The Canadian Journal of Statistics
Vol. 47, No. 1, 2019, Pages 8–26
La revue canadienne de statistique

Predictive assessment of copula models

Elif F. ACAR*, Parisa AZIMAEE, and Md. Erfanul HOQUE

Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

Key words and phrases: Conditional quantiles; copula regression; copula selection; prediction error; vine copulas.

MSC 2010: 62H20; 62J02

Abstract: Copulas are powerful explanatory tools for studying dependence patterns in multivariate data. While the primary use of copula models is in multivariate dependence modelling, they also offer predictive value for regression analysis. This article investigates the utility of copula models for model-based predictions from two angles. We assess whether, where, and by how much various copula models differ in their predictions of a conditional mean and conditional quantiles. From a model selection perspective, we then evaluate the predictive discrepancy between copula models using in-sample and out-of-sample predictions both in bivariate and higher-dimensional settings. Our findings suggest that some copula models are more difficult to distinguish in terms of their overall predictive power than others, and depending on the quantity of interest, the differences in predictions can be detected only in some targeted regions. The situations where copula-based regression approaches would be advantageous over traditional ones are discussed using simulated and real data. *The Canadian Journal of Statistics* 47: 8–26; 2019 © 2018 Statistical Society of Canada

Résumé: Les copules constituent un outil puissant pour expliquer et étudier la structure de dépendance de données multivariées. Bien que les copules soient surtout utilisées dans un but de modèlisation, elles peuvent également servir à faire des prévisions dans le cadre d'une analyse de régression. Les auteurs abordent l'usage des copules dans le cadre de prévisions basées sur un modèle selon deux perspectives. Ils étudient d'abord dans quelles circonstances et à quel point les différents modèles de copules offrent des prévisions différentes pour les moyennes et les quantiles conditionnels. Dans une optique de sélection de modèle, ils évaluent ensuite la différence entre les prévisions pour des données hors échantillon ou non, dans le cas bivarié ou de dimension supérieure. Les résultats suggèrent qu'il est difficile de départager les modèles de copules à partir de leur performance globale, mais que des différences dans les prévisions peuvent être détectées dans certaines régions ciblées, selon la valeur d'intérêt. à l'aide de données réelles et simulées, les auteurs discutent des situations où une approche de régression basée sur les copules s'avère avantageuse par rapport aux méthodes traditionnelles. *La revue canadienne de statistique* 47: 8–26; 2019 © 2018 Société statistique du Canada

1. INTRODUCTION

In many statistical problems the specification of a multivariate distribution is essential to study marginal, conditional and joint features of random variables. According to the seminal result of Sklar's theorem (Sklar, 1959), any multivariate distribution can be represented as a copula function of its marginal distributions, making copulas an attractive tool in multivariate statistical modelling. Specifically, if F is the d-dimensional distribution function of random variables

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author.*
E-mail: elif.acar@umanitoba.ca

$X_1, \ldots, X_d$ with respective univariate marginal distributions $F_1, \ldots, F_d$, then there exists a copula function C: $[0, 1]^d \to [0, 1]$ such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \ . \tag{1}$$

Unlike canonical multivariate distributions, the representation in Equation 1 does not require the random variables to have the same distribution or be of the same type, hence it is very flexible. Furthermore, when F is absolutely continuous, the copula C is unique and fully describes the dependence structure of $X_1, \ldots, X_d$ refined from their marginal distributions.

With these appealing features, copulas have received considerable attention over the past few decades, and have been the subject of much theoretical, methodological and applied work. See, for instance, the monographs of Nelsen (2006) and Joe (1997, 2014), and the references therein.

Copulas have been used primarily for explanatory purposes to describe the dependence structure of random variables. As a result, a large number of bivariate parametric families of copulas (e.g., elliptical, Archimedean) and several new classes of multivariate copula models built hierarchically using bivariate copulas (e.g., vine copulas) have been proposed in the literature. Considering the richness of available copula models, much less attention has been given to their use for prediction purposes.

Motivated by the discussion in Shmueli (2010), which contrasts explanatory and predictive aspects of statistical modelling, this article investigates the predictive utility of copula models. As highlighted in Shmueli (2010), a statistical model that offers a good explanatory insight, in terms of revealing causal or association relationships present in data, may not always possess good predictive power. Conversely, a model tailored for prediction purposes may not always offer an explanation of the nature of the underlying mechanism. Differentiating and (ideally) balancing these two aspects are desirable when suggesting a statistical model.

A number of articles have considered the use of copulas in regression contexts. Sungur (2005) and Crane & van der Hoek (2008) derived conditional mean regression functions for several bivariate copula families, and Leong & Valdez (2005) investigated copula-based predictions for insurance claims. A more formal treatment of copula-based mean regression was provided in Noh, El Ghouch, & Bouezmarni (2013), which presented a semiparametric (SP) inference procedure based on rank-transformed marginal distributions. Following the latter work, Dette, Van Hecke, & Volgushev (2014) discussed the shortcomings of copula regression for non-monotone regression functions and noted that such features can by no means be captured by a copula model with only positive (or only negative) dependence. Their numerical illustrations of this special case indicated that copula-based predictions can fail drastically when the dependence structure is severely misspecified. On the other hand, Cooke, Joe, & Chang (2015) suggested using a Gaussian vine copula model for prediction purposes in higher dimensions, and via an example demonstrated that this approximate model can yield conditional mean predictions that are indistinguishable from those obtained from an optimal regular vine.

On a parallel track, Bouyé & Salmon (2009), Bernard & Czado (2015), Noh, El Ghouch, & Van Keilegom (2015) and Kraus & Czado (2017) investigated the use of copulas for estimating conditional quantiles. While these articles assessed the performance of copula-based conditional quantile estimators relative to other quantile regression methods, the problem of copula model selection and model misspecification was not discussed in detail.

The primary goal of this article is to provide an understanding of whether, where, and by how much various copula models differ in their predictions of a conditional mean and conditional quantiles. Such an understanding would help to identify situations where copula misspecification has a grave or subtle impact on model-based predictions. This, in turn, informs a prediction-based approach to in-sample selection and out-of-sample evaluation of copula models.

These aspects are addressed here in detail, focusing on bivariate copula models. Besides their tractability and ease of interpretation, bivariate copulas are often used as building blocks for

more complex multivariate copula models. Hence, a detailed evaluation of their predictive value and discrepancies facilitates the assessment of the same for vine copula models.

The article is organized as follows. Section 2 introduces the conditional mean and quantile functions for commonly used bivariate copula families, and discusses the potential impact of copula family misspecification on these features using graphical tools. Through numerical studies, Section 3 describes prediction-based copula model selection criteria and evaluates the predictive discrepancy between bivariate copula models, as well as the linear regression model. Section 4 outlines the prediction of the conditional mean in higher-dimensional settings using vine copulas and examines the predictive performance of different vine models via a numerical study and a concrete example. We provide a brief discussion of our findings in Section 5. Selected additional results are reported in the Supplementary Material.

2. PREDICTIVE ASSESSMENT OF BIVARIATE COPULAS

Let Y be a continuous response variable, and X a univariate continuous covariate. Denote by F_Y and F_X the cumulative distribution functions of Y and X, which we assume are continuously differentiable with marginal densities f_Y and f_X. By Equation 1, the joint distribution function of (Y, X) is given by $F(y, x) = C(F_Y(y), F_X(x))$, which admits the joint density function

$$f(y,x) = c(F_Y(y), F_X(x)) f_Y(y) f_X(x),$$

for any real-valued vector (y, x), where $C(u, v) = P(U \leq u, V \leq v)$ is the copula function of (Y, X), with $U = F_Y(Y)$ and $V = F_X(X)$; $c(u, v) = \partial^2 C(u, v)/\partial u \partial v$ denotes the copula density.

These representations lead to specifications of the conditional distribution function $F_{Y|X}$ and the conditional density $f_{Y|X}$ of Y given X in terms of, respectively, the copula C and its density c as

$$\begin{aligned} F_{Y|X}(y \mid X = x) &= P(Y \leq y \mid X = x) = P(F_Y(Y) \leq F_Y(y) \mid F_X(X) = F_X(x)) \\ &= C(U \leq u \mid V = v) = \frac{\partial C(u,v)}{\partial v} = C_{U|V}(u|V = v), \end{aligned}$$

where $u = F_Y(y)$ and $v = F_X(x)$, and

$$f_{Y|X}(y \mid X = x) = f_Y(y) c(F_Y(y), F_X(x)).$$

Any conditional feature of Y given X can be derived using these expressions. In the following, we have a closer look at the conditional mean, variance, and quantile functions for different copula families with the aim of understanding their inherent differences.

2.1. Conditional Mean and Variance Functions under Different Copula Families

The k^{th} moment of the conditional distribution function $F_{Y|X}$ is given by

$$\mathrm{E}(Y^k \mid X = x) = \int_{-\infty}^{\infty} y^k \; c(F_Y(y), F_X(x)) \; dF_Y(y) = \mathrm{E}\{Y^k \; c(F_Y(Y), F_X(x))\},$$

from which one can obtain the conditional mean and variance functions $m(x) = \mathrm{E}(Y|X = x)$ and $\mathrm{Var}(Y|X = x) = \mathrm{E}(Y^2|X = x) - \{\mathrm{E}(Y|X = x)\}^2$, respectively. Analytical expressions for these quantities are available for only a few copula families (Crane & van der Hoek, 2008), hence they are typically evaluated using numerical integration.

The conditional mean and variance functions are displayed in Figure 1 for the most commonly used parametric copula families, that is, the Gaussian, (Student-)t, Clayton, Gumbel and Frank copulas, with standard normal margins. For each family, we considered three levels of the strength of dependence between X and Y, and specified the copula parameter values so

that the corresponding Kendall's tau values were $\tau = 0.2$ (weak dependence), $\tau = 0.5$ (moderate dependence), and $\tau = 0.8$ (strong dependence). The second parameter of the t copula was fixed at $\nu = 4$. These families exhibit very different local dependence patterns, especially in the tails; see Figure 2 for the corresponding contour plots with standard normal margins when $\tau = 0.5$.

Note that the Gaussian copula with a normally distributed response is equivalent to the linear regression model; this equivalence is reflected in its linear conditional mean and constant conditional variance functions. The conditional mean functions for other families are all monotone increasing with slight nonlinearities, some of which are barely distinguishable, for example, t and Gumbel copulas at moderate and high dependence levels.

On the other hand, the conditional variance functions better pronounce the differences among these copulas, especially for covariate values in the tails. The non-constant variance functions for copula families other than Gaussian suggest that copula-based regression models may be more suited for regression analysis under heteroscedasticity. For the Student-t regression models, this observation is well-known (e.g., Spanos, 1994). However, to the best of our knowledge, the heteroscedasticity of conditional variance functions has not been explicitly noted for other copula families. Consistent with its strong central dependence pattern, the Frank copula has a slightly smaller conditional variance for covariate values in the middle range. Similarly, exhibiting a strong lower (upper) tail dependence, the Clayton (Gumbel) copula has a smaller conditional variance for covariate values in the lower (upper) tail, when the dependence is moderate to strong. Interestingly, the opposite conclusion holds for these families under weak dependence. A close investigation of the conditional densities under the five copula families (see Figure S1 in Appendix A of the Supplementary Material) indicates that for the Clayton, Gumbel and t copulas, the level of dependence affects the skewness, and hence the variance, of the underlying conditional density in the region(s) of tail dependence. On the other hand, having no tail dependence the Gaussian and Frank copulas retain symmetry in their conditional densities.

The conditional variance plays an important role in quantifying the impact of copula family misspecification on mean predictions. Consider the working copula model c_W, which may be different than the true copula model c_0. It is evident from Figure 1 that $m_W(x) = \mathrm{E}_W(Y|X = x)$ is biased for the true mean function $m_0(x) = \mathrm{E}_0(Y|X = x)$. In order to assess where, in the range of X, the copula families differ in their mean predictions, we use the conditional expected squared prediction error under the working model, which is

$$\begin{aligned}\mathrm{CEPE}(x; c_W, c_0) &= \mathrm{E}_0[\{Y - m_W(x)\}^2 \mid X = x] \\ &= \int_{-\infty}^{\infty} \left\{ y - \int_{-\infty}^{\infty} y \;\; c_W(F_Y(y), F_X(x)) dF_Y(y) \right\}^2 \times c_0(F_Y(y), F_X(x)) \;\; dF_Y(y).\end{aligned}$$

When the working copula model is the same as the true copula model, that is, $c_W = c_0$, CEPE $(x;c_0, c_0)$ is equal to the conditional variance $\mathrm{Var}(Y|X)$.

The matrix plot in Figure 1 displays the conditional expected prediction error of the ''best'' misspecified copula model c_W^* within each family, relative to that of the true model, that is, $\{\mathrm{CEPE}(x; c_W^*, c_0)/\mathrm{CEPE}(x; c_0, c_0)\}^{1/2}$, where $c_W^* = c_W(\cdot, \cdot; \theta_W^*)$ is closest, in terms of the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951), to the true model. Under each of the five copula families, the values of $\theta_W^* = \arg\min_\theta \mathrm{E}_0[\log\{c_0(uv\theta_0)/c_W(u, v; \theta)\}]$ are tabulated and reported in Table S1 of Appendix A of the Supplementary Material in Kendall's tau scale.

We conclude that the choice of copula family has almost no impact on the expected prediction error when one wishes to make predictions at covariate values in the middle of the range of X. However, for mean predictions at covariate values in the tail, the expected prediction error under a misspecified copula can be inflated by more than 5-fold.

The first row and column of the matrix plot enable us to relate copula-based regression to simple linear regression. When the underlying dependence is Gaussian (the first row of the

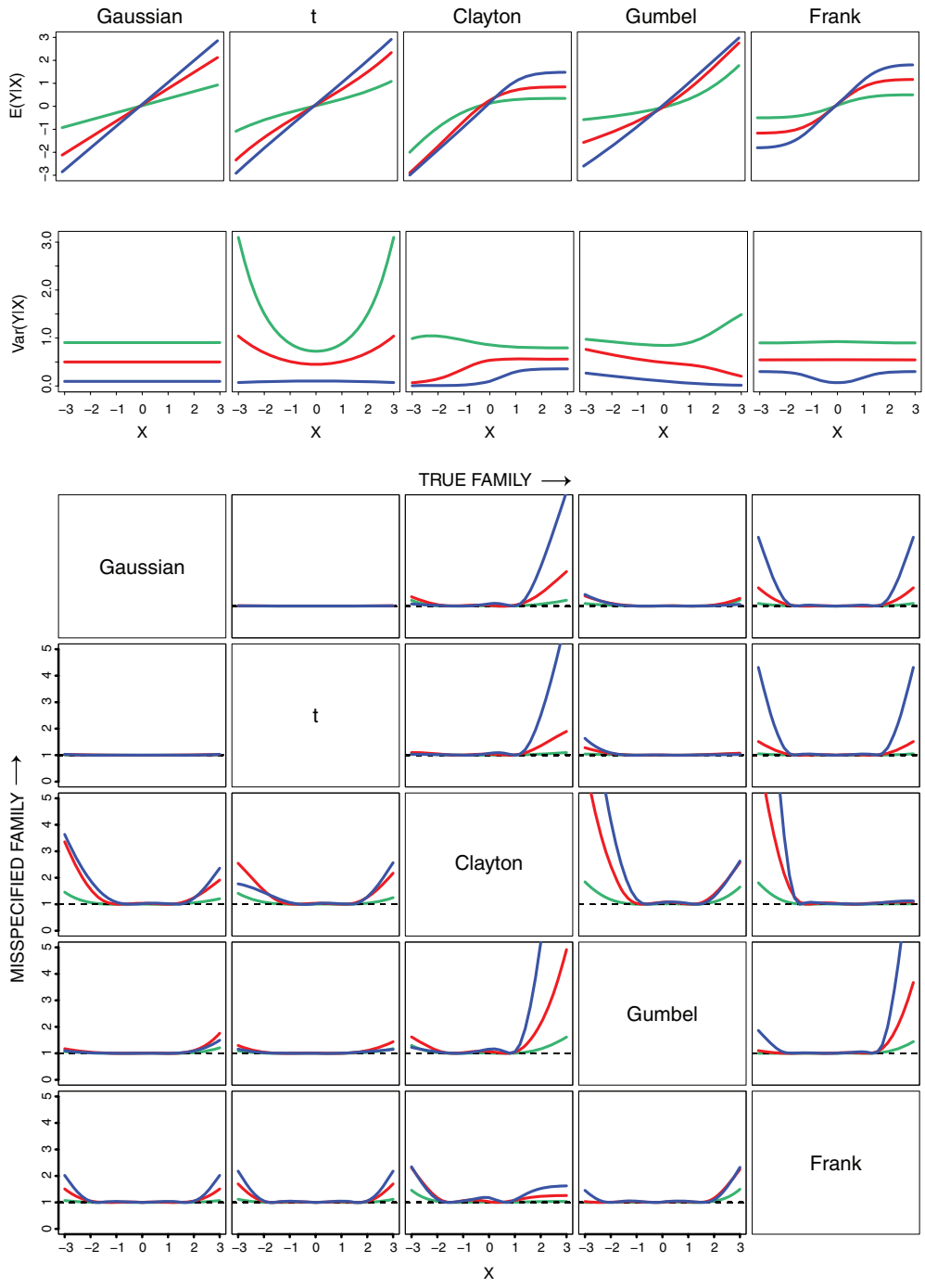

FIGURE 1: Conditional mean, $\mathrm{E}(Y|X)$, and conditional variance, $\mathrm{Var}(Y|X)$, functions of the five copula families, with standard normal margins, at weak ($\tau = 0.2$: green line), moderate ($\tau = 0.5$: red line) and strong ($\tau = 0.8$: blue line) dependence. The matrix plot displays the ratio of conditional expected prediction errors of a misspecified copula, $\mathrm{E}_0[\{Y - \mathrm{E}_W(Y|X)\}^2|X]^{1/2}$, and the true copula, $\mathrm{E}_0[\{Y - \mathrm{E}_0(Y|X)\}^2|X]^{1/2}$, calculated using the tabulated Kendall's tau values in Table S1 of the Supplementary Material at each of the three dependence levels. The diagonal entries specify the true family for each row and a misspecified family for each column.

matrix plot), using a copula-regression model with a misspecified Clayton or Frank copula can result in erratic predictions at very high and/or low covariate values. On the other hand, when the underlying dependence is not Gaussian (the first column of the matrix plot), the simple linear regression model can still be reasonable for some copulas (e.g., Gumbel and t) but would be unreliable for others (e.g., Clayton and Frank).

Here we chose the standard normal distribution for the margins not only for its ease of interpretation but also because of the generality it offers via standardization (z-score) and normalization (e.g., probit transformation) techniques. A natural question to ask is the extent to which these conclusions hold under other (e.g., skewed) marginal distributions. Note that $\mathrm{E}(Y^k|X=x) = \mathrm{E}[Y^k|G^{-1}\{F_X(X)\} = G^{-1}\{F_X(x)\}]$ for any distribution G. Hence, distributions other than normal can be easily accommodated in the covariate, without further calculation, by simply mapping their quantiles to those of the standard normal distribution. For quantile mapping, the uniform scale can be used for the covariate to facilitate interpretations; see Figure S4 in Appendix B of the Supplementary Material.

Since mean regression is not invariant under monotone transformations, a similar conversion is not appropriate for the response variable. One needs to evaluate the conditional features of Y given X, properly accounting for its measurement scale. An assessment in the case of a lognormal response suggests that while the conditional mean and variance functions are inherently different from those of the standard normal, the impact of copula misspecification remains pronounced mostly in similar tail regions of the covariate; see Figures S5 and S6 in Appendix B of the Supplementary Material.

2.2. Conditional Quantiles under Different Copula Families

Unlike the conditional mean function, conditional quantiles enjoy invariance under strictly increasing transformations of both X and Y. The α-level conditional quantile of Y given $X = x$ is given by

$$Q_\alpha(x) = F^{-1}_{Y|X=x}(\alpha) = F^{-1}_Y[C^{-1}_{U|V}\{\alpha \mid V = F(x)\}].$$

A comprehensive study of conditional quantile functions for several bivariate copulas is provided in Bernard & Czado (2015). However, the behaviour of these functions under copula misspecification has not been addressed.

Considering the same five copula families and focusing on the case $\tau = 0.5$, we evaluated the impact of copula misspecification on the calculation of the conditional quantiles at levels $\alpha = 0.10$, 0.50, and 0.90. Figure 2 displays the contour plots and conditional quantile functions of the five copula families with standard normal margins. To assess the accuracy of the α-level conditional quantile $Q_{W;\alpha}(x)$ under a working copula model C_W, we check the conditional percentile it corresponds to under the true model C_0, that is,

$$\begin{aligned} P_\alpha(x; C_W, C_0) &= P\{Y \le Q_{W;\alpha}(x) \mid X = x\} = F_{Y|X}\{Q_{W;\alpha}(x)\} \\ &= C_{0;U|V}[F_Y\{Q_{W;\alpha}(x)\} \mid V = F(x)] \\ &= C_{0;U|V}[C^{-1}_{W;U|V}\{\alpha \mid V = F(x)\} \mid V = F(x)], \end{aligned}$$

which attains the desired α level when $C_W = C_0$.

The matrix plot in Figure 2 shows $P_\alpha(x; C^*_W, C_0)$ for the five copula families at $\alpha = 0.10$, 0.50, 0.90, where $C^*_W = C_W(\cdot, \cdot; \theta^*_W)$ has the smallest KL divergence from the true model within each family. The discrepancies between the curves $P_\alpha(x; C^*_W, C_0)$ and the targeted α levels indicate that copula family misspecification may have a grave impact when one wishes to predict conditional quantiles of the response variable, especially at covariate values in the tail. For instance, when the underlying copula is Gaussian (the first row of the matrix plot), using a t copula may be

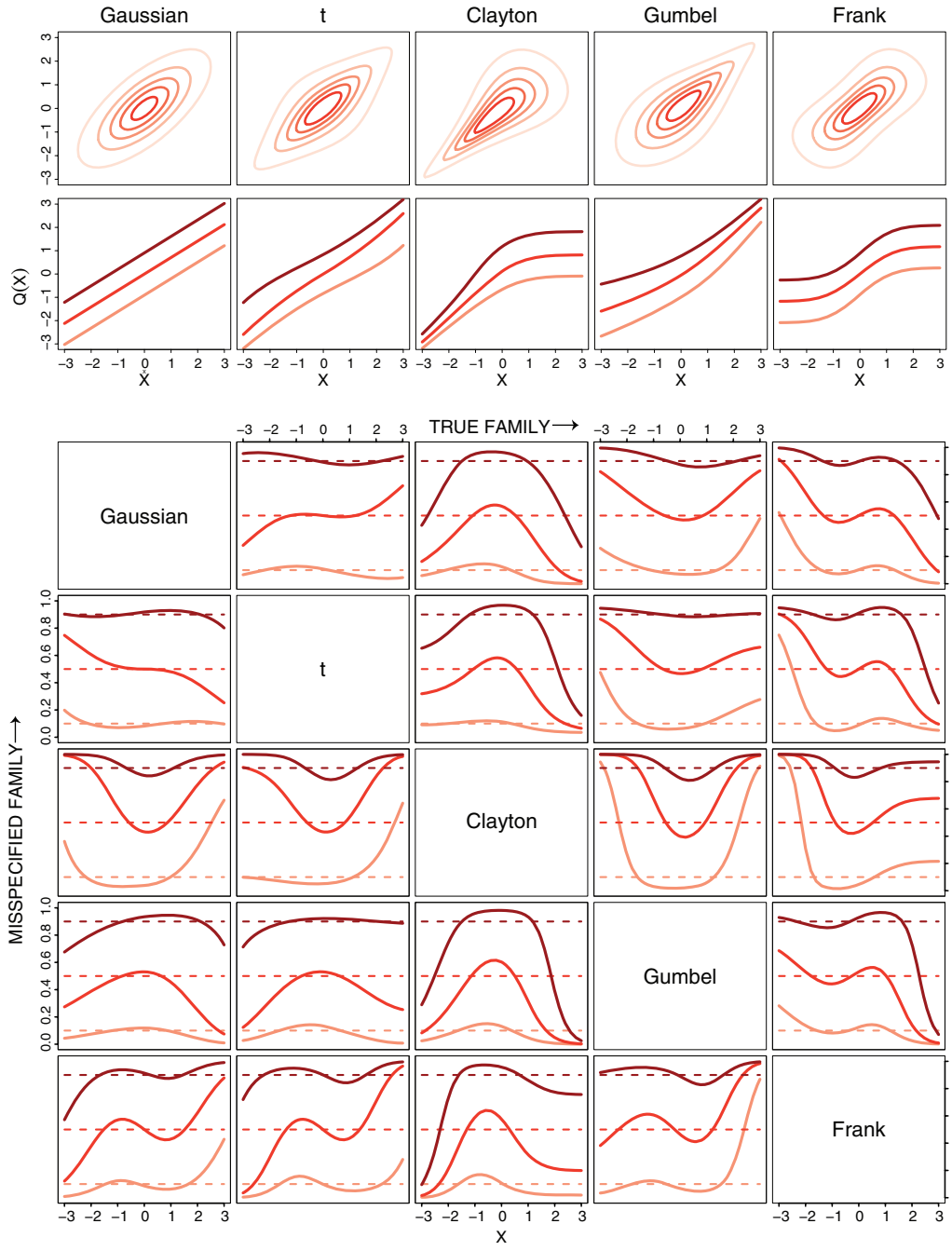

FIGURE 2: Contour plots and conditional quantile functions at $\alpha = 0.1, 0.5, 0.9$ (tint of lines from light to dark) of the five copula families, with standard normal margins, when $\tau = 0.5$. The matrix plot displays the true conditional percentiles (under the true model C_0) of misspecified conditional quantiles (under the misspecified model C_W^*) at $\alpha = 0.1, 0.5, 0.9$ levels (dashed lines). The diagonal entries specify the true family for each row and a misspecified family for each column.

DOI: 10.1002/cjs

acceptable for high (e.g., $Q_{0.90}(x)$), or low (e.g., $Q_{0.10}(x)$), conditional quantiles of the response variable at any x value. However, the conditional median, $Q_{0.50}(x)$, would be considerably underreported for small x and overreported for large x. Similarly, although the Gumbel copula has conditional expected prediction errors that are similar to those of the Gaussian copula in mean regression, using the Gumbel instead of the Gaussian leads to substantial overestimation of the conditional median for both small and large values of x.

Overall, the very different shapes of these curves for different α levels indicate that the impact of copula misspecification depends not only on the dependence patterns of the true and misspecified copulas, but also on the quantile function of interest and the covariate value. Similar comparisons for the cases with $\tau = 0.2$ and 0.8 (see Figures S2 and S3 in Appendix A of the Supplementary Material) further suggest that the discrepancies between the true and misreported conditional quantiles are usually more severe in cases of stronger dependence.

3. PREDICTION-BASED COPULA SELECTION

The question of how to choose a suitable copula family given the data has received a great deal of attention in the literature. Several model selection tools and goodness-of-fit tests have been developed for copula models, but almost all considered the in-sample adequacy of data fit as a suitable criterion; for instance, see the reviews in Berg (2009) and Genest, Rémillard, & Beaudoin (2009). The most commonly used copula model selection tool is the Akaike Information Criterion (AIC) (Akaike, 1974). Despite the increasing interest in using copula models for regression problems, there have been only a few attempts (e.g., Acar, Craiu & Yao, 2011) to use predictions as a tool for selecting a copula family.

In the following, we consider prediction-based copula selection criteria, calculated both purely in-sample and in a cross-validation setting. The selection criteria are evaluated on the accuracy of model identification and, concordantly with Shmueli (2010), on out-of-sample predictive performance. In light of the results in Section 2, the evaluation of model identification accuracy allows us to address whether the true family always yields the best predictive model in finite samples and when a misspecified copula offers a comparable predictive performance. On the other hand, the objective of out-of-sample evaluation is to contrast model selection aiming to find the correct model (e.g., AIC) with model selection aiming to minimize a sum of squares for the purpose of minimizing the prediction error out-of-sample.

3.1. Selection Criteria

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be independent and identically distributed with copula C and marginal distributions F_Y and F_X. Consider a set of candidate copula models $\{C_m(\cdot, \cdot; \theta_m) : m = 1, \ldots, M\}$, from which a selection is to be made based on the best predictive performance. One first fits each of the M copula models to the data. This is typically done in two stages. In the first stage, the marginal distributions F_Y and F_X are estimated either parametrically, by fitting an appropriate univariate distribution, that is, $\widehat{F}_Y(y) = F_Y(y; \widehat{\boldsymbol{\eta}}_1)$ and $\widehat{F}_X(x) = F_X(x; \widehat{\boldsymbol{\eta}}_2)$, or nonparametrically, using the rank transformations $\widehat{F}_Y(y) = \sum_{i=1}^{n} \mathbb{1}(Y_i \leq y)/(n+1)$ and $\widehat{F}_X(x) = \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)/(n+1)$, where $\mathbb{1}(\cdot)$ is the indicator function. In the second stage, the copula parameter(s) θ_m of the m^{th} family is estimated by maximizing the pseudo-copula likelihood. Denote by $\widehat{\theta}_m$ the resulting two-stage parametric estimate, also called the inference functions for margins (IFM) (Joe, 2005), or the two-stage semiparametric estimate (Genest, Ghoudi, & Rivest, 1995; Shih & Louis, 1995). The (in-sample) predictions from each fitted copula model $C_m(\cdot, \cdot; \widehat{\theta}_m)$, for $m = 1, \ldots, M$, are obtained using the conditional expectation formula

$$\widehat{Y}_{i(m)} = \int_{-\infty}^{\infty} y \; c_m(\widehat{F}_Y(y), \widehat{F}_X(X_i); \widehat{\theta}_m) \; d\widehat{F}_Y(y),$$

which can be calculated using numerical integration, or via the approximation (Noh, El Ghouch, & Bouezmarni, 2013)

$$\widehat{Y}_{i(m)} = \frac{1}{n+1}\sum_{j=1}^{n} Y_j \;\; c_m(\widehat{F}_Y(Y_j), \widehat{F}_X(X_i); \widehat{\theta}_m).$$

The selection criterion for in-sample predictive performance is based on the minimum sum of squared prediction errors (SSPE), defined as

$$\underset{m}{\text{argmin}}\ \text{SSPE}(C_m) = \underset{m}{\text{argmin}} \sum_{i=1}^{n} \{Y_i - \widehat{Y}_{i(m)}\}^2.$$

Since in-sample SSPE tends to favour models that overfit the data, we also consider a selection criterion based on cross-validation. Specifically, we used the predicted residual error sum of squares (PRESS) statistic, defined as

$$\underset{m}{\text{argmin}}\ \text{PRESS}(C_m) = \underset{m}{\text{argmin}} \sum_{i=1}^{n} \{Y_i - \widehat{Y}^{(-i)}_{i(m)}\}^2,$$

which is equivalent to leave-one-out cross-validation. The predictions $\widehat{Y}^{(-i)}_{i(m)}$ are calculated from the model $C_m(\cdot,\cdot;\widehat{\theta}^{(-i)}_m)$, which is obtained by leaving out the i^{th} data point. One may also use other validation methods, such as k-fold cross-validation.

The above criteria are defined for conditional mean predictions. When interest focuses on the conditional quantiles of the data, one may consider minimizing the quantile prediction error (QPE) (Koenker & Machado, 1999), defined as

$$\underset{m}{\text{argmin}} \sum_{Y_i \geq \widehat{Y}_{i(m;\alpha)}} \alpha \;\; | \; Y_i - \widehat{Y}_{i(m;\alpha)} \; | + \sum_{Y_i < \widehat{Y}_{i(m;\alpha)}} (1-\alpha) \;\; | \; Y_i - \widehat{Y}_{i(m;\alpha)} \; |,$$

where $\widehat{Y}_{i(m;\alpha)} = \widehat{Q}_{m;\alpha}(X_i)$ is obtained from the fitted model $C_m(\cdot,\cdot;\widehat{\theta}_m)$ for $i = 1, \ldots, n$. This criterion selects a model targeted to Q_α predictions. As in the mean prediction case cross-validation can be used to safeguard model selection from the issue of overfitting.

3.2. Evaluation of Model Identification Accuracy: Simulation Study

We conducted a simulation study to verify the results in Section 2 and to evaluate the prediction-based copula selection criteria in comparison to AIC. We considered the same five copulas, that is, Gaussian, $t_{(4)}$, Clayton, Gumbel and Frank, with weak ($\tau = 0.2$), moderate ($\tau = 0.5$) and strong ($\tau = 0.8$) dependence, and generated 1,000 bivariate samples of size $n = 500$ under each setting, using standard normal margins. For the case when $\tau = 0.5$, we also considered the lognormal scale for the response variable, via the exponential transformation on the same generated data. For brevity, we present the results under moderate dependence and standard normal margins here, and defer those under other settings to Appendix C in the Supplementary Material.

For each generated sample, we fitted all five copula families, using both two-stage parametric (IFM) and two-stage semiparametric (SP) estimation, and obtained the corresponding SSPE and PRESS values. The results obtained using known margins were almost indistinguishable from those under IFM; consequently, we have not reported the former results. Table 1 gives the proportion of times, out of 1,000 replications, that each candidate copula model has been selected using either SSPE or PRESS criteria. We also reported the proportion of times the true copula family has been selected by AIC. Since AIC uses the whole joint distribution to assess a given

model (instead of just the first two moments of the conditional distribution) it offers a much more accurate model selection tool which typically achieves a success rate that exceeds 90%.

From Table 1, we observe that while the Clayton, Frank and Gumbel copulas are easier to identify using a prediction-based selection tool, the Gaussian and t copulas are very difficult to distinguish based on their predictive performance. This is to be expected, due to the distinctive features of the Clayton, Frank and Gumbel copulas, and the close similarity of the Gaussian and t copulas, except in the tail regions. The modest discrepancy between the success rates of SSPE and PRESS indicates that there is no serious overfitting issue, except for a few cases when two-stage semiparametric estimation is employed (e.g., lower detection rates for the Gaussian and t copulas with PRESS). The results under the two estimation methods do not show any obvious pattern when SSPE is used, but using ranks in a cross-validated setting usually results in a lower success rate. Similar conclusions are reached under weak dependence (Table S4), under strong dependence with IFM (Table S5) and in the case of a lognormal response (Table S3). However, when the dependence is strong, predictive-based selection using two-stage semiparametric estimation shows almost no success in distinguishing the Gaussian copula from the t copula when evaluated on SSPE, and the t and Gumbel copulas from the Gaussian copula when evaluated on PRESS. These results suggest that copula misspecification among those choices may not alter predictions when dependence is strong. The minimum quantile prediction error (QPE) criterion exhibited almost the same model selection accuracy as SSPE and PRESS for $\alpha = 0.10$- and 0.90-level conditional quantiles, and showed a slightly better performance for median regression; see Table S2 of Appendix C of the Supplementary Material.

Based on Figure 1, we expect the conditional prediction error of the various copula families to differ mainly at covariate values in the tails. Hence, the SSPE and PRESS criteria can be

TABLE 1: The proportion of times out of 1,000 replications that each candidate copula model has been selected by the criteria using SSPE and PRESS statistic in the case of a normal response and a normal covariate when $\tau = 0.5$; either two-stage parametric (IFM) estimation or two-stage semiparametric (SP) estimation was employed. The last column gives the proportion of times the true copula family is selected by AIC. The numbers in bold represent the success rate.

	SSPE					**PRESS**					
	Gaussian	*t*	Clayton	Gumbel	Frank	Gaussian	*t*	Clayton	Gumbel	Frank	**AIC**
IFM											
Gaussian	**0.529**	0.445	0.000	0.016	0.010	**0.556**	0.418	0.000	0.016	0.010	**0.924**
t	0.273	**0.672**	0.000	0.051	0.004	0.262	**0.678**	0.002	0.053	0.005	**0.993**
Clayton	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	**1.000**
Gumbel	0.017	0.032	0.000	**0.951**	0.000	0.018	0.027	0.000	**0.955**	0.000	**0.996**
Frank	0.017	0.000	0.000	0.002	**0.981**	0.018	0.000	0.000	0.001	**0.981**	**0.990**
SP											
Gaussian	**0.416**	0.556	0.000	0.020	0.008	**0.202**	0.663	0.000	0.095	0.040	**0.895**
t	0.175	**0.768**	0.000	0.053	0.004	0.394	**0.512**	0.000	0.079	0.015	**0.980**
Clayton	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	**1.000**
Gumbel	0.007	0.032	0.000	**0.961**	0.000	0.009	0.034	0.000	**0.956**	0.001	**0.982**
Frank	0.025	0.000	0.000	0.000	**0.975**	0.006	0.002	0.000	0.002	**0.990**	**0.987**

modified to evaluate candidate copula models (trained using all data points) on their predictive performance only in those regions. However, our findings concerning the use of different tail regions (not reported) did not suggest an improvement in the model selection accuracy over the results reported here.

3.3. Evaluation of Out-of-Sample Prediction Performance: Simulation Study

In Table 1, AIC is clearly the best method of identifying the correct model. However, since the true data-generating process is almost never a subset of the candidate models, one might ask whether the success of AIC translates to superior out-of-sample predictive performance. Another question is whether any of the copula models offer added predictive value compared to the use of simple linear regression.

To address these questions, we used the original five sets of 1,000 samples as training data, and generated 1,000 new (test) samples of size $n = 500$ under each copula family. The best AIC and PRESS models were selected in-sample, and then evaluated on the out-of-sample prediction performance. Our performance metric was the prediction error of AIC relative to PRESS, calculated as

$$\left\{ \mathrm{SSPE}^{(\mathrm{R})}_{\mathrm{AIC}} - \mathrm{SSPE}^{(\mathrm{R})}_{\mathrm{BP(S)}} \right\} \Big/ \mathrm{SSPE}^{(\mathrm{R})}_{\mathrm{BP(S)}} \,,$$

where BP(S) stands for best predictive model targeted to the region $X \in \mathrm{S}$, and the identifier (R) indicates that out-of-sample errors were calculated only for the region $X \in \mathrm{R}$. In our evaluations, we considered the regions (both S and R) of all covariate values and of those fall below or above the cut-off values $z = 1.645$ and $z = 1.96$, representing 10% and 5% total tail area under the standard normal distribution. We calculated the relative error of the linear regression model in the same way. Note that the relative error allows us to assess the comparative predictive advantage of different models for each given sample, which would be masked in an overall assessment of prediction errors; see, for instance, Figures S8 and S9 in Appendix C of the Supplementary Material.

The upper panel of Figure 3 displays the error in using the AIC-selected model relative to the BP(S) copula model, across different regions under each copula family, when the two-stage semiparametric estimation is used. Similar assessments for the two-stage parametric estimation and under other settings can be found in Appendix C of the Supplementary Material. For better visibility, the boxplots have been truncated at 100% relative error. The number of truncated samples (if any) is reported above each boxplot.

In almost all settings that we considered, the relative error of AIC was centered around zero, and often within the truncation levels, verifying that AIC is a reasonable tool to select a model for both explanatory and predictive purposes. Only when two-stage semiparametric estimation was used under the t and Gumbel copulas with strong dependence AIC-based model selection yielded considerably inferior predictive performance in out-of-samples compared to a prediction-based model selector, which tended to prefer the Gaussian copula.

The lower panel of Figure 3 displays the error in using the simple linear regression model relative to the BP(S) copula model for the two-stage semiparametric estimator when $\tau = 0.5$. Results for the two-stage parametric estimator and under other settings may be found in Appendix C of the Supplementary Material. For better visibility, the boxplots have been truncated at 200% relative error. In most settings under the Clayton, Gumbel and Frank copulas, the BP(S) copula model has a clear predictive advantage over simple linear regression, especially when the margins are estimated parametrically and/or predictions are targeted to tail regions. This justifies the utility of copulas for prediction when the dependence between response and covariate is asymmetric and/or non-elliptical. Only under the Gumbel copula with strong dependence did

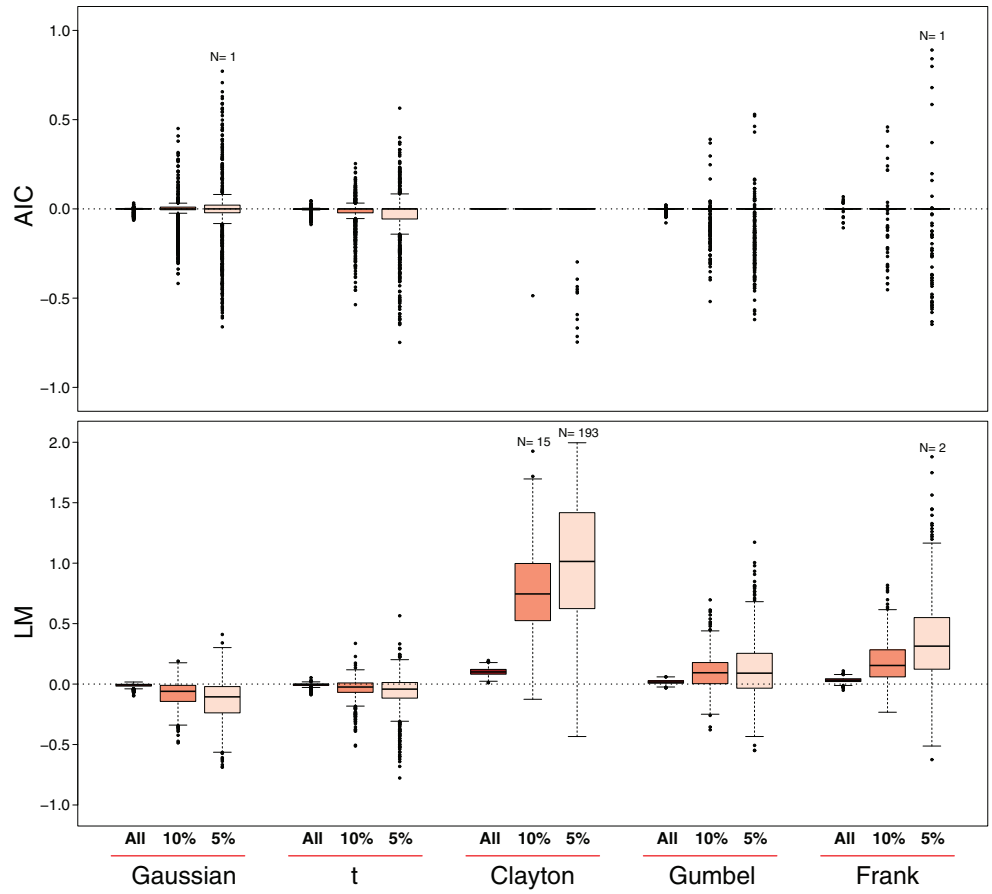

FIGURE 3: Boxplots showing the upper-truncated distributions of the test error rates of an AIC-selected copula model (upper panel) and a linear model (lower panel), relative to the best predictive model under the five copulas (x-axis) with standard normal margins when $\tau = 0.5$. Models under each setting were trained by the two-stage semiparametric (SP) estimation using a set of 1,000 samples of size $n = 500$ and considering all five copula families as well as the linear model. Reported above each copula family name are the relative error rates evaluated on regions of all test covariate values (All) and test extreme covariates determined using the cut-off values $z = 1.645$ (10%) and $z = 1.96$ (5%). The numbers reported above selected boxplots represent the numbers of test samples above the truncation levels 1 (top) and 2 (bottom).

the semiparametric copula regression exhibit inferior out-of-sample prediction performance compared to linear regression. The same result held for the Gaussian and t copulas under this setting. In fact, except for the case with a lognormal response, there was no clear advantage in using copula-based regression when the underlying dependence was elliptical.

4. PREDICTIVE ASSESSMENT OF VINE COPULA MODELS

Having examined in detail the predictive aspects of copula modelling in the bivariate case, we now extend our discussion to multivariate copula models.

A major challenge that comes with the dimensionality of a problem is the construction of a simple yet highly informative model. While traditional multivariate (copula) models such as the Gaussian (copula) have a simple form, as well as a well-developed theory, they may fail to reflect certain data features (e.g., asymmetry). On the other hand, more complex models may be informative but difficult to interpret. Constructed sequentially from bivariate copulas,

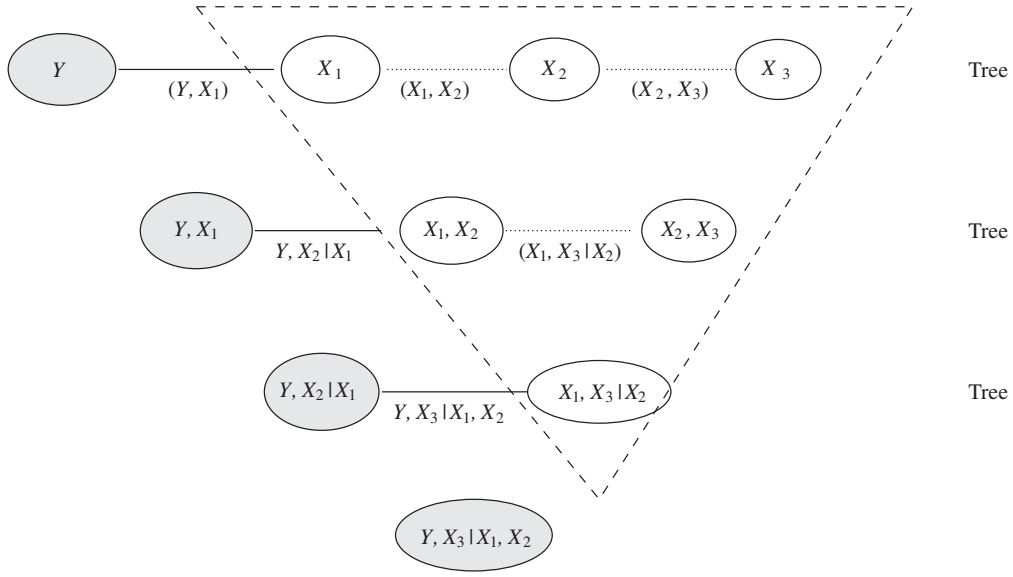

FIGURE 4: A tree representation of a four-variate D-Vine with the response variable in the first node.

vine copula models enjoy both properties. In particular, their ease of interpretation comes from the graphical models, called vines, which give an attractive explanatory tool to unravel the dependence relationships in multivariate data (see Figure 4).

There has been considerable recent research on the construction and properties of vines (Bedford & Cooke, 2002; Kurowicka & Joe, 2010), and accompanying inference methods (Aas et al., 2009; Min & Czado, 2010; Hobæk Haff, 2013). However, the consideration of vine copulas for regression analysis is relatively recent (Cooke, Joe, & Chang, 2015; Kraus & Czado, 2017), with potential impact on future developments.

In this section, we outline the assessment of different vine copula models from a predictive perspective. Despite (and perhaps due to) their richness, available model selection tools for vine copulas are scarce, mostly following the idea of the AIC-based maximum spanning tree algorithm in Dißmann et al. (2013). Hence, the selection of a suitable predictive vine model among all (or a large number of) possible vine constructions is not feasible using the approach outlined in Section 3. Instead, we are obliged to consider a small set of candidate models and compare their predictive abilities.

The difference between two vine copula models is typically assessed using the KL divergence. However, such comparisons may not be suitable for suggesting a predictive vine model in the regression context. On the other hand, sequential model constructions via variable (covariate) selection may better serve the purpose (Kraus & Czado, 2017).

Here, we focus on D-vines of a special form, in which the response variable Y occupies the first (or last) node, followed (or preceded) by the covariate vector $\mathbf{X} = (X_1, \ \ldots, X_p)$. An example of this special vine when $p = 3$ is displayed in Figure 4. The grey nodes represent partial regression relationships, in the order of importance of the covariates, and the white nodes (inside the dashed triangle) describe co-dependencies (i.e., multicollinearity) in $\mathbf{X}$. Each edge connecting two nodes displays the (conditional) bivariate relationship between the node variables (conditioned on the shared node).

Let $U = F_Y(Y)$ and $V_i = F_{X_i}(X_i)$, for $i = 1, 2, 3$. Dropping the arguments, the copula density c of $(Y, \mathbf{X})$ can be factored as

$$c = c_{(v,u_1)} \times c_{(u_1,u_2)} \times c_{(u_2,u_3)} \times c_{(v,u_2|u_1)} \times c_{(u_1,u_3|u_2)} \times c_{(v,u_3|u_1,u_2)}.$$

TABLE 2: D-vine models of the four examples considered in the simulation study, along with the specific objective of model comparison. Reported are the bivariate copula families and Kendall's τ (in parentheses) under each setting.

Example	Tree 1			Tree 2		Tree 3	Objective
	c_{Y,X_1}	c_{X_1,X_2}	c_{X_2,X_3}	$c_{Y,X_2\|X_1}$	$c_{X_1,X_3\|X_2}$	$c_{Y,X_3;X_1,X_2}$	
Gaussian—High	N(0.6)	N(0.4)	N(0.3)	N(0.3)	N(0.1)	N(0.1)	RVine*
Gaussian—Low	N(0.1)	N(0.2)	N(0.3)	N(0.3)	N(0.4)	N(0.6)	RVine*
Mixed—High	G(0.6)	G(0.4)	G(0.3)	C(0.3)	C(0.1)	F(0.1)	LM
Mixed—Low	G(0.1)	G(0.2)	G(0.3)	C(0.3)	C(0.4)	F 0.6)	LM and RVine*

N = Gaussian, C = Clayton, G = Gumbel, F = Frank.

Together with Sklar's theorem (Sklar, 1959), this factorization defines the joint distribution of $(Y, \mathbf{X})$. As we discussed previously in Section 2, the conditional features of Y given $\mathbf{X}$ can be derived from the joint distribution.

In general, the conditional expectation of $Y|\ \mathbf{X} = \mathbf{x}$ is given by

$$\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} y \frac{c(F_Y(y), F_{\mathbf{X}}(\mathbf{x}))}{\int_{-\infty}^{\infty} c(F_Y(y), F_{\mathbf{X}}(\mathbf{x}))dF_Y(y)}\ dF_Y(y),$$

where $F_{\mathbf{X}}$ denotes the p-dimensional distribution of the covariate vector $\mathbf{X}$. In Figure 4, the latter corresponds to the trivariate D-vine displayed inside the dashed triangle. For independent and identically distributed random vectors, one can approximate the conditional expectation by replacing the integrals with sums (Noh, El Ghouch, & Bouezmarni, 2013). However, for non-i.i.d. samples, numerical integration is required.

Once mean predictions are obtained from each (fitted) vine model among a set of candidates, we can evaluate their relative prediction errors using the approach in Section 3. In the following, we briefly illustrate these aspects on both simulated and real data.

4.1. Evaluation of Predictive Performance: Simulation Study

We considered four examples using the four-dimensional D-vine presented in Figure 4. Table 2 summarizes the vine models that consist of the bivariate copula families at each tree, together with their dependence parameters (in Kendall's tau scale). We also list the specific objective under each considered scenario in terms of the primary model being assessed. Specifically, we considered the linear model (LM) and the best regular vine model (RVine*) selected by the algorithm in Dißmann et al. (2013).

The first two examples fall under the Gaussian distribution and, despite multicollinearity, offer a desirable setting for LM. In these examples, we primarily checked whether vine copula regression provided any added value over the (multiple) linear regression model. The latter two examples were constructed using the Gumbel, Clayton and Frank copulas; hence, these examples aim to assess the error in using a linear regression model over vine regression. The models are labelled as ''High'' and ''Low'', reflecting their strength of dependence in the first tree. Since the algorithm for RVine* selects a model that yields the highest dependencies in the first tree, the ''High'' cases define a favourable setting for RVine*, while the ''Low'' cases aim to assess its robustness in predictions in a least favourable setting for the algorithm.

Under each scenario, we generated 1,000 training and 1,000 test samples of size $n = 500$ from the specified models, with standard normal margins. We considered both parametric (with

normal margins) and semiparametric (with ranks) estimation when fitting RVine* and the vine model with the true specification. The latter is included for out-of-sample predictive assessment. The reported results are based on the semiparametric estimation.

For the in-sample predictive assessment of the fitted models under each setting, we used the corresponding data generating D-Vine model as the benchmark. We also compared the test sample predictions with this benchmark. Each fitted model is then evaluated on (i) regions of all covariate values in the test sample, and (ii) regions having at least one extreme covariate value, that is, $|X_i| > z$, where $z = 1.645$ (10%) and 1.96 (5%) for at least one index $i = 1, 2, 3$. We compared the error rates of RVine* and LM, relative to the fitted D-Vine, across these regions under each example. The results under the third example, that is, Mixed−High, are displayed in Figure 5, while those under the other three examples can be found in Appendix D of the Supplementary Material.

Our findings in these investigations suggest that when the underlying distribution is Gaussian, the prediction results under the three fitted models were very similar, and close to those under the true model, and this held regardless of whether the dependencies in the tree structure were ordered from high to low, or from low to high. The comparisons of the relative error rates on test samples further imply that in tail regions of covariates (with at least one extreme value), the linear model outperforms copula-based mean regression, especially in comparison to RVine*. This result may be due to some of the selected copula components in the vine structure being very different from the Gaussian copula. Hence, copula-based regression is not advantageous for prediction purposes in such settings.

Under the vine structures with mixed bivariate copula families, the order of the strength of dependencies in the tree structure affects the predictive performance of RVine* in relation to its distance from the true model. A comparison of the scatterplots in Figure 5 (see also Appendix D of the Supplementary Material) suggests that RVine* may suffer from overfitting, especially in its ideal scenario, where dependencies are the highest in the first tree. The relative error rates are only slightly higher when calculated over the entire covariate region. However, the error inflation can be as high as 80% in the tail regions of the covariates. When the underlying vine has the lowest dependencies in the first tree, both RVine* and LM fail to produce reliable predictions, with error inflation of up to 200% and 400%. The inferior performance of LM under these two examples suggests that copula-based regression offers a much higher predictive accuracy in situations where the data exhibit asymmetric dependence among some pairs of covariates and/or the response.

4.2. Evaluation of Predictive Performance: Example

As a practical illustration of the predictive assessment of vine copulas, we considered the abalone dataset (Blake & Merz, 1998), which contains the age and various physical measurements collected from 4,117 abalones. With its skewed marginal distributions and highly asymmetric dependence structure, this dataset exhibits clear violations of normality, hence is ideal for demonstrating the advantages of using copulas (Ma & Sun, 2008; Hobæk Haff et al., 2016; Liu et al., 2017).

Our aim was to predict the age of abalone, which ranges between 1 and 29, from the physical measurements. We selected a sub-sample of size 800 from the male and female abalones as the training sample, and formed 10 sets of size 200 from the remaining observations to serve as the test samples. Using the training data, we then fit a D-Vine model as specified in Figure 4, where the covariates were ordered based on the magnitudes of their correlations with the response. We also fitted other vine structures (R-Vine, C-Vine and Gaussian C-Vine) as selected by the algorithm in Dißmann et al. (2013). When fitting these models, we used the rank transformation due to the difficulty in specifying an appropriate parametric model for each variable (see Figure S19 in Appendix D of the Supplementary Material). The out-of-sample prediction performance of the fitted vine models were evaluated in comparison to each other and to the linear model.

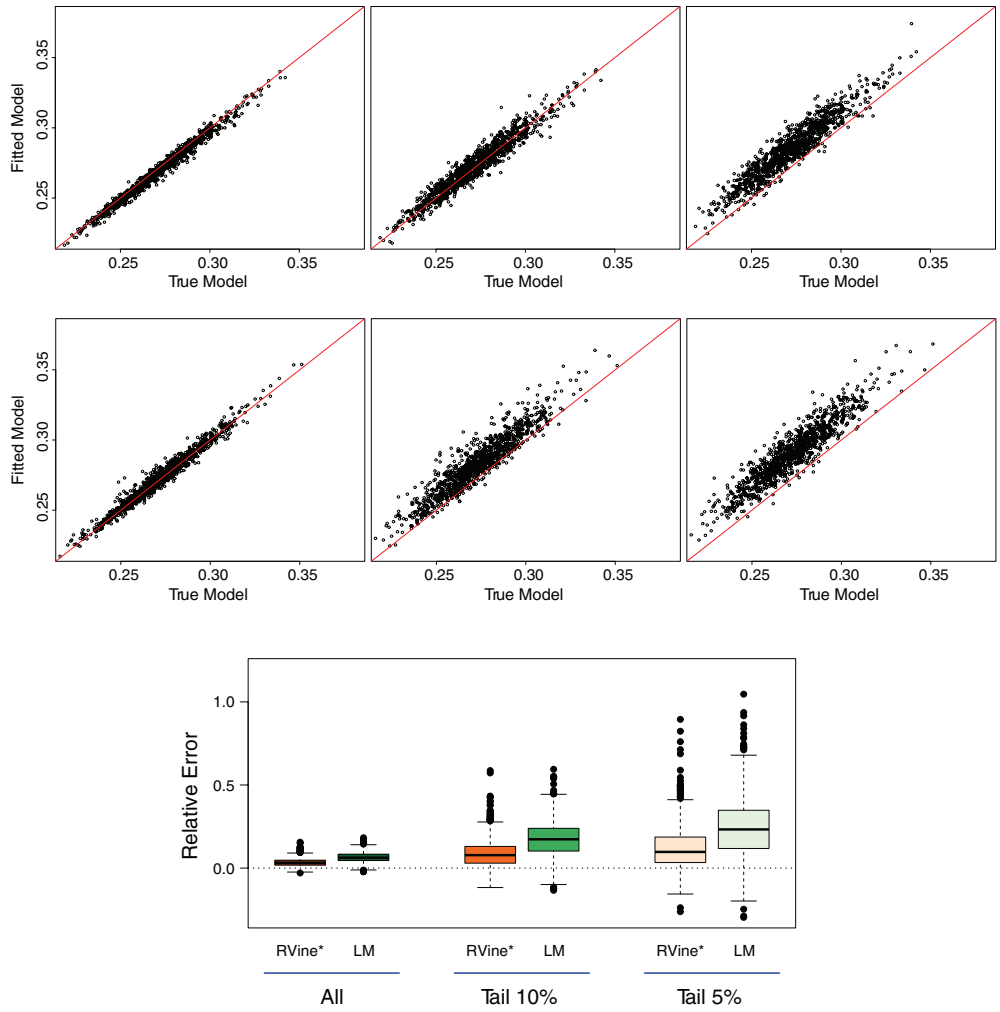

FIGURE 5: Scatterplots of in-sample (upper panel) and out-of-sample (middle panel) mean square prediction errors of the three fitted models, D-Vine (first column), RVine* (second column), and LM (last column) compared to the D-Vine in the Mixed−High example. Boxplots (lower panel) display the distributions of the test error rates of RVine* and LM, relative to the fitted D-Vine. The relative error rates were evaluated on regions of all covariate values (All) and regions having at least one extreme covariate value. The latter are determined based on the cutoff values $z = 1.645$ (10%) and $z = 1.96$ (5%).

We considered all test sample predictions, as well as predictions for regions having at least one extreme covariate value, determined based on the 5% and 10% tail regions of each covariate. Table 3 reports the mean and standard deviation of the mean square prediction errors over 10 test samples. The out-of-sample mean square prediction errors of the considered vine models were almost indistinguishable for these data, and despite the evident asymmetries, only slightly better than the linear model.

5. CONCLUSION

In this article, we have examined the predictive utility of copula models in comparison to their explanatory merits. Our detailed evaluations of the conditional mean and quantiles for various bivariate copula families, using both theoretical and numerical assessments, suggest that some

TABLE 3: Out-of-sample average (over 10 test samples) mean square prediction errors (standard deviations) of D-Vine, RVine*, CVine*, GCVine* (Gaussian C-Vine) and LM, evaluated on regions of all test covariate values (All) and extreme test covariates that were determined using 10% and 5% tail regions.

	D-Vine	**RVine***	**CVine***	**GCVine***	**LM**
Male					
All	11.7274	11.7282	11.7277	11.7275	11.8305
	(0.1622)	(0.1623)	(0.1623)	(0.1621)	(0.1672)
Tail 10%	13.8152	13.8204	13.8286	13.8217	14.2613
	(0.5041)	(0.5029)	(0.5067)	(0.5038)	(0.5451)
Tail 5%	14.8485	14.8437	14.8692	14.8597	15.3352
	(0.7373)	(0.7345)	(0.7477)	(0.7454)	(0.9151)
Female					
All	12.0405	12.0407	12.0415	12.0404	12.1706
	(0.1771)	(0.1769)	(0.1774)	(0.1770)	(0.1824)
Tail 10%	13.8949	13.8966	13.9032	13.8926	14.7032
	(0.5073)	(0.5081)	(0.5170)	(0.5095)	(0.9540)
Tail 5%	14.0193	14.0208	14.0237	14.0160	14.6534
	(0.5236)	(0.5237)	(0.5279)	(0.5292)	(0.5441)

copula families are more difficult to distinguish in terms of their overall predictive power than others, and depending on the quantity of interest, the differences in predictions can be detected only in some extreme events.

In situations where targeted predictions are of interest, such as those associated with extreme covariate regions, we have proposed a prediction-based copula selection approach in contrast to a copula model selected based on overall fit. The issue of overfitting was addressed using the PRESS statistic for validation. We further assessed the out-of-sample predictive performance of AIC as a successful, if not superior, model selector and the linear regression model as the familiar standard. Similar evaluations were performed in higher dimensions using vine copulas.

Our findings confirm the merits of AIC in both model identification and out-of-sample predictions. However, in some cases, for example, for a skewed response and/or when the dependence is strong, a model selected based on prediction criteria can be advantageous. When the dependence structure is elliptical, there is almost no gain in using copula regression for out-of-sample predictions, except for cases where the response variable is highly non-normal. On the other hand, when the dependence between response and covariate is asymmetric and/or non-elliptical, copula regression provides a clear predictive advantage over a linear regression model, especially when predictions are targeted to extreme covariate regions.

In our evaluations we considered both two-stage parametric and two-stage semiparametric estimation when training the models, and observed a slightly inferior out-of-sample predictive performance in the latter case. Under parametric marginal models, test samples are usually placed accurately in their respective percentiles (uniform scale). However, the use of ranks may result in an incorrect recovery of percentiles and hence may lead to erroneous predictions, especially in the tail regions. When the dependence is strong, semiparametric copula regression shows an inferior predictive performance to the linear regression model under the Gaussian, *t* and

Gumbel copulas, which all admit a linear conditional mean function. In the absence of a suitable parametric model, we therefore recommend using semiparametric copula regression with care when dependence between the response and covariates is strong.

Through numerical examples, we investigated situations where a regular vine copula model selected by the algorithm in Dißmann et al. (2013) and traditional multiple linear regression fail to provide reliable mean predictions. Considering the recent interest in the use of (vine) copula models for regression analysis, our evaluations of these situations may inform the design of such tools in the future. In particular, based on our assessments of the tail regions of covariates, robust copula-based regression may offer an interesting avenue for future research.

While copula-based regression has been considered mostly in complete data settings, it can be extended to settings involving censored data. A recent work in this direction is De Backer, El Ghouch, & Van Keilegom (2017), which described semiparametric copula quantile regression. Since copulas are almost indistinguishable in terms of their predictive power except in the tail regions, and since censoring causes loss of information in the tails, we expect that model identification will be a challenging problem and therefore worth investigating.

ACKNOWLEDGEMENTS

The authors thank the guest editor, Mary Thompson, the associate editor and three anonymous referees for their careful review and valuable suggestions, and Martin Lysy for fruitful discussions of the article. Acar acknowledges research support from the Canadian Statistical Sciences Institute (CANSSI) Collaborative Research Team Project and the Natural Sciences and Engineering Research Council of Canada (Grant 435943-2013).

BIBLIOGRAPHY

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

Acar, E. F., Craiu, R. V., & Yao, F. (2011). Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics*, 67, 445–453.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Bedford, T. & Cooke, R. M. (2002). Vines—A new graphical model for dependent random variables. *The Annals of Statistics*, 30, 1031–1068.

Berg, D. (2009). Copula goodness-of-fit testing: An overview and power comparison. *European Journal of Finance*, 5, 675–701.

Bernard, C. & Czado, C. (2015). Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, 138, 104–126.

Blake, C. L. & Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/mlearn/ML.Repository.html.

Bouyé, E. & Salmon, M. (2009). Dynamic copula quantile regressions and tail area dynamic dependence in forex markets. *European Journal of Finance*, 15, 721–750.

Cooke, R. M., Joe, H., & Chang, B. (2015). Vine regression. Resources for the future discussion paper 15-52. http://papers.ssrn.com/abstract=2695063.

Crane, G. & van der Hoek, J. (2008). Conditional expectation formulae for copulas. *Australian & New Zealand Journal of Statistics*, 50, 53–67.

De Backer, M., El Ghouch, A., & Van Keilegom, I. (2017). Semiparametric copula quantile regression for complete or censored data. *Electronic Journal of Statistics*, 11, 1660–1698.

Dette, H., Van Hecke, R., & Volgushev, S. (2014). Some comments on copula-based regression. *Journal of the American Statistical Association*, 109, 1319–1324.

Dißmann, J., Brechmann, E., Czado, C., & Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59, 52–69.

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

Genest, C., Rémillard, B., & Beaudoin, D. (2009). Omnibus goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44, 199–213.

Hobæk Haff, I. (2013). Parameter estimation for pair-copula constructions. *Bernoulli*, 19, 462–491.

Hobæk Haff, I., Aas, K., Frigessi, A., & Lacal, V. (2016). Structure learning in Bayesian networks using regular vines. *Computational Statistics & Data Analysis*, 101, 186–208.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94, 401–419.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman & Hall, London.

Koenker, R. & Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94, 1296–1310.

Kraus, D. & Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis*, 110, 1–18.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86.

Kurowicka, D. & Joe, H. (Eds.) (2010). *Dependence Modeling: Handbook on Vine Copulae*. World Scientific Publishing, Singapore.

Leong, Y. & Valdez, E. A. (2005). Claims prediction with dependence using copula models. Technical Report. University of New South Wales, Sydney.

Liu, H., Ju, Z., Ji, X., Chan, C. S., & Khoury, M. (2017). Fuzzy empirical copula for estimating data dependence structure. In *Human Motion Sensing and Recognition. Studies in Computational Intelligence*, vol. 675. Springer, Berlin, Heidelberg, pp. 123–145.

Ma, J. & Sun, Z. (2008). Dependence structure estimation via copula. https://arxiv.org/abs/0804.4451

Min, A. & Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8, 511–546.

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer-Verlag, New York.

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

Noh, H., El Ghouch, A., & Van Keilegom, I. (2015). Semiparametric conditional quantile estimation through copula-based multivariate models. *Journal of Business & Economic Statistics*, 33, 167–178.

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

Shmueli, G. (2010). To explain or to predict. *Statistical Science*, 25, 289–310.

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

Spanos, A. (1994). On modeling heteroskedasticity: The Student's t and elliptical linear regression models. *Econometric Theory*, 10, 286–315.

Sungur, E. (2005). Some observations on copula regression functions. *Communications in Statistics—Theory and Methods*, 34, 1967–1978.

Received 31 May 2017
Accepted 9 April 2018