Bivariate vine based quantile regression

(ETH-UCPH-TUM Workshop on Graphical Models)

Marija Tepegjozova

m.tepegjozova@tum.de

TU München

Claudia Czado (TU München)

Copulas

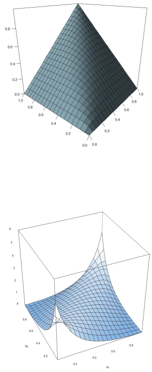
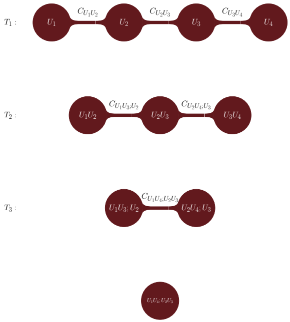


Figure: CDF and PDF of a bivariate copula.

- **Copula**
  - distribution on the unit hypercube
  - uniform margins
- **Sklar's theorem**
  - $F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$
  - probability integral transform(PIT)
  - [Sklar, 1959]
- **Decomposition**
  - conditioning
  - bivariate copulas

Regular vine



$$\mathcal{T} = \{T_1, T_2, T_3\}$$

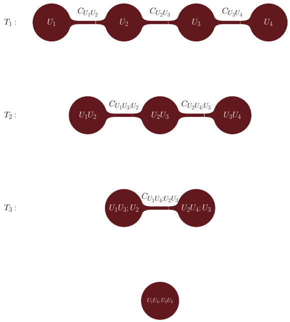
$$\mathcal{B}(\mathcal{T}) = \{C_{U_1U_2}, C_{U_2U_3}, C_{U_3U_4}, C_{U_1U_3;U_2}, C_{U_2U_4;U_3}, C_{U_1U_4;U_2U_3}\}$$

- **Pair Copula Construction**
  - construction of multivariate distributions
  - [Bedford and Cooke, 2002]

- **Regular vine copula**
  - tree sequence
  - pair copulas

Regular vine

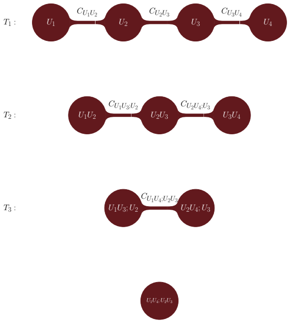
- **Pair Copula Construction**
- **Regular vine copula**
- **Tree sequence conditions**
  - T_1 is a tree with node set $N_1 = U_1, \ldots, U_d$
  - proximity condition
  - for $k \geq 2$, T_k is a tree with node set $N_k = E_{k-1}$ and edge set E_k



$$\mathcal{T} = \{T_1, T_2, T_3\}$$

$$\mathcal{B}(\mathcal{T}) = \{C_{U_1U_2}, C_{U_2U_3}, C_{U_3U_4}, C_{U_1U_3;U_2}, C_{U_2U_4;U_3}, C_{U_1U_4;U_2U_3}\}$$

Regular vine density



$\mathcal{T} = \{T_1, T_2, T_3\}$

$\mathcal{B}(\mathcal{T}) = \{C_{U_1U_2}, C_{U_2U_3}, C_{U_3U_4}, C_{U_1U_3;U_2}, C_{U_2U_4;U_3}, C_{U_1U_4;U_2U_3}\}$

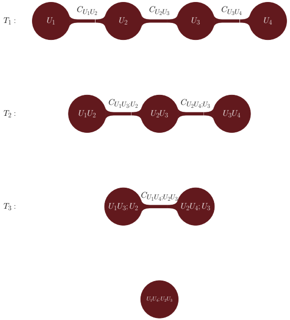
- **Density:**

$$c_{U_1,U_2,U_3,U_4} = c_{U_1U_2} \cdot c_{U_2U_3} \cdot c_{U_3U_4} \cdot c_{U_1U_3;U_2} \cdot c_{U_3U_4;U_3} \cdot c_{U_1U_4;U_2U_3}$$

- **Consequence of Sklar's Theorem**

$$f_{X_1,X_2,X_3,X_4} = f_{X_1} \cdot f_{X_2} \cdot f_{X_3} \cdot f_{X_4} \cdot c_{U_1U_2} \cdot c_{U_2U_3} \cdot c_{U_3U_4} \cdot c_{U_1U_3;U_2} \cdot c_{U_3U_4;U_3} \cdot c_{U_1U_4;U_2U_3}$$

Conditional distributions



$$\mathcal{T} = \{T_1, T_2, T_3\}$$

$$\mathcal{B}(\mathcal{T}) = \{C_{U_1U_2}, C_{U_2U_3}, C_{U_3U_4}, C_{U_1U_3;U_2}, C_{U_2U_4;U_3}, C_{U_1U_4;U_2U_3}\}$$

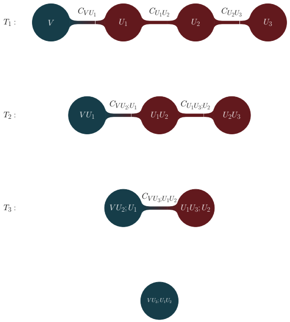
- **Conditional distribution**

$$C_{U_1|U_2,U_3,U_4}$$

can be obtained as a composition of first order derivatives from pair copula densities contained in $\mathcal{B}(\mathcal{T})$

- holds true only if the conditioned variable is a leaf node in every tree of the vine tree sequence
- inverses of first order derivative functions are obtainable

- **Vine based quantile regression**



$$\mathcal{T} = \{T_1, T_2, T_3\}$$

$$\mathcal{B}(\mathcal{T}) = \{C_{VU_1}, C_{U_1U_2}, C_{U_2U_3}, C_{VU_2;U_1}, C_{U_1U_3;U_2}, C_{VU_3;U_1U_2}\}$$

$$q_\alpha(u_1, u_2, u_3) = C^{-1}_{V|U_1,U_2,U_3}(\alpha|u_1, u_2, u_3)$$

- it can be shown that

$$F^{-1}_{Y|X_1,X_2,X_3}(\alpha|\cdot) = F_Y^{-1}\left(C^{-1}_{V|U_1,U_2,U_3}(\alpha|\cdot)\right)$$

- autonomous model building approaches have been proposed
- [Kraus and Czado, 2017] [Tepegjozova et al., 2022]

Motivation for bivariate quantiles and bivariate quantile regression

- In the case of a multivariate response data sets, usually different models are used for modeling each response on the same set of covariates.
- However, the possible interaction or dependence between the responses is disregarded.
- Examples of such data sets are minimum and maximum temperature, minimum and maximum risk values, pressure and volume, and other dependent joint events (more details in [Tepegjozova and Czado, 2022]).

Bivariate unconditional quantiles

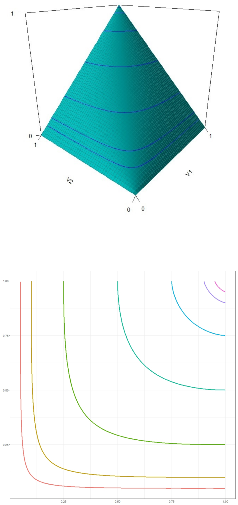


Figure: Bivariate quantile sets of 2-dimensional Gaussian copula with Kendall's tau of 0.50.

- **Bivariate quantiles**
- if U_1 and U_2 are uniformly distributed, their bivariate quantile set is defined as

$$Q_\alpha^U = \{(u_1, u_2) \in [0,1]^2 \; ; \; C_{U_1,U_2}(u_1, u_2) = \alpha\}$$

- given arbitrary distributed and continuous X_1 and X_2, their bivariate quantile set is defined as

$$Q_\alpha^X = \{(x_1, x_2) \in \mathbb{R}^2 \; ; \; F_{X_1,X_2}(x_1, x_2) = \alpha\}$$

- their relation can be described as

$$Q_\alpha^X = \{(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)) \in \mathbb{R}^2 \; ; \; (u_1, u_2) \in Q_\alpha^U\}$$

Bivariate conditional quantiles

- **Copula level**

  given $p+2$ uniformly distributed random variables $V_1, V_2, U_1, \ldots, U_p$, the bivariate quantiles of V_1 and V_2 given $U_1, \ldots, U_p$ are defined as

$$Q_\alpha^V(\mathbf{u}) = \{(v_1, v_2) \in [0,1]^2 \ ; \ C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u}) = \alpha\}$$

- **General case**

  given continuously distributed random variables $Y_1, Y_2, X_1, \ldots, X_p$

$$Q_\alpha^Y(\mathbf{x}) = \{(y_1, y_2) \in \mathbb{R}^2 \ ; \ F_{Y_1,Y_2|\mathbf{X}}(y_1, y_2|\mathbf{x}) = \alpha\}$$

- **Relation**

$$Q_\alpha^Y(\mathbf{x}) = \{(F_{Y_1}^{-1}(v_1), F_{Y_2}^{-1}(v_2)) \in \mathbb{R}^2 \ ; \ C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u}) = \alpha\}$$

Numerical evaluation of bivariate quantiles

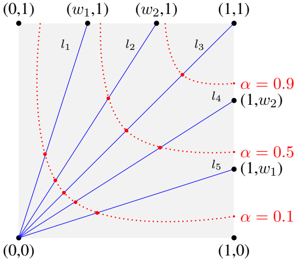
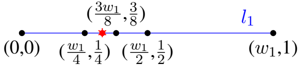
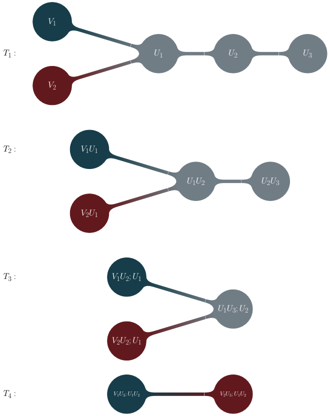


Figure: Graphical representation of the numerical estimation procedure.



Y-vines

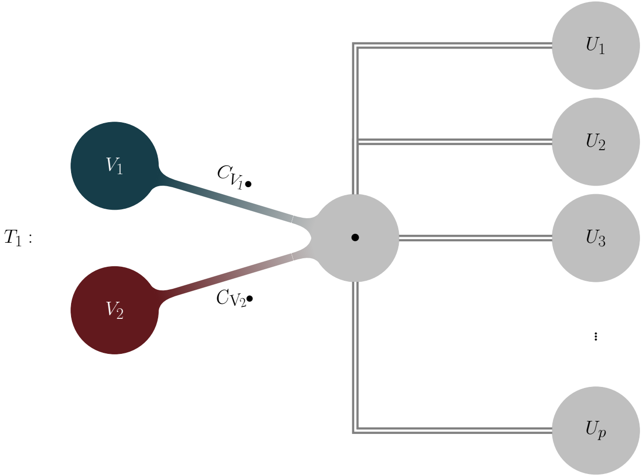


- **Y-vine tree sequence**
  - regular vine copula
  - both response variables are ensured to be leaf nodes in each tree
  - symmetric with respect to response variables
- **Conditional distribution**

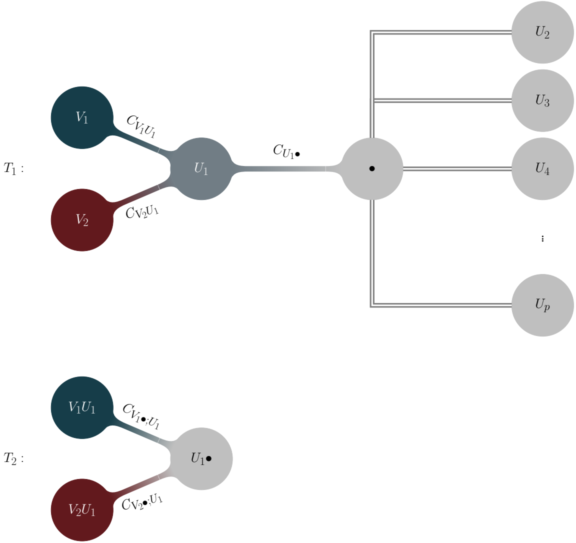
$$C_{V_1,V_2|U_1\dots U_p}$$

  - obtained as univariate integral involving pair copula densities from the Y-vine and the conditional distribution function $C_{V_1|V_2U_1\dots U_p}$

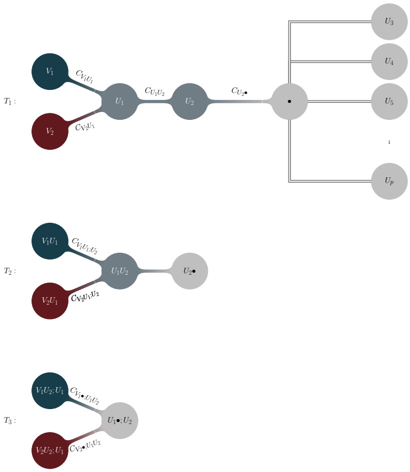
Forward sequential predictor selection



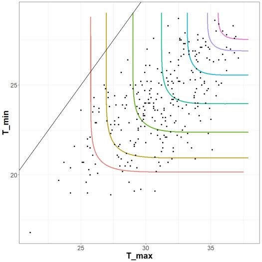
Sequential predictor selection

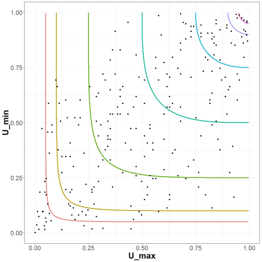


Sequential predictor selection



Data introduction





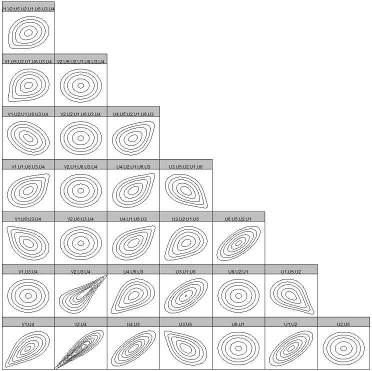
- contains weather data of 25 different stations in the urban area of Seoul
- from 2013 to 2017 data has been collected between June 30th and August 30th
- includes two response variables, next day minimum and maximum temperature
- includes 14 continuous predictor variables

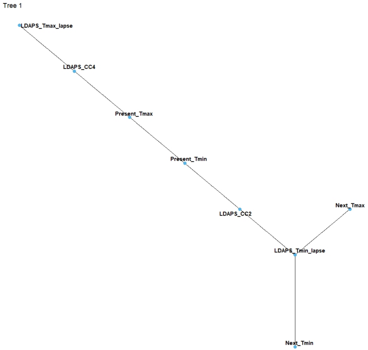
[Dua and Graff, 2017]

Variable name	Description(unit)	Range
Next_Tmax	The next-day maximum air temperature ($^\circ C$)	17.4 to 38.9
Next_Tmin	The next-day minimum air temperature ($^\circ C$)	11.3 to 29.8
Present_Tmax	Maximum air temperature between 0 and 21 h on the present day ($^\circ C$)	20 to 37.6
Present_Tmin	Minimum air temperature between 0 and 21 h on the present day ($^\circ C$)	11.3 to 29.9
LDAPS_RHmin	LDAPS model forecast of next-day minimum relative humidity (%)	19.8 to 98.5
LDAPS_RHmax	LDAPS model forecast of next-day maximum relative humidity (%)	58.9 to 100
LDAPS_Tmax_lapse	LDAPS model forecast of next-day maximum air temperature applied lapse rate ($^\circ C$)	17.6 to 38.5
LDAPS_Tmin_lapse	LDAPS model forecast of next-day minimum air temperature applied lapse rate ($^\circ C$)	14.3 to 29.6
LDAPS_WS	LDAPS model forecast of next-day average wind speed (m/s)	2.9 to 21.9
LDAPS_LH	LDAPS model forecast of next-day average latent heat flux (W/m^2)	-13.6 to 213.4
LDAPS_CC1	LDAPS model forecast of next-day 1st 6-hour split average cloud cover (0-5 h) (%)	0 to 0.97
LDAPS_CC2	LDAPS model forecast of next-day 2nd 6-hour split average cloud cover (6-11 h) (%)	0 to 0.97
LDAPS_CC3	LDAPS model forecast of next-day 3rd 6-hour split average cloud cover (12-17 h) (%)	0 to 0.98
LDAPS_CC4	LDAPS model forecast of next-day 4th 6-hour split average cloud cover (18-23 h) (%)	0 to 0.97
Solar radiation	Daily incoming solar radiation (wh/m^2)	4329.5 to 5992.9

Table: Variable description, the unit of measurement and the range of possible values the considered variables can take.

Fitted Y-vine





Illustration

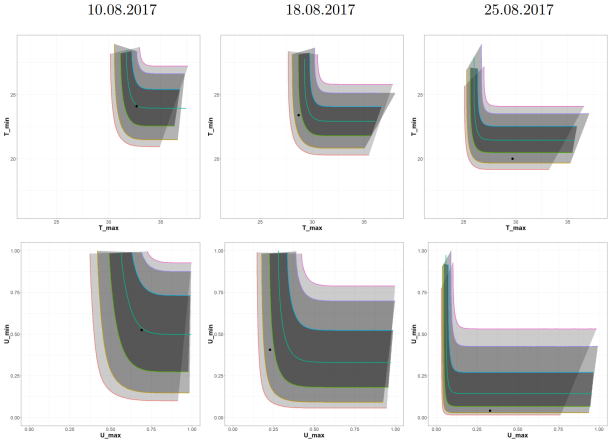


Figure: The plots correspond to the days 10.08.2017, 18.08.2017 and 25.08.2017 (left to right). Shown are estimated conditional quantile curves for $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$ (left bottom to right top) and corresponding 90%, 80% and 50% confidence region (light to dark grey shaded) on each panel. Row 1 are estimates on the x-scale and row 2 is on the u-scale. The black dot is the true value.

Advantages of joint modeling of dependent responses

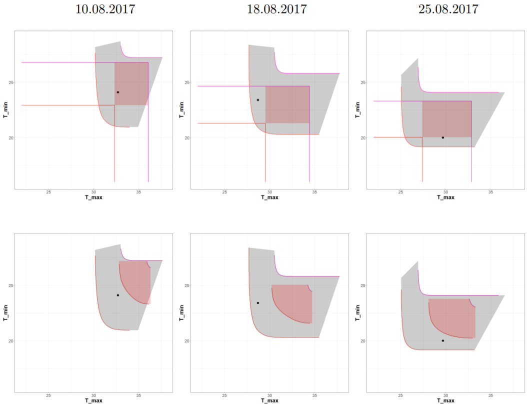


Figure: Shown are conditional bivariate quantile curves $Q^Y_{0.05}(\mathbf{x})$ and $Q^Y_{0.95}(\mathbf{x})$ and the corresponding 90% confidence region $CI^Y_{0.10}$ (grey shaded). Additionally, in row 1 the estimated univariate quantiles for $\alpha = 0.025, 0.975$ for both response variables and the corresponding 90% confidence regions (in red) are shown. In row 2, the bivariate conditional quantiles when the responses are treated as conditionally independent and the associated 90% confidence regions $CI^{Y_1 \perp Y_2 | \mathbf{X}}_{0.10}$ (red shaded) are shown.

References I

Bedford, T. and Cooke, R. M. (2002).
Vines–a new graphical model for dependent random variables.
The Annals of Statistics, 30(4):1031–1068.

Dua, D. and Graff, C. (2017).
UCI machine learning repository.

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

Sklar, M. (1959).
Fonctions de repartition an dimensions et leurs marges.
Publ. inst. statist. univ. Paris, 8:229–231.

Tepegjozova, M. and Czado, C. (2022).
Bivariate vine copula based quantile regression.
arXiv preprint arXiv:2205.02557.

Tepegjozova, M., Zhou, J., Claeskens, G., and Czado, C. (2022).
Nonparametric c-and d-vine-based quantile regression.
Dependence Modeling, 10(1):1–21.

Thank you for your attention!