Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 53 (4) (2024), 1141 – 1157
DOI : 10.15672/hujms.1359072

Research Article

Parametric and semiparametric approaches for copula-based regression estimation

Alam Ali[1], Ashok Kumar Pathak*[1], Mohd Arshad[2]

[1] *Department of Mathematics and Statistics, Central University of Punjab, Bathinda, Punjab, India*
[2] *Department of Mathematics, Indian Institute of Technology Indore, Simrol, Indore, India*

Abstract

Based on the normality assumption on dependent variable, regression analysis is one of the most popular statistical techniques for studying the dependence between response and explanatory variables. However, violation of this assumption in the data makes regression analysis inappropriate in several real life situations. Copula is a powerful tool for modeling multivariate data and have recently been employed in regression analysis. The key concept behind copula-based regression approach is to formulate conditional expectation in terms of copula density and marginal distributions. In this paper, we explore parametric and semiparametric estimations of the copula-based regression function. The maximum likelihood (ML), inference functions for margins (IFM), and pseudo maximum likelihood (PML) techniques are adopted here for estimation purposes. Extensive numerical experiments are performed to illustrate the performance of the proposed copula-based regression estimators under specified and misspecified scenarios of copulas and marginals. Finally, two real data applications are also presented to demonstrate the performance of the considered estimators.

Mathematics Subject Classification (2020). 62H05, 62J05, 62F10

Keywords. Copula-based regression estimation, dependence modelling, regression function, inference function for margins (IFM), semiparametric inference

1. Introduction

For studying dependence between response and explanatory variables, regression analysis is one of the most commonly used statistical techniques. It is widely used for prediction and forecasting. However, in several practical situations the linear regression analysis may give fallacious results as the relationship between variables may be nonlinear, the distribution of the dependent variable could be non-normal, and multicollinearity may be present between explanatory variables.

In general, a statistical model can be presented in the following form

$$Y = h(\mathbf{X}, \boldsymbol{\beta}) + \epsilon, \tag{1.1}$$

*Corresponding Author.
Email addresses: alam2ali1996@gmail.com (A. Ali), ashokiitb09@gmail.com (A.K. Pathak), arshad.iitk@gmail.com (M. Arshad)
Received: 12.09.2023; Accepted: 29.05.2024

where Y is dependent variable, $h : \mathbb{R}^k \to \mathbb{R}$ is a real valued function, $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ are explanatory variables, $\boldsymbol{\beta}$ denotes vector of model parameters, and ϵ is random error term. The function h may take linear and non-linear forms and may assume exponential, logarithmic, polynomial, trigonometric, and Lorenz curve shapes. For a linear function h, Eq. (1.1) leads to

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon,$$

which is commonly known as multiple linear regression. The generalized linear model (GLM) is a natural extension of the classical linear regression models and may overcome several drawbacks associated with it. However, applicability of the GLM models is also limited in practical situations as the distribution of dependent variable is a member of exponential family. For more details one may refer to [23].

Recently, copula functions have been used successfully to study the dependence between response and explanatory variables. Copula is a function that connects marginals to the joint distribution and was first introduced by [30]. Let Y, X_1, X_2,...,X_k be the random variables with marginals $F_Y(y), F_{X_1}(x_1), \ldots, F_{X_k}(x_k)$ and the joint distribution $F(y, x_1, \ldots, x_k)$, respectively. Then Sklar showed that the marginals and the joint distribution are connected through copula C via the relation

$$F(y, x_1, \ldots, x_k) = C(F_Y(y), F_{X_1}(x_1), \ldots, F_{X_k}(x_k)), \quad \forall\ (y, x_1, \ldots, x_k) \in \mathbb{R}^{k+1}, \tag{1.2}$$

where the function $C : [0,1]^{k+1} \to [0,1]$ is termed as a copula. For continuous random variables, the representation is unique. The advantage of the copula is that it enables us to separate marginal effect from dependence structure and is also invariant under strictly monotone transformations of random variables. With these appealing features, copula have been extensively used for theoretical, methodological, and applied work in recent years with applications in diverse areas like finance, biology, economics, engineering, hydrology, and insurance. For describing the dependence among random variables, a large number of parametric families of copulas have been proposed and studied in literature. For a detailed discussion about copula, one can refer the popular monographs of [17, 24] and the references therein.

Despite of the huge families of copulas readily available in literature, only few attempts have been made to study the regression models through copulas in recent years. Leong and Valdez [22] derived the mean regression function for multivariate Gaussian, Student's-t, and FGM families of copulas and utilized them in insurance claims prediction. Sungur [32] introduced an alternative approach to regression analysis using copulas and investigated its key properties. Crane and van der Hoek [6] derived conditional expectation formula for copulas and used them in exchange-rate data. A detailed discussion about the advantage of copula regression over ordinary least squares and generalized linear models is outlined in [26]. de Leon and Wu [7] investigated copula-based regression models for mixed outcomes and applied them to burn injury data. Vellaisamy and Pathak [34] provided a simple characterization for multivariate linear regression copulas. It is well-known that the Gaussian and t-copulas are important members of elliptical families of copulas and play significant role in multivariate modeling. Pitt et al. [28] proposed a Gaussian copula regression model when the dependence between response and explanatory variables is characterized by correlation coefficient. Regression analysis has been using Gaussian copula more and more, but not much work has been done on using t-copula for prediction in the literature. Sheikhi et al. [29] recently explored the idea of heteroscedasticity in relation to regression models that use copulas. Some other important references related to regression analysis via copula include [1, 2, 10, 18, 21]. Noh et al. [25] have recently proposed an important inferential aspect of copula-based regression models. They considered a semiparametric (SP) approach for estimating regression function, in which the copula belongs to a parametic family. The copula parameters are estimated through parametric

technique and the marginals are estimated nonparametrically via rescaled version of empirical distribution. Dette et al. [9] discussed the shortcoming of the estimation proposed by [25]. Bouezmarni et al. [4] addresses the semiparametric estimation of the regression function based on copula for right-censored data. Hamori et al. [14] explored generalized regression models via copula and obtained calibration estimator for the regression curve. A Bayesian approach for regression copulas has been studied by [31].

In the literature, several estimation techniques for copula have been proposed and studied. These methods can be parametric, semiparametric, or nonparametric in nature. For a good source of copula estimation one can refer to [5, 11, 33]. We can develop a range of estimators and estimation methods for copula-based regression functions by taking into consideration various estimation strategies for copula and marginals. Moreover, the development of a new class of copula-based regression estimation techniques is a topic of interest.

The primary aim of this article is to explore parametric and semiparametric estimation for a copula-based regression function and study the performance of these techniques under specified and misspecified scenarios of copulas and marginals. To the best of our knowledge, this study has not been explored in the literature. We consider maximum likelihood (ML), inference function for margins (IFM), and pseudo maximum likelihood (PML) approaches for estimation. A theoretical comparison of the proposed techniques is quite tedious. We consider a numerical comparison based on a well-organized simulation study. We found that under correct specification of copula and marginals, the ML method performs better than IFM and PML with respect to estimated variance, mean-squared error, and relative efficiency. However, in case of misspecified scenario for either copulas or marginals, PML performs better than ML and IFM.

The article is organized as follows: Section 2 introduces a mathematical framework of regression function in terms of copula-based conditional expectation. Regression function for some well-know families of multivariate copulas are presented. In Section 3, we discuss the concept of copula-based regression estimation in brief. Section 4 cares for some numerical experiments under different scenario of marginals and copula dependence and provide a comparative study for the considered estimation techniques. Two real data analyses are reported in Section 5. Finally, a brief discussion about findings of the study are concluded in Section 6.

2. Copula-based regression models

Let Y be a continuous response variable and $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ be explanatory random vector. Let F be the joint distribution of $(Y, \mathbf{X})$. For a Borel measurable function ξ, the conditional expectation of $\xi(Y)$ given $\mathbf{X} = \mathbf{x}$ is given by

$$r(\mathbf{x}) = E(\xi(Y)|\mathbf{X} = \mathbf{x}) = \int \xi(y) \frac{\partial}{\partial y} F_{Y|\mathbf{X}}(y|\mathbf{x}) dy,$$

where $F_{Y|\mathbf{X}}(y|\mathbf{x})$ is conditional distribution of Y given $\mathbf{X}$.

In terms of copula density, the conditional expectation of $\xi(Y)$ given $\mathbf{X} = \mathbf{x}$ is

$$r(\mathbf{x}) = E(\xi(Y)|\mathbf{X} = \mathbf{x}) = \int \xi(y) \frac{c(F_Y(y), \mathbf{F}(\mathbf{x}))}{c_X(\mathbf{F}(\mathbf{x}))} dF_Y(y), \tag{2.1}$$

where $c(v, \mathbf{u}) = c(v, u_1, \ldots, u_k) = \dfrac{\partial^{k+1} C(v, u_1, \ldots, u_k)}{\partial v \partial u_1 \ldots \partial u_k}$ is the density of the copula C, $c_X(\mathbf{u}) = c_X(u_1, \ldots, u_k) = \dfrac{\partial^k C(1, u_1, \ldots, u_k)}{\partial u_1 \ldots \partial u_k}$ is the density of the copula associated with the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, and $\mathbf{F}(\mathbf{x}) = (F_{X_1}(x_1), \ldots, F_{X_k}(x_k))$.

Under the pairwise independence assumption on explanatory variables, Eq. (2.1) leads to

$$r(\mathbf{x}) = E(\xi(Y)|\mathbf{X} = \mathbf{x}) = \int \xi(y)c(F_Y(y), \mathbf{F}(\mathbf{x}))dF_Y(y).$$

It may be noticed that Eq. (2.1) is a general form for the conditional expectation in terms of copulas. One can obtain higher order conditional moments for a suitable choice of function ξ. When $\xi(y) = y$, Eq. (2.1) leads to the conditional mean function on Y on $\mathbf{X}$ which is well-known as a regression of Y on $\mathbf{X}$. Also, with the help of Eq. (2.1), we can deduce conditional variance for Y given $\mathbf{X}$ via relation $\mathrm{Var}\,(Y|\mathbf{X} = \mathbf{x}) = E\left(Y^2|\mathbf{X} = \mathbf{x}\right) - \left(E\left(Y|\mathbf{X} = \mathbf{x}\right)\right)^2$.

Next, we present regression functions for some well-known families of the multivariate copulas. The mathematical derivation of the results for the considered families of copula is either reported in or motivated by [6, 22, 25, 34].

2.1. Regression for elliptical copula

Let the joint distribution of the random vector $(Y, \mathbf{X})$ be determined by a $(k+1)$-dimensional elliptical copula of the form

$$C(v, \mathbf{u}) = \mathcal{H}_\Sigma\left(\mathcal{H}^{-1}(v), \varkappa^{-1}(\mathbf{u})\right),$$

where $\mathcal{H}_\Sigma$ is a multivariate elliptical distribution with correlation matrix Σ, $\mathcal{H}^{-1}$ is quantile function for univariate elliptical distribution, and $\varkappa^{-1}(\mathbf{u}) = (\mathcal{H}^{-1}(u_1), \ldots, \mathcal{H}^{-1}(u_k))$. Multivariate Gaussian and Student's-t copulas are two important members of the multivariate elliptical copula with wide applications in finance and risk management. Then the regression function of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$r(\mathbf{x}) = \int y \frac{\mathbf{h}_\Sigma\left(\mathcal{H}^{-1}(F_Y(y)), \varkappa^{-1}(\mathbf{F}(\mathbf{x}))\right)}{\mathbf{h}_{\Sigma_X}\left(\varkappa^{-1}(\mathbf{F}(\mathbf{x}))\right)} dy,$$

where $\mathbf{h}_\Sigma$ and $\mathbf{h}_{\Sigma_X}$ are joint density of multidimensional elliptical distributions for the vector $(Y, \mathbf{X})$ and $\mathbf{X}$, respectively, and Σ_X is correlation matrix for the vector $\mathbf{X}$. Next, we have the following examples.

Example 2.1. Let the joint distribution of a $(k+1)$-dimensional random vector $(Y, \mathbf{X})$ be determined by a Gaussian copula with correlation matrix

$$\Sigma = \begin{pmatrix} 1 & \boldsymbol{\rho}' \\ \boldsymbol{\rho} & \Sigma_X \end{pmatrix}$$

of the form

$$C(v, \mathbf{u}) = \mathbf{\Phi}_\Sigma\left(\Phi^{-1}(v), \boldsymbol{\varphi}^{-1}(\mathbf{u})\right),$$

where $\mathbf{\Phi}_\Sigma$ is the joint cumulative distribution of a multivariate normal random variable with mean vector zero and correlation matrix Σ, Φ^{-1} is inverse distribution function of standard normal variate, and $\boldsymbol{\varphi}^{-1}(\mathbf{u}) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k))$. Then the regression function of Y given $\mathbf{X} = \mathbf{x}$ is [22, 25]

$$r(\mathbf{x}) = E\left[F_Y^{-1}\left(\Phi\left(\mathbf{u}'\Sigma_X^{-1}\boldsymbol{\rho} + Z\sqrt{1 - \boldsymbol{\rho}'\Sigma_X^{-1}\boldsymbol{\rho}}\right)\right)\right],$$

where $\mathbf{u}' = (\Phi^{-1}(F_{X_1}(x_1)), \ldots, \Phi^{-1}(F_{X_k}(x_k)))$, $\boldsymbol{\rho}' = (\mathrm{corr}(Y, X_1), \ldots, \mathrm{corr}(Y, X_k))$, Z is standard normal variate with cumulative distribution function Φ, and Σ_X is correlation matrix of vector $\mathbf{X}$.

In particular for a bivariate Gaussian copula, if Y and X_1 follow $N(0, 1)$, then regression function of Y given $X_1 = x_1$ is linear of the following form

$$r(x_1) = E(Y|X_1 = x_1) = \rho_1 x_1,$$

where ρ_1 is correlation between Y and X_1 [34].

Example 2.2. Consider a $(k+1)$-dimensional t-copula of the form

$$C(v, \mathbf{u}) = \mathbf{t}_{\nu,\Sigma}\left(t_\nu^{-1}(v), \mathfrak{T}_\nu^{-1}(\mathbf{u})\right),$$

where $\mathbf{t}_{\nu,\Sigma}$ is the multivariate Student's-t distribution with degrees of freedom ν and correlation matrix Σ, and t_ν^{-1} is the inverse distribution function of univariate Student's-t distribution with ν degrees of freedom and $\mathfrak{T}_\nu^{-1}(\mathbf{u}) = (t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_k))$. Let the joint distribution of the random vector $(Y, \mathbf{X})$ be given by t-copula. Then conditional mean function of Y on $\mathbf{X}$ is

$$r(\mathbf{x}) = E\left[F_Y^{-1}\left(t_\nu(\boldsymbol{\rho}'\Sigma_X^{-1}\boldsymbol{u} + \sqrt{\nu(1-\boldsymbol{\rho}'\Sigma_X^{-1}\boldsymbol{\rho})\left(1+\frac{1}{\nu}\boldsymbol{u}'\Sigma_X^{-1}\boldsymbol{u}\right)/(\nu+k)}Z\right)\right],$$

$\mathbf{u}' = (t_\nu^{-1}(F_{X_1}(x_1)), \ldots, t_\nu^{-1}(F_{X_k}(x_k)))$, $\boldsymbol{\rho}' = (\mathrm{corr}(Y, X_1), \ldots, \mathrm{corr}(Y, X_k))$, Z is standard univariate t random variable [22].

It may be verified that if all marginals are identically distributed as a Student's-t with location parameter μ and scale parameter σ, then for bivariate case the regression function of Y given $X_1 = x_1$ is

$$r(x_1) = E(Y|X_1 = x_1) = \rho x_1 + (1-\rho)\mu,$$

where ρ is correlation between Y and X_1.

2.2. Regression for Archimedean copula

Consider a $(k+1)$- dimensional Archimedean copula of the form

$$C(v, \mathbf{u}) = \psi^{-1}\left(\psi(v) + \sum_{i=1}^{k}\psi(u_i)\right),$$

where ψ is a generator function. Let the joint distribution of the random vector $(Y, \mathbf{X})$ be given by $C(v, \mathbf{u})$. Then the regression function of Y on $\mathbf{X} = \mathbf{x}$ is given by

$$r(\mathbf{x}) = E\left[Y\frac{\psi^{-1(k+1)}(\psi(C(F_Y(y), \mathbf{F}(\mathbf{x}))))}{\psi^{-1(k)}(\psi(C_X(\mathbf{F}(\mathbf{x})))}\psi'(F_Y(y))\right],$$

where $\psi^{-1(k+1)}$ and $\psi^{-1(k)}$ represent the $(k+1)$th and kth derivatives of ψ^{-1} and ψ' is first order derivative of ψ.

Example 2.3. Let the distribution of the random vector $(Y, \mathbf{X})$ be determined by the Clayton copula defined by

$$C(v, \mathbf{u}) = \left(v^{-\theta} + \sum_{i=1}^{k}u_i^{-\theta} - (k+1) + 1\right)^{-1/\theta}, \quad \theta \in (-1, \infty).$$

Then regression function of Y on $\mathbf{X}$ is

$$r(\mathbf{x}) = (k\theta+1)\left\{C_X(\mathbf{F}(\mathbf{x}))\right\}^{-(k\theta+1)}\int y\{F_Y(y)\}^{-(\theta+1)}\{C(F_Y(y), \mathbf{F}(\mathbf{x}))\}^{((k+1)\theta+1)}dF_Y(y).$$

When $k = 1$, for the bivariate vector (Y, X_1), we have

$$r(x_1) = (\theta+1)\{F_{X_1}(x_1)\}^{-(\theta+1)}\int y\{F_Y(y)\}^{-(\theta+1)}\{C(F_Y(y), F_{X_1}(x_1))\}^{(2\theta+1)}dF_Y(y).$$

3. Copula-based regression estimation

Eq. (2.1) is a general expression for regression in terms of copulas and can be used as an estimating equation. For $\xi(y) = y$, if $\widehat{F}_Y$ and $\widehat{\mathbf{F}}(\mathbf{x}) = (\widehat{F}_{X_1}(x_1), \ldots, \widehat{F}_{X_k}(x_k))$ are given estimators of F_Y and $\mathbf{F}$, respectively. Also, if $\widehat{c}$ and $\widehat{c}_X$ are some known estimators for densities c and c_X, respectively, then regression function $r(\mathbf{x})$ can be estimated by

$$\widehat{r}(\mathbf{x}) = \int y \frac{\widehat{c}(\widehat{F}_Y(y)), \widehat{\mathbf{F}}(\mathbf{x}))}{\widehat{c}_X(\widehat{\mathbf{F}}(\mathbf{x}))} d\widehat{F}_Y(y). \tag{3.1}$$

In the literature, different techniques are available for estimating the copulas and marginals. It may be noticed that the estimator in Eq. (3.1) may be parametric, semiparametric and nonparametric in nature depending on its component estimation. In general, classical maximum likelihood (ML) estimation is adopted to estimate the copula and marginals parameter simultaneously in parametric method. However, this method is computationally tedious and estimates of the copula parameters may be affected by marginals misspecification. Two stage parametric estimation technique, which is well-known as inference function for margins (IFM) has emerged as alternative of the ML in literature. It is more flexible, easy in computation, and equally efficient as ML method. The two stage semiparametric pseudo maximum likelihood (PML) method proposed by [11] is commonly used in recent years for semiparametric case. First of all, we briefly discuss these estimation techniques here.

Let $\mathbf{X}_i = (X_{1,i}X_{2,i}, \ldots, X_{k,i})$. For a random sample $(Y_i, \mathbf{X}_i)$ of size n from $(Y, \mathbf{X})$. The log-likelihood function $\mathcal{L}$ is given by

$$\mathcal{L} = \sum_{i=1}^n \log c(F_Y(Y_i), F_{X_1}(X_{1,i}), \ldots, F_{X_k}(X_{k,i})) + \sum_{i=1}^n \log f_Y(Y_i) + \sum_{j=1}^k \sum_{i=1}^n \log f_{X_j}(X_{j,i}), \tag{3.2}$$

where f_Y and $f_{X_j}(X_j)$, $j = 1, 2, \ldots, k$ are densities of the random variables Y and X_j's, respectively. Let $C(\cdot, \theta)$ be a parametric copula. Let $\mathcal{C}_0 = \{C(\cdot, \theta) : \theta \in O\}$, where O is an open subset of $\mathbb{R}^k$. Let $Y \sim F_Y(y; \alpha)$, and $X_j \sim F_{X_j}(x_j; \beta_j), j = 1, 2, \ldots, k$. Then, in classical ML estimation technique, the model parameters are estimated by maximizing Eq. (3.2) taking into account the parametric copula and parametric marginals. In IFM method, the marginals parameters α and $\beta_j, j = 1, 2, \ldots, k$ are estimated by using Y_j and $X_{j,1}, \ldots, X_{j,n}$, respectively through the classical ML approach in the first step. Using these estimates for the marginal parameters, an estimate $\widehat{\theta}_n$ of the copula parameter θ is obtained in second step by maximizing pseudo-log-likelihood equation defined by [16,19]

$$\mathfrak{L}(\theta) = \sum_{i=1}^n \log c(F_Y(Y_i; \widehat{\alpha}), F_{X_1}(X_{1,i}; \widehat{\beta}_1), \ldots, F_{X_k}(X_{k,i}; \widehat{\beta}_k); \theta),$$

where $\widehat{\alpha}$ and $\widehat{\beta}_j$ are estimates for α and $\beta_j, j = 1, 2, \ldots, k$, respectively.
However, in semiparametric PML approach, first the marginals are estimated from data nonparametrically through empirical distribution function which is defined by

$$\widehat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^n I(Y_i \le y),$$

where $I\{Y_i \le y\} = 1$ for $Y_i \le y$, and zero otherwise. Similarly, other marginals are estimated via $\widehat{F}_{X_j}(x_j) = \frac{1}{n} \sum_{i=1}^n I(X_{ji} \le x_j), \quad j = 1, \ldots, k$. In the second step, estimate for the copula parameter is obtained by maximizing pseudo-log-likelihood function of the form

$$\mathfrak{L}(\theta) = \sum_{i=1}^n \log c(\widehat{F}_Y(Y_i), \widehat{F}_{X_1}(X_{1,i}), \ldots, \widehat{F}_{X_k}(X_{k,i}); \theta).$$

That is

$$\widehat{\theta}_n = \arg\max_{\theta} \mathfrak{L}(\theta).$$

Based on these estimation techniques for copulas and marginals, a parametric and semiparametric estimators for copula-based regression function $r(\mathbf{x})$ are defined by

$$\widehat{r}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{c(F_Y(Y_i;\widehat{\alpha}), \mathbf{F}(\mathbf{x};\widehat{\boldsymbol{\beta}});\widehat{\theta}_n)}{c_X(\mathbf{F}(\mathbf{x};\widehat{\boldsymbol{\beta}});\widehat{\theta}_n)}$$

and

$$\widehat{r}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{c(\widehat{F}_Y(Y_i), \widehat{\mathbf{F}}(\mathbf{x});\widehat{\theta}_n)}{c_X(\widehat{\mathbf{F}}(\mathbf{x});\widehat{\theta}_n)},$$

where $\mathbf{F}(\mathbf{x};\widehat{\boldsymbol{\beta}}) = (F_{X_1}(x_1;\widehat{\beta}_1), \ldots, F_{X_k}(k_k;\widehat{\beta}_k))$ and $\widehat{\mathbf{F}}(\mathbf{x}) = (\widehat{F}_{X_1}(x_1), \ldots, \widehat{F}_{X_k}(k_k))$ are parametric and nonparametric estimates for $\mathbf{F}(\mathbf{x})$, respectively.

A theoretical comparison for the proposed estimators for the regression function is quite difficult task in a general setup. Therefore, we perform a well-organized numerical scheme to study the behavior of these estimators in the next section.

4. Simulation study

In this section, we undertake some numerical experiments to compare the efficacy of the considered estimation approaches, that is, ML, IFM, and PML estimation for regression function in terms of regression estimated variance (EV), mean squared error (MSE), and relative efficiency under correctly specified and misspecified copulas and marginals. For a large sample, 95% confidence intervals for the regression estimators are also obtained. To carry out a simulation study, we consider the following bivariate families of copulas here:

(a) Gaussian Copula: $C(v,u_1) = \Phi_2\left(\Phi^{-1}(v), \Phi^{-1}(u_1); \rho\right)$.

(b) Ali-Mikhail-Haq (AMH): $C(v,u_1) = vu_1/[1-\theta(1-v)(1-u_1)]$.

(c) Frank Copula: $C(v,u_1) = \frac{-1}{\theta}\left(1 + \frac{(e^{-\theta v}-1)(e^{-\theta u_1}-1)}{(e^{-\theta}-1)}\right)$.

(d) Clayton Copula: $C(v,u_1) = \left(v^{-\theta} + u_1^{-\theta} - 1\right)^{-1/\theta}$.

(e) Plackett Copula: $C(v,u_1) = \frac{[1+(\theta-1)(v+u_1)] - \sqrt{[1+(\theta-1)(v+u_1)]^2 - 4\theta(\theta-1)vu_1}}{2(\theta-1)}$.

(f) Gumbel: $C(v,u_1) = \exp\{-[(-\log v)^\theta + (-\log u_1)^\theta]^{1/\theta}\}$.

(g) Student's-t: $C(v,u_1) = t_{2,\Sigma}\left(t_2^{-1}(v), t_2^{-1}(u_1); \rho\right)$.

These copulas are well-known members of widely used elliptical, Archimedean, and non-Archimedean family of copulas which cover a wide range of dependence. For more detail, one can refer to [17, 24, 29].

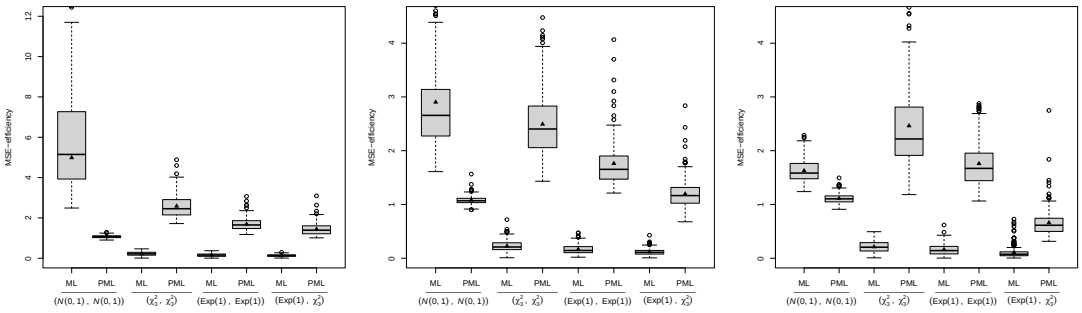
To assess the robustness properties of the considered methods, we consider the following sets of univariate distributions: (i) $X_1 \sim N(0,1)$, $Y \sim N(0,1)$, (ii) $X_1 \sim \chi^2_3$, $Y \sim \chi^2_3$, (df=3), (iii) $X_1 \sim \text{Exp}(1)$, $Y \sim \text{Exp}(1)$, where Exp stands for exponential, and (iv) $X_1 \sim \text{Exp}(1)$, $Y \sim \chi^2_3$. When underlying copula is Gaussian and marginals are normal, it leads to correct specifications in copula and marginals both, whereas each of the other combinations corresponds to misspecification either in copula or in marginals or in both. Let $r^{(i)}(x_1)$ and $r_m^{(i)}(x_1)$ denotes the estimates of regression function and estimated mean regression function of $r(x_1)$ for the ith repeated sample, respectively. Let $r_0(x_1)$ be the true value of regression function, and N denotes the number of repeated samples. Then, we define regression estimated variance (EV)$=N^{-1}\sum\{r^{(i)}(x_1) - r_m^{(i)}(x_1)\}^2$, which measure how far each estimated value of the regression function from their mean of the estimated regression function and mean squared error (MSE)$=N^{-1}\sum\{r^{(i)}(x_1) - r_0(x_1)\}^2$, which is the average of the squares of the errors. The smallest value of EV and MSE of the regression estimator indicates the better performance of the corresponding estimation

approach. Also, the estimated relative MSE-efficiency of method 'M' with respect to IFM method is defined by {Estimated MSE of IFM}/{Estimated MSE of 'M'}, where 'M' stands for ML and PML. The large value of relative MSE-efficiency indicates better performance of the method 'M'. Using regression estimate and standard error (SE), we compute a 95% confidence interval of the form $\widehat{r}(x_1) \pm 1.96$ SE$(\widehat{r}(x_1))$. All the numerical computations were performed using freely available packages in R software (Version 4.2.1). Computations for the numerical experiments considered here in this study are much time consuming, therefore we restrict the number of observations in each sample to 100 and the number of repeated samples to 200 throughout this study.

The main findings of the simulation study are reported in Tables 1-4 and Figure 1-2 below. The EV and MSE of the regression estimators are presented in Table 1 and Table 2. From Table 1, it is quite clear that ML and IFM methods are highly nonrobust under misspecified marginals. For specified marginals ML is robust. In case of specified marginals and misspecified copulas, Table 2 suggest that PML works slightly better over ML and IFM method for AMH, Frank, and Plackett family of copulas. However, the difference between MSE's of these estimators are small, while in case of Clayton, Gumbel, and Student's-t copula, PML performs more robust than ML and IFM. One more thing here, we notice that very large values for EV and MSE in Table 1 and Table 2 are not precise, but we presented them because they successfully reveals the dominating nature of PML over ML and IFM methods. Moreover, Figure 1 shows the relative MSE-efficiency plots of ML and PML methods with respect to IFM method for Gaussian copula with different choice of marginals. Under correct specification of copula and marginals, the ML method performs better over the PML technique as expected and small values of dependence parameter efficiency of ML method is very high as compared to PML and rapidly decreases as dependence parameter increases. In case of misspecified marginals and Gaussian copula, the PML performs better over ML method. Figure 2 display the relative MSE-efficiency plots of regression function for correctly specified normal marginals and misspecified copulas. From Figure 2, we observe that PML performs slightly better than ML for AMH, Frank, Plackett, and Clayton copulas for the lower values of the copula parameters. However, for higher values of the copula parameters, the PML is more efficient than ML when copula is Clayton. As well as, in case of Gumbel and student's-t copula, it can be observed that the PML technique become more efficient as the value of dependence parameter increases. We also get the same observation for other dependence parameter which are not displayed here by the authors to save the space.

Table 1. Estimated variance (EV) and MSE (*) of regression function for Gaussian copula and different marginals.

$\overrightarrow{\text{Margins}}$	$(N(0,1), N(0,1))$			(χ^2_3, χ^2_3)			(Exp(1), Exp(1))			$(\text{Exp}(1), \chi^2_3)$		
ρ	ML	IFM	PML	ML	IFM	PML	ML	IFM	PML	ML	IFM	PML
0.1	0.050	0.300	0.279	102	1.550	0.370	3.58	0.08	0.040	53.6	0.694	0.37
	0.32*	3.690*	3.48*	622*	83.0*	29.9*	39.3*	4.56*	2.86*	499*	53.2*	29.9*
0.2	0.093	0.409	0.387	82.1	1.81	0.440	3.90	0.09	0.040	44.2	0.77	0.44
	0.501*	5.410*	5.130*	602*	91.4*	33.6*	41.5*	5.02*	3.15*	506*	57.0*	33.6*
0.5	0.522	1.100	1.068	165	3.49	0.840	6.12	0.166	0.08	84.8	1.25	0.84
	4.26*	18.43*	17.35*	938*	137*	53.5*	59.7*	7.51*	4.67*	750*	75.5*	53.5*
0.8	2.898	4.143	4.195	355	11.4	2.36	25.5	0.440	0.21	687	2.83	2.36
	48.64*	93.92*	87.47*	2208*	324*	132*	162*	17.3*	10.5*	2327*	125*	133*
0.9	6.898	9.160	8.896	785	25.2	4.94	57.3	0.92	0.45	423	4.00	4.94
	150*	238.0*	216.0*	4591*	637*	262*	334*	33.6*	20.1*	2797*	165*	262*



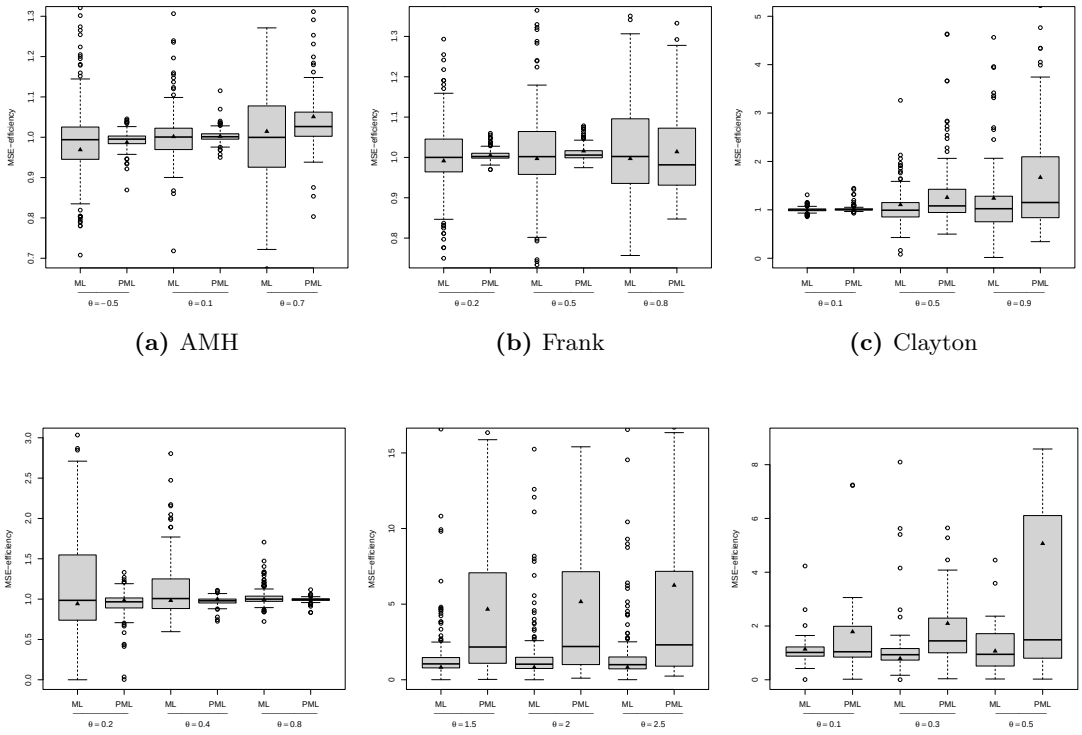
(a) ρ=0.5 **(b)** ρ=0.7 **(c)** ρ=0.9

Figure 1. Relative MSE-efficiency box plots of regression function for ML and PML methods for Gaussian copula and different marginals. Here, triangle dot represents the average value of the relative MSE-efficiency.

Table 2. Estimated variance (EV) and MSE (*) of regression function for different copulas and normal marginals.

↓Copula	θ	ML	IFM	PML	↓Copula	θ	ML	IFM	PML
	-0.50	0.0184	0.0110	0.0108		0.20	0.0092	0.0085	0.0089
		0.2672*	0.2589*	0.2622*			0.2555*	0.2542*	0.2530*
AMH	0.10	0.0101	0.0102	0.0102	Frank	0.50	0.0100	0.0090	0.0091
		0.2589*	0.2575*	0.2569*			0.2613*	0.2603*	0.2590*
	0.70	0.0162	0.0181	0.0140		0.80	0.0107	0.0094	0.094
		0.3502*	0.3556*	0.3342*			0.2678*	0.2672*	0.2635*
	0.20	0.0809	0.0244	0.0219		0.10	0.0063	0.0071	0.0059
		0.4635*	0.5668*	0.5241*			0.5315*	0.5330*	0.5272*
Plackett	0.40	0.0287	0.0127	0.0123	Clayton	0.50	68.2321	56.185	0.0665
		0.6063*	0.5568*	0.5231*			8.4160*	7.7610*	0.8692*
	0.80	0.0115	0.0091	0.0091		0.90	301.98	639.38	0.2310
		0.6159*	0.5627*	0.5251*			17.680*	25.530*	1.4570*
	1.5	84.7388	55.5257	0.7582		0.1	4.0621	7.8002	0.0685
		9.9460*	8.2036*	1.7594*			2.1655*	2.8977*	1.54603*
Gumbel	2.0	225.1018	145.1410	1.3293	Student's-t	0.3	13.9987	4.3279	0.0974
		16.2837*	13.3947*	2.5949*			3.7763*	2.2186*	0.9931*
	2.5	525.4496	342.4594	1.9402		0.5	127.4184	145.3792	0.2280
		24.5084*	20.2524*	3.2423*			11.5866*	12.3601*	2.3925*

Table 3 and Table 4 show an approximate 95% confidence interval for regression estimator based on normal approximation. The lower confidence limit and upper confidence limit are denoted by CL and CU, respectively. The shortest length confidence intervals are obtained under the correct specifications of copula and marginals in ML method in Table 3. However, in the case of misspecified margins, the confidence intervals are wider. For correctly specified normal margins and misspecified copula families, one can easily observe from Table 4 that PML has smaller confidence intervals over ML and IFM methods. We also consider the different choices of copulas and marginals under misspecified scenarios. In such cases, we obtained a highly non-robust result in terms of relative efficiency, EV, and MSE. Therefore, these results are not reported in this paper.



(a) AMH **(b)** Frank **(c)** Clayton

(d) Plackett **(e)** Gumbel **(f)** Student's-t

Figure 2. Relative MSE-efficiency box plots of regression function for ML and PML method for different family of copulas with normal marginals. Here, triangle dot denotes the average value of the relative MSE-efficiency.

Table 3. 95% confidence interval (C.I.) of the regression function for Gaussian copula with different margins.

$\overrightarrow{\text{Margins}}$		$(N(0,1), N(0,1))$		(χ^2_3, χ^2_3)		(Exp(1), Exp(1))		$(\text{Exp}(1), \chi^2_3)$	
↓Techniques	ρ	CL	CU	CL	CU	CL	CU	CL	CU
ML	0.1	-0.0176	0.9375	26.1626	71.8649	-6.3218	18.3074	-22.7536	65.0271
	0.2	-1.7892	0.9939	-25.3768	71.0831	-6.5088	18.8190	-22.6871	65.7470
	0.5	-5.9078	2.2053	-32.3548	88.0380	-7.8361	22.5270	-27.9881	79.6610
	0.8	-20.4271	6.9801	-49.2540	135.440	-13.3071	36.7581	-54.2701	135.331
	0.9	-36.0251	12.144	-71.4060	194.873	-19.2641	52.6513	-55.1848	152.671
IFM	0.1	-5.5362	2.0151	-8.8671	26.9401	-2.1109	6.2819	-7.7068	21.5803
	0.2	-6.7403	2.4056	-9.3073	28.2574	-2.2519	6.5965	-7.3294	22.3445
	0.5	-12.5521	4.3191	-11.4436	34.6142	-2.6919	8.0779	-8.4508	25.7163
	0.8	-28.4889	9.6001	-17.6819	53.0116	-4.0749	12.2838	-10.8992	33.0180
	0.9	-45.4130	15.2136	-24.8489	74.2972	-5.6712	17.1241	-12.5616	38.0271
PML	0.1	-5.3777	1.9651	-5.3057	1.18270	-1.6913	4.9655	-5.3057	16.1826
	0.2	-6.5576	2.3441	-5.6272	17.1547	-1.7668	5.2105	-5.6272	17.1547
	0.5	-12.1753	4.1974	-7.1121	21.6555	-2.1349	6.3606	-7.1121	21.6555
	0.8	-27.4714	9.2832	-11.2121	34.0852	-3.1850	9.5953	-11.2122	34.0852
	0.9	-43.2401	14.5150	-15.7653	47.8759	-4.3842	13.2457	-15.7653	47.8759

Table 4. 95% confidence interval (C.I.) of regression function for different copulas and normal margins.

Techniques →		ML		IFM		PML	
↓Copula	θ	CL	CU	CL	CU	CL	CU
AMH	-0.50	-1.0056	1.0257	-0.9883	1.0112	-0.9985	1.0139
	0.10	-0.9953	0.9985	-0.9953	0.9987	-0.9937	0.9981
	0.70	-1.2129	1.1126	-1.22218	1.1217	-1.1735	1.0982
Frank	0.20	-0.9864	1.0000	-0.9830	0.9986	-0.9795	0.9974
	0.50	-0.9973	1.0114	-0.9953	1.0098	-0.9886	1.0075
	0.80	-1.0098	1.0241	-1.0090	1.0224	-0.9984	1.0191
Plackett	0.20	-1.2578	1.2711	-1.1953	1.1575	-1.2252	1.1950
	0.40	-1.1142	1.1134	-1.0993	1.0891	-1.1171	1.0943
	0.80	-1.0353	1.0244	-1.0349	1.0211	-1.0406	1.0230
Clayton	0.10	-1.0710	1.0176	-1.0749	1.0176	-1.0589	1.0132
	0.50	-17.7370	15.339	-16.0540	13.960	-2.0403	1.3757
	0.90	-37.692	31.825	-53.396	46.947	-3.7436	1.9832
Gumbel	1.5	-15.1716	23.9145	-12.0851	20.1537	-1.3692	5.5451
	2.0	-25.0571	38.9351	-19.8729	32.7660	-2.2200	7.9774
	2.5	-38.8939	57.4200	-31.0095	48.5790	-2.8955	9.8962
Student's-t	0.1	-4.1268	4.3832	-5.5251	5.8625	-1.9799	1.9705
	0.3	-7.2151	7.6253	-4.2281	4.4908	-1.9661	1.9365
	0.5	-23.4022	22.1312	-24.4242	24.1489	-4.7282	4.6739

5. Real data analysis

To illustrate the performance of the considered estimation techniques for copula-based regression functions, we have analyzed two different real data sets here. First of all, we standardized the chosen data sets and performed Anderson-Darling (AD) test for the goodness of fits for the marginals. The AD-test is performed using R-software through 'fitur' package. One of the main challenging tasks is to choose a suitable family of copulas for fitting the data. So, we adopt a reasonable strategy for the selection of copula family and perform the goodness of fit test. Recently, several new methods for the goodness of fit test of copulas have been proposed [3, 12, 13, 20]. Here, we have used the test studied by [12], which is based on the regularized test statistic R_n and is available in R software (see package 'copula' [15]). This test involves PML technique for estimation. The null hypothesis under consideration is $H_0 : C \in \mathcal{C}_0$ where $\mathcal{C}_0$ is a class of parametric families of copulas used in the simulation study. In addition, we also analyze the behavior of the best estimation method based on the predictive performance for real data sets. For prediction performance, we chose the cross-validation (CV) error criteria explored in [1, 25]. The mathematical formula of CV error is given by $\text{CV} = \text{median}_{1\le i\le n}|Y_i - \widehat{r}_{-i}(X_{1i})|$, where $\widehat{r}_{-i}(X_{1i})$ denotes the estimate of $r(x_1)$ from data set $\{(Y_j, X_{1j});\ j \neq i, j = 1, 2, \ldots, n\}$.

5.1. Real data analysis 1

Here we consider the 'bodyfat' data set which contains the percentage of body fat as a response variable along with covariates which represent several physiological measurements related to 252 men [27]. This data set is also available in R-software under package 'mfp'. In the present study, we restrict ourselves to only two variables namely, siri (body fat percent) and ankle circumference (in cm). In this data set, the variable X_1 represents the ankle circumference and Y represents the siri. The scatter plot displayed in Figure 3 shows that the variable Y and X_1 are positively correlated. The correlation between both variables is 0.267.

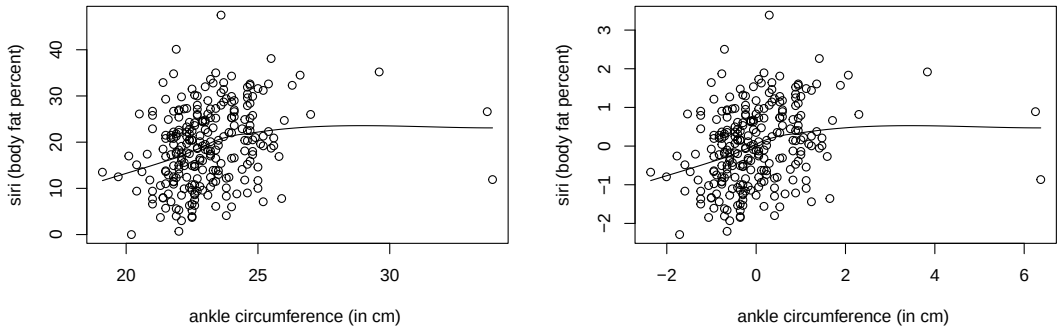


Figure 3. Scatter plot between siri (Y) and ankle circumference (X_1) on original (left) and standardized (right) scale.

From the Anderson-Darling (AD) test for the standardized data, we observed that X_1 and Y both fit normal distribution with p-values 0.8864 and 0.8896, respectively. We also performed the goodness-of-fit test for copula selection for the 'bodyfat' data set. The values of copula parameter estimates, test statistic R_n, and p-values are reported in Table 5 for different families of copulas.

Table 5. Summary statistics for bodyfat data set.

	Gaussian	Clayton	Frank	Plackett	AMH	Gumbel	Student's-t
Estimate($\hat{\theta}$)	**0.3169**	0.5239	1.9011	2.4777	0.6690	1.2356	0.2891
R_n-statistic	**0.2469**	0.3909	0.2793	0.2945	0.3543	0.3749	0.3602
p-value	**0.0744**	0.0045	0.0274	0.0235	0.0025	0.0125	0.0055

Based on the smallest R_n-statistic with the largest p-value, we found that the Gaussian copula provides the best fit to the given data set. As the copula is Gaussian and both marginals are normal, it corresponds to the correct specification scenario. The relative efficiency of the regression estimator with respect to IFM technique is reported in Table 6. The regression EV and MSE for the ML, IFM, and PML techniques are also reported in Table 7. From Table 6 and Table 7, we observe that the ML technique performs better than the IFM and PML methods for the considered data set if the copula and marginals are correctly specified. Moreover, the cross-validation (CV) error of the regression function for ML and PML methods are reported in Table 8. The smallest error corresponding to the Gaussian copula and standard normal marginals affirms the best predictive performance under the ML method. The boxplot of CV error values of the regression function for different families of copulas under ML and PML methods (see Figure 4) also support this study. Overall the copula-based regression function for the Gaussian copula performs better than that for other families of copulas considered in this study.

Table 6. Relative MSE-efficiency (in %) of ML and PML of regression function for bodyfat data.

↓Techniques	Gaussian	Plackett	AMH	Clayton	Frank	Gumbel	Student's-t
ML	**1285.049**	99.9855	100.8865	109.1891	100.3827	91.5162	84.8216
PML	102.4017	101.015	101.2518	270.6355	101.3249	133.5509	556.3656

Table 7. Estimated Variance(EV) and MSE of regression function for bodyfat data set.

Techniques →	ML		IFM		PML	
↓Copula	EV	MSE	EV	MSE	EV	MSE
Gaussian	**0.06060**	**0.53399**	0.18953	6.86204	0.19084	6.70110
AMH	0.00621	0.34200	0.00691	0.34504	0.00705	0.34077
Frank	0.00467	0.27589	0.00542	0.27695	0.00537	0.27333
Plackett	0.00301	0.28249	0.00307	0.28245	0.00299	0.27961
Clayton	3.16875	2.21874	4.18806	2.42262	0.05196	0.89516
Gumbel	0.33395	0.48855	0.27994	0.44710	0.02175	0.33478
Student's-t	1.03439	1.12750	0.86048	0.95637	0.08963	0.17189

Table 8. Cross-validation (CV) error of ML and PML of regression function for bodyfat data.

↓Techniques	Gaussian	Plackett	AMH	Clayton	Frank	Gumbel	Student's-t
ML	**0.128350**	0.229823	0.206144	0.185438	0.229841	0.609933	0.666596
PML	0.161046	0.225764	0.185923	0.188548	0.222643	0.160290	0.239602

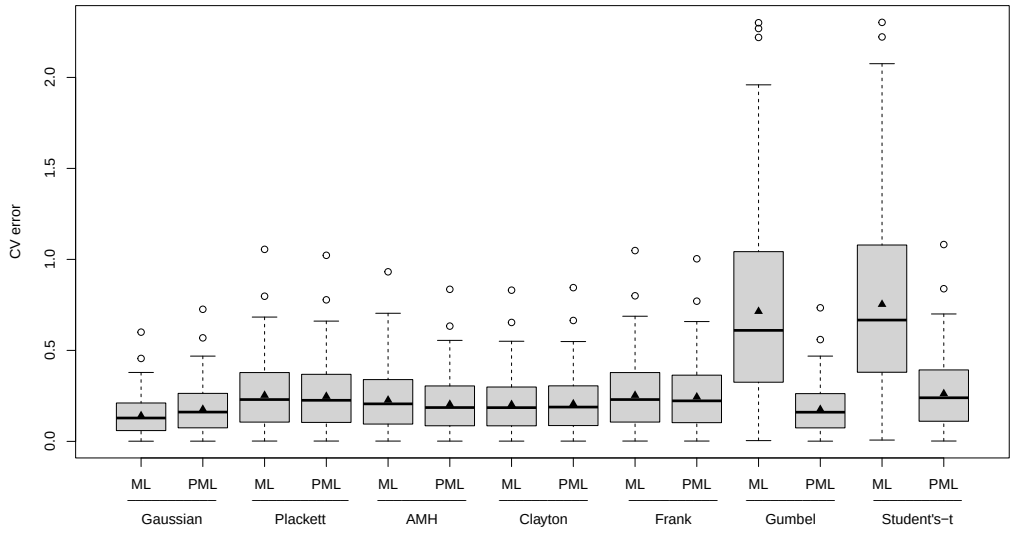


Figure 4. Box plots of the cross-validation (CV) error values of the regression function for ML and PML method. Triangle dots indicate the average value of CV errors.

5.2. Real data analysis 2

Here, we consider 'growth data' studied in [8]. This data measures the national growth of 61 countries around the world during years 1960 to 1985. There are five variables in this data set. GDP per worker growth is considered as the response variable, whereas labor force growth (LFG), equipment investment (EQP), non-equipment investment (NEQ), and the relative GDP gap (GAP) are explanatory variables. We restrict our study to a bivariate case with a response variable GDP per worker growth denoted by Y and explanatory variable LFG denoted by X_1. These variables are negatively correlated with correlation -0.148. The scatter plot of both variables also reveals the same (see Figure 5).

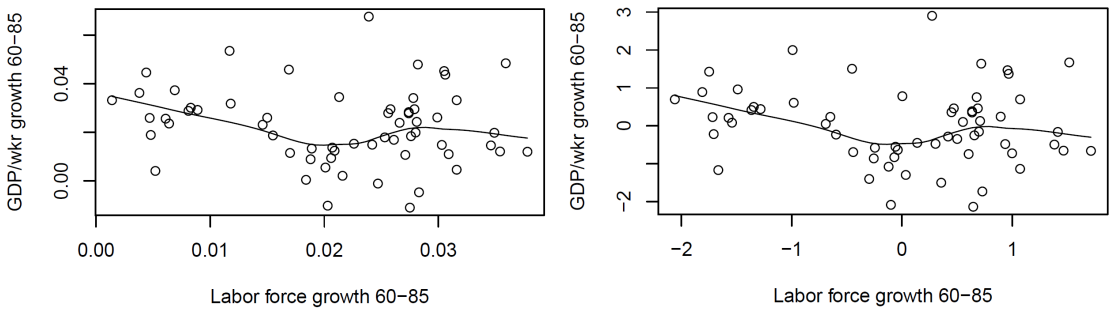


Figure 5. Scatter plot between GDP per worker growth (Y) and labor force growth (LFG) (X_1) on original (left) and standardized (right) scale.

We perform the Anderson-Darling (AD) test for marginals and calculate the goodness of fit-test statistic R_n for the copulas used in the simulation study. Both marginals fit the normal distribution with p-values 0.97 and 0.21, respectively. Table 9 presents the fitted summary statistic along with R_n values of different copulas except the Gumbel copula which yields a negative value of the Kendall's tau and hence does not provide a feasible fit to the given data set. The smallest R_n-statistic with the largest p-value in Table 9 suggest that the Clayton copula gives an acceptable fit compared to other copulas.

Table 9. Summary statistics for the growth data.

	Gaussian	**Clayton**	Frank	Plackett	AMH	Student's-t
Estimate($\hat{\theta}$)	-0.1376	**-0.1460**	-0.7586	0.6953	-0.2991	-0.1023
R_n-statistic	0.0802	**0.0708**	0.0797	0.0781	0.0721	0.0867
p-value	0.0395	**0.1603**	0.0375	0.0495	0.0545	0.0884

Table 10. Relative MSE-efficiency (in %) of ML and PML of regression function for growth data.

↓Techniques	Gaussian	Plackett	AMH	Clayton	Frank	Student's-t
ML	216.1137	99.9292	99.2922	**96.4919**	98.7256	61.4963
PML	105.0046	99.3589	99.8875	**100.6617**	98.9650	100.410

Table 11. Estimated Variance(EV) and MSE of regression function for growth data set.

$\overrightarrow{\text{Techniques}}$	ML		IFM		PML	
↓Copula	EV	MSE	EV	MSE	EV	MSE
Gaussian	0.00980	0.87134	0.18771	1.88309	0.17083	1.79334
AMH	0.01606	0.27026	0.01692	0.26835	0.01678	0.26865
Frank	0.02088	0.30555	0.02085	0.30166	0.02073	0.30481
Plackett	0.02055	0.28303	0.02053	0.28283	0.02030	0.28466
Clayton	0.04225	0.58859	0.02988	0.56794	**0.02325**	**0.56421**
Student's-t	0.54746	0.59776	0.33205	0.36760	0.05402	0.36610

Relative MSE-efficiency, EV, and MSE are reported in Tables 10 and 11. From these values, we observe that the PML technique works good if the copula is misspecified but the marginals are correct. Moreover, Table 12 and Figure 6 show the best prediction performance of the PML technique in terms of the smallest CV error for the Clayton copula. Overall we observe that the copula-based regression function for Clayton copula performs better than other families of copulas for the given data set.

Table 12. Cross-validation (CV) error of ML and PML of regression function for growth data.

↓Techniques	Gaussian	Plackett	AMH	Clayton	Frank	Student's-t
ML	0.018930	0.034897	0.029909	**0.012944**	0.037145	0.017474
PML	0.032690	0.047123	0.037411	**0.002884**	0.048047	0.046862

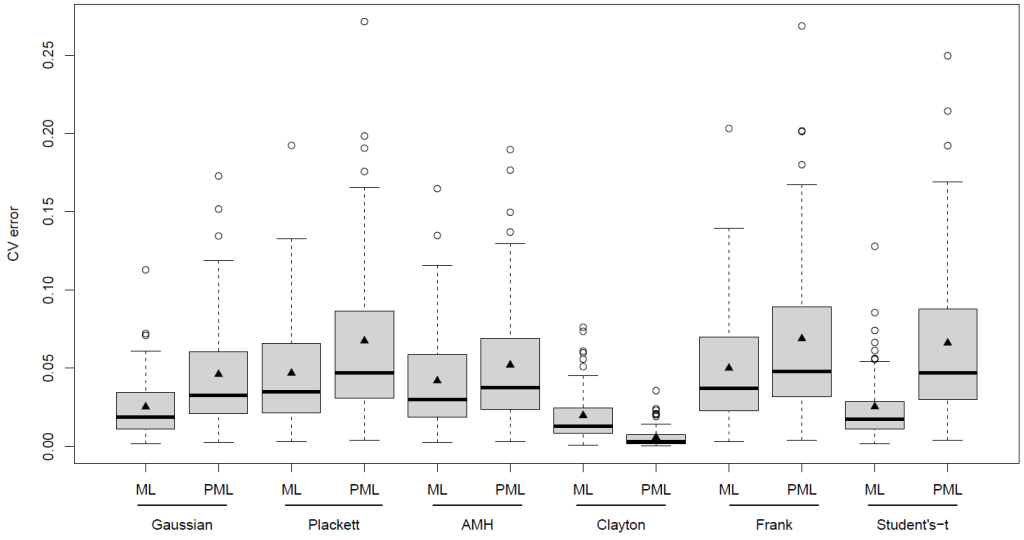


Figure 6. Box plots of the cross-validation (CV) error values of the regression function for ML and PML method where triangle dots indicate the average value of CV errors.

6. Conclusion

In this paper, we employed copulas to regression analysis, in which regression functions are expressed in terms of copula density and marginal distributions. Regression functions for some well-know families of multivariate copulas, namely, Student's-t copula, Gaussian copula, and Archimedian family of copulas are reported. In particular, some bivariate copula-based regression functions are also deduced. Based on a random sample, we have explored parametric and semiparametric estimations of copula-based regression functions. We adopted maximum likelihood (ML), inference function for margins (IFM), and pseudo maximum likelihood (PML) techniques for the copula-based regression estimation. Based on extensive numerical experiments, we study the performance of these techniques under specified and misspecified scenarios of copulas and marginals. Under correct specification of copula and marginals, the ML technique performs better over IFM and PML in terms of estimated variance (EV), mean squared error (MSE), relative MSE-efficiency, relative efficiency, and confidence interval. However, in case of misspecified scenario for either

copulas or marginals, PML technique performs better. In case of misspecified copulas and marginals both, we found that the results are highly non-robust. Finally, to study the performance of the considered estimation techniques for the copula-based regression functions, we analyzed two real data sets. The results of real data analysis also support the findings of our empirical study.

Acknowledgements

The authors are thankful to the editor and referees for their constructive and helpful comments which have significantly improved the article. The first author would like to thank DST Inspire Fellowship [IF190682], Government of India, for providing financial support. A. K. Pathak would like to express his gratitude to Science and Engineering Research Board (SERB), India for financial support under the MATRICS research grant MTR/2022/000796. M. Arshad would like to express his gratitude to the Science and Engineering Research Board (SERB), India for financial support under the Core Research Grant CRG/2023/001230.

Author contributions. All author's contributed equally to this paper.

Conflict of interest statement. There is no conflict of interest.

Funding. No funding.

Data availability. Not applicable.

References

[1] E.F. Acar, P. Azimaee and M.E. Hoque, *Predictive assessment of copula models*, Can. J. Stat. **47** (1), 8-26, 2019.
[2] A. Ahdika, D. Rosadi and A.R. Effendie, *Conditional expectation formula of copulas for higher dimensions and its application*, J. Math. Comput. Sci. **11** (4), 4877-4904, 2021.
[3] D. Berg, *Copula goodness-of-fit testing: An overview and power comparison*, Eur. J. Finance **15** (7-8), 675-701, 2009.
[4] T. Bouezmarni, F. Funke and F. Camirand Lemyre, *Regression estimation based on Bernstein density copulas*, Université de Sherbrooke, Submitted, 2014.
[5] B. Choroś, R. Ibragimov and E. Permiakova, *Copula Estimation*, Copula Theory and Its Applications, Springer, 2010.
[6] G.J. Crane and J. van der Hoek, *Conditional expectation formulae for copulas*, Aust. N. Z. J. Stat. **50** (1), 53-67, 2008.
[7] A.R. de Leon and B. Wu, *Copula-based regression models for a bivariate mixed discrete and continuous outcome*, Stat. Med. **30** (2), 175-185, 2011.
[8] J.B. de Long and L.H. Summers, *Equipment investment and economic growth*, Q. J. Econ. **106** (2), 445-502, 1991.
[9] H. Dette, R. Van Hecke and S. Volgushev, *Some comments on copula-based regression*, J. Am. Stat. Assoc. **109** (507), 1319-1324, 2014.
[10] Ö.K. Erdemir and M. Sucu, *A modified pseudo-copula regression model for risk groups with various dependency levels*, J. Stat. Comput. Simul. **92** (5), 1092-1112, 2022.

[11] C. Genest, K. Ghoudi and L.P. Rivest, *A semiparametric estimation procedure of dependence parameters in multivariate families of distributions*, Biometrika **82** (3), 543-552, 1995.
[12] C. Genest, W. Huang and J.M. Dufour, *A regularized goodness-of-fit test for copulas*, J. SFdS **154** (1), 64-77, 2013.
[13] C. Genest, B. Rémillard and D. Beaudoin, *Goodness-of-fit tests for copulas: A review and a power study*, Insur. Math. Econ. **44** (2), 199-213, 2009.
[14] S. Hamori, K. Motegi and Z. Zhang, *Copula-based regression models with data missing at random*, J. Multivariate Anal. **180**, 1-23, 2020.
[15] M.S. Hofert, I. Kojadinovic, M. Maechler and J. Yan, Package "copula: Multivariate Dependence with Copulas", R package version: 1.1-3, 2023.
[16] J.P. Jaworski, F. Durante, W.K. Hardle and T. Rychlik, *Copula Theory and Its Applications*, **198**, Springer, 2010.
[17] H. Joe, *Dependence Modeling with Copulas*, CRC Press, 2014.
[18] D. Kim and J.M. Kim, *Analysis of directional dependence using asymmetric copula-based regression models*, J. Stat. Comput. Simul. **84** (9), 1990-2010, 2014.
[19] G. Kim, M.J. Silvapulle and P. Silvapulle, *Comparison of semiparametric and parametric methods for estimating copulas*, Comput. Stat. Data. Anal. **51** (6), 2836-2850, 2007.
[20] I. Kojadinovic, J. Yan and M. Holmes, *Fast large-sample goodness-of-fit tests for copulas*, Statist. Sinica **21** (2), 841-871, 2011.
[21] N. Kolev and D. Paiva, *Copula-based regression models: A survey*, J. Statist. Plann. Inference **139** (11), 3847-3856, 2009.
[22] Y.K. Leong and E.A. Valdez, *Claims prediction with dependence using copula models*, Insurance: Mathematics and Economics, 2005.
[23] J.A. Nelder and R.W. Wedderburn, *Generalized linear models*, J. Roy. Statist. Soc. Ser. A **135** (3), 370-384, 1972.
[24] R.B. Nelsen, *An Introduction to Copulas*, 2nd ed., Springer, New York, 2007.
[25] H. Noh, A. El Ghouch and T. Bouezmarni, *Copula-based regression estimation and inference*, J. Amer. Statist. Assoc. **108** (502), 676-688, 2013.
[26] R.A. Parsa and S.A. Klugman, *Copula regression*, Variance **5**, 45-54, 2011.
[27] K.W. Penrose, A. Nelson and A. Fisher, *Generalized body composition prediction equation for men using simple measurement techniques*, Med. Sci. Sports Exerc. **17** (2), 189, 1985.
[28] M. Pitt, D. Chan and R. Kohn, *Efficient Bayesian inference for Gaussian copula regression models*, Biometrika **93** (3), 537-554, 2006.
[29] A. Sheikhi, F. Arad and R. Mesiar, *A heteroscedasticity diagnostic of a regression analysis with copula dependent random variables*, Braz. J. Probab. Stat. **36** (2), 408-419, 2022.
[30] M. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris **8** (3), 229-231, 1959.
[31] M.S. Smith and N. Klein, *Bayesian inference for regression copulas*, J. Bus. Econom. Statist. **39** (3), 712-728, 2021.
[32] E.A. Sungur, *Some observations on copula regression functions*, Comm. Statist. Theory Methods **34** (9-10), 1967-1978, 2005.
[33] H. Tsukahara, *Semiparametric estimation in copula models*, Canad. J. Statist. **33** (3), 357-375, 2005.
[34] P. Vellaisamy and A.K. Pathak, *Copulas and regression models*, J. Indian Statist. Assoc. **52** (1), 113-134, 2014.