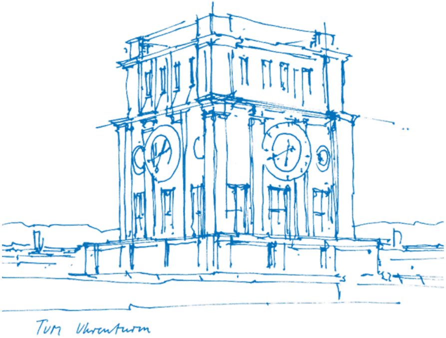
Vine copula structural equation models

Claudia Czado

Applied Mathematical Statistics
Department of Mathematics
Technical University of Munich

Oct.2022



Outline

1 Motivation

2 Pair-copula constructions (PCC) of vine distributions

3 Vine copula based quantile regression models

4 D-vine based structural equation models

5 Analysis of the Sachs Data

6 Summary and outlook

How to allow for non Gaussian behavior in graphical models? (I)

- Consider statistical models on directed acyclic graphs (DAG's) or Bayesian networks (Pearl 1988; Lauritzen 1996)
- For a DAG with Markov properties $X_1, \ldots, X_d$ has density

$$f(x_1, \ldots x_d) = \prod_{j=1}^{d} f(x_j | \boldsymbol{\pi}(X_j) = \boldsymbol{\pi}(x_j)), \tag{1}$$

  where $\boldsymbol{\pi}(X_j)$ is parent set of X_j ($\boldsymbol{\pi}(X_j) = \{X_k : X_k \to X_j\}$).
- Standard graphical models of the form (1) for continuous variables assume joint Gaussianity.
- However some data sets are not easily transformed to joint normality, therefore we want to construct non Gaussian graphical models.

How to allow for non Gaussian behavior in graphical models? (II)

- We can choose in (1) the conditional densities arbitrary and still get a joint density.
- However this does not guarantee a compatible joint distribution (Wang and Ip 2008). But Varin and Vidoni (2005) showed that the conditional specified model (1) minimizes the Kullback-Leibler distances to the conditional distributions.
- So one approach to extend standard Gaussian DAG's is to use other conditional densities.
- Here we will use D-vine regression densities and illustrate it using an experiment from the Sachs data (Sachs et al. 2005).
- Comparison of the D-vine approach will be made to a generalized additive model (GAM) approach.

Structural equation models (SEM) for graphical random variables

- The standard Gaussian DAG model is a linear structural equation model (SEM).
- Assume graph $\mathcal{G}$ with nodes $V = \{X_1, \ldots X_d\}$, edge set E and directed weight adjacency matrix A, i.e. $A_{i,j} \neq 0$ if and only if $(i, j) \in E$.
- Let $\boldsymbol{\epsilon} \sim N_d(\mathbf{0}, \Omega)$, then the linear SEM for $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$

$$\boldsymbol{X} = A^\top \boldsymbol{X} + \boldsymbol{\epsilon}. \tag{2}$$

- This implies that $\boldsymbol{X} \sim N_d(\mathbf{0}, (I - A)\Omega^{-1}(I - A)^\top)$, i.e. $X_j | \boldsymbol{\pi}(X_j) = \boldsymbol{\pi}(x_j)$ is univariate normal.
- Such a factorization is equivalent to the Markov assumption with respect to the graph $\mathcal{G}$ (Lauritzen 1996, Theorem 3.27).

Generalized additive SEM for graphs

- Voorman et al. (2014) assume that

$$X_j|\{X_k, k \neq j\} = \sum_{k \neq j} g_{jk}(X_k) + \epsilon_j \tag{3}$$

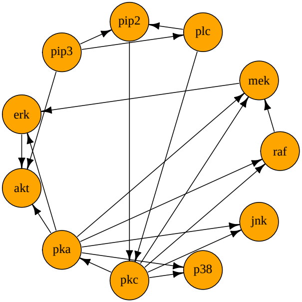
  for zero mean error term ϵ_j and $g_{jk}(\boldsymbol{x}_k) = \Psi_{jk}\boldsymbol{\beta}_{jk}$ where columns of matrix $\Psi_{j,k} \in \mathbb{R}^{n \times d}$ are basis functions.
- Estimation uses a penality approach: Minimize over $\beta_{jk}, 1 \leq j, k \leq d, k \prec j$ (topological order)

$$\frac{1}{2n}||\boldsymbol{x}_j - \sum_{k \prec j} \Psi_{jk}\boldsymbol{\beta}_{jk}||^2 + \lambda \sum_{k \prec j} ||\Psi_{jk}\beta_{jk}||^2 \tag{4}$$

- Penalization parameter λ is chosen to minimize BIC with an appropriately defined degree of freedom (Voorman et al. 2014).

Sachs data

- The Sachs data contains flow cytometry measurements on 11 variables (pip3, plc, pip2, pkc, pka, p38, jnk, raf, mek, erk, akt in topological order) under 14 experimental conditions.
- Concentrate on experiment `cd3cd28_aktinhib` (n=911)
- Consent graph (20 edges) based on all 14 experiments:



More general multivariate models

- Multivariate normal distribution is often the base model.
- However multivariate data exhibit often complex dependency patterns, such as asymmetry and dependence in the extremes not covered by the multivariate normal distribution.
- The copula approach allows separate models for the margins and the dependence.
- Standard classes of multivariate copulas such as Gaussian, Student t and Archimedean copulas are too restrictive
- Vine copulas allow for flexible modeling of (conditional) pairs of variables.
- They can accommodate asymmetric tail behavior and symmetric behavior of variables in a single model.

Outline

What are copulas?

- **Copula:** A d-dimensional copula C is a multivariate distribution on $[0,1]^d$ with uniformly distributed marginals.
- **Copula density function:** $c(u_1, ..., u_d) := \frac{\partial^d}{\partial u_1 ... \partial u_d} C(u_1, ..., u_d)$
- **Theorem (Sklar 1959):**

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$$
$$f(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) f_1(x_1) ... f_d(x_d)$$

for some d-dimensional copula C.

- **Conditional density in $d = 2$**

$$f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2)$$
$$f_{2|1}(x_2|x_1) = c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2)$$

What are these vine copulas?

- Multivariate vine copulas are copulas built out of bivariate copulas.
- A pair copula construction (PCC) is possible through conditioning. Joe (1996) gave a first example.
- Many PCC's are feasible. Bedford and Cooke (2002) introduced a graphical structure to organize them.
- Gaussian vines were analyzed in Kurowicka and Cooke (2006) while ML estimation for Non Gaussian ones started with Aas et al. (2009).

Vine copula resources

- Books:
  - Kurowicka and Joe (eds, 2011): Dependence modeling - Handbook on Vine Copulas
  - Joe (2014): Dependence modeling with copulas
  - Czado (2019): Analyzing dependent data with vine copulas: a practical guide with R
- Reviews:
  - Aas (2016): Pair-copula constructions for financial applications: A review
  - Czado and Nagler (2022): Vine copula based modeling
- Web Resources
  - `vine-copula.org`
  - `en.wikipedia.org/wiki/Vine_copula`

How does this work in 3 dimensions?

Recursion

$$f(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_1(x_1)$$

Using Sklar for $f(x_1, x_2)$, $f(x_2, x_3)$ and $f_{13|2}(x_1, x_3|x_2)$ implies

$$\begin{aligned} f_{2|1}(x_2|x_1) &= c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) \\ f_{3|12}(x_3|x_1, x_2) &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) f_{3|2}(x_3|x_2) \\ &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \end{aligned}$$

Three dimensional pair copula construction

$$\begin{aligned} f(x_1, x_2, x_3) &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) c_{23}(F_2(x_2), F_3(x_3)) \\ &\times c_{12}(F_1(x_1), F_2(x_2)) \times f_3(x_3) f_2(x_2) f_1(x_1) \end{aligned}$$

The copula corresponding to the distribution of (X_1, X_3) given $X_2 = x_2$ is denoted by $c_{13;2}$. Only bivariate copulas and univariate conditional cdf's are used. Can be generalized to d dimensions.

Three data scales

- **x-scale (original i.i.d data vectors):** $(x_{i1}, \dots x_{id})$
- **u-scale (copula data):**

$$(u_{i1}, \dots u_{id}), \text{ where } u_{ij} = F_j(x_{ij}) \; i = 1, \dots, n; j = 1, \dots, d$$

  is the probability integral transform.

- **z-scale (marginal normalized data):**

$$(z_{i1}, \dots z_{id}), \text{ where } z_{ij} = \Phi^{-1}(u_{ij}) \; i = 1, \dots, n; j = 1, \dots, d$$

  and Φ^{-1} quantile function of $N(0, 1)$

Bivariate elliptical copula families

Gaussian copula

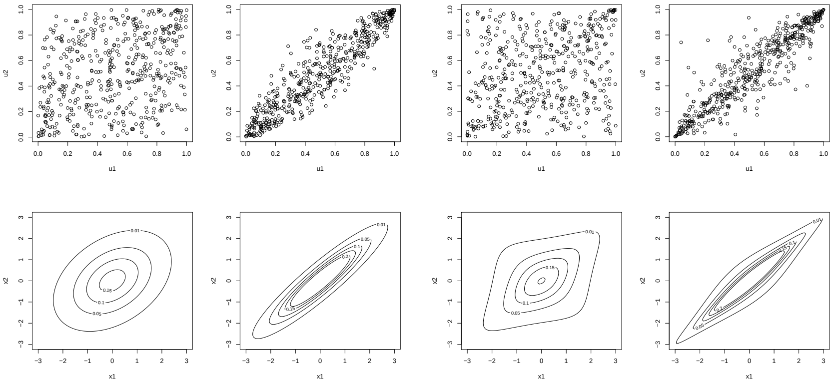
(left $\tau = .25$, right: $\tau = .75$)

symmetric dependence

t-copula with $df = 3$

(left $\tau = .25$, right: $\tau = .75$)

symmetric dependence



Bivariate Archimedean copula families

Gumbel copula

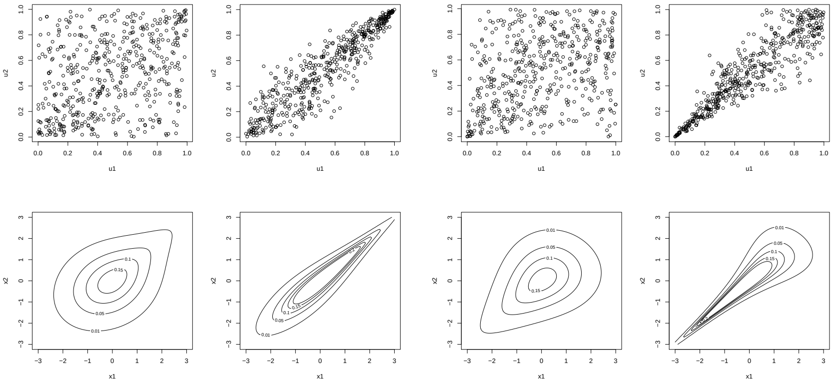
(left $\tau = .25$, right: $\tau = .75$)

upper tail dependent

Clayton copula

(left $\tau = .25$, right: $\tau = .75$)

lower tail dependent



How do vines work in higher dimensions?

- **Which pairs of variables are needed?**
- **What are the conditioning variables?**

How do vines work in higher dimensions?

- **Which pairs of variables are needed?**
- **What are the conditioning variables?**

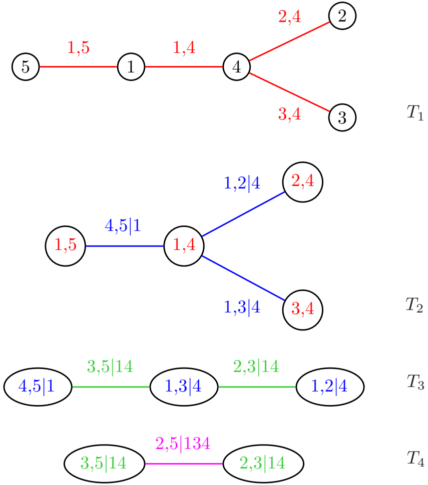
Components of a regular vine $R(\mathcal{V}, \mathcal{C}, \boldsymbol{\theta})$ distribution

1. Tree structure $\mathcal{V}$ of linked trees identifies the pairs of variables and conditioning variables.
2. Parametric bivariate copulas $\mathcal{C} = \mathcal{C}(\mathcal{V})$ for each edge in the tree structure
3. Corresponding parameter value $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathcal{C}(\mathcal{V}))$

- Recursion for conditional distribution functions (Joe 1996):

If $\boldsymbol{v} = (v_j, \boldsymbol{v}_{-j})$ then $F(x|\boldsymbol{v}) = \dfrac{\partial C_{xv_j;\boldsymbol{v}_{-j}}(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))}{\partial F(v_j|\boldsymbol{v}_{-j})}$.

Can we see an example of a tree structure?



Density

$$f = f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5$$
$$\cdot\, c_{14} \cdot c_{15} \cdot c_{24} \cdot c_{34}$$
$$\cdot\, c_{12;4} \cdot c_{13;4} \cdot c_{45;1}$$
$$\cdot\, c_{23;14} \cdot c_{35;14}$$
$$\cdot\, c_{25;134}$$

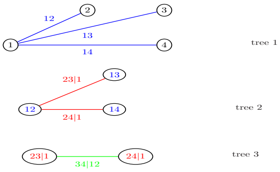
Special regular vines: C and D-vines

C-vine: each tree has unique node connected to $d-j$ edges

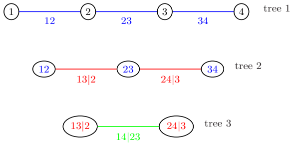
$$f_{1234} = \left[\prod_{i=1}^{4} f_i\right] \cdot c_{12} \cdot c_{13} \cdot c_{14} \cdot c_{23;1} \cdot c_{24;1} \cdot c_{34;12}$$

D-vine: no node is connected to more than 2 edges

$$f_{1234} = \left[\prod_{i=1}^{4} f_i\right] \cdot c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{13;2} \cdot c_{24;3} \cdot c_{14;23}$$



useful for ordering by importance

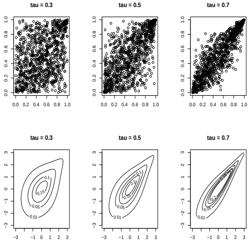


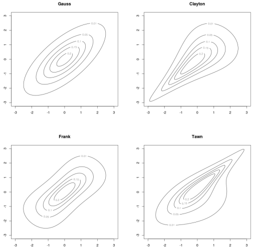
useful for temporal ordering

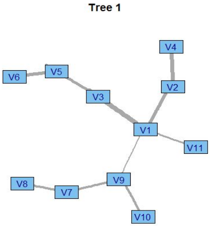
How can we estimate and select PCCs?

Three tasks (Czado et al. (2013))

1. How to estimate the pair copula parameters for a given vine tree structure and pair copula families for each edge?
2. How to choose the pair copula families and estimate the corresponding parameters for a given vine tree structure?
3. How to select and estimate all components of a regular vine?







Task 1: Sequential and ML estimation

Parameters: $\Theta = (\theta_{12}, \theta_{23}, \theta_{13;2})$

Observations: $\{(x_{t1}, x_{t2}, x_{t3}), t = 1, \cdots, T\}$

Sequential estimates:

- Estimate θ_{12} from $\{(x_{t1}, x_{t2}), t = 1, \cdots, T\}$
- Estimate θ_{23} from $\{(x_{t2}, x_{t3}), t = 1, \cdots, T\}$.
- Define pseudo observations

$$\hat{v}_{1|2t} := F(x_{t1}|x_{t2}, \hat{\theta}_{12}) \text{ and } \hat{v}_{3|2t} := F(x_{t3}|x_{t2}, \hat{\theta}_{23})$$

Finally estimate $\theta_{13;2}$ from $\{(\hat{v}_{1|2t}, \hat{v}_{3|2t}), t = 1, \cdots, T\}$.

Maximum likelihood

$$\begin{aligned} L(\Theta|x) &= \sum_{t=1}^{T} [\log c_{12}(x_{t1}, x_{t2}|\theta_{12}) + \log c_{23}(x_{t2}, x_{t3}|\theta_{23}) \\ &+ \log c_{13;2}(F(x_{t1}|x_{t2}, \theta_{12}), F(x_{t3}|x_{t2}, \theta_{23})|\theta_{13;2})] \end{aligned}$$

Task 2: Joint estimation of pair copula families and parameters

- Restrict to a set of bivariate pair copula families and use AIC or Vuong test to select family
- Check for truncation (Brechmann et al. (2012), Nagler et al. (2019)) by using independence copulas in higher trees

Task 3: Sequential treewise selection

- Capture strong pairwise dependencies first.
- Select trees sequentially.
- Give weights to every edge possible and select tree which maximizes the sum of weights.
- Details in Dißmann et al. (2013).

Software/Simulation for vines

- **Software:** `rvinecopulib` (Nagler and Vatter 2021)
- **Simulation of vine copulas:**
  - **Rosenblatt transform:** A sample $u_1, ..., u_d$ from $C_{1,...,d}$ is obtained as follows:

$$\begin{aligned} \textbf{First: } & \text{Sample } w_j \overset{\text{i.i.d.}}{\sim} U[0;1], \quad j = 1, \ldots, d \\ \textbf{Then: } & u_1 := w_1 \\ & u_2 := C^{-1}_{2|1}(w_2|u_1) \\ & \vdots \\ & u_d := C^{-1}_{d|d-1,\ldots,1}(w_d|u_{d-1}, \ldots, u_1). \end{aligned}$$

  - So we need conditional distributions associated with d dimensional vine copulas

h functions and univariate cond. copula cdf's

Indices: $r : s := (r, r+1, \ldots, s)$, sets: $\mathbf{x}_{r:s} = (x_r, \ldots, x_s)$
h functions:

$$
\begin{aligned}
h_{1|d;2:(d-1)}(u_1|v_d) &:= \frac{\partial}{\partial v_d} C_{1d;2:(d-1)}(u_1, v_d) \\
h_{d|1;2:(d-1)}(u_d|v_1) &:= \frac{\partial}{\partial v_1} C_{1d;2:(d-1)}(v_1, u_d)
\end{aligned}
$$

Recursion for univariate conditional copula cdf's

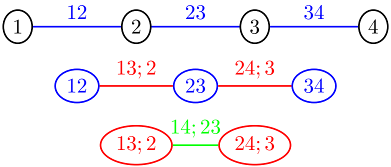
$$C_{1|2:d}(u_1|\mathbf{u}_{2:d}) = h_{1|d;2:(d-1)}(C_{1|2:(d-1)}(u_1|\mathbf{u}_{2:(d-1)})|C_{d|2:(d-1)}(u_d|\mathbf{u}_{2:(d-1)}))$$

$$C_{d|2:(d-1)}(u_d|\mathbf{u}_{2:(d-1)}) = h_{d|2;3:(d-1)}(C_{d|3:(d-1)}(u_d|\mathbf{u}_{3:(d-1)})|C_{2|3:(d-1)}(u_2|\mathbf{u}_{3:(d-1)}))$$

Univariate conditional distributions of D-vine copulas

Restrict to D-vine copulas, but extension to R-vines possible

$$
\begin{aligned}
c_{1234} = \;& c_{12} \cdot c_{23} \cdot c_{34} \\
& \cdot c_{13;2} \cdot c_{24;3} \\
& \cdot c_{14;23}
\end{aligned}
$$



Univariate conditional copula cdf of first node

$C_{1|2:4}(u_1|\mathbf{u}_{2:4}) =$
$h_{1|4;2,3}(h_{1|3;2}(h_{1|2}(u_1|u_2)|h_{3|2}(u_3|u_2))|h_{4|2;3}(h_{4|3}(u_4|u_3)|h_{2|3}(u_2|u_3)))$
where $h_{i|j;D}(u|v) := \partial C_{ij;D}(u,v)/\partial v$.

Outline

Vine copula based (quantile) regression

- Paper: Kraus, D., and Czado, C. (2017). D-vine copula based quantile regression. Computational Statistics & Data Analysis, 110, 1-18.
- Software: Nagler. T. (2022). vinereg: D-Vine Quantile Regression. R package version 0.8.1. `https://CRAN.R-project.org/package=vineregNagler`


Conditional quantiles in a D-vine copula

- Express the univariate conditional copula cdf $C_{v|1:m}(\cdot|\mathbf{u}_{1:m})$ for fixed conditioning values $\mathbf{u}_{1:m}$ using h functions.
- Denote by $Q_{v|1:m}(\cdot|\mathbf{u}_{1:m})$ the quantile function corresponding to $C_{v|1:m}(\cdot|\mathbf{u}_{1:m})$ for fixed $\mathbf{u}_{1:m}$.
- For continuous pair copulas we have

Conditional quantiles

$$Q_{v|1:m}(\alpha|\mathbf{u}_{1:m}) := C^{-1}_{v|1:m}(\alpha|\mathbf{u}_{1:m}) \text{ for } \alpha \in (0,1).$$

- Use the inverses of the h function to recursively invert the univariate conditional cdf $C_{v|1:m}(\cdot|\mathbf{u}_{1:m})$ to obtain the corresponding conditional quantile $Q_{v|1:m}(\alpha|\mathbf{u}_{1:m})$.

Copula quantile regression

	Original scale	Copula scale
Node variable	$X \sim F_X$	$V := F_X(X)$
Parent variables	$\mathbf{S} = (S_1, \ldots, S_m)$	$\mathbf{U} := (U_1, \ldots, U_m)$
		where $U_k := F_{S_k}(S_k)$

Copula quantile regression:

$$F^{-1}_{X|S_1,\ldots S_m}(\alpha|\mathbf{s}) = F_X^{-1}\left(C^{-1}_{V|U_1,\ldots U_m}(\alpha|u_1,\ldots u_m)\right)$$

For DAG context:

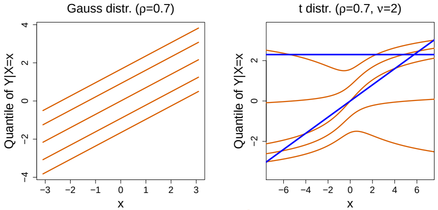
- X corresponds to a particular node
- $S_1, \ldots S_m$ corresponds to parent variables of that node.
- Conditional quantiles $C^{-1}_{V|U_1,\ldots U_m}$ are based on a D-vine copula.

Linear and copula quantile regression

Linear quantile regression

$$F_{X|S}^{-1}(\alpha|\mathbf{s}) = \beta_0(\alpha) + \sum_{k=1}^{m} \beta_k(\alpha) s_k$$

- Linearity assumption often violated $\rightarrow$ quantile crossing



- Linear quantiles only occur with Gaussian dependence (Bernard and Czado 2015).
- No quantile crossing occurs in copula quantile regression

D-vine copula quantile regression estimation (Kraus and Czado 2017)

- Given i.i.d. data $(\mathbf{x}, \mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_m))$ of sample size n
- Create pseudo copula data $(\hat{\mathbf{v}}, \hat{\mathbf{u}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m))$ by estimating univariate cdf's and applying the probability integral transform
- We use kernel smoothing estimators for the cdf of all variables since we need $\hat{F}_x^{-1}(\alpha)$ later.
- Over all D-vines $\mathcal{D}$ with D-vine ordering $\mathcal{V}$, pair copula families $\mathcal{B}(\mathcal{V})$, parameter set $\Theta(\mathcal{B}(\mathcal{V}))$ and i th pseudo copula values $(\hat{v}_i, \hat{\mathbf{u}}^i)$ maximize

conditional log-likelihood based on D-vine $\mathcal{D}$

$$cll(\hat{\mathbf{v}}, \hat{\mathbf{u}}; \mathcal{D}) := \sum_{i=1}^{n} \ln(c_{v|u_1,\ldots u_m}(\hat{v}_i|\hat{u}_{i1}, \ldots \hat{u}_{im}; \mathcal{D}))$$

- conditional copula density $c_{v|u_1,\ldots,u_m}$ is analytically available.

Selection of D-Vine copula quantile regression models (Kraus and Czado 2017)

- There are $m!$ possible D-vine orderings, these are too many.
- So we follow a forward selection of the parent nodes:
  - Start with the parent node, which has the largest cll value, call this variable 1, so we have ordering $v \leftrightarrow 1$ with cll_{max} and pair copula family c_{v1} and parameter θ_{v1}.
  - For each remaining parent node w, determine $cll(w)$ based on D-vine ordering $v \leftrightarrow 1 \leftrightarrow w$, families and parameters, choose the one with largest $cll(w)$.
  - If $cll(w) > cll_{max}$ then call w variable 2 and consider ordering $v \leftrightarrow 1 \leftrightarrow 2$ with cll_{max}, otherwise stop.
  - Continue until cll_{max} cannot be improved.
- This gives a ranking of the parent nodes by importance.

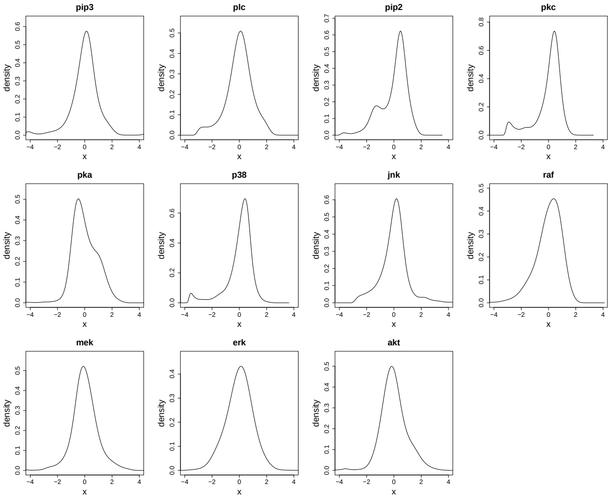
Outline

D-vine SEM

- Assume that a DAG for data at hand is given.
- Use a D-vine regression model for each $f(x_j|\boldsymbol{\pi}(X_j) = \boldsymbol{\pi}(x_j))$
- Since this involves a forward selection of the parent nodes, the starting DAG can be reduced if parent nodes are not selected.
- This D-vine SEM is approximately compatible with the specified conditional distributions for each node as long as selected pair copulas are a good approximation for the integral of certain conditional pair copulas.

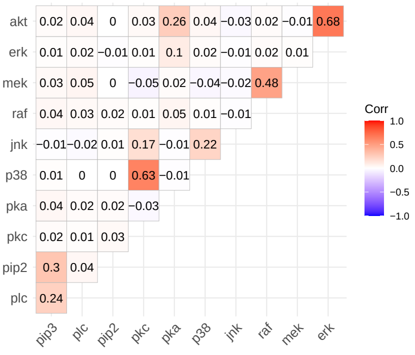
Outline

Marginal exploration



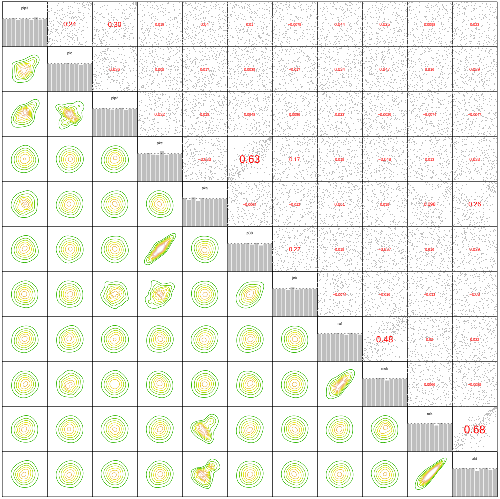
Many non-normal margins.

Pairwise exploration (I)



Higher estimated Kendall's τ between (pip3,plc),(pip3,pip2),(pkc,p38),(pkc,jnk), (pka,akt),(pka,erk),(p38,jnk),(raf,mek),(erk,akt)

Pairwise exploration (II)

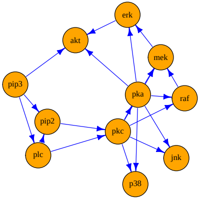


Graphical analysis of Sachs data

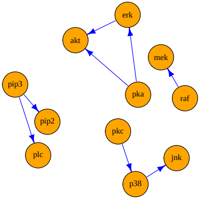
- **Data preparation**: all variables logarithmized and standardized
- **Gaussian SEM model** (5 edges selected):
  - uses R package `sparsebn` based on Fu and Zhou (2013)
  - uses concave penalization of Aragam and Zhou (2015)
- **GAM SEM model:** (8 edges selected)
  - uses R package `spacejam` based on Voorman et al. (2014)
  - uses topological order of consent graph
  - selects using minimal BIC with cubic polynomials as basis
- **D-vine SEM model** (10 edges selected):
  - uses R package `vinereg` based on Kraus and Czado (2017) with non parametric margins
  - Starting from consent graph
  - Removes edges, when parent nodes are not selected using BIC

Chosen DAG's for the Sachs Data

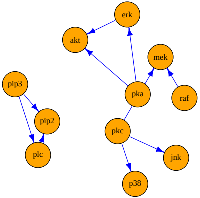
Consent DAG



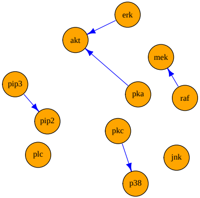
GAM DAG



D−vine DAG



Gauss DAG



Comparison of chosen edges

- **Gaussian SEM** (5 edges):
  - pip3 $\rightarrow$ pip2, pkc $\rightarrow$ p38, pka $\rightarrow$ akt, erk $\rightarrow$ akt, raf $\rightarrow$ mek
- **GAM SEM** (8 edges):
  - pip3 $\rightarrow$ pip2, pip3 $\rightarrow$ plc, pkc $\rightarrow$ p38, p38 $\rightarrow$ jnk, pka $\rightarrow$ akt, pka $\rightarrow$ erk, erk $\rightarrow$ akt, raf $\rightarrow$ mek
  - Extra edges compared to Gauss SEM in blue.
- **D-vine SEM** (10 edges):
  - pip3 $\rightarrow$ pip2, pip3 $\rightarrow$ plc, plc $\rightarrow$ pip2, pkc $\rightarrow$ jnk, pkc $\rightarrow$ p38, pka $\rightarrow$ akt, pka $\rightarrow$ mek, pka $\rightarrow$ erk, erk $\rightarrow$ akt, raf $\rightarrow$ mek
  - Extra edges compared to Gauss SEM in green.
- **D-vine SEM versus GAM SEM**:
  - D-vine SEM does not have edge p38 $\rightarrow$ jnk of GAM SEM, but has edges plc $\rightarrow$ pip2, pkc $\rightarrow$ jnk and pka $\rightarrow$ mek.

Chosen pair copulas in D-vine SEM (I)

- **Node plc with parent node pip3**

edge	family	parameters	tau	loglik
(plc, pip3)	bb8	1.70 , 0.98	0.25	100

- **Node pip2 with parents pip3 and plc**

edge	family	parameters	tau	loglik
(pip2, pip3)	bb7	1.50, 0.30	0.31	142
(pip3, plc)	bb8	1.70, 0.98	0.25	100
(pip2, plc; pip3)	nonpar		-0.10	144

- **Node p38 with parent node pkc**

edge	family	parameters	tau	loglik
(p38, pkc)	bb1	0.29 , 2.31	0.62	549

- **Node jnk with parent node pkc**

edge	family	parameters	tau	loglik
(jnk, pkc)	gauss	0.26 , —	0.17	33

Pair copulas in D-vine SEM (II)

- **Node mek with parents raf and pkc**

edge	family	parameters	tau	loglik
(mek, raf)	gauss	0.69, —	0.48	291
(raf, pkc)	ind	—, —	0.00	0
(mek, pkc; raf)	gauss	-0.11,—	-0.07	6

- **Node erk with parent node pka**

edge	family	parameters	tau	loglik
(erk, pka)	nonpar	— , —	0.13	187

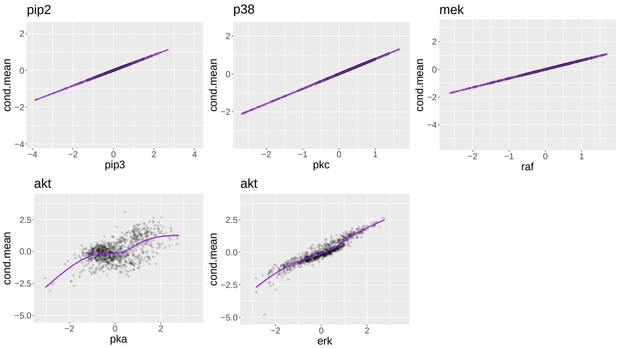
- **Node akt with parents erk and pka**

edge	family	parameters	tau	loglik
(akt, erk)	gumbel	3.00, —	0.67	663
(erk, pka)	nonpar	—, —	0.13	187
(akt, pka; erk)	bb8	2.50, 0.87	0.33	135

Fitted conditional means

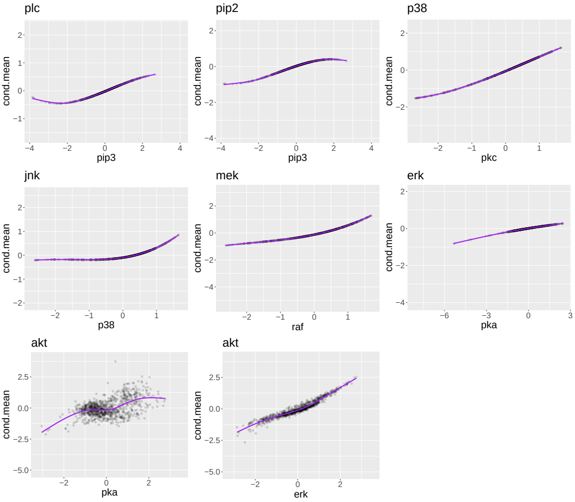
- For the Gaussian SEM and the GAM SEM the fits of the conditional means of each observations are computed and then non linearly smoothed over all observations (**purple** line in plots)
- For D-vine SEM the fitted conditional medians, 10% conditional quantiles and 90% conditional quantiles for all observations are calculated and non linearly smoothed. The fitted quantiles of the D-vine SEM can serve as 80% confidence interval.
- We plot the conditional means or quantiles as function of each parent variable separately.

Conditional means using Gauss SEM



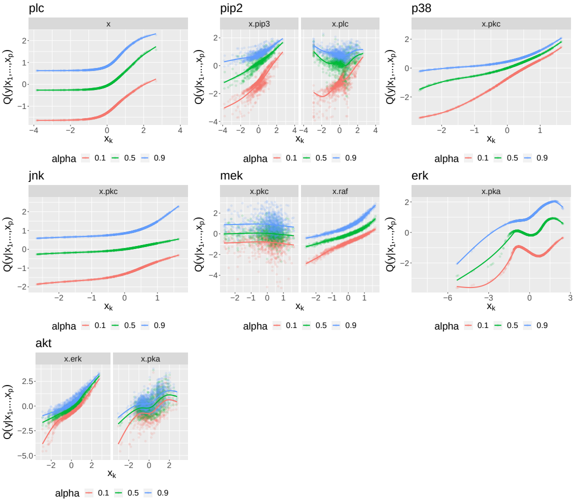
- **purple** smooths are mostly linear as postulated by model
- Only `akt` has two parents (`pka,erk`)

Conditional means using GAM SEM



non linear conditional means effects, but no confidence intervals available

Conditional means using D-vine SEM



non linear conditional means effects with confidence intervals

Summary of Sachs analysis results

- Gaussian DAG is not appropriate for this experimental setting of the Sachs data
- GAM SEM does not allow for confidence intervals, while the D-Vine SEM does.
- D-Vine SEM selects more edges compared to the other two methods.
- More complex marginal conditional median effects are seen in D-vine SEM compared to the mean effects in the GAM SEM.
- Many non Gaussian pair copulas are needed for the D-vine SEM.

Outline

Summary and outlook

- D-vine SEM's useful tool to identify and analyze non Gaussian graphical data
- Extension to R-vine based SEM's (start with Chang and Joe (2019)) and/or discrete variables (start with Panagiotelis et al. (2012)) are possible.
- Develop forward and backward selection algorithms of parents.
- Bauer et al. (2012) use R-vine based pairwise conditional dependence tests within the PC algorithm, while Müller and Czado (2018) look at sparse R-vine DAG's Tepegjozova and Czado (2022) developed more suitable Y-vine structure to model bivariate conditional distributions. Can be utilized for identifying DAG's from data.
- Higher dimensional case studies are needed.

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

Aragam, B. and Q. Zhou (2015).
Concave penalized estimation of sparse gaussian bayesian networks.
The Journal of Machine Learning Research 16(1), 2273–2328.

Bauer, A., C. Czado, and T. Klein (2012).
Pair-copula constructions for non-Gaussian DAG models.
Canadian Journal of Statistics 40, 86–109.

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

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

Brechmann, E., C. Czado, and K. Aas (2012).
Truncated regular vines in high dimensions with application to financial data.
Canadian Journal of Statistics 40, 68–85.

Chang, B. and H. Joe (2019).
Prediction based on conditional distributions of vine copulas.
Computational Statistics & Data Analysis 139, 45–63.

Czado, C., S. Jeske, and M. Hofmann (2013).
Selection strategies for regular vine copulae.
Journal de la Société Francaise de Statistique 154, 174–191.

Dißmann, J., E. Brechmann, C. Czado, and D. Kurowicka (2013).
Selecting and estimating regular vine copulae and application to financial returns.
Computational Statistics and Data Analysis 52(1), 52–59.

Fu, F. and Q. Zhou (2013).
Learning sparse causal gaussian networks with experimental intervention: regularization and coordinate descent.
Journal of the American Statistical Association 108(501), 288–300.

Joe, H. (1996).
Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters.
In L. Rüschendorf and B. Schweizer and M. D. Taylor (Ed.), *Distributions with Fixed Marginals and Related Topics*.

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

Kurowicka, D. and R. Cooke (2006).
Uncertainty analysis with high dimensional dependence modelling.
Chichester: Wiley.

Lauritzen, S. L. (1996).
Graphical Models (1st ed.).
Oxford, England: University Press.

Morales-Nápoles, O. (2011).
Counting vines.
In D. Kurowicka and H. Joe (Eds.), *Dependence Modeling: Vine Copula Handbook*, pp. 189–218. World Scientific Publishing Co.

Müller, D. and C. Czado (2018).
Representing sparse gaussian dags as sparse r-vines allowing for non-gaussian dependence.
Journal of Computational and Graphical Statistics 27(2), 334–344.

Nagler, T., C. Bumann, and C. Czado (2019).
Model selection in sparse high-dimensional vine copula models with an application to portfolio risk.
Journal of Multivariate Analysis 172, 180–192.

Nagler, T. and T. Vatter (2021).
rvinecopulib: High Performance Algorithms for Vine Copula Modeling.
R-project.
R package version 0.6.1.1.1.

Panagiotelis, A., C. Czado, and H. Joe (2012).
Pair copula constructions for multivariate discrete data.
Journal of the American Statistical Association 107, 1063–1072.

Pearl, J. (1988).
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann Publishers Inc.

Sachs, K., P. Omar, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan (2005).
Causal protein signaling networks derived from multiparameter single-cell data.
Science 308, 523–529.

Sklar, A. (1959).
Fonctions dé repartition á n dimensions et leurs marges.
Publ. Inst. Stat. Univ. Paris 8, 229–231.

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

Varin, C. and P. Vidoni (2005).
A note on composite likelihood inference and model selection.
Biometrika 92(3), 519–528.

Voorman, A., A. Shojaie, and D. Witten (2014).
Graph estimation with joint additive models.
Biometrika 101(1), 85–101.

Wang, Y. J. and E. H. Ip (2008).
Conditionally specified continuous distributions.
Biometrika 95(3), 735–746.