Predictive Comparison of Vine Copula Models

University of Manitoba

Md. Erfanul Hoque and Elif F. Acar

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

Introduction

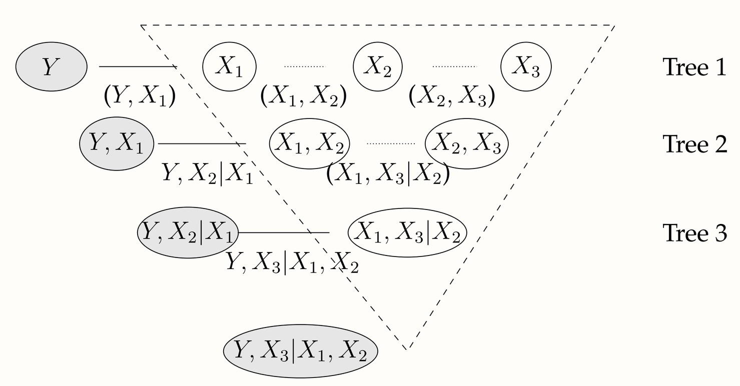
- Vine copulas are a popular tool for the flexible and tractable specification of high dimensional joint densities for representing multivariate data.
- There are many possible vine constructions: D-Vines, C-Vines and more general R-Vines.
- In recent years, vine copula models have been considered in regression contexts to predict conditional mean and conditional quantiles of a variable of interest given the other variables.

Objectives:

- Assess the predictive performance of different vine copula models.
- Compare the predictive utility of vine copula regression over classical regression models.

Vine Copula Regression

- Tree representation of a four-variate D-vine with response variable as the first node:



- The conditional expectation of $Y \mid \mathbf{X} = \mathbf{x}$:

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

- It can be approximated by replacing the integrals with sums [Noh et al., 2013]:

$$\hat{E}(Y_i \mid \mathbf{X} = \mathbf{x}) = \sum_{j=1}^{n} Y_j \frac{c(\hat{F}_Y(Y_j), \hat{F}_{\mathbf{X}}(\mathbf{x}))}{\sum_{j=1}^{n} c(\hat{F}_Y(Y_j), \hat{F}_{\mathbf{X}}(\mathbf{x}))}.$$

- **Two-stage sequential estimation**:
  1. estimate the marginal distributions F_Y and F_j of Y and X_j, $j = 1, \ldots, d$, either parametrically or nonparametrically, using the rank transformations.
  2. sequentially estimate the copula parameters of pair copulas in the vine construction.
- **Model selection** is a challenging problem in vine copula models. The AIC-based maximum spanning tree algorithm (Dißmann algorithm) is commonly used.

Simulation Study

Simulation Settings

- Generate 1000 training and 1000 test samples, of size $n = 500$ with standard normal margins, under each D-Vine scenario.

Example	Tree 1			Tree 2		Tree 3
	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 (0.81)	N (0.59)	N (0.45)	N (0.45)	N (0.16)	N (0.16)
Gaussian – Low	N (0.16)	N (0.31)	N (0.45)	N (0.45)	N (0.59)	N (0.81)
Mixed – High	G (2.50)	G (1.67)	G (1.43)	C (0.86)	C (0.22)	F (0.91)
Mixed – Low	G (1.11)	G (1.25)	G (1.43)	C (0.86)	C (1.33)	F (7.93)
Mixed – Non-simplified	G (2.5)	G (1.75)	G (1.50)	C ($\theta_1(x_1)$)	C ($\theta_2(x_2)$)	F ($\theta_3(x_1, x_2)$)

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

where $\theta_1(x_1) = \exp[\log(0.7) + 0.4X_1]$, $\theta_2(x_2) = \exp[\log(0.3) + 0.6x_2]$, $\theta_3(x_1, x_2) = 0.5 + 5x_1 + (-2)x_2$.

- "High" and "Low" cases reflect the strength of dependence in the first tree.
- We consider both parametric (with $N(0,1)$ margins) and semi-parametric (with ranks) estimation.
- In predictive comparisons, we use the data generating D-vine model as the benchmark.
- The reported results are based on the semi-parametric approach.
- We calculate mean square prediction errors for in-sample and out-of-sample conditional mean predictions.
- Scatterplots show the mean square prediction errors of the three fitted models: D-Vine (first column), RVine* (second column), LM (last column) in **Mixed–High** example.

In-sample Predictive Performance

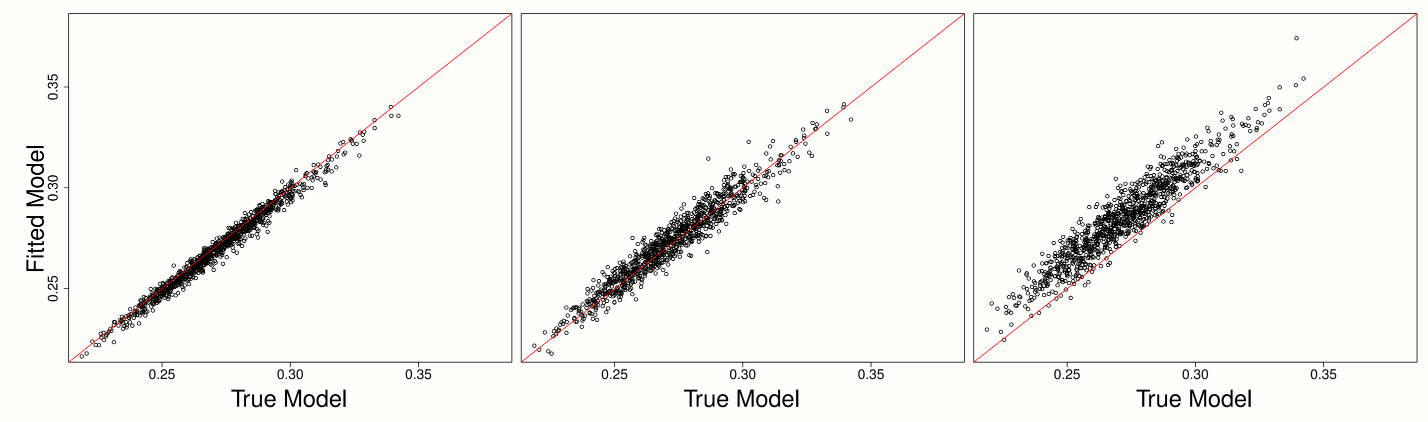


Table 1: Average in-sample mean square prediction errors of all models for each scenarios

Example	True	Fitted	RVine*	LM
Gaussian – High	0.299	0.297	0.295	0.297
Gaussian – Low	0.267	0.267	0.266	0.265
Mixed – High	0.272	0.270	0.273	0.285
Mixed – Low	0.267	0.271	0.311	0.387
Mixed – Non-simplified	0.217	0.217	0.269	0.290

RVine* is the best regular vine model selected by Dißmann algorithm

Out-of-sample Predictive Performance

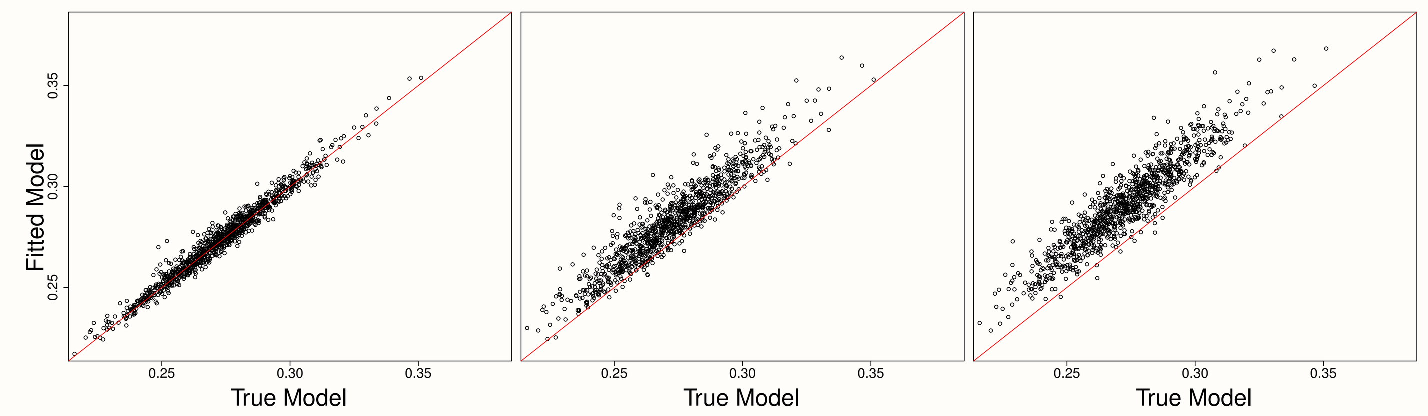


Table 2: Average out-of-sample mean square prediction errors of all models for each scenarios

Example	True	Fitted	RVine*	LM
Gaussian – High	0.299	0.300	0.303	0.302
Gaussian – Low	0.270	0.272	0.274	0.270
Mixed – High	0.272	0.274	0.283	0.291
Mixed – Low	0.274	0.277	0.328	0.397
Mixed – Non-simplified	0.216	0.220	0.230	0.296

RVine* is the best regular vine model selected by Dißmann algorithm

Conclusions

- When the underlying distribution is Gaussian, mean prediction results are very similar for the three fitted models, and close to those under the true model.
- The order of strength of dependencies in the tree structure affects the predictive performance of RVine* in relation to its distance from the true model.
- RVine* and LM may have over-fitting problem under some cases.
- When the simplifying assumption is violated, RVine* and linear model have inferior predictive performance compared to true or fitted non-simplified vine models.

Data Examples

Example 1: Body Fat Data

- Dataset contains the percent of body fat and various physical measurements on 252 men.
- **Response**: percentage of body fat (ranges from 0 to 40.10 with mean 18.85)

  Covariates: BMI and five body circumferences: Neck, Chest, Abdomen, Hip, Thigh.
- After identifying 4 outliers, we consider 200 as a training sample and 48 as test sample.
- We compare the out-of-sample prediction performance of different vine models.

Table 3: Out-of-sample mean square prediction errors for Body Fat data

	R-vine	C-vine	D-vine	Gaussian C-vine	LM
Training data	19.113	19.112	19.114	19.114	19.115
Test data	20.495	20.477	20.490	20.470	20.466

Example 2: Abalone data

- Dataset contains the age of abalone and various physical measurements on 4117 abalone.
- **Response**: Rings (ranges from 3 to 29 with mean 11)

  Covariates: Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight.
- We consider male and female abalones and select a subset of size 1000 from each group, with 880 in the training sample and 120 in the test sample.
- We compare the out-of-sample prediction performance of different vine structures with the prediction under linear model.

Table 4: Out-of-sample mean square prediction errors of vine prediction under different models for Abalone data

		R-vine	C-vine	D-vine	Gaussian C-vine	LM
Male	Training data	10.114	10.114	10.114	10.114	10.215
	Test data	10.476	10.486	10.476	10.477	10.606
Female	Training data	10.417	10.417	10.417	10.417	10.529
	Test data	10.563	10.567	10.563	10.562	10.665

Conclusions

- **Body Fat**: Vine copula regression does not offer significant improvement in mean predictions of body fat over linear regression.
- **Abalone**: Vine copula regression performs better in predicting the age of male and female abalone over linear regression.

References

[1] Noh H., El Ghouch A. and Bouezmarni T. [2013]; *Copula-Based Regression Estimation and Inference*, Journal of American Statistical Association, 108, 676 - 688.

[2] Cooke R. M., Joe H. and Chang, B. [2015]; *"Vine Regression," Resources for the Future* Discussion Paper.

[3] Kraus D. and Czado C. [2017]; *D-vine copula based quantile regression*, Computational Statistics & Data Analysis, 110.

[4] Dißmann J., Brechmann E., Czado C. and Kurowicka D. [2013]; *Selecting and estimating regular vine copulae and application to financial returns*, Computational Statistics & Data Analysis, 59, 52 - 69.

[5] Aas K., Czado C., Frigessi A. and Bakken H. [2009]; *Pair-copula constructions of multiple dependence*, Insurance: Mathematics and Economics, 44, 182 - 198.

Acknowledgement

We would like to thank the Faculty of Science and Department of Statistics of the University of Manitoba for the financial support. Also we would like to thank Centre de recherches mathmatiques (CRM) for the financial support to attend the workshop.

Contact information: `hoqueme@myumanitoba.ca`