"Copula-Based Regression Estimation and Inference"

Noh, Hohsuk ; El Ghouch, Anouar ; Bouezmarni, Taoufik

ABSTRACT

We investigate a new approach to estimating a regression function based on copulas. The main idea behind this approach is to write the regression function in terms of a copula and marginal distributions. Once the copula and the marginal distributions are estimated, we use the plug-in method to construct our new estimator. Because various methods are available in the literature for estimating both a copula and a distribution, this idea provides a rich and flexible family of regression estimators. We provide some asymptotic results related to this copula-based regression modeling when the copula is estimated via profile likelihood and the marginals are estimated nonparametrically. We also study the finite sample performance of the estimator and illustrate its usefulness by analyzing data from air pollution studies.

CITE THIS VERSION

Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Copula-Based Regression Estimation and Inference

Hohsuk Noh a, Anouar El Ghouch a & Taoufik Bouezmarni b

a Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain la Neuve, 1348, Belgium

b Département de Mathématiques, Université de Sherbrooke, Sherbrooke, Quebec, J1K 2R1, Canada

Accepted author version posted online: 29 Mar 2013. Published online: 01 Jul 2013.


To link to this article: http://dx.doi.org/10.1080/01621459.2013.783842

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Copula-Based Regression Estimation and Inference

Hohsuk NOH, Anouar El GHOUCH, and Taoufik BOUEZMARNI

We investigate a new approach to estimating a regression function based on copulas. The main idea behind this approach is to write the regression function in terms of a copula and marginal distributions. Once the copula and the marginal distributions are estimated, we use the plug-in method to construct our new estimator. Because various methods are available in the literature for estimating both a copula and a distribution, this idea provides a rich and flexible family of regression estimators. We provide some asymptotic results related to this copula-based regression modeling when the copula is estimated via profile likelihood and the marginals are estimated nonparametrically. We also study the finite sample performance of the estimator and illustrate its usefulness by analyzing data from air pollution studies.

KEY WORDS: Dependence modeling; Profile likelihood; Semiparametric regression; Vine copula.

1. INTRODUCTION

Let \( X = (X_1, \ldots, X_d)^\top \) be a random vector of dimension \( d \geq 1 \) and \( Y \) be a random variable with continuous cumulative distribution functions (cdf’s) \( F_1, \ldots, F_d \) and \( F_0 \), respectively. \( Y \) is our response variable and \( X \) is our set of covariates. We denote the density of \( X_j \) by \( f_j \) and that of \( Y \) by \( f_0 \). For a given \( x = (x_1, \ldots, x_d)^\top \), we will use \( F(x) \) as a shortcut for \((F_1(x_1), \ldots, F_d(x_d))\). From the inspiring work of Sklar (1959), the cdf of \((Y, X)\) evaluated at \((y, x)\) can be expressed as \( C(F_0(y), F(x)) \), where \( C \) is the copula distribution of \((Y, X)\), which is the function from \([0, 1]^d \) to \([0, 1]\) defined by \( C(u_0, u_1, \ldots, u_d) = P(U_0 \leq u_0, U_1 \leq u_1, \ldots, U_d \leq u_d) \), where \( U_0 = F_0(Y) \) and \( U_j = F_j(X_j), \ j = 1, \ldots, d \). This is nothing but a joint distribution function with marginals that are uniform over the unit interval \([0, 1]\). Copulas enable us to model the dependence between variables separately from their marginal distributions and by specifying a copula we summarize all the dependencies between margins. See Nelsen (2006) for more about this subject. From the definition of a copula function, the conditional density of \( Y \) given \( X = x \) is given by

\[
f_0(y)\frac{c(F_0(y), F(x))}{c_x(F(x))},
\]

where \( c(u_0, u) \equiv c(u_0, u_1, \ldots, u_d) = \partial^{d+1}C(u_0, u_1, \ldots, u_d)/\partial u_0 \partial u_1 \ldots \partial u_d \) is the copula density corresponding to \( C \) and \( c_x(u) \equiv c_x(u_1, \ldots, u_d) = \partial^d C(1, u_1, \ldots, u_d)/\partial u_1 \ldots \partial u_d \) is the copula density of \( X \). Obviously, the conditional mean, \( m(x) \), of \( Y \) given \( X = x \) can be written as

\[
m(x) = \mathbb{E}(Yw(F_0(Y), F(x))) = \frac{e(F(x))}{c_x(F(x))},
\]

where \( w(u_0, u) = c(u_0, u)/c_x(u) \) and

\[
e(u) = \mathbb{E}(Yc(F_0(Y), u)) = \int_0^1 F_0^{-1}(u_0)c(u_0, u)du_0.
\]

The equality (1) shows that, given the marginals, one can obtain the mean regression function relating \( Y \) to \( X \) directly from the copula density, or equivalently the copula distribution of \((Y, X^\top)^\top\). It also implies that the conditional mean is just a weighted mean with weights induced by the unknown “conditional” copula function \( w \) defined above. This relation is not new and has already been applied in Sun- gur (2005), Leong and Valdez (2005), and Crane and Van Der Hoek (2008) to compute the mean regression function corresponding to several well-known copula families (Gaussian, t, Farlie–Gumbel–Morgenstern (FGM), Iterated FGM, Archimedean, etc.) with single \((d = 1)\) and multiple covariate(s). To illustrate the idea, we briefly cite two examples:

- If the copula density of \((Y, X_1)\) belongs to the FGM family with a parameter \( \theta \), that is, \( c(u_0, u_1) = 1 + \theta(1 - 2u_0)(1 - 2u_1) \), then we have

\[
m(x_1) = \mathbb{E}(Y) + \theta(2F_1(x_1) - 1) \int F_0(y)(1 - F_0(y))dy.
\]

A similar formula holds for the multiple covariate case; see Leong and Valdez (2005).

- If the copula of \((Y, X^\top)^\top\) is Gaussian with correlation matrix

\[
\Sigma_{Y,X} = \begin{bmatrix} 1 & \rho^\top \\ \rho & \Sigma_X \end{bmatrix},
\]

then we have

\[
m(x) = \mathbb{E}[F_0^{-1}(\Phi(u^\top \Sigma_X^{-1} \rho + \sqrt{1 - \rho^\top \Sigma_X^{-1} \rho} Z))],
\]

where \( u = (\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d)))^\top, \ Z \sim \mathcal{N}(0, 1) \), and \( \Phi \) is the cdf of a standard normal distribution.

Hohsuk Noh is postdoctoral researcher, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain la Neuve 1348, Belgium (E-mail: word5810@gmail.com). Anouar El Ghouch is Professor, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain la Neuve 1348, Belgium (E-mail: anouar.elghouch@uclouvain.be). Taoufik Bouezmarni is Assistant Professor, Département de Mathématiques, Université de Sherbrooke, Sherbrooke, Quebec J1K 2R1, Canada (E-mail: taoufik.bouezmarni@usherbrooke.ca). H. Noh’s research is supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy) and the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 203650. A. El Ghouch’s research is supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and the contract “Projet d’Actions de Recherche Concertées” (ARC) 11/6-039 of the “Communauté Française de Belgique,” granted by the “Académie Universitaire Louvain.” The authors thank two anonymous referees, the associate editor, and the coeditor for their valuable suggestions that have significantly improved the article.
Note that in the single covariate case, we have $c_X(u) \equiv c_X(u_1) = 1$ for all $u_1 \in [0, 1]$. In such a case, the weight function $w$ coincides with the copula density $c$ and (1) reduces to $m(x) = e(F(x)) = E(Y | c(F_0(Y), F(x)))$. Also, if the covariates are mutually independent, then $c_X(u) = 1$ and $m(x)$ coincides with $e(F(x))$. In other words, $e(F(x))$, the numerator of $m(x)$ in (1), is the mean regression function of $Y$ given $X$ assuming independence between the covariates or, equivalently, assuming that the conditional density of $Y | X$ is $f_0(y)(F_0(y), F(x))$. Thus, in term of copulas, the mean regression function is a ratio of a numerator that only captures the mean dependence between $Y$ and $X$ and a denominator that captures the dependence within $X$.

The equality (1) can also be used as an estimating equation. In fact, if $\hat{\nu}$, $\hat{F}_0$, and $\hat{F}_j$ are any given estimators for $w$, $F_0$, and $F_j$, respectively, then $m$ can obviously be estimated by

$$\hat{m}(x) = \int_{-\infty}^{\infty} y \hat{\nu}(\hat{F}_0(y), \hat{F}(x))d\hat{F}_0(y),$$

(5)

where $\hat{F} = (\hat{F}_1(x_1), \ldots, \hat{F}_d(x_d))^T$. To the best of our knowledge, such an approach has never been proposed or investigated in the literature, neither in the single nor in the multiple covariate case. To estimate $w$, one needs an estimator for the copula densities $c$ and $c_X$. The copula density $c_X$ can be obtained from $c$ by integration. In fact,

$$c_X(u) = \int_{-\infty}^{\infty} f_0(y)c(F_0(y), u)dy = \int_0^1 c(u_0, u)du_0.$$  

(6)

Therefore, given an estimator $\hat{c}$ for $c$, one can easily estimate $c_X$ using the plug-in method and then estimate $m$ by (5).

Since, in the literature, there are many different methods available for estimating a copula and a cdf, $\hat{m}(x)$ defines a large new class of interesting estimators. Depending on the method of estimating the components in (5), $\hat{m}(x)$ can be a nonparametric or a semiparametric or a fully parametric estimator. For example, using nonparametric estimators for $c$, $F_0$, and $F_j$, $j = 1, \ldots, d$, leads to a fully nonparametric estimator. Nonparametric methods for estimating $c$ include kernel-smoothing estimators (Gijbels and Mielenzuk 1990; Charpentier, Fermanian, and Scaillet 2006; Chen and Huang 2007; Omelka, Gijbels, and Veraverbeke 2009), Bernstein estimator (Bouezmarni, Rombouts, and Taamouti 2010), and spline estimator (Kauermann et al. in press) to name a few. In spite of the great flexibility, nonparametric methods are typically affected by the curse of dimensionality and come with the difficult problem of selecting a good smoothing parameter. On the other hand, imposing a parametric structure on both the copula and marginal distributions can lead to severely biased and inconsistent (fully parametric) estimators in case of misspecification. For these reasons and to avoid such problems as much as possible, we consider here a semiparametric approach where the copula is modeled parametrically but the marginal distributions are modeled nonparametrically. As will be shown in the next sections, the proposed method has many interesting properties both from a theoretical and a practical point of view. In particular, the asymptotic properties are easy to obtain, the numerical calculations can be done directly using existing software packages and, unlike many semiparametric methods, no iteration procedure is needed to guarantee consistency. Also, the asymptotic variance can be estimated without any extra complications.

The plan of the article is as follows. Section 2 presents the general theoretical framework of the method with the necessary notations and assumptions. In Section 3, we identify the asymptotic representation of the proposed estimator in the single and multiple covariate cases. From this representation, we establish the asymptotic distribution of the estimator. Further, we propose and study an estimator of its asymptotic variance. In Section 4, we investigate theoretical properties of the estimator under misspecification. In Section 5, we present some numerical simulations designed to confirm the theoretical results and to compare the performance of our estimators with that of some competitors. Finally, in Section 6, we analyze data from air pollution studies to illustrate the usefulness of the proposed estimator. All the proofs appear in the Appendix.

2. THEORETICAL BACKGROUND

Let $X_i = (X_{1,i}, \ldots, X_{d,i})^T$ and let $(Y_i, X_i)$, $i = 1, \ldots, n$, be an independent and identically distributed (iid) sample of $n$ observations generated from the distribution of $(Y, X)$. Clearly, the shape and the performance of our estimator $\hat{m}$ in (5) will depend heavily on the methods of estimation for $c$, $F_0$, and $F_j$. In this work, $F_0$ is estimated by a rescaled empirical distribution function:

$$\hat{F}_{0,n}(y) = \frac{1}{n+1} \sum_{i=1}^{n} I(Y_i \leq y).$$

Estimating the other cdf’s $F_j$, $j = 1, \ldots, d$, can also be done in the same way via $\hat{F}_{1,n}$. However, this results in a nonsmooth estimate $\hat{m}(x)$ as is illustrated in Figure 1, where we show the resulting estimator using $\hat{F}_{1,n}$ in the univariate case. To get a more visually attractive regression curve or surface, one should smooth the empirical cdf’s $\hat{F}_{1,n}, \ldots, \hat{F}_{d,n}$. The simple way to do that is to use a kernel-smoothing method. Let $k(\cdot)$ be a function, which is a symmetric probability density function and $h \equiv h_n \to 0$ be a bandwidth parameter. Then, a kernel-smoothing estimator of $F_j$ is given by

$$\hat{F}_{j,n}(x) = \frac{1}{n} \sum_{i=1}^{n} K \left(\frac{x - X_{j,i}}{h}\right),$$

where $K(x) = \int_{-\infty}^{x} k(t)dt$. The methods of estimating the marginals $F_1, \ldots, F_d$ do not bring out any difference in asymptotic behavior of the estimator $\hat{m}(x)$ as long as they satisfy the following assumption:

**Assumption A.** For a given point of interest $x \in \mathbb{R}^d$,

$$\hat{F}_{j,n}(x_j) = n^{-1} \sum_{i=1}^{n} I(X_{j,i} \leq x_j) + o_p(n^{-1/2}), \quad j = 1, \ldots, d.$$

Note that the kernel-smoothing estimator $\hat{F}_{j,n}$ satisfies Assumption A if $nh^d \to 0$. Evidently, this assumption also holds true for the rescaled empirical cdf $\hat{F}_{j,n}$.

We examined the effect of the methods of estimating the marginal distributions on the regression fit in an extensive simulation study and came to the conclusion that both estimators ($\hat{F}_{j,n}$ and $\hat{F}_{j,n}$) show more or less similar performance. Here,
we present one example to illustrate this. We generate \((Y, X_1, \ldots, X_l)\) using a Gaussian copula according to DGP S.a described in Section 5. To assess the effect of each method, we consider the distribution of \(X\) and \(\hat{\theta}\) via the cross-validation (CV) method. Compared with the kernel-smoothing estimate with the mean square optimal bandwidth, that is, the kernel-smoothing estimate with the mean square optimal bandwidth gives better results if the optimal bandwidth is used. However, when the bandwidth is chosen by the data, its performance is similar to that of the nonsmooth estimator \(\hat{F}_{1,n}\).

Figure 1, which shows the boxplots of the empirical integrated squared errors of \(\hat{m}\) with \(\hat{F}_{1,n}, \hat{F}_{1,n}^{\text{opt}}\), and \(\hat{F}_{1,n}^{\text{CV}}\), respectively (from left to right).

Table 1 shows the IMSE together with the IBIAS and the IV AR squared errors (ISEs), also supports this observation. Based on that, in all our numerical simulations (see Section 5), we use the rescaled empirical cdf for its simplicity.

As mentioned in the introduction, in this work we adopt a semiparametric approach. For the theoretical analysis, we assume that the true copula density \(c\) belongs to a certain parametric family, which is known. However, in our simulations we consider situations where such a copula family is unknown and should be chosen adaptively using the data. We denote the copula family to which \(c\) belongs by \(C = \{ c(\cdot; \theta) \mid \theta \in \Theta \}\), where \(\Theta\) is a compact subset of \(\mathbb{R}^p\). Define \(\hat{\theta}_n\) to be the true (but unknown) copula parameter, which lies in the interior of \(\Theta\) so that \(c(\cdot; \hat{\theta}_n)\) coincides with \(c(\cdot; \theta_0)\). We will restrict our interest to an estimator \(\hat{\theta}_n\) of \(\theta_0\), which satisfies the following assumption.

Assumption B.

\[
\hat{\theta}_n - \theta_0 = n^{-1} \sum_{i=1}^{n} \eta_i + o_p(n^{-1/2}),
\]

where \(\eta_i = \eta(U_{0,i}, U_i; \theta_0)\) is a \(p\)-dimensional random vector such that \(\mathbb{E} \eta = 0\) and \(\mathbb{E} \eta^\top \eta < \infty\) and \(U_i = (U_{1,i}, \ldots, U_{d,i})^\top\).

Many existing estimators for \(\theta_0\) in the literature satisfy this assumption. One promising estimator among them is the (semiparametric) maximum pseudo-likelihood (PL) estimator \(\hat{\theta}_n^{\text{PL}}\), which is defined as the maximizer of

\[
L(\theta) = \sum_{i=1}^{n} \log c(\hat{F}_0(n,Y_i), \hat{F}_n(X_i); \theta),
\]

where \(\hat{F}_n(X_i) = (\hat{F}_{1,n}(X_{1,i}), \ldots, \hat{F}_{d,n}(X_{d,i}))^\top\). \(\hat{\theta}_n^{\text{PL}}\) was studied by Genest, Ghoudi, and Rivest (1995), Silvapulle, Kim, and Silvapulle (2004), Tsukahara (2005), and Kojadinovic and Yan (2011). Kojadinovic and Yan (2011) compared \(\hat{\theta}_n^{\text{PL}}\) with two method-of-moment estimators and found that the PL estimator...
performs best overall in terms of mean squared error. A similar conclusion was drawn in Silvapulle, Kim, and Silvapulle (2004) regarding the comparison of $\hat{\theta}_{n}^{PL}$ with two other estimators based on the maximum likelihood. Assumption B holds for $\hat{\theta}_{n}^{PL}$ whenever the score function of the density $c$ satisfies the conditions (A.1)–(A.5) given in Tsukahara (2005). For the PL estimator, the function $\eta$ is given by

$$J(U_0, U; \theta) = J^{-1}(\theta) \times K(U_0, U; \theta),$$

where

$$J(\theta) = \int_{[0,1]^{d+1}} \left( \frac{\partial^2}{\partial \theta \partial \theta} \log c(u_0, u; \theta_0) \right) dC(u_0, u; \theta)$$

and $K(U_0, U; \theta)$ is a $p$-dimensional vector whose $k$th element is

$$\frac{\partial}{\partial \theta_k} \log c(U_0, U; \theta) + \sum_{j=0}^{d} \int_{[0,1]^{d+1}} (I(U_j \leq v_j) - v_j) \times \frac{\partial^2}{\partial \theta_k \partial v_j} \log c(v_0, v; \theta) dC(v_0, v; \theta).$$

Remark 1. Tsukahara (2005) stated the conditions under which the asymptotic normality of $\hat{\theta}_n$ holds and does not give explicitly the iid representation in (8). However, to prove the asymptotic normality, Tsukahara (2005) made use of the general theory of Z-estimators, see, for example, sec. A.10 of Bickel (1998). PL explicitly the iid representation in (8). However, to prove the asymptotic normality, Tsukahara (2005) made use of the general theory of Z-estimators, see, for example, sec. A.10 of Bickel et al. (1993) and chap. 3.3 of van der Vaart and Wellner (1996).

Now we are ready to present the main results. We will show the asymptotic iid representation of the proposed estimator, which implies that the estimator follows a normal distribution asymptotically.

### 3. MAIN RESULTS

#### 3.1 Single Covariate Case ($d = 1$)

First, consider the simpler case where there is only one covariate $X_1$. In this case,

$$m(x_1) = e(F_1(x_1); \theta_0) = \mathbb{E}[Y c(F_0(Y), F_1(x_1); \theta_0)],$$

can be estimated by

$$\hat{m}(x_1) = \hat{e}(\hat{F}_{1,n}(x_1); \hat{\theta}_n).$$

The following theorem gives an asymptotic iid representation of this estimator. Its proof is given in the Appendix.

**Theorem 1.** Assume that $\hat{F}_{1,n}()$, $\hat{\theta}_n$, and $c()$ satisfy Assumptions A, B, and C, respectively. If $\mathbb{E}Y^2 < \infty$, then we have

$$\sqrt{n}(\hat{m}(x_1) - m(x_1)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ I(X_{1,i} \leq x_1) - F_1(x_1) \right] \hat{\eta} e(F_1(x_1); \theta_0)$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(Y_i \leq x_1) - F_0(x_1) c(F_0(x_1), F_1(x_1); \theta_0) dy + \frac{1}{\sqrt{n}} \eta^T \hat{e}(F_1(x_1); \theta_0) + o_p(1).$$

Theorem 1 implies that $\sqrt{n}(\hat{m}(x_1) - m(x_1))$ follows asymptotically a normal distribution with mean 0 and variance $\sigma^2 = \text{var}(E_i(\theta_0))$, where $E_i(\theta_0)$ is the term within the bracket in Theorem 1. By the plug-in principle, a natural estimator of $\sigma^2$ is given by

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (\hat{F}_{1,i}(x_1) - F_1(x_1))^2,$$

where

$$\hat{F}_{1,i}(x_1) = \frac{1}{n} \sum_{i=1}^{n} I(X_{1,i} \leq x_1).$$

The sketch of the proof for the consistency of $\hat{\sigma}^2$ with the necessary assumptions is given in the Appendix. Additionally, we check the consistency empirically in Section 5 by investigating the validity of the confidence interval for $\hat{m}(x_1)$ based on $\hat{\sigma}^2$.

#### 3.2 Multiple Covariate Case ($d \geq 2$)

In the general case ($d \geq 2$), the regression function is given by

$$m(x) = \frac{e(F(x); \theta_0)}{c(F(x); \theta_0)}.$$

Estimating the numerator of $m(x)$ can be done, as in the single covariate case, by

$$\hat{\hat{e}}(\hat{F}_{n}(x); \hat{\theta}_n) = n^{-1} \sum_{i=1}^{n} Y_i c(\hat{F}_{0,n}(Y_i), \hat{F}_{n}(x); \hat{\theta}_n),$$

where

$$\hat{F}_{n}(x) = (\hat{F}_{1,n}(x_1), \ldots, \hat{F}_{d,n}(x_d)).$$

Following the proof of Theorem 1, one can easily check that under Assumptions A, B, and C, we have

$$\hat{\hat{e}}(\hat{F}_{n}(x); \hat{\theta}_n) - e(F(x); \theta_0) = n^{-1} \sum_{i=1}^{n} E_i(x; \theta_0) + o_p(n^{-1/2}).$$

(11)
Table 2. IBIAS, IVAR, and IMSE (×1000) of \( \hat{m} \) and \( \tilde{m} \)

<table>
<thead>
<tr>
<th>IBIAS</th>
<th>IVAR</th>
<th>IMSE</th>
<th>IBIAS</th>
<th>IVAR</th>
<th>IMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{m} )</td>
<td>0.172</td>
<td>1.567</td>
<td>1.737</td>
<td>1.178</td>
<td>1.963</td>
</tr>
<tr>
<td>( \tilde{m} )</td>
<td>6.071</td>
<td>15.807</td>
<td>17.377</td>
<td>11.178</td>
<td>19.639</td>
</tr>
</tbody>
</table>

where

\[
E_i(x; \theta_0) = \sum_{j=1}^{d} \left( I(X_{j,i} \leq x_j) - F_j(x_j) \right) \partial_j c(F(x); \theta_0) \\
- \int (I(Y_i \leq y) - F_0(y)) c(F_0(y), F(x); \theta_0) dy \\
+ \eta^\top \hat{c}(F(x); \theta_0).
\]

From the Equation (6), we consider two estimators for the denominator of (10), which are

\[
\hat{c}_X(\tilde{F}_n(x); \tilde{\theta}_n) = \int_0^1 c(u_0, \tilde{F}_n(x); \tilde{\theta}_n) du_0 \\
\hat{c}_X(\tilde{F}_n(x); \tilde{\theta}_n) = n^{-1} \sum_{i=1}^{n} c(\tilde{F}_{i,n}(Y_i), \tilde{F}_n(x); \tilde{\theta}_n).
\]

These two estimators lead to two different estimators for \( m(x) \), \( \hat{m}(x) \), and \( \tilde{m}(x) \), respectively. However, the difference between \( \hat{m}(x) \) and \( \tilde{m}(x) \) is negligible asymptotically (see the remark below). In our simulation studies, we have found that in many cases the estimator \( \hat{m}(x) \) is preferable to \( \tilde{m}(x) \) in terms of finite sample performance. To illustrate this remark, we briefly present one small Monte Carlo study designed to compare these two estimators. Using \( N = 1000 \) random samples generated from DGP M.a \( (d = 2) \) described in Section 5, we compute the empirical IMSE, IBIAS, and IVAR for \( \hat{m} \) and \( \tilde{m} \). From Table 2, we see that the latter clearly performs better than the former, both in terms of bias and variance. This observation is also confirmed by Figure 2, which shows the boxplots of the empirical ISEs of the two estimators. From this observation, we only consider \( \hat{m}(x) \) hereafter.

To state the main result about \( \hat{m}(x) \), first note that the asymptotic representation of \( \hat{c}_X(\tilde{F}_n(x); \tilde{\theta}_n) \) follows by using similar arguments as in the proof of Theorem 1. In fact, under Assumptions A, B, and C, one can easily check that we have

\[
\hat{c}_X(\tilde{F}_n(x); \tilde{\theta}_n) - c_X(F(x); \theta_0) = n^{-1} \sum_{i=1}^{n} D_i(x; \theta_0) + o_p(n^{-1/2}),
\]

where

\[
D_i(x; \theta_0) = \sum_{j=1}^{d} \left( I(X_{j,i} \leq x_j) - F_j(x_j) \right) \partial_j c(F(x); \theta_0) + \eta^\top \hat{c}(F(x); \theta_0).
\]

Remark 2. This representation implies that up to \( o_p(n^{-1/2}) \), \( \hat{c}_X(\tilde{F}_n(x); \tilde{\theta}_n) \) is asymptotically equivalent to \( \int_0^1 c(u_0, \tilde{F}_n(x); \tilde{\theta}_n) du_0 \). This also means that the difference between \( \hat{m}(x) \) and \( \tilde{m}(x) \) is negligible up to the first-order approximation.

Finally, combining (11) with (12) leads to our main result.

**Theorem 2.** Assume that every component of \( F_q \) satisfies Assumption A, \( \tilde{\theta}_n \) and \( c(\cdot) \) satisfy Assumptions B and C, respectively. If \( \mathbb{E}[Y^2] < \infty \), then we have

\[
\sqrt{n}(\hat{m}(x) - m(x)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{c_X(F(x); \theta_0)} \\
\times [E_i(x; \theta_0) - m(x)D_i(x; \theta_0)] + o_p(1).
\]

As in the univariate case, the asymptotic normality of \( \sqrt{n}(\hat{m}(x) - m(x)) \) can be deduced from Theorem 2. The asymptotic variance of \( \sqrt{n}(\hat{m}(x) - m(x)) \) is given by

\[
\text{var} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{c_X(F(x); \theta_0)} [E_i(x; \theta_0) - m(x)D_i(x; \theta_0)] \right),
\]

can be estimated using the plug-in method. As a consequence, one can easily construct pointwise confidence intervals for \( m \). The validity of this approach is investigated in Section 5.

**4. CONSIDERATION OF MISSPECIFICATION**

If the true copula family is known, the proposed estimator performs well as will be shown in the simulation study. However, in practice such information is not available and the copula shape needs to be selected using the data. In any selection procedure,
there is the possibility that the wrong copula family will be selected and using a misspecified copula model for our method will lead to an inconsistent estimator of \(m(x)\). In this section, we are interested in the following question: what is the effect (cost) of using a misspecified copula model on the resulting regression estimator? To answer this question, assume that \(C = \{c(\cdot; \theta), \theta \in \Theta\}\) is a parametric family of copula densities under consideration. In the previous sections, we assumed that the copula family is well-specified, that is, there exists the true parameter \(\theta_0\) such that \(c(\cdot; \theta_0)\) coincides with the true copula density, \(c(\cdot)\), of \((Y, X)\). Under a misspecified model, such \(\theta_0\) may not exist. However, even in such a situation, we can define the pseudo-true parameter \(\theta^*\) to be the unique minimum within the set \(\Theta\) of

\[
I(\theta) = \int_{[0,1]^{d+1}} \ln \left( \frac{c(u_0, u)}{c(u_0, u; \theta)} \right) dC(u_0, u).
\]

Here, \(I(\theta)\) is the classical Kullback–Leibler information criterion expressed in terms of copula densities instead of the traditional densities. Following the lines of the proof for Theorem 2 with \(\theta^*\) instead of \(\theta_0\), we are able to describe the asymptotic behavior of \(\hat{m}(x)\) under misspecification in the following theorem.

**Theorem 3.** Assume that every component of \(\bar{F}_n\) satisfies Assumption A and \(\hat{\theta}_n\) and \(c(\cdot)\) satisfy Assumptions B and C, respectively, with \(\theta^*\) instead of \(\theta_0\). If \(E[Y^2] < \infty\), then we have

\[
\sqrt{n} (\hat{m}(x) - m(x; \theta^*)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{1}{x \left(F_X(x); \theta^*\right)} \left[ E(X_i | F_X(x); \theta^*) - m(x; \theta^*) \right] \right\} + o_p(1),
\]

where \(m(x, \theta^*)\) is the mean regression function under the assumption that the joint distribution of \((Y, X)\) is \(C(F_Y(y), F_X(x); \theta^*)\).

**Remark 3.** Let

\[
\hat{\theta}_n = \arg \min_\theta \sum_{i=1}^n \log c \left( \tilde{F}_{0,n}(Y_i), \tilde{F}_{n}(X_i); \theta \right)
\]

be the maximum PL estimator. By the classical maximum likelihood theory under misspecification, see White (1982), and following the proof of Theorem 1 in Tsukahara (2005), we have verified that \(\hat{\theta}_n\) satisfies Assumption B with \(\eta\) as given by (9) but with \(\theta^*\) instead of \(\theta_0\).

Clearly, this theorem reduces to Theorem 2 when the copula family is well specified. Additionally, the result implies that under the misspecification, the estimator \(\hat{m}(x)\) is still asymptotically normal but is biased. More specifically, a misspecified copula brings out a bias in the estimation of \(m(x)\), which is (asymptotically) nothing but the difference between \(m(x)\), the true regression function, and \(m(x; \theta^*)\) its best approximation (in terms of likelihood) among the regression function family \(m(x; \theta) = E[Y | c(F_Y(y), F_X(x); \theta)] \) with \(\theta \in \Theta\). As a result, it is important to choose a rich and flexible copula family to make the difference \(m(x) - m(x; \theta^*)\) small. More discussion about this will be given in the next section.

### 5. SIMULATIONS

The objective of this section is first to check whether the asymptotic theory for \(\hat{m}(x)\) works both when the copula is well specified (Theorem 2) and when the copula is misspecified (Theorem 3). The second objective is to compare our semiparametric estimator with some competitors both when the true copula family is known and when the copula family and its parameters are adaptively selected using the data. To this end, we consider the following data generating procedures (DGPs):

- **DGP S.a (F₀(Y), F₁(X₁)) ~ Gaussian copula with parameter \(\rho_1; Y \sim N(\mu_Y, \sigma_Y^2)\).**
  - (i) The resulting regression function is \(m(x_1) = \mu_Y + \sigma_Y \rho_1 \Phi^{-1}(F_1(x_1))\), where \(\Phi\) is the cdf of a standard normal distribution.
  - (ii) \(X_1\) is generated from \(N(\mu_{x_1}, \sigma_{x_1}^2)\).
- **DGP S.b (F₀(Y), F₁(X₁)) ~ FGM copula with parameter \(\theta; Y \sim N(\mu_Y, \sigma_Y^2)\).**
  - (i) The resulting regression function is \(m(x_1) = \mu_Y - \frac{\rho}{\sigma_x} \sigma_Y + 2 \frac{\rho}{\sigma_x} \sigma_Y F_1(x_1)\).
  - (ii) \(X_1\) is generated from the cdf \(F_{X_1}(x_1) = 1 - \exp(-\exp(x_1))\).
- **DGP S.c (F₀(Y), F₁(X₁)) ~ Student’s \(t\) copula with parameters \(\rho\) and \(df; Y \sim N(\mu_Y, \sigma_Y^2)\).**
  - (i) The resulting regression function is
    \[
    m(x_1) = \mu_Y + \mathbb{E}[(\sigma_Y \Phi^{-1}(\Phi_{df}(\rho \theta)) + T \sqrt{df} (1 - \rho^2)(1 + a^2/df)/(df + 1))],
    \]
    where \(a = \Phi_{df}^{-1}(F_{X_1}(x_1))\). \(\Phi_{df}\) is the cdf of a univariate Student’s \(t\) random variable with degrees of freedom \(df\) and \(T\) is a univariate Student’s \(t\) random variable with degrees of freedom \(df + 1\).
  - (ii) \(X_1\) is generated from \(N(\mu_{x_1}, \sigma_{x_1}^2)\).
- **DGP S.d (F₀(Y), F₁(X₁)) ~ Clayton copula with parameters \(\delta; Y \sim N(\mu_Y, \sigma_Y^2)\).**
  - (i) The resulting regression function is
    \[
    m(x_1) = \mu_Y + \mathbb{E} \{ \sigma_Y \Phi^{-1}(T^{-1/\delta}) \},
    \]
    where \(T \sim f_T(t) = (1/\delta + 1)(1 + \xi)^{1/\delta+1}/(t + \xi)^{2/\delta+1}\) for \(t > 1\) and \(\xi = F_{X_1}(x_1)^{-\delta} - 1\).
  - (ii) \(X_1\) is generated from \(N(\mu_{x_1}, \sigma_{x_1}^2)\).
- **DGP M.a (F₀(Y), F₁(X₁), \ldots, F_d(X_d)) ~ Gaussian copula with correlation matrix \(\Sigma = [\rho_{ij} \Sigma_{X_i}^{-1}]\), where \(\rho\) is a \(d\)-dimensional vector; \(Y \sim U(0, 1)\).**
  - (i) The resulting regression function is
    \[
    m(x) = \Phi \left( \sum_{j=1}^d a_j \frac{\phi^{-1}(F_j(x_j))}{\sqrt{2 - \rho^2 a}} \right),
    \]
    where \(a = (a_1, \ldots, a_d)^T \equiv \Sigma_{X_i}^{-1} \rho\).
  - (ii) \(X_j\) is generated from \(N(\mu_{x_j}, \sigma_{x_j}^2)\), \(j = 1, \ldots, d\).
that we use for each DGP are given in Table 3. All computations that satisfies mean of the distribution of $X$ is close to the nominal level (1 − $\alpha$) of the regression function $m(x)$ calculated according to Theorem 2. ECP means the proportion of confidence intervals that contain the true value of the regression function $m(x)$. By seeing whether the ECPs are close to the nominal level (1 − $\alpha$), we can check that the proposed estimator $\hat{m}(x)$ is asymptotically normal. Also, we can validate the asymptotic iid parameter representation of $\hat{m}(x)$ and the appropriateness of the proposed variance estimator $\hat{m}(x)$. We calculate the estimator $\hat{m}(x)$ from $N = 1000$ random samples of size $n = 50, 100, 200, 400$. This is done for different values of significance level $\alpha$ and points of interest $x$. Only selected and representative results are shown to save space.

5.1 Verifying the Asymptotic Results

To verify the established asymptotic results, we draw Quantile-Quantile (Q-Q) plots of $\hat{m}(x)$ and calculate empirical coverage probabilities (ECPs) of the (1 − $\alpha$)-confidence intervals of $m(x)$ calculated according to Theorem 2. ECP means the proportion of confidence intervals that contain the true value of the regression function $m(x)$. By seeing whether the ECPs are close to the nominal level (1 − $\alpha$), we can check that the proposed estimator $\hat{m}(x)$ is asymptotically normal. Also, we can validate the asymptotic iid representation of $\hat{m}(x)$ and the appropriateness of the proposed variance estimator $\hat{m}(x)$. We calculate the estimator $\hat{m}(x)$ from $N = 1000$ random samples of size $n = 50, 100, 200, 400$. This is done for different values of significance level $\alpha$ and points of interest $x$. Only selected and representative results are shown to save space.

5.1.1 When the Copula is Well Specified. Table 4 presents the ECP of the 95% confidence intervals for $m(x)$ with data generated from DGP S.a, DGP S.b, and DGP M.a (d = 2). We calculate the ECP, when $x$ is an interior point ($x = \mu_X$, the mean of the distribution of $X$) and a boundary point (a point that satisfies $P(\|X - \mu_X\| > \|x - \mu_X\|) = 0.1$, where $\|\cdot\|$ is the Euclidean norm). As seen in Table 4, even for a small sample size, the ECPs are generally quite close to their nominal confidence level. This demonstrates the validity of our iid representation and the consistency of our asymptotic variance estimator.

Figure 3, which shows the Q-Q plots for our estimator with DGP S.b when $n = 50$, also confirms this finding. These plots clearly indicate the accuracy of the normal approximation to the asymptotic distribution of $\hat{m}(x)$. To estimate the copula parameter, we make use of the R package copula (see Yan 2007).

5.1.2 When the Copula is Misspecified. To verify the asymptotic behavior of our estimator under misspecification, we generate data from Student’s $t$ copula and Clayton copula according to DGP S.c and DGP S.d, respectively, but in our estimation procedure we use a Gaussian copula. To see how the degree of misspecification influences our estimator, we vary the degrees of freedom ($\nu$) of the Student’s $t$ copula in [3, 5, 100]. Since the difference between the best approximation $m(x_1, \rho^*)$ and the true value $m(x_1)$ depends on the distance between the given copula family $C = \{c(\cdot; \theta), \theta \in \Theta\}$ and the true copula $C(\cdot)$, we expect to see the estimator $\hat{m}(x_1)$ concentrating more around $m(x_1)$ as the $\nu$ of the true $t$-copula increases based on our established theory. The “pseudo”-true regression function is $m(x_1, \rho^*) = 1 + \rho^* x_1$ with $\rho^* = 0.583, 0.590, 0.596$ for the $t$-couples ($\nu = 3, 5, 100$, respectively, and $\rho^* = 0.503$ for the Clayton copula.

Figure 4 shows the boxplots of the estimators $\hat{m}(x_1)$ obtained from 1000 random samples of size 200. As stated in Theorem 3, we see that the observed values are symmetrically distributed around the pseudo-true parameter $m(x_1, \rho^*)$ instead of true parameter $m(x_1)$. The difference between these two quantities corresponds exactly to the asymptotic bias. This bias decreases as the “distance” between the true and the used copula decreases.

In case of the $t$-copula, our estimator becomes consistent as $\nu$ increases. Clearly, the selection of an appropriate copula family is important for guaranteeing good performance of our methodology considering the case of the Clayton copula.

![Table 3. Parameters of the copula and of the marginal distributions for each DGP](image)

<table>
<thead>
<tr>
<th>DGP</th>
<th>Copula parameter</th>
<th>Marginal parameter</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.a</td>
<td>$\rho_1 = 0.6$</td>
<td>$\mu_X = 0, \mu_Y = 1$</td>
<td>$1 + 0.6\xi_1$</td>
</tr>
<tr>
<td>S.b</td>
<td>$\theta = 0.8$</td>
<td>$\mu_Y = 0, \sigma_Y = 1$</td>
<td>$0.8/\sqrt{\pi} - 1.6/\sqrt{\pi} \exp(-\exp(x_1))$</td>
</tr>
<tr>
<td>S.c</td>
<td>$\rho = 0.6, df = 3, 5, 100$</td>
<td>$\mu_X = 0, \mu_Y = 1$</td>
<td>No simple form</td>
</tr>
<tr>
<td>S.d</td>
<td>$\delta = 1$</td>
<td>$\mu_X = 0, \mu_Y = 1$</td>
<td>No simple form</td>
</tr>
<tr>
<td>M.a (d = 2)</td>
<td>$\Sigma = \begin{pmatrix} 1 &amp; -0.5 &amp; 0.9 \ -0.5 &amp; 1 &amp; -0.4 \ 0.9 &amp; -0.4 &amp; 1 \end{pmatrix}$</td>
<td>$\mu_X = \mu_X = 0$</td>
<td>$\Phi(-0.154\xi_1 + 0.771\xi_2)$</td>
</tr>
<tr>
<td>M.a (d = 3)</td>
<td>$\Sigma = \begin{pmatrix} 1 &amp; 0.23 &amp; 0.90 &amp; 0.67 \end{pmatrix}$</td>
<td>$\mu_X = \mu_X = 0$</td>
<td>$\Phi(-0.3\xi_1 + 0.9\xi_2 + 0.3\xi_3)$</td>
</tr>
</tbody>
</table>

The parameters of the copula and of the marginal distributions that we use for each DGP are given in Table 3. All computations are done with R (R Development Core Team 2010).

![Table 4. Coverage probabilities of the proposed confidence interval for $m(x), \alpha = 0.05$](image)

<table>
<thead>
<tr>
<th>DGP</th>
<th>$x$</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interior</td>
<td>S.a</td>
<td>0.00</td>
<td>0.943</td>
<td>0.938</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>S.b</td>
<td>-0.58</td>
<td>0.953</td>
<td>0.952</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td>M.a (d = 2)</td>
<td>(0.00,0.00)</td>
<td>0.946</td>
<td>0.943</td>
<td>0.937</td>
</tr>
<tr>
<td>Boundary</td>
<td>S.a</td>
<td>-1.64</td>
<td>0.907</td>
<td>0.933</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td>S.b</td>
<td>-2.49</td>
<td>0.968</td>
<td>0.957</td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td>M.a (d = 2)</td>
<td>(1.53,1.53)</td>
<td>0.946</td>
<td>0.945</td>
<td>0.946</td>
</tr>
</tbody>
</table>
5.2 Robustness and Comparison With Other Methods

In this section, we compare our semiparametric estimator both with semiparametric and nonparametric competitors. We consider four estimators for comparison.

- $\hat{m}_{tc}$: our estimator when the true copula family is used.
- $\hat{m}_{uc}$: our estimator when the copula density family is adaptively selected using the data (see the explanation below).
- $\hat{m}_{ll}$: local linear estimator with the bandwidth selected by CV using the R package \textit{np} (see Hayfield and Racine 2008).
- $\hat{m}_{si}$: single index regression estimator based on a two-stage estimation method; the single index coefficients are first estimated by the R package \textit{dr} using the slice inverse regression method (see Li 1991) and then the link function is estimated in the same way as for $\hat{m}_{ll}$.

When no information is available about the true copula density, we should choose an appropriate copula family for our estimator $\hat{m}_{uc}$. This step may be difficult especially when the number of covariates is large. The reason is that the set of high-dimensional copulas available in the literature is limited to very special and restrictive copula families such as elliptical copulas and Archimedean copulas. For this reason, the strategy that we advocate and adopt here is to make use of the recent available work about the simplified pair-copula decomposition. The main idea is to decompose a multivariate copula to a cascade of bivariate copulas so that we can take advantage of the relative simplicity of bivariate copula selection and estimation. To be more specific, we briefly describe such an approach for the case of a three-variate vector $X = (X_1, X_2, X_3)^T$. By applying Sklar’s theorem recursively, one can write, for example,

$$c(F_0(y), F_1(x_1), F_2(x_2), F_3(x_3)) = c(x_1, F_2(x_2), F_3(x_3)c_{01}(F_0(y), F_1(x_1)) \times c_{02|1}(F_0(y|x_1), F_2(x_2|x_1|x_1) \times c_{03|12}(F_0(y|x_1, x_2), F_3(x_3|x_1, x_2)|x_1, x_2),$$

where $c_{01}, c_{02|1}$ and $c_{03|12}$ are the copula densities associated with the distributions of $(Y, X_1)$, $(Y, X_2)|X_1$, and $(Y, X_3)|(X_1, X_2)$.
respectively. Similarly, $c_x$ can be, for example, decomposed as
$$c_x(F_1(x_1), F_2(x_2), F_3(x_3)) = c_{12}(F_1(x_1), F_2(x_2))c_{23}(F_2(x_2), F_3(x_3))c_{13}(F_1(x_1), x_2) 	imes F_3(x_3|x_2|x_2).$$

If we assume that all the conditional copulas depend on the conditioning variables only through the conditional distributions, for example, $c_{02}(F_0(y|x_1), F_2(x_2|x_1)) = c_{02}(F_0(y|x_1), F_2(x_2|x_1))$, then it leads to the so-called simplified pair-copula decomposition. Because any bivariate copula density can be used as a building block for this decomposition, the simplified pair-copula decomposition provides high flexibility and the ability to cover a wide range of complex dependencies. Hobæk Haff, Aas, and Frigessi (2010) and Stoeber, Joe, and Czado (2012) discussed the conditions under which such a simplification is possible and found that this is not a severe restriction in many situations. For more about this decomposition, see the recent book by Kurowicka and Joe (2010).

In our simulation, for our estimator $\hat{m}_{uc}$, we choose one decomposition (R-vine structure) for the data and then choose the pair-copulas independently among 10 candidate copulas: two are elliptical (Gaussian and Student’s t) and eight are Archimedean (Clayton, Gumbel, Frank, Joe, Clayton–Gumbel, Joe–Gumbel, Joe–Clayton, Joe–Frank) using the R package VineCopula; see Dißmann et al. (2011). As a selection criterion for bivariate copula, we use the Akaike information criterion (AIC).

**Remark 4.** As noted by the referees, instead of using the classical AIC, which lacks theoretical justification in the context of pseudo-maximum likelihood, it is possible to use the copula information criterion (CIC); see Grønneberg and Hjort (2011) and Grønneberg (2011). However, the CIC does not exist for some popular copula families such as Gumbel and Joe copulas and is complicated to compute even when it exists. Compared with the CIC, the AIC is easy to compute and has been shown to perform well for copula selection in R-vine framework; see Brechmann (2010, chap. 5).

**5.2.1 When the Simplified Pair-Copula Assumption Holds.** To generate data that satisfy our copula assumptions, we consider DGP M.a (d = 2) and M.a (d = 3). As a reference case, we consider the nonlinear least-square estimator $\hat{m}_{ls}$ with the true link function. We use the R package nlrtr to calculate $\hat{m}_{ls}$. For details, we refer to Ritz and Streibig (2008). Table 5 shows the IMSE of each estimator. As expected, the least-square estimator $\hat{m}_{ls}$ gives the best performance. The difference in IMSE between $\hat{m}_{ls}$ and our copula estimators decreases as the sample size increases. Since additional variability comes into the estimation from the selection of copula family for each bivariate copula, $\hat{m}_{uc}$ has larger IMSE than $\hat{m}_{uc}$ but the difference is relatively moderate. The reason for the modest difference is that the Gaussian copula admits R-vine decompositions with Gaussian pair-copulas (see Hobæk Haff, Aas, and Frigessi 2010) and the AIC works reasonably well. We observe that the estimator $\hat{m}_{uc}$ does better than the local linear estimator on the whole and especially at boundary regions (the details are not shown here).

The same remarks remain valid in three-covariate case. In this case, it is important to note that the difference between $\hat{m}_{uc}$ and $\hat{m}_{uc}$ becomes bigger than in two-covariate case especially when the sample size is small. This is because the number of bivariate copulas to be selected and estimated by the data in the decomposition increases with the number of the covariates (six instead of three in our case). However, when $d = 3$, $\hat{m}_{uc}$ still remains significantly better than the local linear estimator, which is known to suffer from the curse of dimensionality. Finally, it is notable that our estimator $\hat{m}_{uc}$ is even better than the single index estimator $\hat{m}_{si}$ overall. The reason for it may be that the uncertainty stemming from the copula choice is relatively smaller than that coming from the estimation of the single index coefficients when the dimension of the covariates is relatively small.

**5.2.2 When the Simplified Pair-Copula Assumption Does not Hold.** In this section, we investigate how our estimator performs when the true copula does not belong to the class of R-vines under consideration. For such purpose, we generate the data from DGP M.b, DGP M.c, and DGP M.d, which are based on the regression model $Y = m(X) + \sigma \epsilon$, where $X$ is multivariate normal with mean 0 and $\text{cov}(X,i, X,j) = 0.5^{i-j}$ and $\epsilon \sim \mathcal{N}(0, 1)$ independent of $X$. Specification of each DGP is given in Table 6.

In these cases, to our knowledge, the true copula densities cannot fit into any simplified pair-copula decomposition with

<table>
<thead>
<tr>
<th>DGP</th>
<th>$m(x)$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M.b</td>
<td>$\psi(-0.3X_1 + 0.9X_2 + 0.3X_3)$</td>
<td>0.1</td>
</tr>
<tr>
<td>M.c</td>
<td>$\sqrt{2[X_1 - X_2 + 0.5] + (-0.5X_1 + 1)(0.1X_2^3)}$</td>
<td>2</td>
</tr>
<tr>
<td>M.d</td>
<td>$(-2X_1 + X_2 - 4X_3 + 3X_4 + X_5 + 2X_6)/\sqrt{35}$</td>
<td>0.5</td>
</tr>
</tbody>
</table>

between $\hat{m}_{ls}$ and our copula estimators. For details, we refer to Ritz and Streibig (2008). Table 5 shows the IMSE of each estimator. As expected, the least-square estimator $\hat{m}_{ls}$ gives the best performance. The difference in IMSE between $\hat{m}_{ls}$ and our copula estimators decreases as the sample size increases. Since additional variability comes into the estimation from the selection of copula family for each bivariate copula, $\hat{m}_{uc}$ has larger IMSE than $\hat{m}_{uc}$ but the difference is relatively moderate. The reason for the modest difference is that the Gaussian copula admits R-vine decompositions with Gaussian pair-copulas (see Hobæk Haff, Aas, and Frigessi 2010) and the AIC works reasonably well. We observe that the estimator $\hat{m}_{uc}$ does better than the local linear estimator on the whole and especially at boundary regions (the details are not shown here).

The same remarks remain valid in three-covariate case. In this case, it is important to note that the difference between $\hat{m}_{uc}$ and $\hat{m}_{uc}$ becomes bigger than in two-covariate case especially when the sample size is small. This is because the number of bivariate copulas to be selected and estimated by the data in the decomposition increases with the number of the covariates (six instead of three in our case). However, when $d = 3$, $\hat{m}_{uc}$ still remains significantly better than the local linear estimator, which is known to suffer from the curse of dimensionality. Finally, it is notable that our estimator $\hat{m}_{uc}$ is even better than the single index estimator $\hat{m}_{si}$ overall. The reason for it may be that the uncertainty stemming from the copula choice is relatively smaller than that coming from the estimation of the single index coefficients when the dimension of the covariates is relatively small.

**5.2.2 When the Simplified Pair-Copula Assumption Does not Hold.** In this section, we investigate how our estimator performs when the true copula does not belong to the class of R-vines under consideration. For such purpose, we generate the data from DGP M.b, DGP M.c, and DGP M.d, which are based on the regression model $Y = m(X) + \sigma \epsilon$, where $X$ is multivariate normal with mean 0 and $\text{cov}(X,i, X,j) = 0.5^{i-j}$ and $\epsilon \sim \mathcal{N}(0, 1)$ independent of $X$. Specification of each DGP is given in Table 6.

In these cases, to our knowledge, the true copula densities cannot fit into any simplified pair-copula decomposition with
the bivariate copulas in our candidate list. Consequently, the performance of \( \hat{m}_{uc} \) is affected by the misspecification caused by assuming the pair-copula decomposition and arising in the bivariate copula selection. Table 7 shows that on the whole our estimator \( \hat{m}_{uc} \) has better performance than the nonparametric estimator \( \hat{m}_{si} \) especially when the sample size is small and the number of covariates is large. In the cases of DGP M.b and M.d, where the true model is a single-index model, our estimator \( \hat{m}_{si} \) performs similarly to the estimator \( \hat{m}_{si} \) but with \( \hat{m}_{si} \) having a slight advantage. This is natural because the estimator \( \hat{m}_{si} \) directly exploits the information about the true regression structure. However, under DGP M.c where the single index assumption is not correct, our estimator \( \hat{m}_{uc} \) seems to work better than the other estimators \( \hat{m}_{si} \) and \( \hat{m}_{ll} \).

6. REAL DATA ANALYSIS

To illustrate our method, we analyze data from air pollution studies. The data consist of measurements of daily ozone concentration \((Y = \log(\text{ozone}))\), solar radiation \((X_1 = \text{rad})\), daily maximum temperature \((X_2 = \text{temp})\), and wind speed \((X_3 = \text{wind})\) on \( n = 111 \) days from May to September 1973 in New York. To estimate \( m(X_1, X_2, X_3) = \mathbb{E}(Y|X_1, X_2, X_3) \), we consider the following:

(LL) Local linear estimator with the bandwidth selected by CV.

(LS) Least-square estimator \( \hat{\beta}_0 + \hat{\beta}_1 \text{rad} + \hat{\beta}_2 \text{temp} + \hat{\beta}_3 \text{wind} \).

(AD) Additive model estimator \( \hat{g}_1(\text{rad}) + \hat{g}_2(\text{temp}) + \hat{g}_3(\text{wind}) \).

(SI) Single index model estimator \( \hat{g}(\beta_1 \text{rad} + \beta_2 \text{temp} + \beta_3 \text{wind}) \).

Table 8. Cross-validation error of each estimator for ozone data

<table>
<thead>
<tr>
<th></th>
<th>(LS)</th>
<th>(AD)</th>
<th>(SI)</th>
<th>(CO)</th>
<th>(LL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV1</td>
<td>0.3183</td>
<td>0.2917</td>
<td>0.4465</td>
<td>0.2768</td>
<td>0.2815</td>
</tr>
<tr>
<td>CV2</td>
<td>0.2174</td>
<td>0.2203</td>
<td>0.3446</td>
<td>0.2065</td>
<td>0.1701</td>
</tr>
</tbody>
</table>

Model (LS) was considered by Crawley (2005), who found that the quadratic model fitted the data well. (AD), (SI), and (CO) are semiparametric models. We use a smoothing spline to fit the AD using R package gam and a local linear estimator to fit the SI. Finally, as an evaluation measure of each estimator, we compute the following two versions of CV error:

\[
CV_1(\hat{m}) = \text{median}_{1 \leq i \leq n} |Y_i - \hat{m}_{-i}(X_i)|
\]

\[
CV_2(\hat{m}) = \sum_{i=1}^{n} (Y_i - \hat{m}_{-i}(X_i))^2,
\]

where \( \hat{m}_{-i}(X_i) \) is the estimate of \( m(X_i) \) from the dataset \( \{Y_j, X_j\}; j \neq i, j = 1, \ldots, n \). \( CV_1 \) is a robustified version of the standard CV error \( CV_2 \).

Table 8 suggests that all the estimators except (SI) show more or less similar performances with our method having a slight advantage. Clearly, the single index structure is not appropriate for this data and the additive structure seems to be more adequate. In terms of the CV criteria, our estimator shows very good performance. The residual plots between \( X_{3,i} \) (wind) and \( Y_i - \hat{m}_{-i}(X_i) \) for (AD), (SI), and (CO), see Figure 5, also support these remarks. We observe that the residuals from (SI) show a decreasing trend as \( X_3 \) (wind) increases while the residuals from (AD) and (CO) show a random pattern. This example clearly shows the flexibility of our estimator and its ability to adapt to the underlying regression structure of the data.

7. CONCLUSION AND FUTURE WORKS

We proposed a new semiparametric estimator of a regression function based on copulas. The estimator is obtained by modeling the joint distribution of the response and its covariates via a parametric copula family and the marginals via nonparametric methods. This method is flexible, easy to implement and seems to be less influenced by the curse of dimensionality. However, this advantage is attained at the price of a model risk in the
copula modeling. Here, we empirically checked that such a model risk was small but some theoretical analysis about the additional risk stemming from the model selection step is needed. Additionally, since copula has some advantages in modeling tail dependence, it would be interesting to see whether our copula regression framework benefits from those advantages in the estimation when the data have a specific tail dependence.

APPENDIX: TECHNICAL DETAILS

Lemma 1. Assume that $e$ satisfies Assumption C. If (i) $\mathbb{E}[|Y|] < \infty$, (ii) $\hat{F}_{i,n}(Y_i) - F_i(Y_i)$ is $O_p(n^{-1/2})$, and (iii) $\hat{\theta}_n - \theta_0$ is $O_p(n^{-1/2})$, then we have

$$\hat{\theta}_n = n^{-1} \sum_{i=1}^{n} Y_i F_0(Y_i) F_i(x_1; \theta_0) + V_{1,n} + V_{2,n} + V_{3,n}, \quad (A.1)$$

where

$$V_{1,n} = n^{-1} \sum_{i=1}^{n} Y_i \partial_c(F_0(Y_i), F_i(x_1; \theta_0)),$$

$$V_{2,n} = n^{-1} \sum_{i=1}^{n} Y_i \hat{F}_{i,n}(Y_i) \partial_c(F_0(Y_i), F_i(x_1; \theta_0)),$$

$$V_{3,n} = n^{-1} \sum_{i=1}^{n} Y_i \hat{F}_{i,n}(Y_i) \partial_c(F_0(Y_i), \hat{F}_{i,n}(x_1; \theta_0)).$$

Proof. Using the first-order Taylor expansion, we have

$$\hat{\theta}_n = n^{-1} \sum_{i=1}^{n} Y_i c(F_0(Y_i), F_i(x_1; \theta_0)), \quad (A.2)$$

which leads to $R_n = o_p(n^{-1/2})$. Thus, we know that

$$V_{1,n} = n^{-1} \sum_{i=1}^{n} Y_i (\hat{F}_{i,n}(Y_i) - F_0(Y_i)) \partial_c(F_0(Y_i), F_i(x_1; \theta_0)) + o_p(n^{-1/2}). \quad (A.2)$$

Now, we turn to the second term $V_{2,n}$. Following the same arguments as for $V_{1,n}$ with Assumption (ii), we have

$$V_{2,n} = n^{-1} \sum_{i=1}^{n} Y_i \hat{F}_{i,n}(x_1) \partial_c(F_0(Y_i), F_i(x_1; \theta_0)) + o_p(n^{-1/2}). \quad (A.3)$$

In the last equality, we use the fact that $n^{-1} \sum_{i=1}^{n} Y_i \partial_c(F_0(Y_i), F_i(x_1; \theta_0))$ converges in probability to $\mathbb{E}(Y \partial_c(F_0(Y), F_i(x_1; \theta_0))) = \partial_c(F_i(x_1; \theta_0)).$ Similarly, using Assumption (iii), the last term $V_{3,n}$ can be expressed as

$$V_{3,n} = \mathbb{E}(\hat{\theta}_n - \theta_0)^\top \hat{\epsilon}(\hat{\theta}_n, \hat{\theta}_0) + o_p(n^{-1/2}). \quad (A.4)$$

Recollceting the elements (A.2), (A.3), (A.4), and (A.1) gives the claimed result.

Proof of Theorem 1

Put $\hat{V}_n = n^{-1} \sum_{i=1}^{n} Y_i \hat{F}_{i,n}(Y_i) \partial_c(F_0(Y_i), F_i(x_1; \theta_0)).$ This is a $V$-statistic with the kernel

$$h(Y_i, x) = \begin{cases} \frac{1}{2} \left[ Y_i (I(Y_i \leq y) - F_0(y)) \partial_c(F_0(y), F_i(x_1; \theta_0) 
+ Y_i (I(Y_i > y) - F_0(y)) \partial_c(F_0(y), F_i(x_1; \theta_0)) \right] \end{cases}.$$

By Assumption C and (i), we have $\mathbb{E}[h^2(Y, Y')] < \infty$. Therefore, using Lemma 5.7.3 and Theorem 5.3.2 in Serfling (1980) and the fact that $\mathbb{E}h(Y, Y') = 0$, we get that

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \lambda(Y_i, F_i(x_1; \theta_0)) + o_p(n^{-1/2}), \quad (A.5)$$

where

$$\lambda(u_0, u_1; \theta) = \int_y \{ I(y \leq u_0) - F_0(y) \} \partial_c(F_0(y), u_1; \theta) f_0(y) dy.$$

Using integration by part and the fact that $\mathbb{E}[|Y|] < \infty$, it is easy to check that

$$\lambda(u_0, u_1; \theta) = \int y \{ c(F_0(y), u_1; \theta) f_0(y) dy - u_0 c(F_0(u_0), u_1; \theta) 
- \int (I(y \leq u_0) - F_0(y)) c(F_0(y), u_1; \theta) dy.$$

This result together with Lemma 1, Assumptions A and B concludes the proof.

Sketch of the Proof for the Consistency of $\hat{\sigma}^2$

Here, we sketch the proof for the single parameter ($p = 1$) and single covariate ($d = 1$) case. The proof for the general case can be done in a similar way. To prove the consistency, we mainly need to prove that

$$n^{-1} \sum_{i=1}^{n} \hat{E}_i^2(\hat{\theta}_n) = n^{-1} \sum_{i=1}^{n} \hat{E}_n^2(\hat{\theta}_n) + o_p(1) \quad (A.6)$$

$$= n^{-1} \sum_{i=1}^{n} \hat{E}_i^2(\hat{\theta}_n) + o_p(1). \quad (A.7)$$
This holds because Assumption (C) implies

\[
\sum_{i=1}^{n} Y_i \frac{c(F_0(Y_i), F_1(x_i); \theta_0)}{\partial \theta} + O_{_p}(n^{-1/2}) = O_{_p}(1).
\]

Following the similar arguments for the proof of Lemma 1 with (C1) and (C2), one can easily show that

\[
\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \frac{c(F_0(Y_i), F_1(x_i); \theta_0)}{\partial \theta} + O_{_p}(n^{-1/2})
\]

To prove the inequality (A.6), it is enough to show that

\[
\sum_{i=1}^{n} \left( \hat{E}_i - \hat{E}_i(\theta_0) \right)^2 = O_{_p}(1). \tag{9}
\]

This holds because Assumption (C3) implies

\[
\sum_{i=1}^{n} \left( \frac{1}{\theta_n} Y_n \frac{c(F_0(Y_i), F_1(x_i); \theta_n)}{\partial \theta} \right)^2 = \frac{1}{\theta_n^2} \sum_{i=1}^{n} M^2(F_0(Y_i), F_1(x_i)) + O_{_p}(1).
\]

The last equality follows by applying Theorem 6 in Fermanian, Radulović, and Wegkamp (2004) to the multivariate rank order statistics

\[
n^{-1} \sum_{i=1}^{n} M^2(F_0(Y_i), F_1(x_i)).
\]

Similarly to (A.6), to prove (7), it is enough to show that

\[
n^{-1} \sum_{i=1}^{n} \left( \hat{E}_i - E_i(\theta_0) \right)^2 = O_{_p}(1). \tag{8}
\]

The second equality holds because, by Assumption C, the triangle inequality, and Glivenko–Cantelli Theorem, we have

\[
\int |Y_i \leq y | - F_{0,y}(y) - \mathcal{E}_i(\theta_0) dy = O_{_p}(1).
\]

As for the first equality, from the definition of \(\eta(u_0, u_1, \theta)\) in (9), we only need to show that

\[
n^{-1} \sum_{i=1}^{n} \left( \frac{1}{\theta_n} Y_n \frac{c(F_0(Y_i), F_1(x_i); \theta_n)}{\partial \theta} \right)^2 = O_{_p}(1).
\]

This can be done by mimicking the arguments that Genest, Ghoudi, and Rivest (1995) used to prove the consistency of their estimator for the asymptotic variance.

REFERENCES


