Biometrics 000, 000–000 DOI: 000
000 0000

Dependence Calibration in Conditional Copulas: A Nonparametric Approach

Elif F. Acar *, Radu V. Craiu, and Fang Yao

Department of Statistics, University of Toronto

100 St. George Street, Toronto, Ontario M5S 3G3, Canada

**email:* elif@utstat.toronto.edu

Summary: The study of dependence between random variables is a mainstay in Statistics. In many cases the strength of dependence between two or more random variables varies according to the values of a measured covariate. We propose inference for this type of variation using a conditional copula model where the copula function belongs to a parametric copula family and the copula parameter varies with the covariate. In order to estimate the functional relationship between the copula parameter and the covariate, we propose a nonparametric approach based on local likelihood. Of importance is also the choice of the copula family that best represents a given set of data. The proposed framework naturally leads to a novel copula selection method based on cross-validated prediction errors. We derive the asymptotic bias and variance of the resulting local polynomial estimator, and outline how to construct pointwise confidence intervals. The finite sample performance of our method is investigated using simulation studies and is illustrated using a subset of the Matched Multiple Birth data.

Key words: Copula parameter; Copula selection; Covariate adjustment; Local likelihood; Local polynomials; Prediction error.

1. Introduction

Understanding dependence is an important, yet challenging, task in multivariate statistical modeling. One often needs to specify a complex joint distribution of random variables to have a complete view of the dependence structure. The challenge of constructing such multivariate distributions can be significantly reduced if one uses a copula model to separate the marginal components of a joint distribution from its dependence structure. Sklar's theorem (1959) is central to the theoretical foundation needed for the use of copulas as it states that a multivariate distribution can be fully characterized by its marginal distributions and a copula, i.e., a multivariate distribution function having uniform $[0, 1]$ marginals.

In what follows, we focus on the bivariate case only for simplicity, the arguments being extendable to more than two dimensions. Let Y_1 and Y_2 be continuous random variables of interest with joint distribution function H and marginal distributions F_1 and F_2, respectively. Sklar's theorem ensures the existence of a unique copula $C : [0, 1]^2 \to [0, 1]$, which satisfies $H(y_1, y_2) = C(F_1(y_1), F_2(y_2))$, for all $(y_1, y_2) \in \mathbb{R}^2$.

In the last twenty years copulas have been widely used in a variety of applied work. We refer the reader to Embrechts et al. (2002), Cherubini et al. (2004) and Frees and Valdez (1998) for applications specific to finance and insurance. In survival analysis, Clayton (1978), Shih and Louis (1995), Wang and Wells (2000) and the monograph by Hougaard (2000) present the copula techniques to model multivariate time-to-event data and competing risks. As a direct result of their wide applicability, a large number of parametric families of copulas, typically indexed by a real-valued parameter θ, have been proposed in the literature to represent different dependence patterns. While the copula family describes the functional form, within each family it is the copula parameter, θ, which controls the strength of the dependence. A comprehensive introduction on copulas and their properties can be found in Nelsen (2006) and the connections between various copulas and dependence concepts are discussed in detail

by Joe (1997). If a parametric form is assumed for the copula function, estimation can be achieved using maximum likelihood estimation for the single copula parameter (Joe, 1997; Genest et al., 1995). Alternatively, estimation can be performed fully nonparametrically by using kernel estimators (Fermanian and Scaillet, 2003; Chen and Huang, 2007).

Although copulas have been in use in the applied statistical literature for more than twenty years, the covariate adjustment for copulas has been considered only recently. The extension of Sklar's theorem for conditional distributions (Patton, 2006) allows us to adjust for covariates. For instance, if, in addition to Y_1 and Y_2, we have information on a covariate X, then the influence of X on the dependence between Y_1 and Y_2 can be modeled by the conditional copula $C(\cdot \mid X)$, which is the joint distribution function of $U_1 \equiv F_{1|X}(Y_1 \mid x)$ and $U_2 \equiv F_{2|X}(Y_2 \mid x)$ given $X = x$, where $Y_i \mid X = x$ has cdf $F_{i|X}(. \mid x)$, $i = 1, 2$. Patton (2006) showed that for each x in the support of X, the joint conditional distribution is uniquely defined by

$$H_X(y_1, y_2 \mid x) = C(F_{1|X}(y_1 \mid x), F_{2|X}(y_2 \mid x) \mid x), \quad \text{for all } (y_1, y_2) \in \mathbb{R}^2. \tag{1}$$

Conditional copulas have been used mostly in the context of financial time series to allow for time-variation in the dependence structure via likelihood inference for ARMA models (Patton, 2006; Jondeau and Rockinger, 2006; Bartram et al., 2007).

The main contribution of the current work is to provide a nonparametric procedure to estimate the functional relationship between the copula parameter and the covariate(s). Our motivation is to relax the parametric assumptions about this relationship. Since a parametric model will only find features in the data that are already incorporated *a priori* in the model, parametric approaches might not be adequate if the dependence structure does not fall into a preconceived class of functions.

In Section 4 we investigate the impact of gestational age on the dependence between twin birth weights using a subset of the Matched Multiple Birth Data Set. Among the twin

live births, we consider those who were delivered between 28–42 weeks of gestation and in which both twins survived for the first year of life. Our initial investigation, shown in Figure 1, indicates a relatively stronger dependence between the birth weights (in grams) of the preterm (28–32 weeks) and post–term (38–42 weeks) twins compared to the twins delivered at term (33–37 weeks). This suggests the need of such nonparametric approach as an exploratory tool for detecting the underlying functional relationship between the copula parameter and the covariate.

[Figure 1 about here.]

Smoothing methods for function estimation have been substantially studied for various problems. In this paper we use the local polynomial framework (see Fan and Gijbels, 1996, for a comprehensive review) for the covariate adjusted copula estimation via local likelihood-based models (Tibshirani and Hastie, 1987). In practice, all inferential methods for copulas must be accompanied by a strategy to select among a number of copula families the one that best approximates the data at hand. Choosing an appropriate family of copulas to fit a given set of data is challenging and has recently attracted considerable interest. Some methods for copula selection include goodness-of-fit tests based on the empirical copula (Durrleman et al., 2000), on the Kendall process (Genest and Rivest, 1993; Wang and Wells, 2000; Genest et al., 2007), and on kernel density estimation (Fermanian, 2005; Craiu and Craiu, 2008). Our estimation procedure naturally leads to a novel copula selection method based on cross–validated prediction error. Besides being data-adaptive, the proposed selection criterion makes comparisons across copula families possible due to its general applicability.

The paper is organized as follows. In Section 2, we present the proposed estimation procedure, discuss aspects related to copula selection, and derive the asymptotic bias and variance of the nonparametric estimator used for constructing pointwise confidence bands. Section 3 contains our simulation studies and in Section 4 we use the Matched Multiple Birth

Data Set to investigate the influence of the gestational age on the structure of dependence between the twin birth weights. Discussion and conclusions are presented in Section 5. Technical details and additional simulated and real data examples can be found in the Web Appendix.

2. Methodology

2.1 *Proposed method*

Let Y_1 and Y_2 be continuous variables of interest and X be a continuous variable that may affect the dependence between Y_1 and Y_2. We consider the model (1) with the conditional density $h_X(Y_1, Y_2 \mid x; \theta, \alpha_1, \alpha_2)$ in which our main interest lies in the conditional copula parameter θ, while the conditional marginal densities $f_{1|X}$ and $f_{2|X}$ are characterized by α_1 and α_2, respectively,

$$h_X(y_1, y_2 \mid x; \theta, \alpha_1, \alpha_2) = f_{1|X}(y_1 \mid x; \alpha_1)\, f_{2|X}(y_2 \mid x; \alpha_2)\, c(u_1, u_2 \mid x; \theta, \alpha_1, \alpha_2),$$

where $u_i = F_{i|X}(y_i \mid x; \alpha_i), i = 1, 2$ and $c(u_1, u_2 \mid x; \theta, \alpha_1, \alpha_2)$ is the conditional copula density. Here, we impose a minimal requirement that the parameters that govern the marginals are different from the copula parameter. It is easy to see that such a requirement is not restrictive, for instance, in a regression setting, marginals may correspond to mean effects and the copula to covariance structure. Hence, the estimation can be performed in two-stages, first for the marginal parameters and then for the copula. Then, by replacing the estimates $\hat{F}_{1|X}(y_1 \mid x)$ and $\hat{F}_{2|X}(y_2 \mid x)$ in (1), we can estimate the functional form of the copula parameter.

Since our main focus is on the dependence structure, we assume that the conditional marginal distributions $F_{1|X}$ and $F_{2|X}$ are known, and consider the following model

$$(U_{1i}, U_{2i}) \mid X_i \sim C(u_{1i}, u_{2i} \mid \theta(x_i)\,),$$

where $\theta(x_i) = g^{-1}(\eta(x_i)), \quad i = 1, \ldots, n.$

Here, $g^{-1} : \mathbb{R} \to \Theta$ is the known inverse link function, which ensures that the copula

parameter has the correct range, and η is the unknown *calibration function* to be estimated. The term *calibration* emphasizes that the level of dependence is adjusted for the covariate effect on the copula parameter. Analogous to the generalized linear models one needs to choose an appropriate link function, since there is no guarantee that the estimate of θ is in the correct parameter range for the particular copula family under consideration. For instance, for the Clayton copula family $\theta \in (0, \infty)$, so we use $g^{-1}(t) = \exp(t)$. As long as the link function is monotone the choice is irrelevant, since inference is invariant to monotone transformations of θ.

We begin with a classical polynomial functional form to motivate the proposed nonparametric strategy. If the relationship between θ and X falls into a pre-specified class of functions, for instance, the polynomials up to degree p, we may estimate the calibration function $\eta(\cdot)$ via maximum likelihood estimation. More specifically, we write $\eta(X) = \sum_{j=0}^{p} \tilde{\beta}_j X^j$, and estimate $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ by maximizing $\mathcal{L}(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \ln c(U_{1i}, U_{2i} \mid g^{-1}(\tilde{\beta}_0 + \tilde{\beta}_1 X_i + \cdots + \tilde{\beta}_p X_i^p))$.

However, for most copula families, the function $\eta(\cdot)$ is not necessarily well approximated by a pre-conceived polynomial model. Moreover, in contrast to classical regression, the form of the calibration function $\eta(\cdot)$ characterizing the underlying dependence structure is more difficult to discern by simply inspecting the data. Therefore, a nonparametric approach for estimating such a latent function is more needed here than it is in the classical regression context.

We adopt the local polynomial framework (Fan and Gijbels, 1996) within the local likelihood formulation (Tibshirani and Hastie, 1987) as follows. Assume η has $(p+1)^{\text{th}}$ continuous derivatives at an interior point x. For data points X_i in the neighborhood of x, we approximate $\eta(X_i)$ by a Taylor expansion of polynomial of degree p,

$$\eta(X_i) \;\approx\; \eta(x) + \eta'(x)(X_i - x) + \cdots + \frac{\eta^{(p)}(x)}{p!}(X_i - x)^p \;\equiv\; \boldsymbol{x}_{i,x}^T \boldsymbol{\beta},$$

where $\boldsymbol{x}_{i,x} = (1, X_i - x, \ldots, (X_i - x)^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ with $\beta_\nu = \eta^{(\nu)}(x)/\nu!$. In our implementations we use the commonly adopted local linear fit, i.e. $p = 1$, as in Fan and Gijbels (1996).

The contribution of each data point $(U_{1i}, U_{2i}) \mid X_i$ in a neighborhood of x to the local likelihood is given by $\ln c(U_{1i}, U_{2i} \mid \ g^{-1}(\boldsymbol{x}_{i,x}^T \boldsymbol{\beta}))$. The weighted sum of contributions forms the conditional local log-likelihood

$$\mathcal{L}(\boldsymbol{\beta}, x, p, h) = \sum_{i=1}^{n} \ln c(U_{1i}, U_{2i} \mid g^{-1}(\boldsymbol{x}_{i,x}^T \boldsymbol{\beta}))\, K_h(X_i - x),$$

where h is a bandwidth controlling the size of the local neighborhood and $K_h(\cdot) = 1/h\ K(\cdot/h)$ with K a kernel function assigning weights to the data points in a certain local "window". In our implementations we use the commonly adopted Epanechnikov kernel, $K(z) = 3/4(1 - z^2)_+$, where the subscript "+" denotes the positive part.

The local maximum likelihood estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ is thus obtained by solving the following estimating equation,

$$\nabla \mathcal{L}(\boldsymbol{\beta}, x) = \frac{\partial \mathcal{L}(\boldsymbol{\beta}, x, p, h)}{\partial \boldsymbol{\beta}} = 0. \tag{2}$$

The numerical solution for (2) is found via the Newton-Raphson iteration

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - [\nabla^2 \mathcal{L}(\boldsymbol{\beta}_m, x)]^{-1} \nabla \mathcal{L}(\boldsymbol{\beta}_m, x), \quad m = 0, 1, \ldots\ ,$$

where $\nabla \mathcal{L}$ denotes the score vector and $\nabla^2 \mathcal{L}$ the hessian matrix (expressions of $\nabla \mathcal{L}$ and $\nabla^2 \mathcal{L}$ can be found in the Web Appendix). One can then obtain the estimator for $\eta^{(\nu)}(x)$, $\nu = 0, \ldots, n$, where of particular interest is $\hat{\eta}(x) = \hat{\beta}_0(x)$. Finally, the copula parameter is estimated at covariate value x by applying the inverse link function

$$\hat{\theta}(x) = g^{-1}(\hat{\eta}(x)).$$

2.2 *Model tuning*

For practical implementation, we distinguish two aspects of dependence in copula models, which are the level of dependence within the copula function and, more importantly, the

functional dependence characterized by the copula family. We shall deal with both facets of the model tuning corresponding to selection of the smoothing bandwidth and the copula family.

Various methods for bandwidth selection exist in the literature, including cross-validation techniques, plug-in methods, and others. Since our estimation procedure is based on the local copula likelihood, the *leave-one-out* cross–validated local likelihood serves as a natural choice for the bandwidth selection.

Let $\hat{\theta}_h(\cdot)$ denote the estimate of the copula parameter function depending on a bandwidth parameter h. For each $1 \leqslant i \leqslant n$, we leave out the data point (U_{1i}, U_{2i}, X_i) and use the remaining data $\{U_{1j}, U_{2j}, X_j, j \neq i\}$ to obtain $\hat{\theta}_h^{(-i)}(X_i)$, the estimate of the copula parameter θ at X_i. The estimates obtained by leaving out the i^{th} data point, are then used to build the objective function depending on the bandwidth parameter,

$$\mathcal{B}(h) = \sum_{i=1}^{n} \ln c(U_{1i}, U_{2i} \mid \hat{\theta}_h^{(-i)}(X_i)). \tag{3}$$

The optimum bandwidth h^* is the one that maximizes (3). In practice, when the underlying function has spiky features and/or the covariate design is highly nonuniform, one may use a nearest neighbor type of variable bandwidth selection that chooses a suitable proportion of data points contributing to each local estimate in a manner similar to (3).

One can see that the above general principle of the cross–validated likelihood does not apply to the selection of copula family, as the scale of likelihoods varies across families. It is necessary to characterize the goodness-of-fit using different families on a comparable benchmark. Here we propose the cross–validated prediction of each response variable based on the other in a symmetric fashion. One can certainly modify the proposed criterion below if not both variables are of equal interest.

Suppose we have a (finite) set of candidate families $\mathfrak{C} = \{\mathcal{C}_q : q = 1, \ldots Q\}$, from which we want to choose the one that represents best the data at hand. For the qth copula family the

bandwidth selection process yields the optimal bandwidth h_q^*. For each left-out sample point (U_{1i}, U_{2i}, X_i), we obtain the estimate for the conditional copula's parameter $\hat{\theta}_{h_q^*}^{(-i)}$, which, in turn, leads to the *best candidate model* from the qth family, $C_q(U_{1i}, U_{2i} \mid \hat{\theta}_{h_q^*}^{(-i)}(X_i))$, with $i = 1, \ldots, n, \quad q = 1, \ldots, Q$. We use the conditional expectation formula to measure the predictive ability for each of the candidate models. Within family $\mathcal{C}_q$, the *best conditional prediction* for U_{1i} is

$$\widehat{E}_q^{(-i)}(U_{1i} \mid U_{2i}, X_i) = \int_0^1 U_1 \; c_q(U_1, U_{2i} \mid \hat{\theta}_{h^*}^{(-i)}(X_i)) \; dU_1.$$

Then, the cross-validated prediction error (CVPE) is used to define the model selection criterion

$$\text{CVPE}(\mathcal{C}_q) = \sum_{i=1}^{n} \Big\{ (U_{1i} - \widehat{E}_q^{(-i)}(U_{1i} \mid U_{2i}, X_i))^2 + (U_{2i} - \widehat{E}_q^{(-i)}(U_{2i} \mid U_{1i}, X_i))^2 \Big\}. \qquad (4)$$

The copula family $\mathcal{C}_q$ which yields the minimum CVPE($\mathcal{C}_q$) value is selected. This criterion can be justified as follows. If we denote M_0 as the true copula family and M the working copula family, then the first part in (4) normalized by $1/n$ is an approximation of $E_{M_0}[(U_1 - E_M[U_1|U_2, X])^2|U_2, X]$ which is minimized when the model M is correctly specified, i.e. $M = M_0$. A similar result holds for the second part in (4).

2.3 *Asymptotic properties*

Before presenting the main results, we shall introduce some notation. Let $f_X(\cdot) > 0$ be the density function of the covariate X. Denote the moments of K and K^2 respectively by $\mu_j = \int t^j K(t)dt$ and $\nu_j = \int t^j K^2(t)dt$, and write the matrices $S = \big(\mu_{j+\ell}\big)_{0 \leqslant j,\ell \leqslant p}$, $S^* = \big(\nu_{j+\ell}\big)_{0 \leqslant j,\ell \leqslant p}$, the $(p+1) \times 1$ vectors $\boldsymbol{s}_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$, as well as the unit vector $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$. For simplicity, we use $\ell\,(\theta, U_1, U_2) = \ln c(U_1, U_2 \mid \theta)$ for the log-copula density and denote its first and second derivatives with respect to θ by $\ell'(\theta, U_1, U_2) = \partial\ell(\theta, U_1, U_2)/\partial\theta$ and $\ell''(\theta, U_1, U_2) = \partial^2\ell(\theta, U_1, U_2)/\partial\theta^2$, respectively. For a fixed point x lying in the interior of the support of f_X, define $\sigma^2(x) = -E\{\; \ell''\; (g^{-1}(\eta(x)), U_1, U_2) \;\mid\; X = x\}$.

For our derivations, we require the assumptions given in the Appendix. The assumption (A1) is to ensure that the copula density satisfies the first and second order Bartlett identities. Further discussion on condition (A1) for certain copula families can be found in Hu (1998) and Chen and Fan (2006). The mild regularity conditions in (A2) are commonly adopted in nonparametric regression.

Typically, an odd order polynomial fit is preferred to an even order fit in local polynomial modeling, as the latter induces a higher asymptotic variance (see Fan and Gijbels, 1996, for details). Therefore, we consider only the odd order fits in the asymptotic expressions for the conditional bias and variance. The following theorem summarizes the main results, denoting the collection of covariate/design variables $\{X_1, \ldots, X_n\}$ by $\mathbb{X}$, while the technical details are deferred to the Web Appendix.

THEOREM 1: *Assume that (A1) and (A2) hold, $h \to 0$ and $nh \to \infty$ as $n \to \infty$, for an odd order local polynomial fit of degree p,*

$$Bias(\hat{\eta}(x) \mid \mathbb{X}) = \boldsymbol{e}_1^T S^{-1} \; \boldsymbol{s}_p \; \frac{\eta^{(p+1)}(x)}{(p+1)!} \; h^{p+1} + o_p(h^{p+1}),$$

$$Var(\hat{\eta}(x) \mid \mathbb{X}) = \frac{1}{nh \; f(x) \; [(g^{-1})'(\eta(x))]^2 \; \sigma^2(x)} \; \boldsymbol{e}_1^T S^{-1} \; S^* \; S^{-1} \; \boldsymbol{e}_1 + o_p(\frac{1}{nh}).$$

As a direct corollary of Theorem 1, we obtain the asymptotic conditional bias and variance of the copula parameter estimator, $\hat{\theta}(x) = g^{-1}(\hat{\eta}(x))$.

COROLLARY 1: *Assume that conditions of Theorem 1 holds, then*

$$Bias(\hat{\theta}(x) \mid \mathbb{X}) = \boldsymbol{e}_1^T S^{-1} \; \boldsymbol{s}_p \; \frac{\eta^{(p+1)}(x)}{g'(\theta(x)) \; (p+1)!} \; h^{p+1} + o_p(h^{p+1}), \tag{5}$$

$$Var(\hat{\theta}(x) \mid \mathbb{X}) = \frac{1}{nh \; f(x)\sigma^2(x)} \; \boldsymbol{e}_1^T S^{-1} \; S^* \; S^{-1} \; \boldsymbol{e}_1 + o_p(\frac{1}{nh}). \tag{6}$$

One can use the result of Corollary 1 to approximate the bias and variance of the estimated copula parameter. The unknown quantity $\sigma^2(x)$ in the variance expression can be

approximated by

$$\hat{\sigma}^2(x) = -\int_0^1 \int_0^1 \ell''\left(\hat{\theta}(x), U_1, U_2\right)\ c(U_1, U_2 \mid \hat{\theta}(x))\ dU_1\ dU_2. \tag{7}$$

The approximate $100(1-\alpha)\%$ pointwise confidence bands for the copula parameter are given by,

$$\hat{\theta}(x) - \widehat{b}(x) \pm z_{1-\alpha/2}\ \widehat{V}(x)^{1/2}, \tag{8}$$

where $\widehat{b}(x)$ and $\widehat{V}(x)$ are the estimated bias and variance based on (5) and (6), and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)^{th}$ quantile of the standard normal distribution. In practice, estimating the bias can be difficult due to unknown higher order derivatives (see Fan and Gijbels, 1996, for further discussion on bias correction). Alternatively, when variability plays a dominant role in (8), one may use a smaller bandwidth to bring the bias down to negligible levels (Fan and Zhang, 2000).

In practice, the estimation of the conditional marginal distributions may have an impact on the inference of the copula parameter. If the marginal distributions can be adequately characterized by a parametric model, as in the example considered in Section 4, the $\sqrt{n}$-convergence rate is negligible compared to the nonparametric convergence rate, i.e., the additional variability due to the estimation of the conditional marginals can be ignored.

In the case when the marginal conditional distributions are estimated nonparametrically, the rate of convergence is of the same order as that of the copula estimator. It is thus difficult to analytically assess the two sources of uncertainty in a unified manner. A theoretical investigation of this scenario is of substantial interest, but requires a different framework. In practice, one viable way to incorporate the uncertainty from nonparametrically estimated marginals is to bootstrap the raw data and calculate the quantile-based bootstrap confidence bands. We illustrate this approach in Figure 7 in the Web Appendix using the twin data set. It is not surprising that, due to the uncertainty in the nonparametric marginals, the bootstrap bands are wider than the asymptotic ones using (8). Thus in the absence of an

adequate parametric marginal model, we suggest the use of the raw bootstrap approach. Nevertheless, the construction of more computationally efficient and theoretically justified inference procedures is of great importance and constitutes a central topic of our future efforts.

3. Simulation Study

This section illustrates the finite sample performance of the local copula parameter estimation and the copula selection method. We consider the Clayton, Frank and Gumbel families of copulas in simulations. These choices cover a wide range of situations as the Clayton and Gumbel copulas are known to exhibit strong and weak lower tail dependence, respectively, while the Frank copula is symmetric and shows no tail dependence. Basic properties of these copulas can be found in the Appendix.

The inverse link functions are chosen as $g^{-1}(t) = \exp(t)$ for the Clayton copula, $g^{-1}(t) = t$ for the Frank copula and $g^{-1}(t) = \exp(t) + 1$ for the Gumbel copula, so that the resulting copula parameter estimates will be in the correct range.

We generate the data $\{(U_{1i}, U_{2i} \mid X_i) : i = 1, 2, \ldots, n\}$ from the Clayton copula under each of the following models:

(1) Linear calibration function : $\eta(X) = 0.8X - 2$

$$(U_1, U_2) \mid X \sim C(u_1, u_2 \mid \ \theta = \exp(0.8X - 2)\ ) \quad \text{where } X \sim \text{Uniform}(2, 5),$$

(2) Quadratic calibration function: $\eta(X) = 2 - 0.3(X - 4)^2$

$$(U_1, U_2) \mid X \sim C(u_1, u_2 \mid \ \theta = \exp(2 - 0.3(X - 4)^2)\ ) \quad \text{where } X \sim \text{Uniform}(2, 5).$$

We first generate the covariate values X_i from Uniform $(2, 5)$. Then, for each $i = 1, 2, \ldots, n$, we obtain the copula parameter, $\theta_i = \exp(\eta(x_i))$, imposed by the given calibration and link functions, and simulate the pairs $(U_{1i}, U_{2i}) \mid X_i$ from the Clayton copula with the parameter θ_i. The true copula parameter varies from 0.67 to 7.39 in the linear calibration model, and

from 2.22 to 7.39 in the nonlinear one. Under each model, we conduct experiments with sample sizes $n = 200$ and $n = 500$, each replicated $m = 100$ times. Additional simulation results in which the true underlying model is based on the Frank copula are included in the Web Appendix.

To estimate the copula parameter, we perform the local linear estimation, with $p = 1$, under each family, as well as the parametric estimation in which η is assumed to be linear in X. By performing the estimation also under the Frank and Gumbel families, we investigate the impact of the copula (mis)specification. We compare our proposed approach with the parametric estimation when the underlying calibration function is correctly specified, as in the first model, and when it is misspecified, as in the second one. For the bandwidth parameter, we consider 12 candidate values, ranging from 0.33 to 2.96, equally spaced on a logarithmic scale. All results reported are based on the local estimates at the chosen optimum bandwidth, which is given by the cross-validated likelihood criterion (3).

Since the accuracy of the calibration estimation is not directly comparable across different copula families, we convert the copula parameters to a common scale provided by the Kendall's tau measure of association, a widely used approach in copula inference (Trivedi and Zimmer, 2007). The population version of Kendall's tau can be expressed in terms of a conditional copula function

$$\tau_C(x) = 4 \int_0^1 \int_0^1 C(u_1, u_2 \mid x) dC(u_1, u_2 \mid x) \; - 1.$$

For each of the three families, we report the connection between θ and Kendall's tau in the Appendix. Table 1 displays the Monte Carlo estimates of the integrated Mean Square Error (IMSE) along with the integrated square Bias (IBIAS^2) and integrated Variance (IVAR),

$$\text{IBIAS}^2(\hat{\tau}) = \int_{\mathcal{X}} \big(\mathbb{E}(\hat{\tau}(x)) - \tau(x) \big)^2 \, dx, \quad \text{IVAR}(\hat{\tau}) = \int_{\mathcal{X}} \mathbb{E}\left[\big(\hat{\tau}(x) - \mathbb{E}(\hat{\tau}(x)) \big)^2 \right] dx,$$
$$\text{IMSE}(\hat{\tau}) = \int_{\mathcal{X}} \mathbb{E}\left[(\hat{\tau}(x) - \tau(x))^2 \right] dx \; = \; \text{IBIAS}^2(\hat{\tau}) + \text{IVAR}(\hat{\tau}).$$

[Table 1 about here.]

From Table 1 we see that the parametric estimation performs better under the linear calibration model when it is using the correct functional form. However, when there is no known parametric model for the calibration function, the proposed nonparametric approach better captures the covariate effect on the copula parameter. The results also show that the performance of the estimation deteriorates as the properties of the used copula family significantly depart from the true one. Since the Frank copula, having no tail dependence, is closer to the Clayton copula than is the Gumbel copula, it yields better results in all simulation scenarios compared to Gumbel.

We also construct the approximate 90% pointwise confidence bands for the copula parameter under the correctly selected Clayton family, where the bias is much smaller than the variance as noticed in Table 1. We thus use half of the optimum bandwidth to assess $\sigma^2(x)$ in (7) while further reducing the bias to a negligible level (Fan and Zhang, 2000). To be consistent, for each Monte Carlo sample, the confidence intervals obtained for the copula parameter are converted to the Kendall's tau scale. Figure 2 displays the confidence bands for the Kendall's tau, averaged over 100 Monte Carlo samples ($n = 200$). For comparison, we present the Monte Carlo based confidence bands obtained from 100 estimates of Kendall's tau, which agree well with our proposal.

[Figure 2 about here.]

For the copula selection, we calculate the cross-validated prediction errors (4) and evaluate the performance by counting the number of times the Clayton copula is selected. The results show that we successfully identify the true copula family, 91% (n=200) and 99% (n=500) of the times under the linear calibration model; and 97% (n=200) and 100% (n=500) of the times under the quadratic calibration model.

4. Data Example: Gestational age-specific birth weight dependence in twins

We now apply our proposed method to a subset of the Matched Multiple Birth Data Set. The data containing all twin births in the United States from 1995 to 2000 enables detailed investigation of twin gestations. In our application, we consider the twin live births in which both babies survived their first year of life with mothers of age between 18 and 40. Of interest is the dependence between the birth weights of twins (in grams), denoted by BW_1 and BW_2, respectively. The gestational age, GA, is an important factor for prenatal growth and is therefore chosen as the covariate. We consider a random sample of 30 twin live births for each gestational age (in weeks) between 28 to 42.

The scatterplot and histograms of the birth weights, for $n = 450$ twin pairs, are given in Figure 3(a), from which the marginals are seen to be fitted well by parametric cubic regression models with normal noise, shown in Figure 3(c) and 3(d). The resulting coefficients, all significant at 1% level, are plugged into $U_j = \Phi((\mathrm{BW}_j - \hat{\mu}_j(\mathrm{GA}))/\hat{\sigma}_j)$ to transform the response variables to uniform scale, where $\hat{\mu}_j$ are the cubic fit and $\hat{\sigma}_j$ are the estimated standard errors and Φ is the c.d.f. of $N(0,1)$. Figure 3(b) gives the scatterplot and histograms of the transformed random variables U_j, $j = 1, 2$, which support the parametric fit of marginal models considered here.

[Figure 3 about here.]

We then estimate the calibration function using the proposed nonparametric model with local linear fit ($p = 1$), under the Clayton, Frank and Gumbel families. For comparison, we also perform the parametric estimation with a constant and a linear form, respectively. The results converted to the Kendall's tau scale are given in Figure 4. In all cases, our approach yields a nonlinear pattern, with highest strength of dependence in pre-term and post-term twins which can not be detected by the parametric linear fit.

[Figure 4 about here.]

In the local linear estimation, the optimum bandwidths are chosen as 5.52, 4.04, and 5.52, for the Clayton, Frank and Gumbel copulas, respectively. We constructed approximate 90% confidence intervals under each family using half of these bandwidth values. As seen in Figure 4, the parametric estimates are not within the confidence bands, suggesting that the linear model may not be adequate. Our copula selection method chooses the Frank family, having the minimum cross–validated prediction error (4) with value 47.05, followed by the Clayton copula with 47.36 and the Gumbel copula with 47.40.

5. Discussion

We propose a conditional copula approach to model the strength and type of dependence between two responses. In order to capture the relationship between the strength of the dependence and a measured covariate, we allow the copula parameter to change according to a *calibration function* of the covariate. Statistical inference for the calibration function is obtained using local polynomial estimation. The methodology proposed here leads to a novel conditional copula selection procedure, in which we use prediction accuracy to select, via cross validation, among a number of copula families, the one that best approximates the data at hand.

Our simulation study conveys that i) the nonparametric estimator of the calibration function is flexible enough to capture non-linear patterns and ii) the copula selection procedure performs well in the cases studied so far. The data analysis reveals a gestational age specific dependence pattern in the birth weights of twins that, to our knowledge, has not been detected before and may be of scientific interest.

Although in this paper we focus our attention on bivariate copulas, it is possible to extend our method to more general multivariate copulas. In addition, if more covariates are of potential interest for the conditional copula model, then we recommend a careful selection

of the variables prior to estimation of the calibration function, as the computational cost increases significantly with each covariate added to the model. We are currently working on extending the framework of the work presented here by considering mixtures of conditional copulas in order to increase the spectrum of applications that can be tackled using our approach.

Supplementary Materials

A Web Appendix containing technical details (referenced in Sections 2.1 and 2.3), additional simulated and real data examples is available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors would like to thank the co-editor, the associate editor and the two referees for their careful review and constructive comments that have helped significantly improve the quality and presentation of the paper. E.F. Acar, R.V. Craiu and F. Yao were partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Bartram, S., Taylor, S., and Wang, Y. (2007). The euro and european financial market dependence. *Journal of Banking and Finance* **31,** 1461–1481.

Chen, S. X. and Huang, T.-M. (2007). Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics-Revue Canadienne De Statistique* **35,** 265–282.

Chen, X. H. and Fan, Y. Q. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics* **130,** 307–335.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons, Hoboken, NJ.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Craiu, R. V. and Craiu, M. (2008). On the choice of parametric families of copulas. *Advances and Applications in Statistics* **10,** 25–40.

Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000). Which Copula is the right one? Technical report, Groupe de Recherche Opèrationelle, Crédit Lyonnais.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In *RISK Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66. Chapman & Hall, London, 1st ed edition.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* **27,** 715–731.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis* **95,** 119–152.

Fermanian, J.-D. and Scaillet, O. (2003). Nonparametric estimation of copulas for time series. *Journal of Risk* **5,** 25–54.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2,** 1–25.

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

Genest, C., Rémillard, B., and Beaudoin, D. (2007). Goodness-of-fit tests for copulas: A review and a power study. *Insurance Mathematics and Economics* .

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate

Archimedean copulas. *Journal of the American Statistical Association* **88,** 1034–1043.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer-Verlag, New York.

Hu, H. L. (1998). *Large sample theory of pseudo-maximum likelihood estimates in semiparametric models*. PhD thesis, University of Washington.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Jondeau, E. and Rockinger, M. (2006). The copula-garch model of conditional dependencies: An international stock market application. *Journal of International Money and Finance* **25,** 827–853.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer series in statistics. Springer, New York, 2nd ed edition.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review* **47,** 527–556.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51,** 1384–1399.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* **8,** 229–231.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82,** 559–567.

Trivedi, P. K. and Zimmer, D. M. (2007). *Copula Modeling: An Introduction for Practitioners*. Now Publisher Inc, Hanover, MA.

Wang, W. and Wells, M. T. (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* **95,** 62–76.

Appendix

Regularity assumptions

Let $N(x)$ denote the neighborhood of an interior point x.

(A1) $\ell'(\theta(x), U_1, U_2)$ and $\ell''(\theta(x), U_1, U_2)$ exist and are continuous on $N(x) \times (0,1)^2$, and can be bounded by integrable functions of u_1 and u_2 in $N(x)$.

(A2) The functions $f_X(\cdot)$, $\eta^{(p+2)}$, $g''(\cdot)$ and $\sigma^2(\cdot)$ are continuous in $N(x)$, and $\sigma^2(x') \geqslant c$ for $x' \in N(x)$ and some $c > 0$. Without loss of generality, the kernel density $K(\cdot)$ has a compact support $[-1, 1]$.

Copula families used in our implementations

The Clayton family has the copula function

$$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}, \quad \theta \in (0, \infty), \quad \text{with Kendall's } \tau = \frac{\theta}{\theta + 2}.$$

The Frank family has the copula function

$$C(u_1, u_2) = -\frac{1}{\theta} \ln \Big\{ 1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \Big\}, \quad \theta \in (-\infty, \infty) \setminus \{0\},$$

with Kendall's $\tau = 1 + \frac{4}{\theta}[D_1(\theta) - 1]$, where $D_1(\theta) = \frac{1}{\theta} \int_0 \theta \frac{t}{e^t - 1} dt$ is the Debye function.

The Gumbel family has the copula function

$$C(u_1, u_2) = \exp \Big\{ -[(-\ln u_1)^\theta (-\ln u_2)^\theta]^{\frac{1}{\theta}} \Big\}, \quad \theta \in [1, \infty), \quad \text{with Kendall's } \tau = 1 - \frac{1}{\theta}.$$

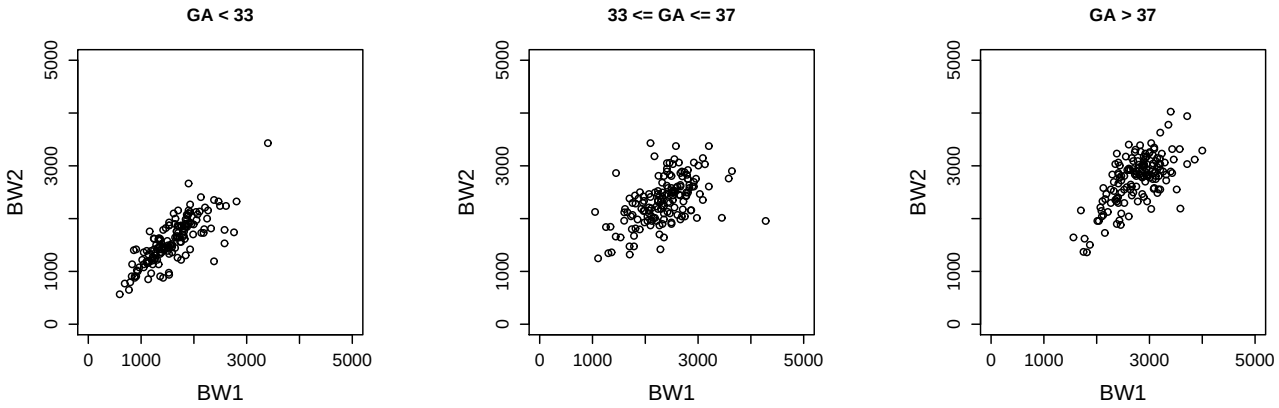

Figure 1. Scatterplots of the twin birth weights at different gestational ages.

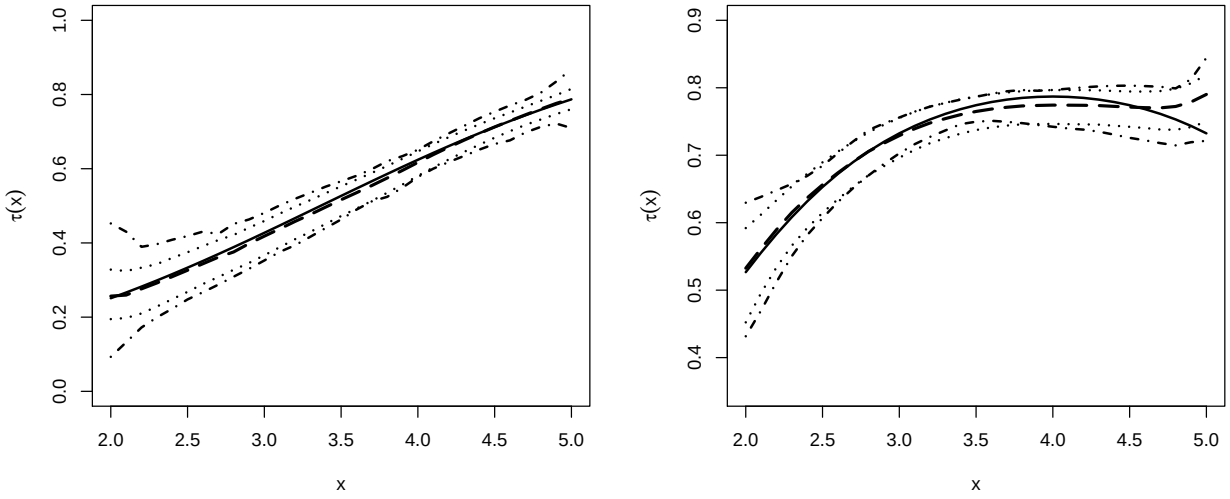

Figure 2. Twin data: 90% confidence intervals for the Kendall's tau under the Clayton copula with the linear (left panel) and quadratic (right panel) calibration models: truth (solid line), averaged local linear estimates (dashed line), approximate confidence intervals (dotted line), and Monte Carlo confidence intervals (dotdashed line).

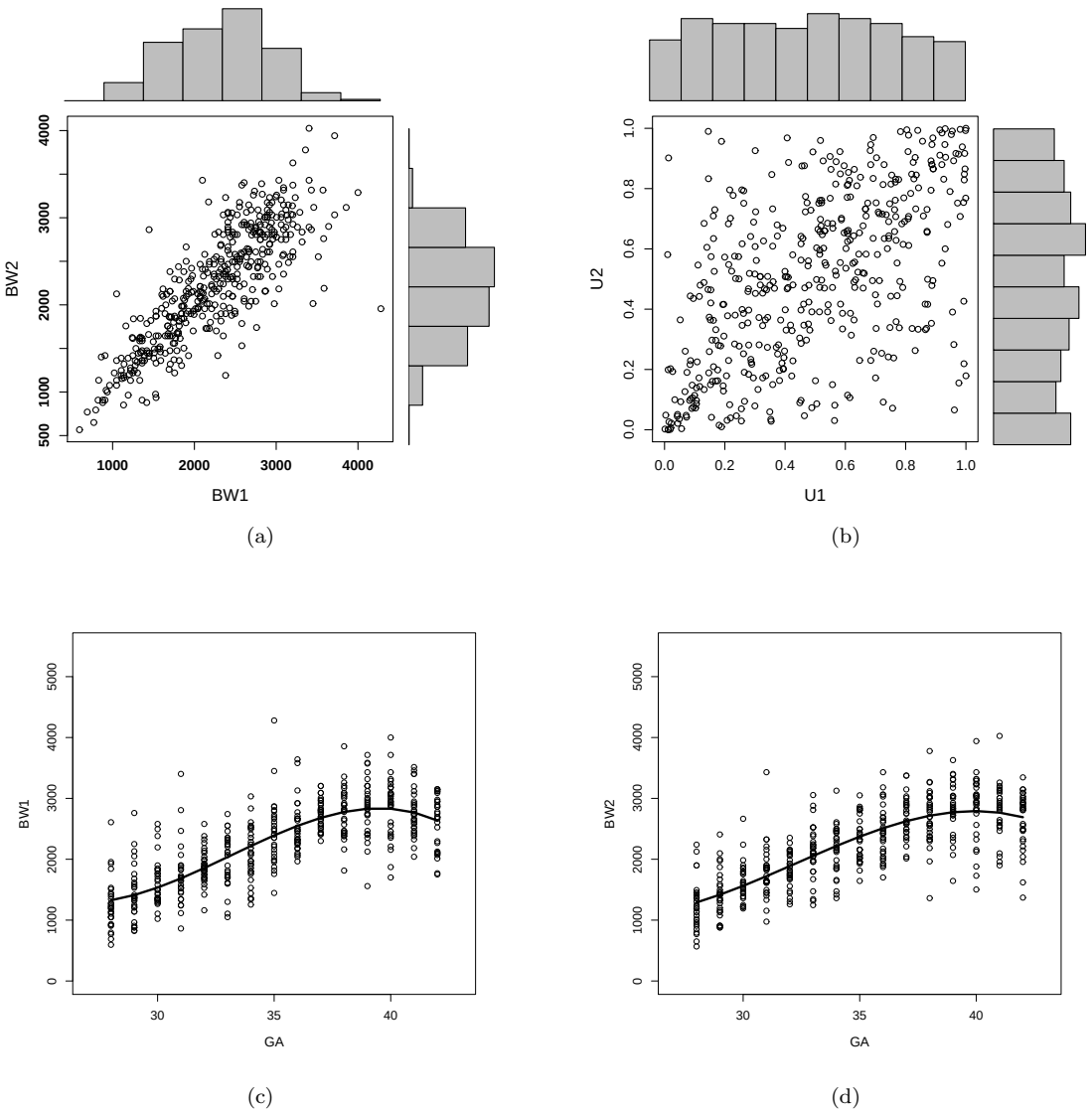

Figure 3. Histograms and scatterplots for: (a) the twin birth weights, (b) the marginal distributions of the transformed variables (U_1, U_2). Scatterplots with the parametric cubic regression fits for: (c) gestational age and BW_1, (d) gestational age and BW_2.

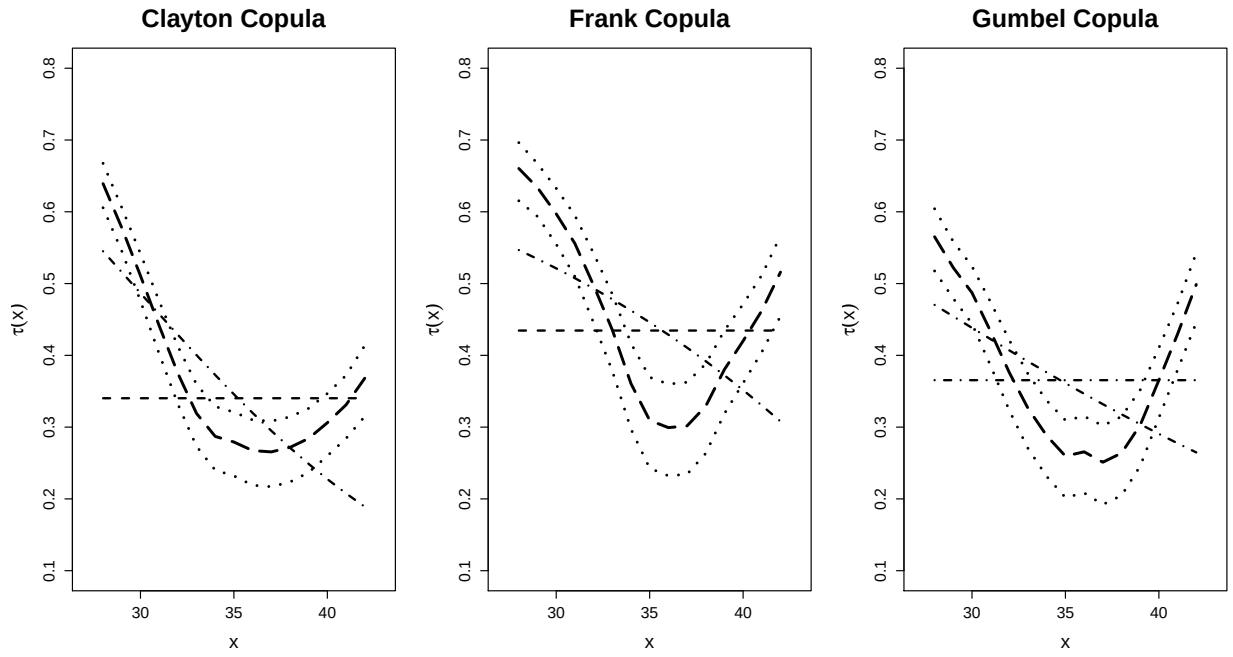

Figure 4. Twin data: the Kendall's tau estimates under three copula families: global constant estimate (dashed line), global linear estimates (dotdashed line), local linear estimates at the optimum bandwidth (longdashed line), and an approximate 90% confidence interval (dotted lines).

Table 1

Integrated Squared Bias, Integrated Variance and Integrated Mean Square Error (multiplied by 100) of the Kendall's tau estimator. The last column shows the averages of the bandwidths h^* selected using (3).

		Linear Calibration Model						
		Parametric estimation			Local estimation			
	n	IBIAS^2	IVAR	IMSE	IBIAS^2	IVAR	IMSE	h^*
Clayton	200	0.024	0.305	0.329	0.017	0.553	0.570	2.183
	500	0.018	0.136	0.154	0.020	0.228	0.248	2.283
Frank	200	0.490	0.596	1.086	0.044	0.963	1.007	1.848
	500	0.479	0.265	0.744	0.060	0.437	0.497	1.644
Gumbel	200	3.739	1.115	4.660	3.704	1.716	5.389	2.095
	500	3.499	0.429	3.928	3.510	0.556	4.066	2.385
		Quadratic Calibration Model						
		Parametric estimation			Local estimation			
	n	IBIAS^2	IVAR	IMSE	IBIAS^2	IVAR	IMSE	h^*
Clayton	200	0.414	0.129	0.543	0.040	0.288	0.328	1.392
	500	0.423	0.046	0.469	0.027	0.113	0.140	0.910
Frank	200	0.324	0.276	0.600	0.123	0.504	0.627	1.779
	500	0.357	0.090	0.447	0.114	0.188	0.302	1.238
Gumbel	200	4.914	0.696	5.610	4.808	1.301	6.109	1.977
	500	4.862	0.246	5.108	4.761	0.497	5.258	1.676