CONDITIONAL COPULAS, ASSOCIATION MEASURES AND THEIR APPLICATIONS

IRÈNE GIJBELS[1], NOËL VERAVERBEKE[2] AND MAREK OMELKA[3]

[1] Department of Mathematics and Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven, Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium;
[2] Center for Statistics, Hasselt University, Agoralaan -building D, B-3590 Diepenbeek, Belgium;
[3] Jaroslav Hájek Center for Theoretical and Applied Statistics, Charles University Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic.

July 7, 2009

Abstract. Of major interest in statistics is the study of dependencies between variables. One way to model a dependence structure is through the copula function which is a mean to capture the dependence structure in the joint distribution of the variables. Association measures such as Kendall's tau or Spearman's rho can be expressed as functionals of the copula. The dependence structure between two variables can be highly influenced by a covariate, and it is of real interest to know how this dependence structure changes with the value taken by the covariate. This motivates the need for introducing conditional copulas, and the associated conditional Kendall's tau and Spearman's rho association measures. After the introduction and motivation of these concepts in this paper we propose two nonparametric estimators for a conditional copula and discuss them. We then derive nonparametric estimates for the conditional association measures. A key issue is that these measures are now looked at as functions in the covariate. We investigate the performances of all estimators via a simulation study which also includes a data-driven algorithm for choosing the smoothing parameters. The usefulness of the methods is illustrated on two real data examples.

Keywords and phrases: Asymptotic bias; asymptotic variance; conditional copula; conditional Kendall's tau; conditional Spearman's rho; empirical estimation; global and local bandwidths; local dependencies; smoothing.

1. Introduction

Suppose we observe a three-dimensional vector $(Y_1, Y_2, X)^\mathsf{T}$ and our main interest is in the relationship of $(Y_1, Y_2)^\mathsf{T}$. If one ignores the variable X (called the covariate in the sequel),

*This work was supported by the IAP Research Network P6/03 of the Belgian State (Belgian Science Policy). This work was started while Marek Omelka was a postdoctoral researcher at the Katholieke Universiteit Leuven and the Universiteit Hasselt within the IAP Research Network. The support of Project LC06024 is also highly appreciated. The first author gratefully acknowledges support from the GOA/07/04-project of the Research Fund KULeuven.

then it is quite common to characterize the degree of dependence of $(Y_1, Y_2)^{\mathsf{T}}$ by just one number, usually the Pearson correlation coefficient. If one prefers nonparametric measures, one uses for example Kendall's tau or Spearman's rank correlation coefficient. On the other hand if one wants to capture the whole dependence structure of $(Y_1, Y_2)^{\mathsf{T}}$, one uses a copula function.

But often the variable X is a confounding factor and one has to incorporate it into the analysis, otherwise the true relationship of $(Y_1, Y_2)^{\mathsf{T}}$ is distorted. To adjust for the influence of the variable X, the most straightforward way is to use a partial correlation coefficient (either Pearson's or a rank based one) of $(Y_1, Y_2)^{\mathsf{T}}$ given X. But this adjustment may not answer all scientific questions. For instance it seems to be natural to ask whether the relationship of $(Y_1, Y_2)^{\mathsf{T}}$ is the same for 'small' as well as 'large' values of X.

Let us illustrate this with an example. Suppose we have data on life expectancies at birth ('average lengths of lives') at different countries and the interest is in the relationship of the life expectancies of males (Y_1) and females (Y_2). Then a natural question is whether this relationship is different in poor and rich countries. Let us take e.g. gross domestic product (GDP) per capita (X) as a proxy for the economic welfare of a country. Then, mathematically speaking, the question is about the relationship of $(Y_1, Y_2)^{\mathsf{T}}$ conditionally upon the given value of the covariate $X = x$ and whether this relationship changes with the values of x. As will be seen later the dependence structure of $(Y_1, Y_2)^{\mathsf{T}}$ given $X = x$ is fully described by a function which we will call a conditional copula. In the following we are interested in estimating that function.

Denote the joint and marginal distribution functions of $(Y_1, Y_2)^{\mathsf{T}}$, conditionally upon $X = x$, as

$$H_x(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2 \,|\, X = x),$$
$$F_{1x}(y_1) = P(Y_1 \leq y_1 \,|\, X = x), \quad F_{2x}(y_2) = P(Y_2 \leq y_2 \,|\, X = x).$$

If F_{1x} and F_{2x} are continuous, then according to Sklar's theorem (see e.g. Nelsen (2006)) there exists a unique copula C_x such that

$$H_x(y_1, y_2) = C_x(F_{1x}(y_1), F_{2x}(y_2)). \tag{1}$$

From equation (1) we see that the conditional copula C_x fully describes the conditional dependence structure of $(Y_1, Y_2)^{\mathsf{T}}$ given $X = x$ and it depends in a general way on the covariate value x.

To the best of our knowledge, the area of conditional copula estimation is up to this moment almost completely unexplored. Our research extends the work on conditional distribution estimation (see. e.g. Stute (1986), Yu and Jones (1998) and Hall et al. (1999)). Moreover, as conditional copulas can be used to construct conditional measures of dependence (e.g. conditional Kendall's tau), our work also complements the methodology of partial rank correlation coefficients introduced in Kendall (1942).

The paper is organised as follows. In Section 2 we suggest two nonparametric estimators of C_x, which will be used to analyze two real data sets in Section 3. The suggested estimators will be further investigated and compared in a simulation study in Section 4.

2. Estimating the conditional copula

To estimate the conditional copula C_x it is convenient to invert Sklar's theorem in (1) which enables to express C_x as

$$C_x(u_1, u_2) = H_x(F_{1x}^{-1}(u_1), F_{2x}^{-1}(u_2)), \quad (u_1, u_2) \in [0,1]^2, \tag{2}$$

where $F_{1x}^{-1}(u) = \inf\{y : F_{1x}(y) \geq u\}$ is the conditional quantile function of Y_1 given $X = x$ and F_{2x}^{-1} is the conditional quantile function of Y_2 given $X = x$.

Now suppose that we observe independent identically distributed three-dimensional vectors $(Y_{11}, Y_{21}, X_1)^\mathsf{T}, \ldots, (Y_{1n}, Y_{2n}, X_n)^\mathsf{T}$ from the cumulative distribution function $H(y_1, y_2, x)$. Based on the sample of observations we have the following empirical estimator for $H_x(y_1, y_2)$:

$$H_{xh}(y_1, y_2) = \sum_{i=1}^{n} w_{ni}(x, h_n)\, \mathbb{I}\{Y_{1i} \leq y_1, Y_{2i} \leq y_2\}, \tag{3}$$

where $\{w_{ni}(x, h_n)\}$ is a sequence of weights that smooth over the covariate space (see Section 2.2) and $h_n > 0$ is a bandwidth going to zero as the sample size increases. Here $\mathbb{I}\{A\}$ denotes the indicator of an event A. In view of (2) a straightforward estimator of the copula function $C_x(u_1, u_2)$ $(0 \leq u_1, u_2 \leq 1)$ is given by

$$\begin{aligned} C_{xh}(u_1, u_2) &= H_{xh}\left(F_{1xh}^{-1}(u_1), F_{2xh}^{-1}(u_2)\right) \\ &= \sum_{i=1}^{n} w_{ni}(x, h_n)\, \mathbb{I}\{Y_{1i} \leq F_{1xh}^{-1}(u_1), Y_{2i} \leq F_{2xh}^{-1}(u_2)\}, \end{aligned} \tag{4}$$

where F_{1xh} and F_{2xh} are corresponding marginal distribution functions of H_{xh}.

Although the copula estimator C_{xh} given by (4) seems very natural, since it mimics the structure of the true copula C_x given in (2), a closer inspection of the estimator points to some potential pitfalls of it. For instance suppose that Y_1 and Y_2 are conditionally independent given $X = z$, but that their conditional distributions are stochastically increasing with z. Then, intuitively speaking, larger values of Y_1 will occur together with larger values of Y_2 purely because of the same trend in the covariate z creating an artificial dependence.

This intuition was also confirmed by Monte Carlo experiments in which we observed that the estimator C_{xh} may be severely biased, if any of the conditional marginal distributions changes with the value of the covariate $X = x$. We also observed that this bias can be reduced to a great extent, if we are able to remove the effect of the covariates on the marginals. Further recall that the copula function is invariant to increasing transformations. Thus if one knew F_{1X}, F_{2X} it would be advisable to base the estimator C_{xh} on the observations $\{(U_{1i}, U_{2i})^{\mathsf{T}}, i = 1, \ldots, n\}$ where

$$(U_{1i}, U_{2i})^{\mathsf{T}} = (F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}))^{\mathsf{T}}, \tag{5}$$

whose marginal distributions are uniform (for each $i = 1, \ldots, n$).

Unfortunately, we usually do not know the theoretical conditional marginal distribution functions (F_{1X_i}, F_{2X_i}), but we can estimate them in the same way as we estimate F_{1x} and F_{2x}, that is

$$\begin{aligned} F_{1X_i g_1}(y) &= \sum_{j=1}^{n} w_{nj}(X_i, g_{1n})\, \mathbb{I}\{Y_{1j} \le y\}, \\ F_{2X_i g_2}(y) &= \sum_{j=1}^{n} w_{nj}(X_i, g_{2n})\, \mathbb{I}\{Y_{2j} \le y\}, \end{aligned}$$

where $g_1 = \{g_{1n}\} \searrow 0$ and $g_2 = \{g_{2n}\} \searrow 0$.

This leads to the following procedure. First, transform the original observations to reduce the effect of the covariate by

$$(\tilde{U}_{1i}, \tilde{U}_{2i})^{\mathsf{T}} = (F_{1X_i g_1}(Y_{1i}), F_{2X_i g_2}(Y_{2i}))^{\mathsf{T}}, \quad i = 1, \ldots, n. \tag{6}$$

Second, use the transformed observations $(\tilde{U}_{1i}, \tilde{U}_{2i})^{\mathsf{T}}$ in a similar way as the original observations, and construct

$$\tilde{C}_{xh}(u_1, u_2) = \tilde{G}_{xh}\left(\tilde{G}_{1xh}^{-1}(u_1), \tilde{G}_{2xh}^{-1}(u_2)\right), \tag{7}$$

where

$$\tilde{G}_{xh}(u_1, u_2) = \sum_{i=1}^{n} w_{ni}(x, h_n)\, \mathbb{I}\{\tilde{U}_{1i} \le u_1, \tilde{U}_{2i} \le u_2\},$$

and $\tilde{G}_{1xh}$ and $\tilde{G}_{2xh}$ are its corresponding marginals.

The asymptotic properties of the estimators C_{xh} and $\tilde{C}_{xh}$ are studied in Veraverbeke et al. (2009). The main result states that provided the bandwidths h_n, g_{n1}, g_{n2} satisfy (for $j = 1, 2$)

$$h_n = O(n^{-1/5}), \quad \sqrt{n\, h_n}\, g_{jn}^2 = O(1), \quad \tfrac{h_n}{g_{jn}} = O(1), \quad n \min(h_n, g_{1n}, g_{2n}) \to \infty, \tag{8}$$

and some other regularity conditions hold, then there is no price in terms of asymptotic bias or variance that we pay for substituting the unknown $(U_{1i}, U_{2i})^{\mathsf{T}}$ with the estimates $(\tilde{U}_{1i}, \tilde{U}_{2i})^{\mathsf{T}}$. Moreover, comparing the estimators C_{xh} and $\tilde{C}_{xh}$ we see that (for the same bandwidth h_n) both estimators have the same asymptotic variance. But the asymptotic bias of the estimator $\tilde{C}_{xh}$ consists only of those terms of the asymptotic bias of C_{xh} that do not include the partial derivatives of the conditional marginal distribution functions F_{1x} and F_{2x} with respect to x.

Remark 1. The aim of the transformation (6) is to remove the effect of the covariate X on the marginal distributions. For this reason we use the nonparametric estimators of the conditional distribution functions. Of course, if we can assume a parametric model for the influence of the covariate on the marginals, then it is advisable to use this model. Although it does not change asymptotic properties of the estimator, it may stabilize the finite sample properties. For example, in many practical situations it may be simply sufficient to replace the original observations $(Y_{1i}, Y_{2i})^{\mathsf{T}}$ with the estimated residuals from simple linear regressions, where Y_1 and respectively Y_2 are regressed on the covariate X.

2.1. **Conditional measures of association.** In many situations we would like to quantify the degree of dependence by only one number. In nonparametric settings Kendall's tau and Spearman's rho are probably the most widely used. In the following we use the conditional copula methodology to express and estimate conditional versions of those measures of dependence.

2.1.1. *Kendall's tau.* For random variables $(Y_1, Y_2)^{\mathsf{T}}$ Kendall's tau is defined as

$$\tau = 2\,P\big((Y_1 - Y_1')(Y_2 - Y_2') > 0\big) - 1,$$

where $(Y_1', Y_2')^{\mathsf{T}}$ is an independent copy of the random vector $(Y_1, Y_2)^{\mathsf{T}}$. It is well known (see e.g. Nelsen (2006)) that if C is the copula for the vector $(Y_1, Y_2)^{\mathsf{T}}$, then τ may be expressed as

$$\tau = 4 \iint C(u_1, u_2)\, dC(u_1, u_2) - 1.$$

This leads immediately to an expression for the population version of the conditional Kendall's tau of $(Y_1, Y_2)^{\mathsf{T}}$ given $X = x$

$$\tau(x) = 4 \iint C_x(u_1, u_2)\, dC_x(u_1, u_2) - 1, \tag{9}$$

where C_x is the appropriate conditional copula. The interpretation of the conditional Kendall's tau is

$$\tau(x) = 2\,P\big((Y_1 - Y_1')(Y_2 - Y_2') > 0 \,|\, X = X' = x\big) - 1,$$

where $(Y_1', Y_2', X')^{\mathsf{T}}$ is an independent copy of the random vector $(Y_1, Y_2, X)^{\mathsf{T}}$.

The most straightforward way to estimate the conditional Kendall's tau is to replace the unknown quantity C_x in (9) with the estimate C_{xh} to get

$$\hat{\tau}_n^I(x) = 4 \iint C_{xh}(u_1, u_2)\, dC_{xh}(u_1, u_2) - 1. \tag{10}$$

Although expression (10) is convenient for exploring asymptotic properties of the estimator, in finite samples we have a slightly better experience with the formula

$$\hat{\tau}_n(x) = \frac{4}{1 - \sum_{i=1}^n w_{ni}^2(x, h_n)} \sum_{i=1}^n \sum_{j=1}^n w_{ni}(x, h_n)\, w_{nj}(x, h_n)\, \mathbb{I}\{Y_{1i} < Y_{1j}, Y_{2i} < Y_{2j}\} - 1, \tag{11}$$

which mimics the formula for (unconditional) Kendall's tau estimation

$$\hat{\tau}_n = \frac{4}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\{Y_{1i} < Y_{1j}, Y_{2i} < Y_{2j}\} - 1.$$

Further it may be shown that $\hat{\tau}_n(x)$ is asymptotically equivalent to $\hat{\tau}_n^I(x)$ up to order $O_P(\frac{1}{n\,h_n})$ (see Veraverbeke et al. (2009)).

2.1.2. *Spearman's rho.* As the unconditional version of Spearman's rho may be expressed as $\rho = 12 \iint C(u_1, u_2)\, du_1\, du_2 - 3$, the population conditional version is thus given by $\rho(x) = 12 \iint C_x(u_1, u_2)\, du_1\, du_2 - 3$, which may be estimated as

$$\hat{\rho}_n(x) = 12 \iint C_{xh}(u_1, u_2)\, du_1\, du_2 - 3 = 12 \sum_{i=1}^{n} w_{ni}(x, h_n)(1 - \hat{U}_{1i})(1 - \hat{U}_{2i}) - 3.$$

For interpretations of Spearman's rho see Nelsen (2006).

2.2. **Some common choices of weights.** For the weights many common choices are provided such as these listed below (where X_i may be taken fixed or random). Assuming that the support of X is a bounded interval (without loss of generality we take it to be $[0, 1]$), let $X_{1:n} \le \ldots \le X_{n:n}$ be the ordered sample of $X_1, \ldots, X_n$, and put $X_{0:n} = 0$ and $X_{n+1:n} = 1$. With slight abuse of notation R_i will denote the rank of X_i among $X_1, \ldots, X_n$.

- Nadaraya-Watson (see Nadaraya (1964) or Watson (1964))

$$w_{ni}(x, h_n) = \frac{K(\frac{X_i - x}{h_n})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h_n})}.$$

- Local linear [LL] (see e.g p. 20 of Fan and Gijbels (1996))

$$w_{ni}(x, h_n) = \frac{\frac{1}{n\,h_n} K(\frac{X_i - x}{h_n}) \Big(S_{n,2} - \frac{X_i - x}{h_n} S_{n,1}\Big)}{S_{n,0}\, S_{n,2} - S_{n,1}^2},$$

where

$$S_{n,j} = \frac{1}{n\,h_n} \sum_{i=1}^{n} \left(\frac{X_i - x}{h_n}\right)^j K\left(\frac{X_i - x}{h_n}\right), \qquad j = 0, 1, 2.$$

- Priestley-Chao (see Priestley and Chao (1972))

$$w_{ni}(x, h_n) = \frac{X_{R_i:n} - X_{R_i-1:n}}{h_n} K\left(\frac{X_{R_i-1:n} - x}{h_n}\right).$$

- Gasser-Müller (see Gasser and Müller (1979))

$$w_{ni}(x, h_n) = \frac{1}{h_n} \int_{S_i}^{T_i} K(\tfrac{z - x}{h_n})\, dz,$$

where $T_i = (1 - \beta)\, X_{R_i:n} + \beta\, X_{R_i+1:n}$, $S_i = (1 - \beta)\, X_{R_i-1:n} + \beta\, X_{R_i:n}$, and $\beta \in [0, 1]$.

- h_n-nearest-neighbourhood (see Yang (1981))

$$w_{ni}(x,h_n) = \tfrac{1}{n\,h_n} K\left(\tfrac{F_n(X_i)-F_n(x)}{h_n}\right), \qquad \text{where} \quad F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq z\}.$$

2.3. **Bandwidth selection.** A crucial point of smoothing methods is the bandwidth selection. The proposed estimator $\tilde{C}_{xh}$ requires to choose three bandwidths – g_{1n}, g_{2n} and h_n. To the best of our knowledge the problem of bandwidth choice in our context has not been investigated yet. In this paper we adopted the idea of Gasser et al. (1991), which was further extended in Brockmann et al. (1993).

The main idea of the bandwidth selection rule may be summarized as follows. From the results of Veraverbeke et al. (2009) we can deduce that the asymptotic mean squared errors of the estimators C_{xh} and $\tilde{C}_{xh}$ are given by

$$\text{AMSE}(C_{xh}(u_1,u_2)) = \frac{V_x(u_1,u_2)}{n\,h_n} + h_n^4\, b_x^2(u_1,u_2), \tag{12}$$

$$\text{AMSE}(\tilde{C}_{xh}(u_1,u_2)) = \frac{V_x(u_1,u_2)}{n\,h_n} + h_n^4\, \tilde{b}_x^2(u_1,u_2), \tag{13}$$

where V_x is an asymptotic variance function (common for both C_{xh} and $\tilde{C}_{xh}$) and b_x, $\tilde{b}_x$ are asymptotic bias functions. Provided we know these functions, we can theoretically compute a bandwidth that minimizes the asymptotic mean squared error of the estimator C_{xh} ($\tilde{C}_{xh}$) for a given (x, u_1, u_2). Let us denote h_B and h_V pilot bandwidths that are used to estimate the functions $b_x(u_1,u_2)$ ($\tilde{b}_x(u_1,u_2)$) and $V_x(u_1,u_2)$ respectively.

The algorithm for selecting the bandwidth may be summarized as follows (details are available from the authors upon request).

0. Let h_V be an initial value for the bandwidth;
1. Put $h_B = 2\,\hat{\sigma}\,h_V\,n^{1/10}$, where $\hat{\sigma}$ stands for the interquantile range of the observad values of the covariate X;
2. Using h_B estimate the function $b_x(u_1,u_2)$ ($\tilde{b}_x(u_1,u_2)$) and using h_V estimate $V_x(u_1,u_2)$;
3. Find h^* that minimizes the estimated asymptotic mean squared error given by (12) (or (13)) with respect to h_n;
4. Put $h_V = h^*$ and go to 1, unless a convergence or a maximum number of iteration steps is reached. Otherwise, go to 5.
5. Return the current value h_V as the chosen bandwidth.

The above general procedure describes how to obtain a local bandwidth for a given (u_1,u_2) at a given value of the covariate $X = x$. If one is interested in estimating the whole copula function, it makes sense to integrate the expression (12) (or (13)) with respect to (u_1,u_2) and then the suggested procedure gives a bandwidth that is minimizing an estimated asymptotic mean integrated squared error. Further, a global in x bandwidth is obtained by integrating the expression given in (12) (or (13)) over the covariate space.

The algorithm described above can be used directly to find a bandwidth for the estimator C_{xh}. For the suggested estimator $\tilde{C}_{xh}$ we need to run the algorithm three times. First, we use its univariate adaptation on the problem of estimating F_{1x} and F_{2x} to find g_1 and g_2. Then with the help of g_1 and g_2 and equation (6) we calculate $(\tilde{U}_{1i}, \tilde{U}_{2i})$ that are subsequently used in finding the bandwidth h for the copula estimation. In the sequel we refer to this method as *the plug-in bandwidth choice.*

Remark 2. It is quite common that the main effect of the covariate X is on the mean functions of the conditional distributions. In that case we can try to find the appropriate g_1 and g_2 by employing bandwidth selection rules suggested for nonparametric regression. Similarly as in Yu and Jones (1998) we can argue that it seems reasonable to multiply the bandwidth suggested for nonparametric regression by two.

3. Real data examples

In the following we use LL weights introduced in Section 2.2 together with the triweight kernel $K(x) = \frac{35}{32}\,(1-x^2)^3\,\mathbb{I}\{|x| \leq 1\}$.

3.1. **Life expectancies at birth.** Recall the example analyzing the relationship of the life expectancies at birth of males (Y_1) and females (Y_2). From the World Factbook of the Central Intelligence Agency (CIA) we retrieved a data set consisting of life expectancies and the gross domestic product (GDP) in USD per capita (X) for 222 countries. Scatterplots of this data set are in Figure 1. We see that life expectancies of males and females are strongly correlated

Figure 1. Life expectancy data

giving (unconditional) Kendall's tau equal to 0.86.

Further in Figure 1 (b) it is observed that the life expectancies seem to be increasing with GDP per capita - using $\log_{10}$ transformation of GDP which is quite common in this context. There are several ways to incorporate the information about GDP into the analysis. For instance we may be interested in the relationship of life expectancies when the effect of GDP is removed. In nonparametric settings this may be answered by Kendall's partial correlation coefficient, suggested in Kendall (1942), which equals 0.78 here.

A different scientific question is whether the strength of the relationship of the life expectancy of males and females is the same for poor and rich countries. We will report results for four different methods of estimation. The estimator computed through (11) (which is tied to the estimator C_{xh}) will be denoted as `tau1`. If we replace $(Y_{1i}, Y_{2i})^{\mathsf{T}}$ with $(\tilde{U}_{1i}, \tilde{U}_{2i})^{\mathsf{T}}$ we get the estimator tied to $\tilde{C}_{xh}$ and we will refer to it as `tau2`. Further as the scatterplot in Figure 1(b) suggests a quadratic relationship of life expectancy to $\log_{10}$(GPD), we try to replace the original observations $(Y_{1i}, Y_{2i})^{\mathsf{T}}$ with residuals coming from fitting a linear model with polynomial of order two of $\log_{10}$(GDP) to life expectancies through least squares regression. The estimators resulting from this adjustment will be called `tau1-lm` and `tau2-lm`.

The estimates of Kendall's tau are plotted for different values of GDP in Figure 2.

Plots (a)–(c) correspond to a fixed bandwidth, while in (d) the (local) plug-in bandwidth rule (see Section 2.3) is used. For the estimators based on $\tilde{C}_{xh}$ we need to specify also the bandwidths g_{1n} and g_{2n}. For simplicity of implementation we used 'lokern' which is a library available for the R computing environment (see R Development Core Team (2008)) and which implements the ideas of bandwidth choice in nonparametric regression as introduced in Gasser et al. (1991) and Brockmann et al. (1993). If the interest is in the conditional Kendall's tau just at a few points, then we may use locally adaptive bandwidths g_{1n} and g_{2n} for each of the points of interest. But if the interest is in the overall curve, we decided to use a global bandwidth for all of the points to avoid the resulting curve to be too wiggly.

Comparing the curves of the estimates several points may be noticed.

- The main message is that the conditional Kendall's tau decreases from about 0.85 (for countries with about $10^3 = 1000$ USD of GDP per capita) to 0.70 (for countries with about $10^{4.5} \doteq 31\,628$ USD of GDP per capita).
- Adjusting for the obvious trend in the covariate makes the estimates less wiggly, which is in particular true for the estimators based on C_{xh}. For the estimators tied to $\tilde{C}_{xh}$ the effect of adjusting is minor and it makes a noticeable difference only when the effective sample size $(n\,h_n)$ is small (and near borders).
- The estimator `tau1` consistently produces bigger estimates of the conditional Kendall's tau for higher values of bandwidths. Comparing `tau1` with `tau1-lm` we see that this is partially corrected by removing the trend of life expectancy when regressed on $\log_{10}$(GPD).

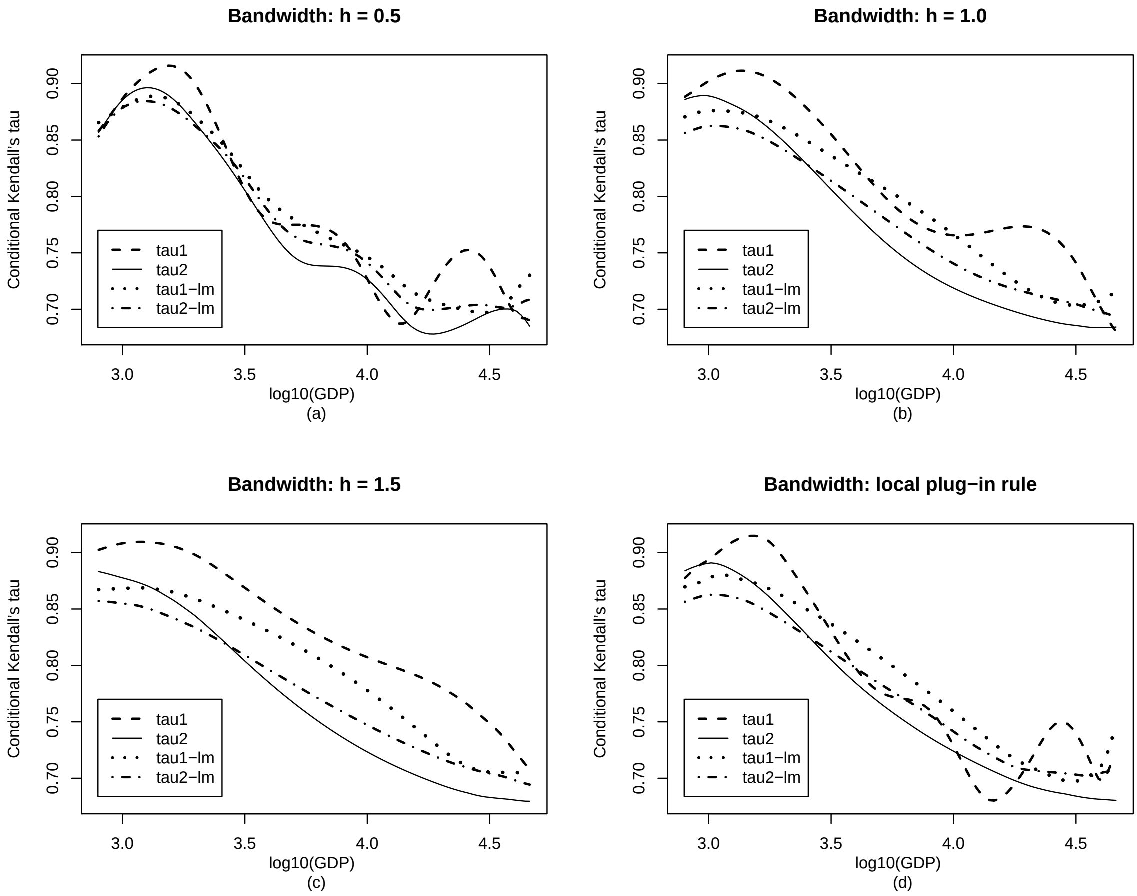

FIGURE 2. Estimated conditional Kendall's tau for life expectancy data.

Note also that unless we know the model generating the data, it is extremely difficult to judge what is a too wiggly curve for given data. In contrast to nonparametric regression (with one variable) we cannot make a scatterplot and try to judge by eye what is a reasonable fit for our data.

3.2. **Soil contamination.** The following data set gives several soil characteristics from 119 locations in the vicinity of a former lead smelter in the city of Příbram (Czech Republic). Industrial activity has contaminated soil with metals like As (arsenic), Cd (cadmium), Pb (lead), Zn (zinc) and others. Researchers were interested to find out the relationship between the amount of metals present in the soil and microbial characteristics of the soil such as biomass, dehydrogenase and soil respiration which could serve as indicators of soil quality. In the following we will concentrate on the amount of `Zn` and the microbial activity `dehydro`.

It is quite natural to expect that the more amount of metal in the soil the lower level of microbial activity. Contrary to that intuition the (unconditional) Kendall's tau is slightly

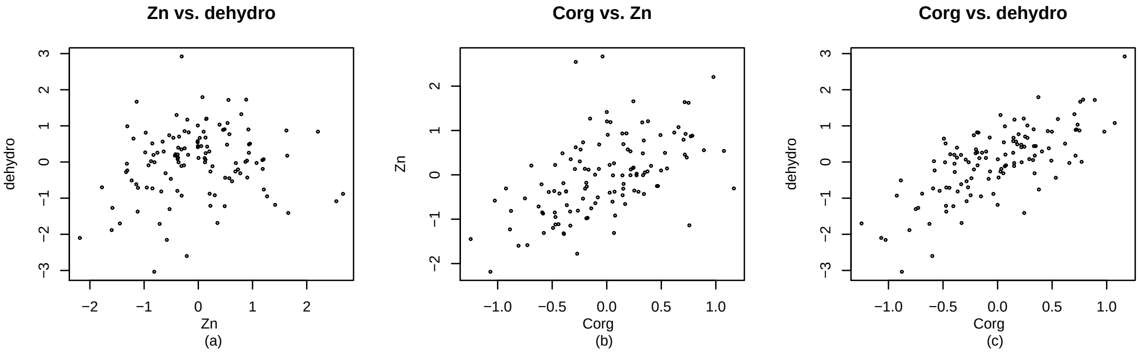

FIGURE 3. Příbram data

above zero (0.09). A partial explanation for this may be deduced from the scatterplots of `Zn`, `dehydro` and the quantity of organic material `Corg` that can be found in Figure 3. It may be surprising to see a strong positive correlation of `Zn` and `Corg`. The researchers explain this by the fact that areas closer to that former factory have not been used for agriculture or any other economical activity. That is why the bigger amount of the organic material `Corg` together with higher contamination are observed. Thus it is sensible to estimate the relationship of `Zn` and `dehydro` for the soils with the same value of `Corg`. Kendall's partial tau of `Zn` and `dehydro` adjusted for `Corg` equals -0.13 and seems to be more in agreement with our intuition.

Another option to incorporate the variable `Corg` in the analysis is to apply the methodology of Section 2. The same estimators as in the previous example are employed. The only difference is that the adjustment for the covariate made before computation of the estimator `tau1-lm` and `tau2-lm` is through a simple linear (and not quadratic) relationship. One again may notice that the estimator `tau1`, which is the only one not trying to remove in any way the effect of the covariate on the marginals, produces rather different results than the other estimators. It seems likely that this estimator overestimates the true conditional Kendall's tau of `Zn` and `dehydro` because of the same trend these variables follow with `Corg`.

Note that the association between `Zn` and `dehydro` seems to be changing with the value of `Corg` and ranges from slightly positive to negative values. This may be a very useful information when `dehydro` is considered as a response variable and the interest is in building a parametric model with the help of covariates `Zn` and `Corg`.

These two examples clearly motivate the interest in studying concepts such as conditional copulas and conditional association measures.

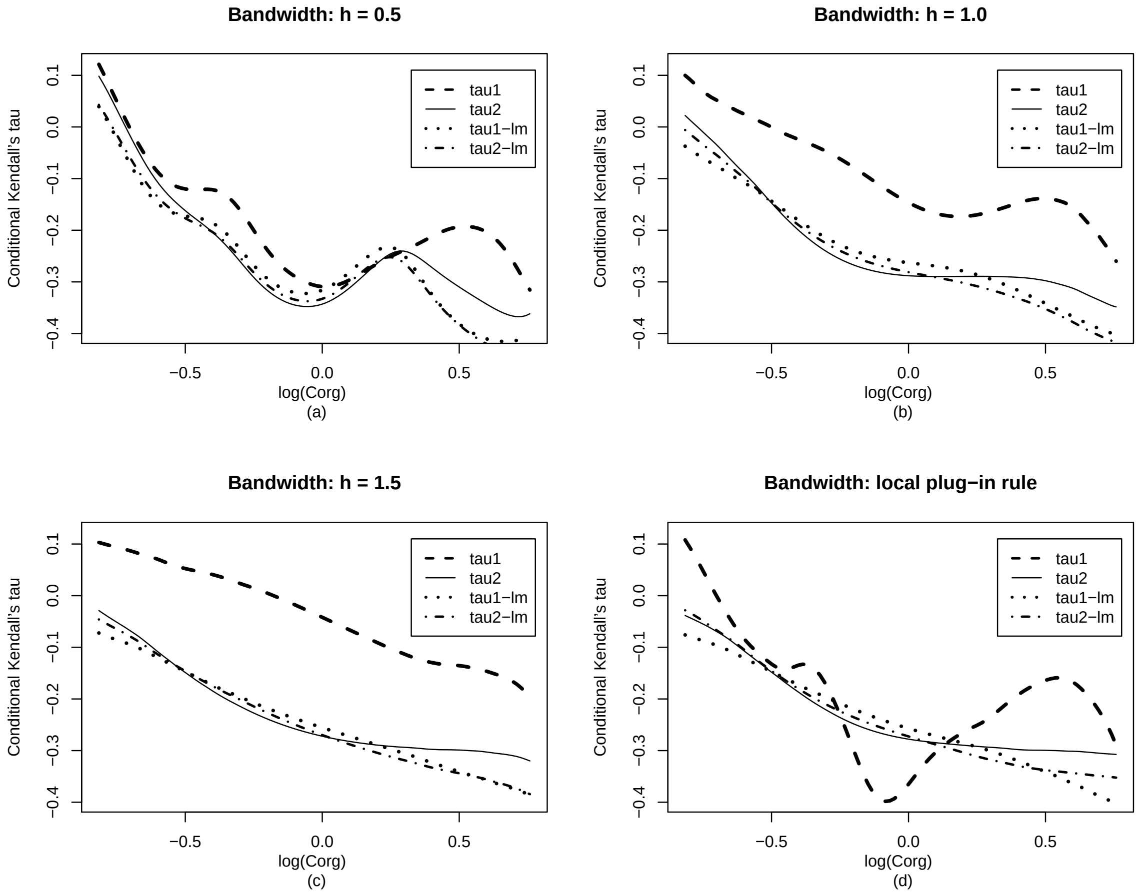

FIGURE 4. Conditional Kendall's tau for Příbram data.

4. Simulation study

To complement the real data examples of the previous section as well as the theoretical comparison of $\tilde{C}_{xh}$ and C_{xh} done in Veraverbeke et al. (2009), we provide here a simulation study to illustrate the finite sample performance of both estimators.

In the following we compare the estimators $\tilde{C}_{xh}$ and C_{xh} in two ways. First, we compare the behaviour of these estimators when the bandwidth h is held fixed and putting $g_1 = g_2 = h$. Second, we compare the performance of the estimators when the plug-in bandwidth selection rule of Section 2.3 is used.

4.1. **Copula estimation.**

In this application we are interested in estimation of a copula as a function on $[0,1]^2$. The performance of the estimators is evaluated using the average (over all simulations) of the integrated squared error

$$\int_0^1 \int_0^1 \left[\hat{C}_{xh}(u_1,u_2) - C_x(u_1,u_2)\right]^2 \, du_1 \, du_2 \, ,$$

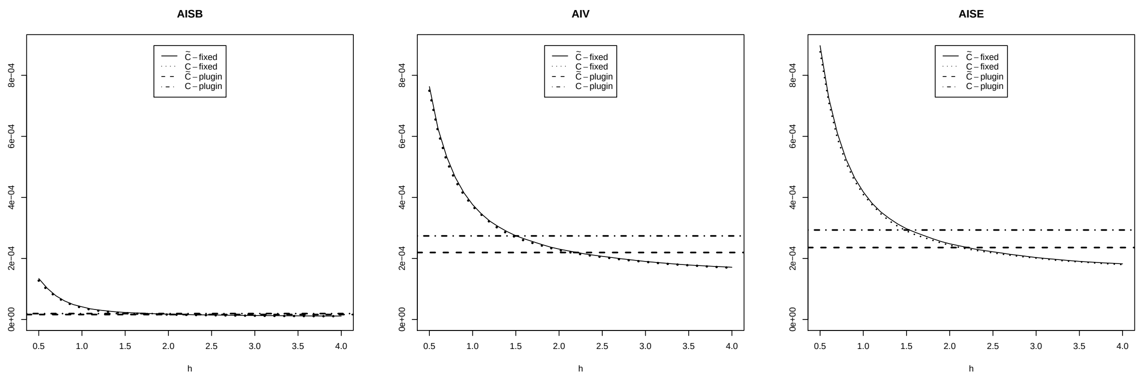

FIGURE 5. Copula estimation; Model 1, $\mu_1(z) = 1$, $\mu_2(z) = 1$, $\rho = 1$.

where $\hat{C}_{xh}$ stands either for C_{xh} or $\tilde{C}_{xh}$.

To illustrate our main findings we report results for the following setup: the covariate is supposed to be standard normal and we are interested in the point $X = 1$. The copula which joins the margins is a Frank copula with the parameter depending on the value of the covariate $X = z$ as $\theta(z) = 5 + \rho \sin(\frac{(z-1)\pi}{6})$. This results into Kendall's tau equal to 0.46 for $z = 1$. The margins are taken normal with unit variances and mean functions $\mu_1(z)$ and $\mu_2(z)$. The considered models are given in Table 1.

Model	mean functions		parameter ρ
1 / 2	$\mu_1(z) = 1$	$\mu_2(z) = 1$	1 / 5
3 / 4	$\mu_1(z) = 1$	$\mu_2(z) = \sin(z-1)$	1 / 5
5 / 6	$\mu_1(z) = \sin(z-1)$	$\mu_2(z) = \sin(z-1)$	1 / 5
7 / 8	$\mu_1(z) = \cos(z-1)$	$\mu_2(z) = \sin(z-1)$	1 / 5

TABLE 1. Simulation models.

Models 1 and 2 represent situations where the covariate does not influence conditional marginal distributions; in Models 3 and 4 only one of the marginals is affected; while in Models 5 and 6 both marginals are stochastically increasing with z; finally in Models 7 and 8 the marginals are affected in different directions. The two values of ρ represent the situations when there is a mild ($\rho = 1$) or strong effect ($\rho = 5$) of the covariate on the conditional dependence structure.

Further, the sample size is $n = 200$ and the number of generated samples is 1 000.

The results are to be found in Figures 5–12, where the average of the integrated squared bias (AISB), the average of the integrated variance (AIV) and the average of the integrated squared error (AISE) are plotted as functions of the bandwidth h. The solid curve shows the result for the estimator $\tilde{C}_{xh}$ (`C̃-fixed`) with $g_1 = g_2 = h$ and the dotted curve for the estimator C_{xh} (`C-fixed`). The dashed and dotdashed horizontal lines represent the values of AISB, AIV and AISE when the plug-in bandwidth choice is used for the estimator $\tilde{C}_{xh}$

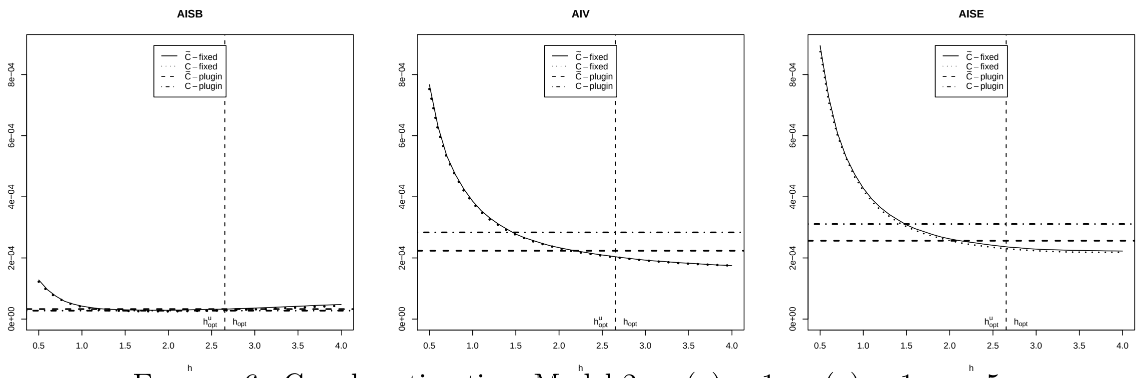

FIGURE 6. Copula estimation; Model 2, $\mu_1(z) = 1$, $\mu_2(z) = 1$, $\rho = 5$.

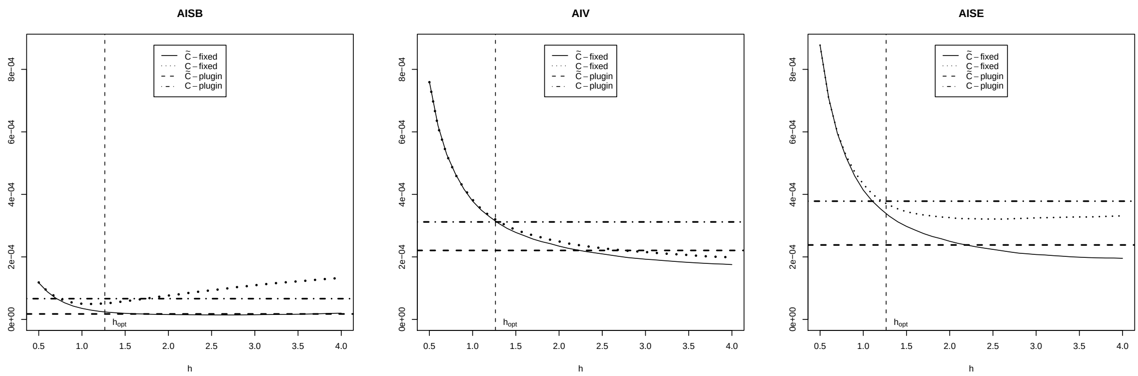

FIGURE 7. Copula estimation; Model 3, $\mu_1(z) = 1$, $\mu_2(z) = \sin(z - 1)$, $\rho = 1$.

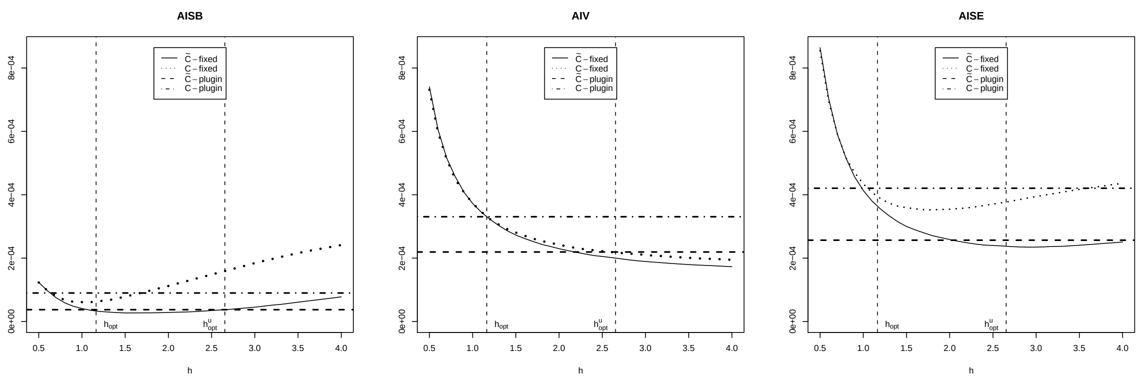

FIGURE 8. Copula estimation; Model 4, $\mu_1(z) = 1$, $\mu_2(z) = \sin(z - 1)$, $\rho = 5$.

(`C̃-plugin`) and C_{xh} (`C-plugin`) respectively. Finally, the vertical dashed lines indicate the asymptotically optimal values of bandwidths: h_{opt} for C_{xh} and $h^{\text{u}}_{\text{opt}}$ for $\tilde{C}_{xh}$.

Models 1 and 2 represent situations when the distributions of the marginals are independent of the covariate. From Figures 1 and 2 we see that the performance of the estimators C_{xh} and $\tilde{C}_{xh}$ is effectively the same for both mild or strong effect of the covariate on the dependence structure.

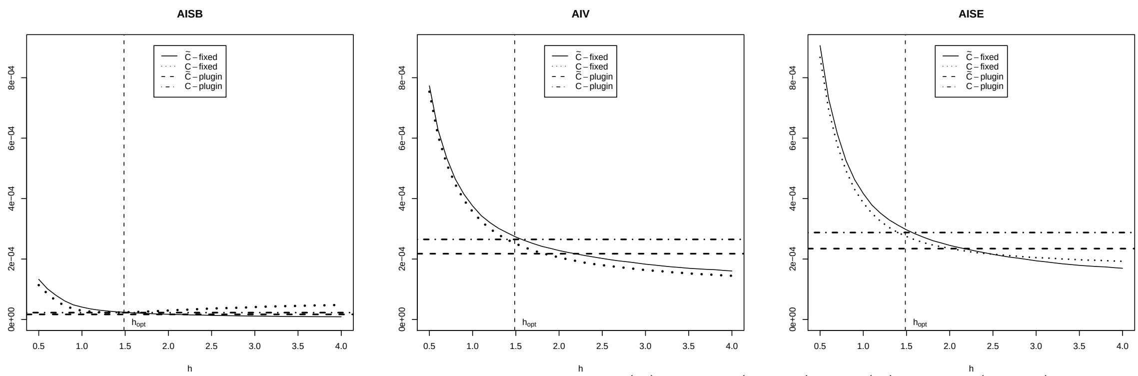

FIGURE 9. Copula estimation; Model 5, $\mu_1(z) = \sin(z-1)$, $\mu_2(z) = \sin(z-1)$, $\rho = 1$.

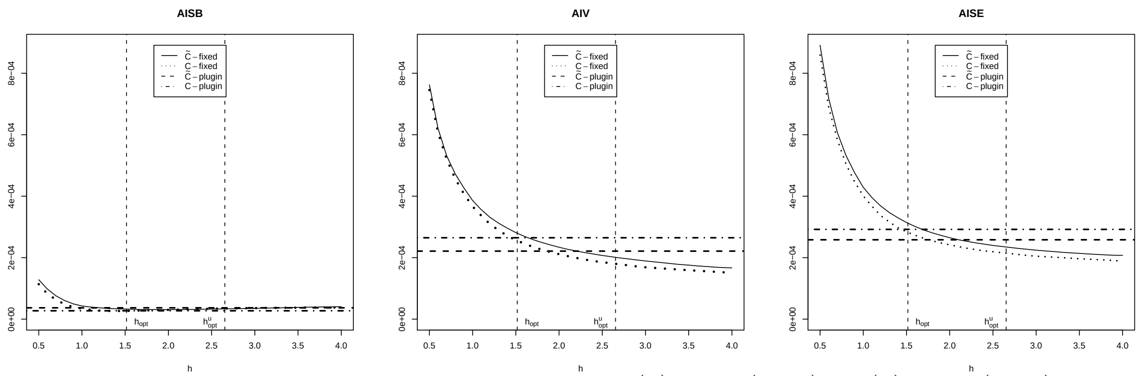

FIGURE 10. Copula estimation; Model 6, $\mu_1(z) = \sin(z-1)$, $\mu_2(z) = \sin(z-1)$, $\rho = 5$.

On the other hand Models 3–8 stand for situations when the dependence of the marginal distributions on the covariate may introduce a substantial bias in the estimation of the conditional copula. We see that both estimators are comparable for bandwidths which are smaller than the bandwidth minimizing the asymptotic mean integrated squared error of C_{xh} (indicated by the vertical line h_{opt}) but for larger bandwidths $\tilde{C}_{xh}$ usually has a substantially better performance. Also with plug-in choice for the bandwidth $\tilde{C}_{xh}$ works better.

The conclusion of the above paragraph does not hold completely in Model 5 and 6 where the estimators C_{xh} and $\tilde{C}_{xh}$ are very comparable for fixed as well as plug-in bandwidths, although the conditional marginal distributions change with the value of the covariate. While for the sample size $n = 500$ the estimator $\tilde{C}_{xh}$ becomes clearly preferable to the estimator C_{xh} for Model 5 (results not shown here), there is only a very slight preference for $\tilde{C}_{xh}$ in Model 6. Further, comparing the results of Model 6 with the results of Model 5 (either for $n = 200$ or $n = 500$, the latter not presented here) we see that the increase of the influence of the covariate on the conditional dependence structure in Model 6 has almost no influence on the bias function, which is in contrast to the bias functions seen in other pairs of Models ($1 \leftrightarrow 2$;

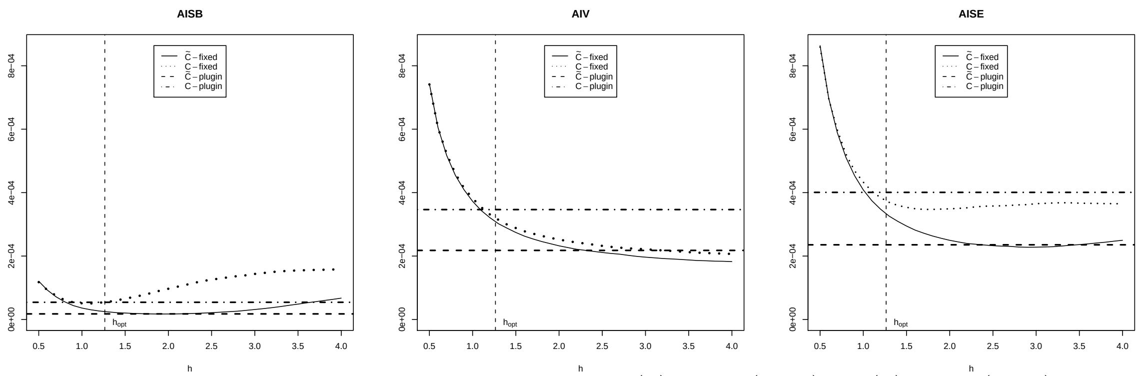

FIGURE 11. Copula estimation; Model 7, $\mu_1(z) = \cos(z-1)$, $\mu_2(z) = \sin(z-1)$, $\rho = 1$.

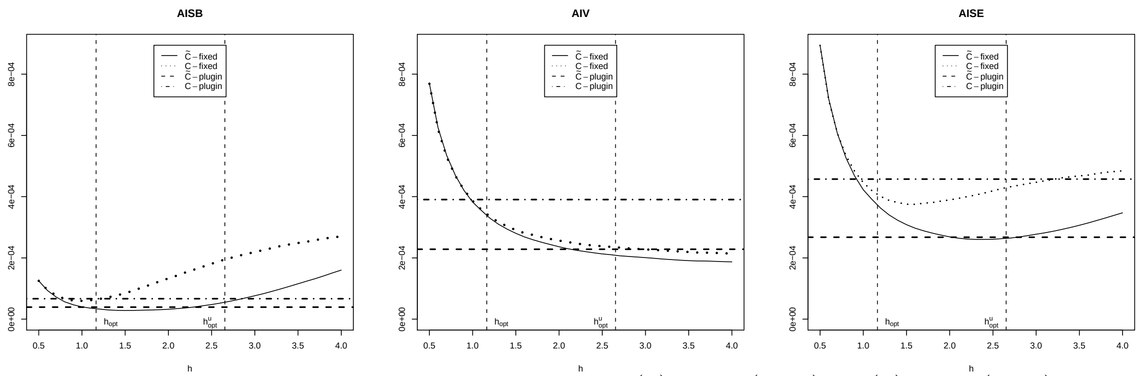

FIGURE 12. Copula estimation; Model 8, $\mu_1(z) = \cos(z-1)$, $\mu_2(z) = \sin(z-1)$, $\rho = 5$.

$3 \leftrightarrow 4$; $7 \leftrightarrow 8$) differing only by the parameter ρ. This indicates that the biases coming from the effects of the covariate on the dependence structure and on the marginals cancel out to some extent.

As the presented results are confirmed with the results for sample size $n = 500$ we can summarize as follows:

- $\tilde{C}_{xh}$ is (in comparison to C_{xh}) quite safe to use and it mostly improves substantially upon C_{xh} if the effect of the covariate on the marginals is not negligible;
- C_{xh} might be slightly preferable if the covariate does not influence marginals distributions or if (by a lucky coincidence) the effect of the covariate on the conditional marginal distributions helps to suppress the effects of the covariate on the conditional dependence structure. For details see Veraverbeke et al. (2009).

4.2. **Kendall's tau.** Although we found that the results on estimation of the entire copula function C_x strongly speak in favor of $\tilde{C}_{xh}$, we are interested whether these findings carry

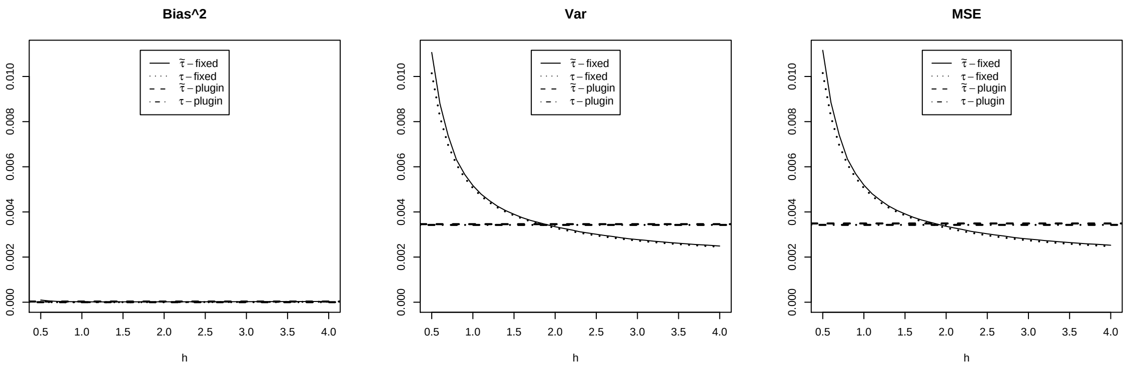

FIGURE 13. Kendall's tau estimation; Model 1, $\mu_1(z) = 1$, $\mu_2(z) = 1$, $\rho = 1$.

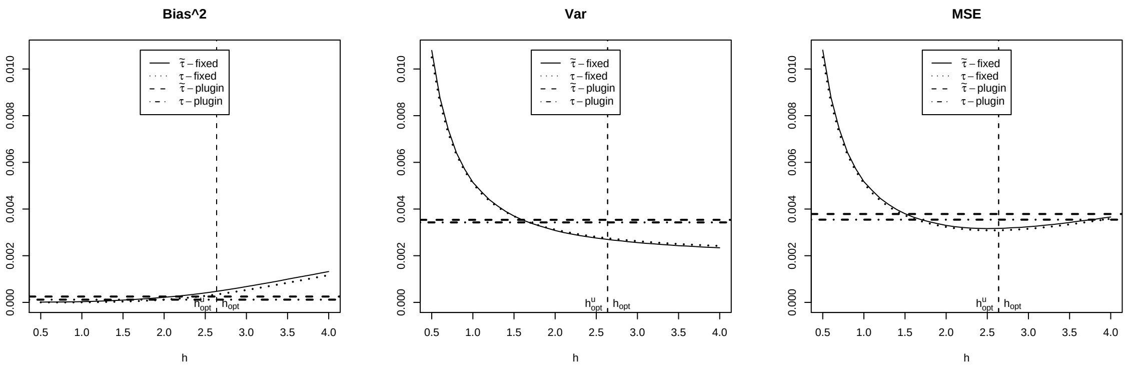

FIGURE 14. Kendall's tau estimation; Model 2, $\mu_1(z) = 1$, $\mu_2(z) = 1$, $\rho = 5$.

over to functionals of C_x. In this application we investigate Kendall's tau, which was already introduced in Section 2.1.1.

In a small simulation study we compared the performance of the estimator of Kendall's tau given by (11) when applied to

[A] the original observations $(Y_{1i}, Y_{2i})^{\mathsf{T}}$, $i = 1, \ldots, n$; (τ-`fixed` and τ-`plugin`);

[B] the transformed 'uniform' alike observations $(\tilde{U}_{1i}, \tilde{U}_{2i})$, $i = 1, \ldots, n$. ($\tilde{\tau}$-`fixed` and $\tilde{\tau}$-`plugin`);

Note that the estimator resulting from [A] is up to some finite sample corrections equivalent to $4 \iint C_{xh}\, dC_{xh} - 1$ and the one resulting from [B] is first order equivalent to $4 \iint \tilde{C}_{xh}\, d\tilde{C}_{xh} - 1$.

We here use the same setting as in Section 4.1. As the findings are analogous to the results of that section we report them only for Models $1, 2, 4, 6$ and 8. These can be found in Figures 13–17 that use the same conventions as Figures 5–12. The only difference is that instead of AISB, AIV and AISE we simply plot bias squared, variance and mean squared error of the estimators.

The findings here are in a close agreement with these for copula estimation. Comparing the results of copula and Kendall's tau estimation it can be noted that Kendall's tau is a

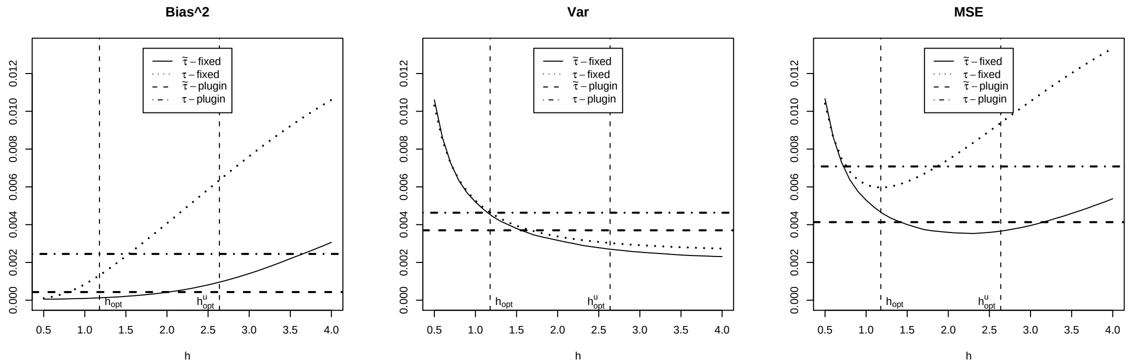

FIGURE 15. Kendall's tau estimation; Model 4, $\mu_1(z) = 1$, $\mu_2(z) = \sin(z-1)$, $\rho = 1$.

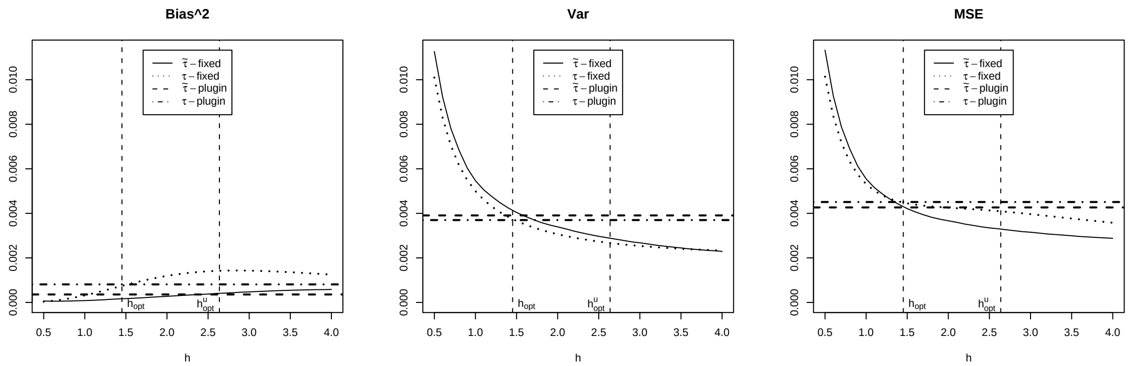

FIGURE 16. Kendall's tau estimation; Model 6, $\mu_1(z) = \sin(z-1)$, $\mu_2(z) = \sin(z-1)$, $\rho = 5$.

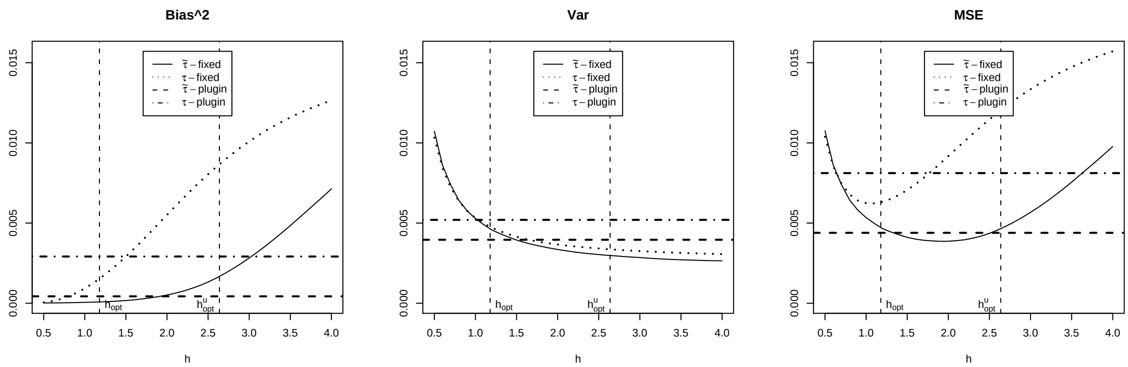

FIGURE 17. Kendall's tau estimation; Model 8, $\mu_1(z) = \cos(z-1)$, $\mu_2(z) = \sin(z-1)$, $\rho = 5$.

functional of a copula, whose estimation is very sensitive to bias properties of an underlying copula estimator.

References

Brockmann, M., Gasser, T., and Hermann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *J. Amer. Statist. Assoc.*, 88:1302–1309.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall.

Gasser, T., Kneip, A., and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.*, 86:643–652.

Gasser, T. and Müller, H.-G. (1979). Kernel estimates of regression functions. In Gasser, T. and Rosenblatt, M., editors, *Lecture Notes in Mathematics 757*, pages 23–68. Springer, New York.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.*, 94:154–163.

Kendall, M. G. (1942). Partial rank correlation. *Biometrika*, 32:277–283.

Nadaraya, E. A. (1964). On estimating regression. *Theor. Prob. Appl.*, 9:141–142.

Nelsen, R. B. (2006). *An Introduction to Copulas.* Springer, New York. Second edition.

Priestley, M. B. and Chao, T. M. (1972). Non-parametric function fitting. *J. Royal Statistical Society. Series B.*, 34:385–392.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Stute, W. (1986). Conditional empirical processes. *Ann. Statist.*, 14:638–647.

Veraverbeke, N., Omelka, M., and Gijbels, I. (2009). Estimation of a conditional copula and association measures. Submitted.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā A*, 26:359–372.

Yang, S.-S. (1981). Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.*, 76:658–662.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.*, 93:228–237.