Some comments on copula-based regression

Holger Dette, Ria Van Hecke, Stanislav Volgushev

Ruhr-Universität Bochum

Fakultät für Mathematik

44780 Bochum

Germany

Abstract

In a recent paper Noh et al. (2013) proposed a new semiparametric estimate of a regression function with a multivariate predictor, which is based on a specification of the dependence structure between the predictor and the response by means of a parametric copula. This comment investigates the effect which occurs under misspecification of the parametric model. We demonstrate by means of several examples that even for a one or two dimensional predictor the error caused by a "wrong" specification of the parametric family is rather severe, if the regression is not monotone in one of the components of the predictor. Moreover, we also show that these problems occur for all of the commonly used copula families and we illustrate in several examples that the copula-based regression may lead to invalid results even when flexible copula models such as vine copulae (with the common parametric families) are used in the estimation procedure.

Keywords: curse of dimensionality, semiparametric inference, copulae, pairwise copulae, vine copulae

1 Introduction

It is well known that nonparametric regression estimates suffer from the curse of dimensionality if the dimension of the predictor is large. In this case a regression function cannot be estimated with reasonable accuracy and several authors have proposed methods to avoid this problem. A common feature of all publications in this direction consists in additional structural or parametric assumptions regarding the unknown regression function, such as additivity [see Stone (1985)], tree-based models [Hastie et al. (2001)] or single index models [Ichimura (1993)]. In a recent paper Noh et al. (2013) introduced a novel semiparametric estimate of the regression function in a nonparametric regression model with a high-dimensional predictor. Roughly speaking, these authors propose to model the dependency structure between the response and the predictor by a parametric copula family in order to obtain estimates of the regression function which converge with a parametric rate. The authors demonstrate (theoretically and empirically) that the resulting estimates have nice properties if the parametric copula family has been chosen correctly. This note is devoted to a careful investigation of the properties of the copula-based regression estimate in the case where the copula family is misspecified. More precisely, our aim is to investigate the kinds of regression dependence that can be described by commonly used copula models.

Let Y and $\mathbf{X} = (X_1, \dots, X_d)^T$ be a real and d-dimensional random variable ($d \geq 1$), respectively, and denote by $F_Y, F_1, \dots, F_d$ the cumulative distribution functions of Y and the margins of $\mathbf{X}$, which will be assumed as differentiable throughout this note. The corresponding densities are denoted by $f_Y, f_1, \dots, f_d$. The famous Sklar's theorem [Sklar (1959)] shows that the joint distribution function F of the vector $(Y, \mathbf{X}^T)^T$ can be represented as

$$F(y, \mathbf{x}^T) = C(F_Y(y), F_1(x_1), \dots, F_d(x_d)),$$

where $(y, \mathbf{x}^T)^T = (y, x_1, \dots, x_d)^T$ and C is the copula. Noh et al. (2013) showed that the mean regression function $m(x_1, \ldots, x_d) = \mathbb{E}[Y|\mathbf{X} = (x_1, \ldots, x_d)]$ can be represented as

$$(1) \qquad m(x_1, \ldots, x_d) = \int_{-\infty}^{\infty} y \frac{c(F_Y(y), F_1(x_1), \ldots, F_d(x_d))}{c_{\mathbf{X}}(F_1(x_1), \ldots, F_d(x_d))} dF_Y(y),$$

where $c = \frac{\partial^{d+1}}{\partial y \partial x_1 \ldots \partial x_d} C$ denotes the density of the copula C corresponding to the vector $(Y, \mathbf{X}^T)^T$, and

$$c_{\mathbf{X}}(u_1, \ldots, u_d) = \int_{-\infty}^{\infty} c(F_Y(y), u_1, \ldots, u_d) dF_Y(y)$$

denotes the copula density corresponding to the vector $\mathbf{X}$. In order to avoid the curse of dimensionality in the estimation of the regression function m these authors propose to use a semi-parametric estimate using a parametric copula family, say $\{c_\theta | \ \theta \in \Theta\}$ for the copula density c in (1) and to estimate the unknown marginal distributions separately. More precisely, if $(Y_1, \mathbf{X}_1^T)^T, \ldots, (Y_n, \mathbf{X}_n^T)^T$ denotes a sample of independent identically distributed observations with copula C and marginal distribution functions $F_Y, F_1, ..., F_d$, Noh et al. (2013) suggest to estimate the marginal distributions F_Y and F_j non-parametrically by $\hat{F}_Y(y) = \frac{1}{n+1} \sum_{i=1}^n I(Y_i \leq y)$ and $\hat{F}_j(x) = \frac{1}{n+1} \sum_{i=1}^n I(X_{ij} \leq x)$, respectively (here we use the notation $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^T$) and to estimate the parameter $\boldsymbol{\theta}$ of the parametric copula family by a pseudo–maximum likelihood method, that is

$$\hat{\boldsymbol{\theta}}_{PL} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \sum_{i=1}^n \log c(\hat{F}_Y(Y_i), \hat{F}_1(X_{i1}), \ldots, \hat{F}_d(X_{id}); \boldsymbol{\theta})$$

[see Genest et al. (1995) or Tsukahara (2005)]. The final estimate of the regression function m is then defined by

$$(2) \qquad \begin{aligned} \hat{m}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{PL}) &= \int_{-\infty}^{\infty} y \frac{c(\hat{F}_Y(y), \hat{F}_1(x_1), \ldots, \hat{F}_d(x_d); \hat{\boldsymbol{\theta}}_{PL})}{\int_{-\infty}^{\infty} c(\hat{F}_Y(u), \hat{F}_1(x_1), \ldots, \hat{F}_d(x_d); \hat{\boldsymbol{\theta}}_{PL}) d\hat{F}_Y(u)} d\hat{F}_Y(y) \\ &= \frac{1}{n+1} \sum_{i=1}^n Y_i \frac{c(\hat{F}_Y(Y_i), \hat{F}_1(x_1), \ldots, \hat{F}_d(x_d); \hat{\boldsymbol{\theta}}_{PL})}{\frac{1}{n+1} \sum_{j=1}^n c(\hat{F}_Y(Y_j), \hat{F}_1(x_1), \ldots, \hat{F}_d(x_d); \hat{\boldsymbol{\theta}}_{PL})}. \end{aligned}$$

In the case of a one-dimensional covariate, i.e. $d = 1$, we have $c_{X_1} \equiv 1$ and thus the estimate simplifies to

$$\hat{m}(x) = \frac{1}{n+1} \sum_{i=1}^{n} Y_i c(\hat{F}_Y(Y_i), \hat{F}_1(x); \hat{\boldsymbol{\theta}}_{PL}). \tag{3}$$

Noh et al. (2013) demonstrate that the estimator defined in (2) avoids the problem of the curse of dimensionality. More precisely, they show that $\hat{m}(\mathbf{x})$ is a $\sqrt{n}$-consistent and asymptotically normal distributed estimate if the parametric copula model has been specified correctly. On the other hand, under misspecification of the copula structure, it is shown that the statistic $\hat{m}(\mathbf{x})$ estimates the quantity

$$m(\mathbf{x}; \boldsymbol{\theta}^*) = \int_{-\infty}^{\infty} y \frac{c(F_Y(y), F_1(x_1), \ldots, F_d(x_d); \boldsymbol{\theta}^*)}{c_{\mathbf{X}}(F_1(x_1), \ldots, F_d(x_d); \boldsymbol{\theta}^*)} dF_Y(y), \tag{4}$$

where $\boldsymbol{\theta}^*$ is the minimum of the function

$$I(\boldsymbol{\theta}) = \int_{[0,1]^{d+1}} \log\Big(\frac{c(u_0, \ldots, u_d)}{c(u_0, \ldots, u_d; \boldsymbol{\theta})}\Big) dC(u_0, \ldots, u_d), \tag{5}$$

and C, c denote the 'true' copula and copula density that generated the data, respectively. As it was pointed out by Noh et al. (2013), the quantity $m(\mathbf{x}; \boldsymbol{\theta}^*)$ does in general not coincide with the true regression function $m(\mathbf{x})$. Consequently there exists a bias if the parametric copula has been misspecified, but no further evidence regarding the kinds of regression functions which can be estimated well (i.e. for which this bias is small) is given. Overall, one might hope that the commonly used parametric copula models are flexible enough to model a rich variety of regression dependencies. In the following section however we will demonstrate that this is not the case and that the quality of the estimate (2) under misspecification of the parametric copula depends heavily on the specific structure of the unknown regression function m. In particular we show that for non-monotone regression functions these estimates are in fact not reliable. We will also demonstrate that model selection from a class of the commonly used copula families by information type criteria (in the case $d = 1$) or the application of more flexible copula families such as vine copulae in the case $d \geq 2$ [see Aas et al. (2009)] does not solve these problems. As soon as the regression is not monotone in one of the components of the explanatory variable the copula-based regression estimate and the true regression function show substantially different qualitative features.

2 Inference under misspecification - examples

All presented results are based on a sample of size $n = 100$. For the sake of brevity we restrict ourselves to the case $d = 1$ and $d = 2$, where the problems of misspecification of the parametric copula family are already very visible and the arguments are more transparent. We expect that for high dimensional predictors these problems are even more severe.

We start our investigation with the one-dimensional regression model

$$Y_i = (X_i - 0.5)^2 + \sigma\varepsilon_i, \quad i = 1, \dots, n, \tag{6}$$

where the explanatory variable X_i is uniformly distributed on the interval $[0, 1]$ and the errors are normally distributed with mean 0 and variance $\sigma^2 = 0.01$. We present in Figure 2 ”typical” simulated data from this model with the corresponding copula regression estimates, where we use the t, Frank copula in the left and middle panel and a mixture of two normal copulas in the right panel (here the mixing proportion is also estimated from the data) and as a first conclusion we note that none of these choices yields a reasonable estimate of the regression function. In fact we considered all copulae from the following list {“amh copula”,“independence copula”, “Gaussian copula”, ”t-copula”, “Clayton copula”, “Gumbel copula”, “Frank copula”, “Joe copula”, “Clayton-Gumbel copula”, “Joe-Gumbel copula”, “Joe-Clayton copula”, “Joe-Frank copula”} together with corresponding rotations. No copula mentioned above reproduces the structure of the regression function in the resulting estimate.

Insert Figure 1 and 2 about here

This observation can be explained by the fact that none of the available parametric copula models for the vector (Y, X_1) yields a non-monotone regression function

$$m(x_1) = \mathbb{E}[Y|X_1 = x_1] \; = \int_{-\infty}^{\infty} yc(F_Y(y), F_1(x_1))dF_Y(y). \tag{7}$$

In Figure 1 this function is exemplarily displayed for various commonly used parametric copula models (and different parameters). Other results, which are not displayed here for the sake of brevity, show similar features. We observe that all of the commonly used parametric copula models lead to a monotone regression in (7). As a consequence we point out that model selection (for example by the AIC criterion) from a large class of commonly used parametric copula models will not improve the performance of the estimate.

Remark 2.1

(1) Obviously, by definition of a copula, there exists a copula model corresponding to the model (6). However, our numerical investigations indicate that this copula cannot be well approximated by any of the commonly proposed parametric copula models.

(2) The application of alternative estimates for the parameter of the copula does not lead to a significant improvement of the situation. For example investigations for the L^2-type estimator defined by

$$\hat{\boldsymbol{\theta}}_{L^2} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \hat{m}(X_i; \boldsymbol{\theta})\right)^2 , \tag{8}$$

with

$$\hat{m}(x; \hat{\boldsymbol{\theta}}_{L^2}) = \frac{1}{n+1} \sum_{i=1}^{n} Y_i c(\hat{F}_Y(Y_i), \hat{F}_1(x); \hat{\boldsymbol{\theta}}_{L^2}) \tag{9}$$

yield a picture very similar to the results presented here (these results are not displayed for the sake of brevity).

Insert Figure 3 about here

Obviously the observations of the previous paragraph carry over to higher dimensional predictors if the regression is not monotone in one of the predictors. To demonstrate this, we exemplarily briefly consider the two-dimensional regression model

$$Y_i = m(X_{i1}, X_{i2}) + \sigma\varepsilon_i, \quad i = 1, \dots, n, \tag{10}$$

with regression function

$$m(x_1, x_2) = (x_1 - 0.5)^2 + (x_2 - 0.5)^2, \tag{11}$$

where the sample size is again $n = 100$ and X_{i1} and X_{i2} are independent uniformly distributed on the interval $[0, 1]$. Some results for the Gaussian, Gumbel and t copula are displayed in Figure 3, and we observe the same problems as in the one-dimensional case. The considered parametric copula families are simply not flexible enough and the resulting estimate is unable to reflect the curvature of the regression function.

In order to investigate more flexible classes of parametric copula models we briefly consider the concept of vine copulae, which is based on a decomposition of the copula density into a product of (conditional) bivariate copula densities according to a carefully chosen so called R-vine structure. The bivariate copula densities are then chosen from parametric copula families by applying a model selection criterion. It has been argued by several authors [see e.g. Aas et al. (2009)] that the resulting vine copula obtained by this pair-copula decomposition admits a flexible modeling of the dependency structure in the case of multiple covariates. Here we briefly investigate if this concept can be used to obtain improved copula-based regression estimates in the case $d \geq 2$. The pair-copula decomposition gives us the possibility to model the copula density c in different ways by first selecting an R-Vine structure and then choosing the pair-copulae independently from a set of parametric copula families. To implement this approach, we used the *R*-package *VineCopula* to select the vine structure as well as the parametric pair copulae with the corresponding estimated parameters by the Akaike information criterion [see Dißmann et al. (2011) for more details]. Here, the parametric pair copulae were selected from the set {“independence copula”, “Gaussian copula”, ”t-copula”, “Clayton copula”, “Gumbel copula”, “Frank copula”, “Joe copula”, “Clayton-Gumbel copula”, “Joe-Gumbel copula”, “Joe-Clayton copula”, “Joe-Frank copula”} with corresponding rotations.

In Figure 4 we display a typical situation for model (10) with regression functions given in (11) as well as,

$$m(x_1, x_2) = (x_1 - 0.5)^2 - (x_2 - 0.5)^2. \tag{12}$$

The sample size is again $n = 100$ and the variance is $\sigma^2 = 0.01$. We observe that in all cases the copula-based regression method does not yield estimates which reflect the qualitative behavior of the regression function. These results show that even the rather large family of vine-copulae does not reproduce the *regression* structures imposed by the models (11) and (12).

Insert Figure 4 and 5 about here

3 Conclusions

In this note we have studied some properties of a semiparametric copula-based regression estimate which has been recently proposed in Noh et al. (2013), and, more broadly, the types of regression dependence that can be obtained from commonly used copula families. Our simulations (not shown in this note) confirmed that the approach of Noh et al. (2013) is attractive if the dependency structure of the data can be specified accurately. On the other hand – if the true copula structure has been misspecified – the approach often does not yield reliable estimates of the regression function. If the regression function is not monotone, copula-based regression estimates do not reproduce the qualitative features of the regression function. This property does not depend on the specific misspecified copula model but can be observed for all of the commonly used parametric copula families. For a high-dimensional predictor the situation is even worse. Moreover, we also show that for high-dimensional predictors more flexible models as vine copulae (based on the commonly used parametric models) will not improve the properties of the estimator. The reason for these problems consists in the fact that (for $d = 1$) all commonly used parametric copula families produce a regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ which is monotone in the explanatory variable $\mathbf{X}$. As a consequence non-monotone features of the regression function cannot be reproduced by the copula-based regression estimate. In Figure 5 we display level sets of a copula density corresponding to some non-monotone regression functions. We observe that these differ substantially from the sets of all commonly used parametric copula densities. Future research is necessary to develop more flexible parametric copula models reflecting these structures.

Acknowledgements. The authors would like to thank Martina Stein, who typed parts of this manuscript with considerable technical expertise. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1, C1, A7) of the German Research Foundation

(DFG). We are also grateful to C. Czado for interesting discussion about the concept of vine copulae and to Axel Bücher and Anouar El Ghouch for comments on a preliminary version of this manuscript.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance, Mathematics and Economics*, 44:182–198.

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2011). Selecting and estimating regular vine copulae and application to financial returns. Technical report, Technische Universität München.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Springer Series in Statistics*. Springer, New York.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58:71–120.

Noh, H., El Ghouch, A., and Bouezmarni, T. (2013). Copula-based regression estimation and inference. *Journal of the American Statistical Association*, 108:676–688.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33:357–375.

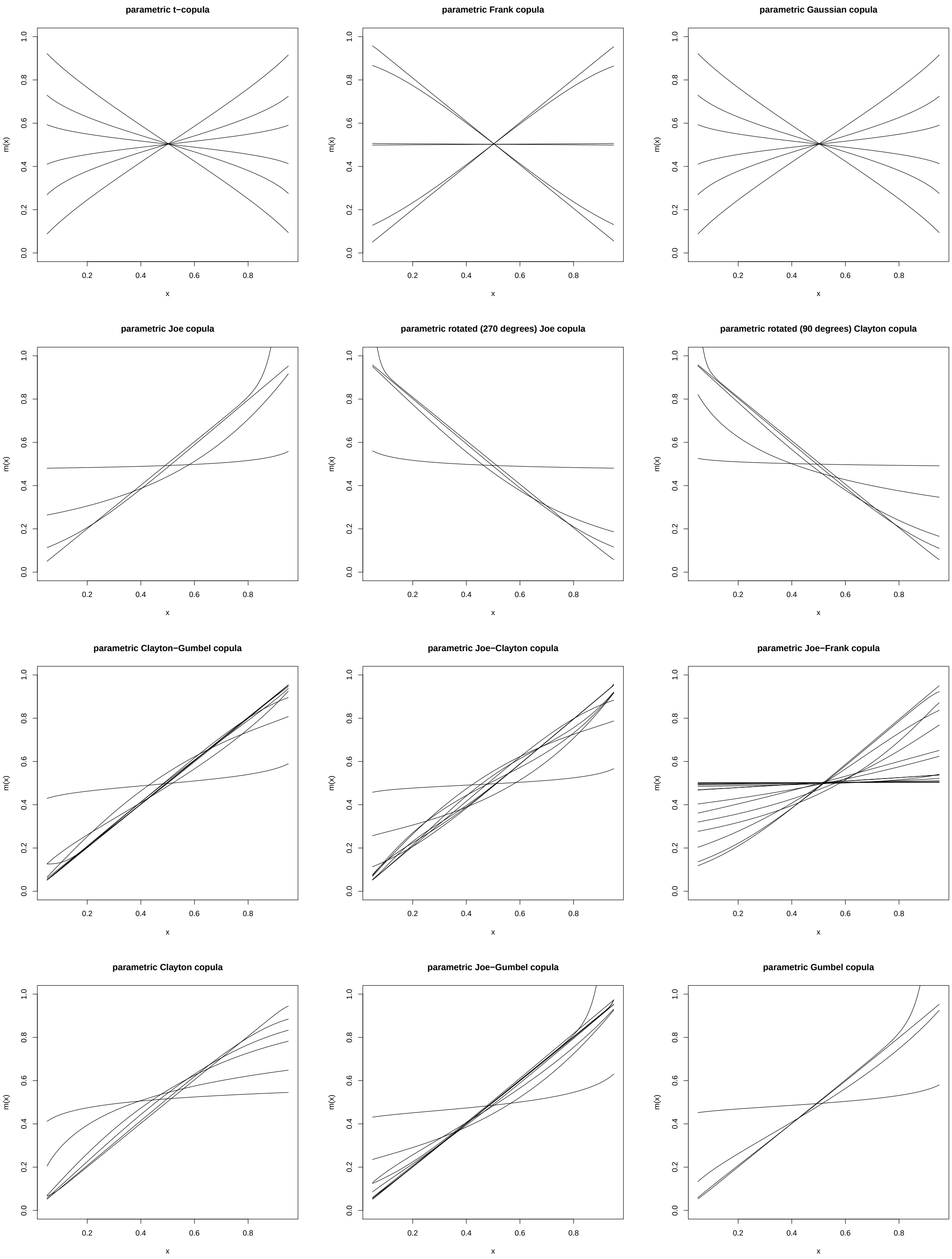

Figure 1: *The function (7) for commonly used parametric copula families (different parameters are used in each figure)*

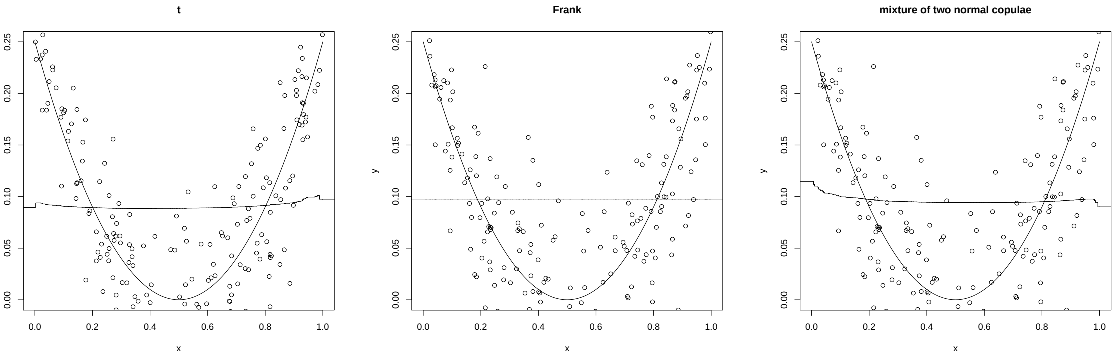

Figure 2: *Copula based regression estimates of the regression function in model* (6)*. The t-copula (left panel), Frank copula (middle panel) and a mixtures of two normal copulae (right panel) are used in the estimate* (2)*.*

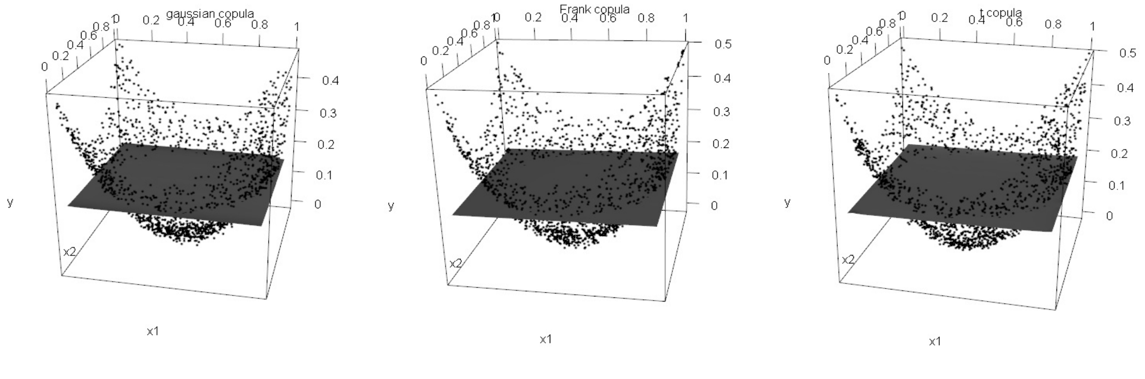

Figure 3: *Copula based regression estimates of the two-dimensional regression function* (11)*. The copula in the estimate* (2) *is chosen as Gaussian copula (left panel), Frank copula (middle panel) and t-copula (right panel).*

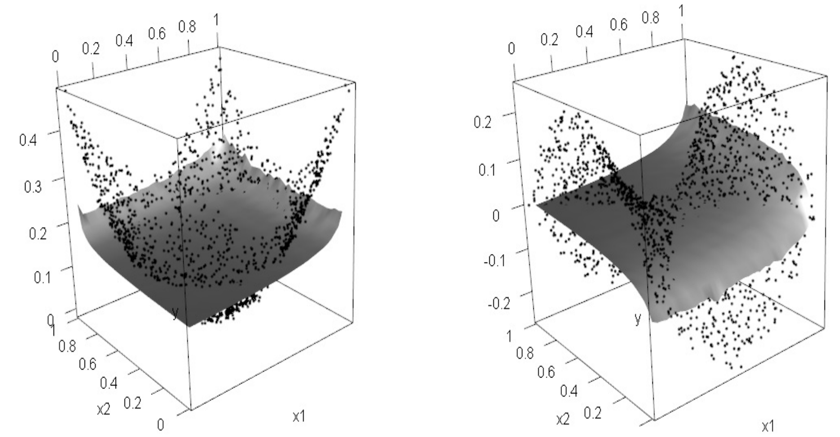

Figure 4: *Copula based regression estimates of the regression function in model* (11) *(left panel), and* (12) *(right panel). A vine copula selected by the AIC criterion has been used in the estimate* (2).

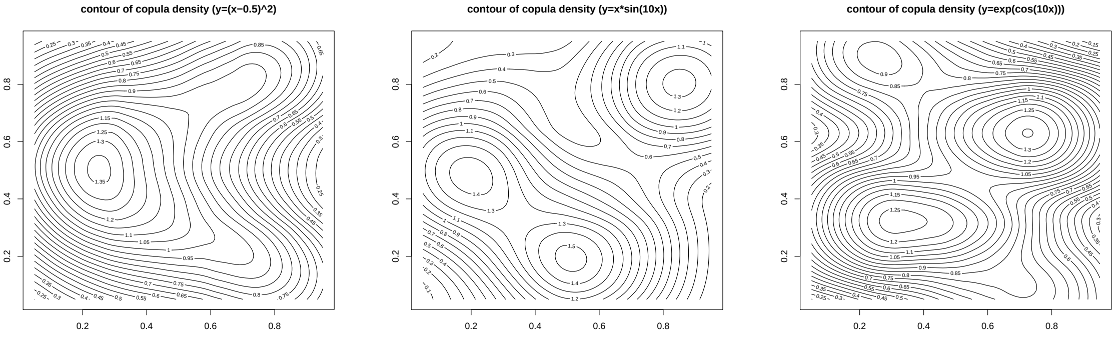

Figure 5: *Simulated contour plots of the copula density corresponding to the one-dimensional regression model* (6) *(left panel), the function* $m(x) = x\sin(10x)$ *(middle panel) and the function* $m(x) = \exp(\cos(10x))$ *(right panel) with Gaussian errors.*