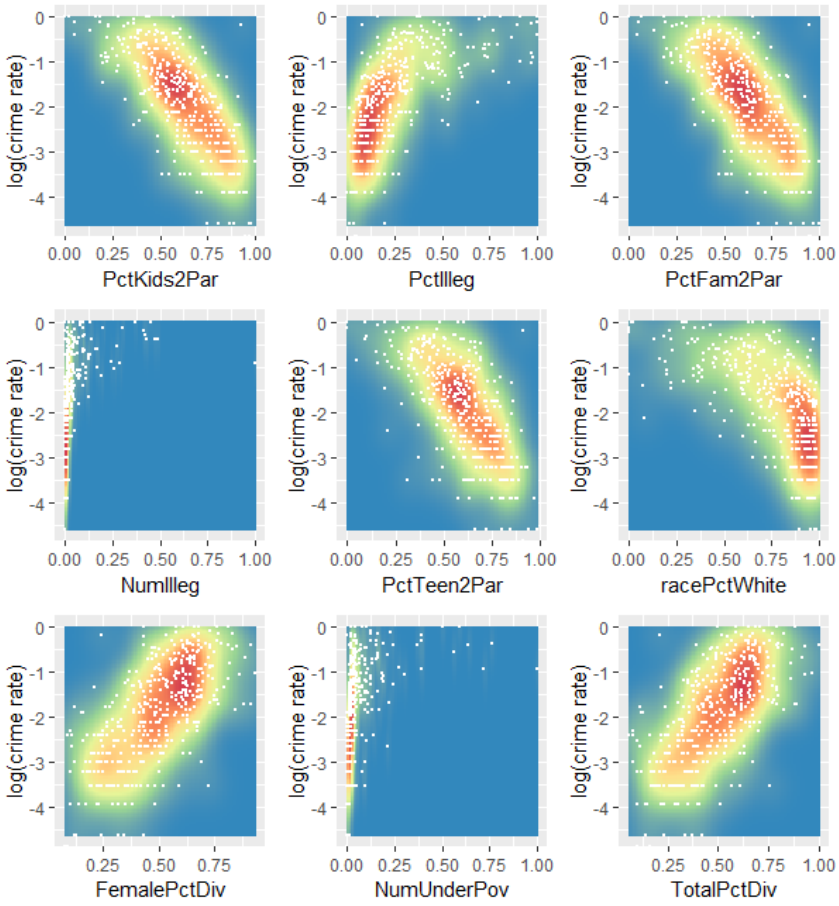


Figure 7. Scatter plot and kernel density estimation of community crime rate and other variables.

6. Discussion

This paper showed that, compared with Spearman's ρ or Kendall's τ, the measures of functional relationship could not only measure the strength of a relationship but also indicate the direction of a

possible functional relationship. We provide a novel method to estimate the measures of functional relationships. The simulation results showed that they have fairly good accuracy.

Although MFD can quantify the strength of functional dependence, it doesn't suggest any specific form of the function. So one possible application of this measure is in variable selection. We use MFD to filter out the less correlated variables, then use parametric or nonparametric methods to construct a predicting model. In the community crime data example, we showed that MFD could detect nonlinear relationship, but as for how many variables should be retained, in other words, how to set up the threshold of MFD in variable selection is a question that needs to be discussed and may involve some subjective opinions. After the desired number of variables are chosen, people may use either parametric or non-parametric methods to set up the model.

Acknowledgments

This work was supported in part by the Education Department of Jiangxi Province under Grant GJJ190253 and Grant GJJ190259.

Conflict of interest

The authors have no conflicts of interest to declare.

References

1. H. O. Lancaster, Dependence, measures and indices of, In: *Encyclopedia of statistical sciences*, 1982.
2. K. F. Siburg, P. A. Stoimenov, A measure of mutual complete dependence, *Metrika*, **71** (2010), 239–251.
3. S. Tasena, S. Dhompongsa, A measure of multivariate mutual complete dependence, *Int. J. Approx. Reason.*, **54** (2013), 748–761.
4. W. F. Darsow, E. T. Olsen, Norms for copulas, *Int. J. Math. Math. Sci.*, **18** (1995), 576296.
5. H. Dette, K. F. Siburg, P. A. Stoimenov, A copula-based non-parametric measure of regression dependence, *Scand. J. Stat.*, **40** (2013), 21–41.
6. Q. S. Shan, T. Wongyang, T. H. Wang, S. Tasena, A measure of mutual complete dependence in discrete variables through subcopula, *Int. J. Approx. Reason.*, **65** (2015), 11–23.
7. M. Sklar, Fonctions de répartition à n dimensions et leurs marges, *Publ. inst. statist. univ. Paris*, **8** (1959), 229–231.
8. G. Kauermann, C. Schellhase, D. Ruppert, Flexible copula density estimation with penalized hierarchical b-splines, *Scand. J. Stat.*, **40** (2013), 685–705.
9. C. Genest, E. Masiello, K. Tribouley, Estimating copula densities through wavelets, *Insur. Math. Econ.*, **44** (2009), 170–181.
10. M. Omelka, I. Gijbels, N. Veraverbeke, Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing, *Ann. Statist.*, **37** (2009), 3023–3058.

11. A. Charpentier, J. D. Fermanian, O. Scaillet, The estimation of copulas: Theory and practice, *Copulas: From theory to application in finance*, 2007, 35–60.
12. S. X. Chen, Beta kernel estimators for density functions, *Comput. Statist. Data Anal.*, **31** (1999), 131–145.
13. G. Geenens, A. Charpentier, D. Paindaveine, Probit transformation for nonparametric kernel estimation of the copula density, *Bernoulli*, **23** (2017), 1848–1873.
14. A. Majdara, S. Nooshabadi, Nonparametric density estimation using copula transform, bayesian sequential partitioning, and diffusion-based kernel estimator, *IEEE T. Knowl. Data En.*, **32** (2019), 821–826.
15. F. E. Harrell, C. E. Davis, A new distribution-free quantile estimator, *Biometrika*, **69** (1982), 635–640.
16. S. X. Chen, Beta kernel smoothers for regression curves, *Stat. Sinica*, **10** (2000), 73–91.
17. T. Nagler, *Kernel methods for vine copula estimation*, München: Universi at Munchen, 2014.
18. A. W. Van Der Vaart, J. A. Wellner, *Weak convergence and empirical processes*, Springer, 1996.
19. W. F. Darsow, B. Nguyen, E. T. Olsen, Copulas and markov processes, *Illinois J. Math.*, **36** (1992), 600–642.
20. T. Nagler, kdecopula: An R package for the kernel estimation of copula densities, 2016, arXiv: 1603.04229.
21. X. Han, Z. L. Wang, M. Xie, Y. H. He, Y. Li, W. Z. Wang, Remaining useful life prediction and predictive maintenance strategies for multi-state manufacturing systems considering functional dependence, *Reliab. Eng. Syst. Safe.*, **210** (2021), 107560.
22. Y. H. He, Z. X. Chen, Y. X. Zhao, X. Han, D. Zhou, Mission reliability evaluation for fuzzy multistate manufacturing system based on an extended stochastic flow network, *IEEE T. Reliab.*, **69** (2019), 1239–1253.
23. D. Dua, C. Graff, UCI machine learning repository, Irvine, CA: University of California, school of information and computer acience. Available from: https://archive.ics.uci.edu/ml/index.php.

AIMS Press