C- and D-vine based quantile regression

Marija Tepegjozova
m.tepegjozova@tum.de
TU München

Claudia Czado (TU München)
Gerda Claeskens (KU Lueven)
Jing Zhou (KU Lueven)

Motivation

- Usual drawbacks of the standard models for quantile regression, are distributional assumptions, quantile crossings and misspecification of the tail dependencies.
- We would like to offer more flexible method that can overcome many drawbacks.
- Vine copulas might be the solution, as they offer highly flexible modeling of high dimensional dependence structures.

Copulas

Definition

A $d-$dimensional copula C is a multivariate distribution function on the $d-$dimensional unit hypercube $[0,1]^d$ with uniformly distributed marginals.

Theorem (Sklar's Theorem)

Let $\mathbf{X} := (X_1, ..., X_d)^T$ be a $d-$dimensional random vector with joint distribution function F and marginal distribution functions F_i, $i = 1, ..., d$, then the joint distribution function can be expressed as

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)).$$

Vine based quantile regression

The conditional quantile function for $\alpha \in (0, 1)$ and a continuous response variable Y given the outcome of some predictor variables $X_1, ..., X_p$ for some number of predictors $p \geq 1$ is

$$q_\alpha (x_1, ..., x_p) := F^{-1}_{Y|X_1,...,X_p} (\alpha | X_1 = x_1, ..., X_p = x_p) .$$

To scale the variables to the d-dimensional hypercube, we define the probability integral transformed variables as:

$$V := F_Y (Y) \text{ and } U_j := F_{X_j} (X_j) .$$

Now we can rewrite the conditional quantile function as

$$q_\alpha(x_1, \dots, x_p) = F_Y^{-1}\left(C^{-1}_{V|U_1,\dots,U_p}(\alpha|u_1, \dots u_p)\right).$$

Assuming the margins F_Y, F_{X_j}, for $j = 1, \dots, p$ are known, to obtain an estimate of the conditional quantile function q_α we only need to estimate the copula $C_{V,U_1,\dots,U_p}$.

- Estimating multivariate distributions is a very complex problem.
- Thus we only use bivariate copulas or the pair copulas, to construct multivariate distributions using conditioning.
- This method is known as pair copula construction (or PCC).

Drawable (D-) vine copula

Definition

A regular vine tree sequence $\mathcal{V} = (T_1, \ldots, T_{d-1})$ is called a D-vine tree sequence if it holds that each tree T_i has degree less or equal to two.

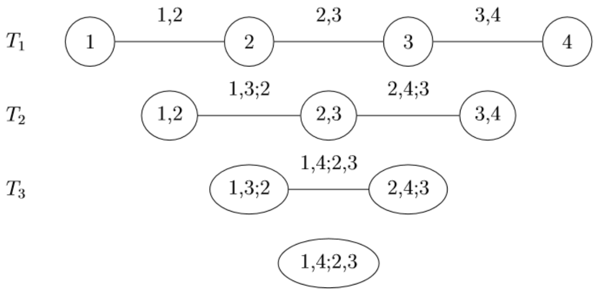


Figure: Four dimensional D-vine

Order of a D-vine

Definition

A D-vine $\mathcal{C}$ has order $\mathcal{O}_D(\mathcal{C}) = \left(U_0, U_{i_1}, \ldots, U_{i_p}\right)$, if U_0 is the first node of T_1 and U_{i_k} is the $(k+1)-$th node of T_1.

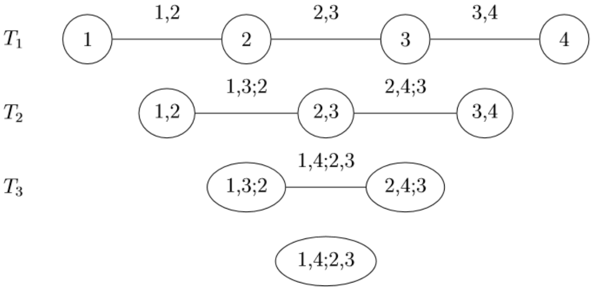


Figure: D-vine with order $\mathcal{O}_D(\mathcal{C}) = (1, 2, 3, 4)$

Canonical (C-) vine copula

Definition

A regular vine tree sequence $\mathcal{V} = (T_1, ..., T_{d-1})$ is called $C-$vine tree sequence if in each tree T_i there is one node n such that it has degree $d-i$. That node is called the root node of tree T_i.

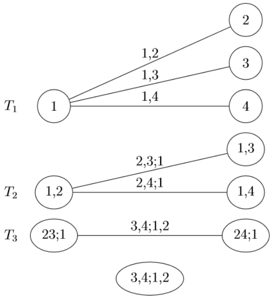


Figure: Four dimensional C-vine

Order of a C-vine

Definition

A C-vine $\mathcal{C}$ has order $\mathcal{O}_C(\mathcal{C}) = \left(U_0, U_{i_1}, \ldots, U_{i_p}\right)$, if U_{i_1} is the root node in T_1, $U_{i_2}U_{i_1}$ is the root node in T_2, and $U_{i_k}U_{i_{k-1}}|U_{i_1}, \ldots, U_{i_{k-2}}$ is the root node in T_k for $k = 3, \ldots, p-1$.

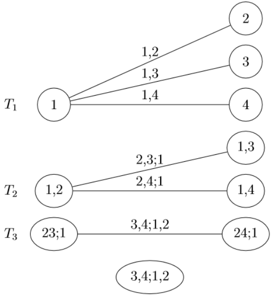


Figure: C-vine with order $\mathcal{O}_C(\mathcal{C}) = (4, 1, 2, 3,)$

Intermediate summary

- Data set containing n independent observations from the $(p+1)$-dimensional random vector $(V, U_1, \ldots, U_p)$.
- Biggest challenge is to estimate the $(p+1)$-dimensional copula.
- We limit the copula to being either a C- or D-vine.
- The problem simplifies to estimating the optimal order of predictors.

Goal: Find the optimal order of the covariates, given that the response V has to be a leaf and has to take the first place in the order of the fitted C- or D-vine model.

$$\mathcal{O}\left(\mathcal{C}^{*}\right) = \left(V, U_{t_1}, \ldots, U_{t_p}\right)$$

Idea: Starting with an initial order containing only the response, we sequentially update the order by adding predictors based on a fit measure.

Fit measure: Let $\mathcal{C}$ be a C- or D-vine with order $(V, U_1, \ldots, U_p)$. Additionally, assume that we are given n observations $\mathbf{v}$ and $\mathbf{u}_j$ for $j = 1, \ldots, p$. Then the conditional log likelihood function is given as

$$cll\left(\mathcal{C}, \mathbf{v}, (\mathbf{u}_1, \ldots, \mathbf{u}_p)\right) = \sum_{i=1}^{n} \log c_{V|U_1,\ldots,U_p}\left(v^{(i)} | u_1^{(i)}, \ldots, u_p^{(i)}\right). \quad (1)$$

Sequential forward selection of predictors

- At the beginning of step r the current optimal order is

$$\left(V, U_{t_1}, \dots, U_{t_{r-1}}\right),$$

  with $U_{t_1}, \dots, U_{t_{r-1}}$ being the predictors chosen at steps $1, \dots, r-1$ respectively.
- The r-th predictor is chosen from the candidate set K, containing k remaining predictors with the highest absolute partial correlation measure with the response.

- To determine which predictor to choose we consider the two-step ahead models with orders of the form

$$\mathcal{O}(\mathcal{C}) = \left(V, U_{t_1}, \ldots, U_{t_{r-1}}, U_c, U_j\right),$$

  - $U_c \in K$,
  - U_j comes from the set of remaining predictors not included in the model.

- Choose the r-th predictor from K as the predictor which corresponds to the two-step ahead model with the highest conditional loglikelihood.

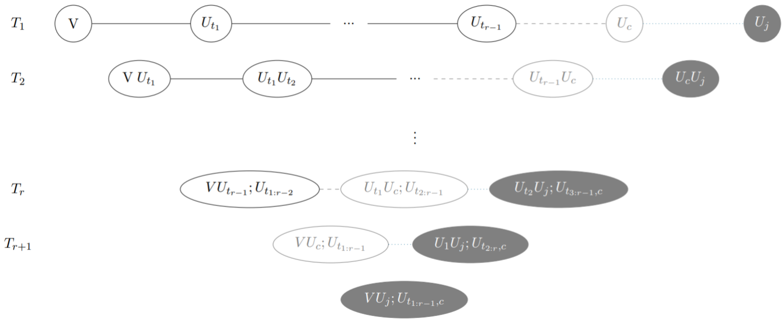
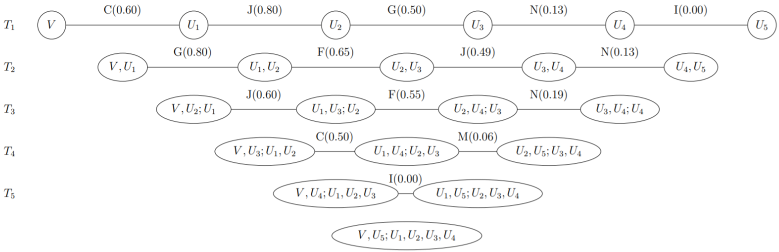


Figure: Graphical representation of a D-vine model at step r.

Example



Input: Six dimensional data set

$$\left(v^{(i)}, u_1^{(i)}, u_2^{(i)}, u_3^{(i)}, u_4^{(i)}, u_5^{(i)}\right)^T, \quad i = 1, \ldots, 500,$$

sampled from $(V, U_1, U_2, U_3, U_4, U_5)^T$ which follows a six dimensional D-vine copula distribution.

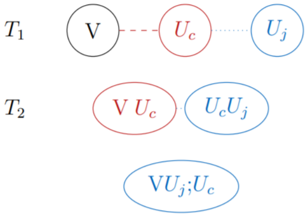
Step 1: Candidate selection

$\hat{\tau}_{VU_1}$	$\hat{\tau}_{VU_2}$	$\hat{\tau}_{VU_3}$	$\hat{\tau}_{VU_4}$	$\hat{\tau}_{VU_5}$
0.62	**0.71**	0.39	0.37	0.17

Table: Estimated Kendall's tau values.

$\Rightarrow$ Candidate set for Step 1: $K = \{U_1, U_2\}$.

Step 1: Predictor selection

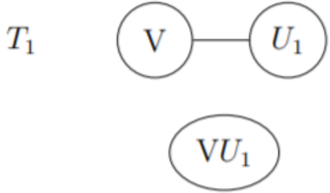


Candidate U_1 ($c = 1$)			
Order	c_{VU_c}	$c_{VU_j;U_c}$	Conditional log-lik
$V - U_1 - U_2$	322.58	592.46	915.04
$V - U_1 - U_3$	322.58	33.35	355.93
$V - U_1 - U_4$	322.58	5.44	328.02
$V - U_1 - U_5$	322.58	8.34	330.92

Candidate U_2 ($c = 2$)			
Order	c_{VU_c}	$c_{VU_j;U_c}$	Conditional log-lik
$V - U_2 - U_1$	327.78	123.97	451.75
$V - U_2 - U_3$	327.78	12.25	340.03
$V - U_2 - U_4$	327.78	6.24	334.02
$V - U_2 - U_5$	327.78	0	327.78

$\Rightarrow$ Selected candidate: U_1.

Fitted model after Step 1



Order of fitted model: $O = (V, U_1)$ and $cll = 322.58$.

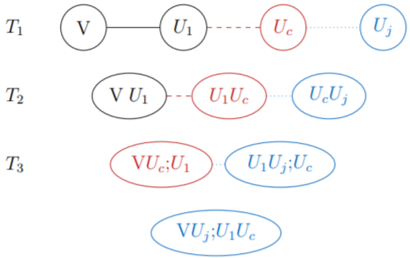
Step 2: Candidate selection

$\hat{\rho}_{V,U_2;U_1}$	$\hat{\rho}_{V,U_3;U_1}$	$\hat{\rho}_{V,U_4;U_1}$	$\hat{\rho}_{V,U_5;U_1}$
0.68	-**0.35**	0.13	0.12

Table: Estimated partial correlations .

$\Rightarrow$ Candidate set for Step 2: $K = \{U_2, U_3\}$.

Step 2: Predictor selection

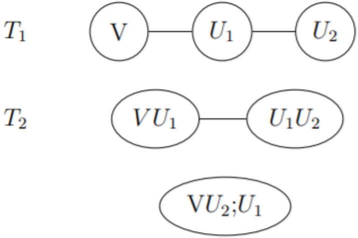


Candidate U_2 $(c = 2)$			
Order	$c_{VU_c;U_1}$	$c_{VU_j;U_1,U_c}$	Conditional log-lik
$V - U_1 - U_2 - U_3$	592.46	347.88	1262.91
$V - U_1 - U_2 - U_4$	592.46	38.00	953.04
$V - U_1 - U_2 - U_5$	592.46	0	915.04

Candidate U_3 $(c = 3)$			
Order	$c_{VU_c;U_1}$	$c_{VU_j;U_1,U_c}$	Conditional log-lik
$V - U_1 - U_3 - U_2$	33.35	303.66	659.59
$V - U_1 - U_3 - U_4$	33.35	10.09	366.02
$V - U_1 - U_3 - U_5$	33.35	7.95	363.88

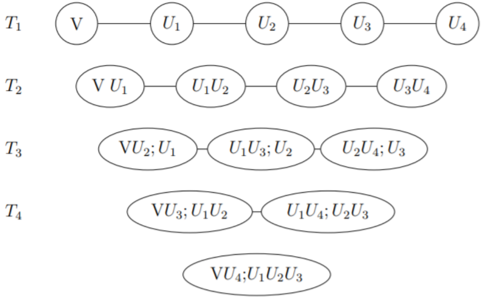
$\Rightarrow$ Selected candidate: U_2.

Fitted model after Step 2



Order of fitted model: $O = (V, U_1, U_2)$ and $cll = 915.04$.

Output



Order of optimal model: $O = (V, U_1, U_2, U_3, U_4)$ and $cll = 1460.82$.

References I

Czado, C. (2019).
Analyzing Dependent Data with Vine Copulas.
Lecture Notes in Statistics, Springer.

Tepegjozova, M. (2019).
D- and C-vine quantile regression for large data sets.
Masterarbeit, Technische Universität München, Garching b. München.

Zhou, J. (2020).
High dimensional quantile regression: model averaging and composite estimation.
Dissertation, KU Leuven, Leuven.