Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

www.nr.no

Pair-copula constructions: even more flexible than copulas

Workshop on Copulas and Extremes

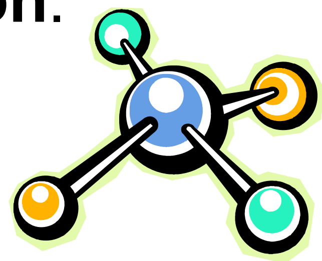
Grenoble, France, November 19th, 2013

Kjersti Aas
Norwegian Computing Center

Joint work with:
Claudia Czado, Ingrid Hobæk Haff, Arnoldo Frigessi, Daniel Berg, Eike C. Brechmann

Pair-copula construction

- While there is a multitude of bivariate copula, the class of multivariate copulae is still quite restricted.
- Hence, if the dependency structures of different pairs of variables in a multivariate problem are very different, not even the copula approach will allow for the construction of an appropriate model.
- In this talk I will describe an extension to the state-of-the-art theory of copulas, modelling multivariate data using a so-called **pair-copula construction**.



Copula

- The Sklar's theorem states that every multivariate distribution F with marginals $F_1(x_1),\ldots,F_n(x_n)$ can be written as

$$F(x_1,\ldots,x_n) = C\left(F_1(x_1),\ldots,F_n(x_n)\right)$$

for some appropriate n-dimensional copula C.

- Using the chain rule, for an absolutely continuous joint distribution F with strictly increasing, continuous marginal distribution functions $F_1,\ldots F_n$ it holds that

$$f(x_1,\ldots,x_n) = c_{1\cdots n}(F_1(x_1),\ldots F_n(x_n))\prod_{i=1}^{n} f_i(x_i)$$

for some n-variate copula density $c_{1\cdots n}(\cdot)$.

Pair-copula constructions (I)

- For two random variables X_1 and X_2 we have

$$f(x_1|x_2) \quad = \quad c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1)$$

- Further, for three random variables X_1, X_2 and X_3 we have

$$f(x_1|x_2, x_3) \quad = \quad c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot f(x_1|x_2)$$

- It follows that for every j we have

$$f(x|\boldsymbol{v}) = c_{xv_j|\boldsymbol{v}_{-j}}(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})) \cdot f(x|\boldsymbol{v}_{-j})$$

Pair-copula construction (II)

By combining the two results:

$$f(x_1,\ldots,x_n) = f(x_n)\cdot f(x_{n-1}|x_n)\ldots\cdot f(x_2|x_3,\ldots,x_n)\cdot f(x_1|x_2,\ldots,x_n)$$

and

$$f(x|\boldsymbol{v}) = c_{xv_j|\boldsymbol{v}_{-j}}(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))\cdot f(x|\boldsymbol{v}_{-j}),$$

we may derive a decomposition of $f(x_1,\ldots,x_n)$ that only consists of marginal distributions and bivariate copulae.

We denote a such decomposition a pair-copula construction (PCC)

Joe (1996) was the first to give a probabilistic construction of multivariate distribution functions based on pair-copulas, while Aas et. al. (2009) were the first to set the PCC in an inferential context.

PCC in three dimensions

- A pair-copula construction of a three-dimensional density is given by:

$$
\begin{aligned}
& f(x_1, x_2, x_3) \\
= \quad & f(x_1) \cdot f(x_2) \cdot f(x_3) \\
\cdot \quad & c_{12}\left(F(x_1), F(x_2)\right) \cdot c_{23}\left(F(x_2), F(x_3)\right) \\
\cdot \quad & c_{13|2}\left(F(x_1|x_2), F(x_3|x_2)\right).
\end{aligned}
$$

Special case: Trivariate normal distribution

If the marginal distributions are standard normal, and c_{12}, c_{23} and $c_{13|2}$ are bivariate Gaussian copula densities, the resulting distribution is trivariate standard normal.

PCC in five dimensions

- A possible pair-copula construction for a five-dimensional density is:

$$
\begin{aligned}
& f(x_1, x_2, x_3, x_4, x_5) \\
= \; & f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot f(x_4) \cdot f(x_5) \\
\cdot \; & c_{12}(F(x_1), F(x_2)) \cdot c_{23}(F(x_2), F(x_3)) \cdot c_{34}(F(x_3), F(x_4)) \cdot c_{45}(F(x_4), F(x_5)) \\
\cdot \; & c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot c_{24|3}(F(x_2|x_3), F(x_4|x_3)) \cdot c_{35|4}(F(x_3|x_4), F(x_5|x_4)) \\
\cdot \; & c_{14|23}(F(x_1|x_2, x_3), F(x_4|x_2, x_3)) \cdot c_{25|34}(F(x_2|x_3, x_4), F(x_5|x_3, x_4)) \\
\cdot \; & c_{15|234}(F(x_1|x_2, x_4, x_3), F(x_5|x_2, x_4, x_3)) \, .
\end{aligned}
$$

- There are as many as 480 different such constructions in the five-dimensional case, 23,040 in the 6-dimensional case and 2,580,480 in the 7-dimensional case……….

Vines

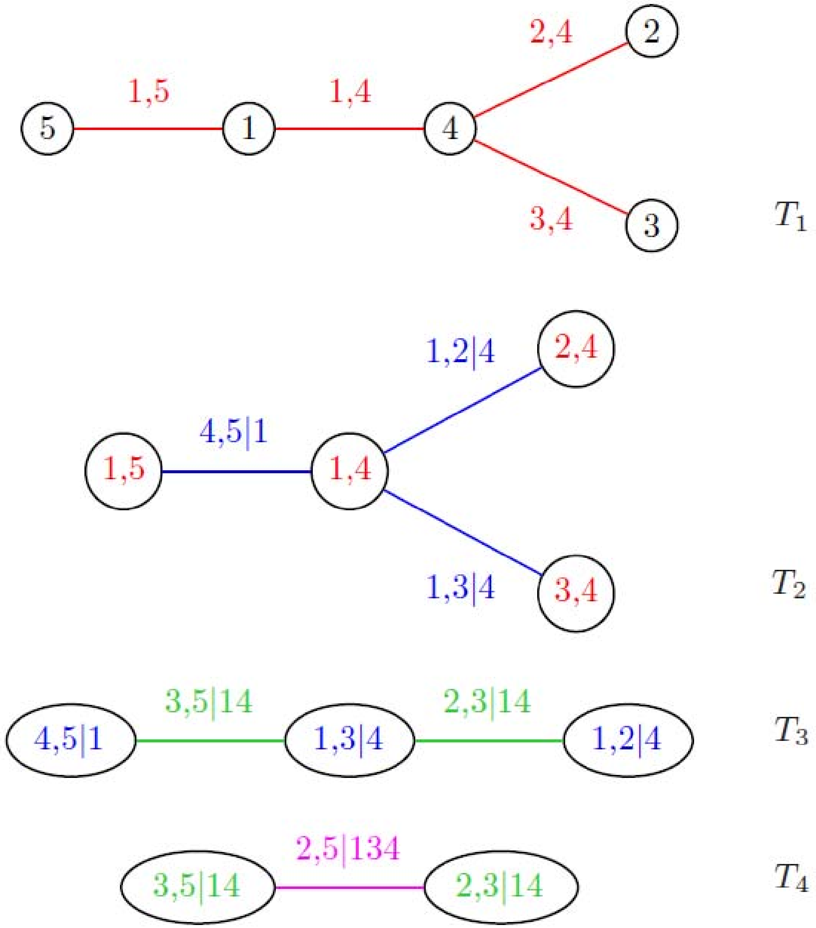
- Hence, for high-dimensional distributions, there are a significant number of possible pair-copula constructions.
- To help organising them, *Bedford and Cooke* (2001) introduced graphical models denoted **regular vines (R-vines)**.

Regular vine (Bedford and Cooke 2002)

A regular vine is a sequence of $d-1$ linked trees where:

- Tree T_1 is a tree on nodes 1 to d.
- Tree T_j has $d+1-j$ nodes and $d-j$ edges.
- Edges in tree T_j become nodes in tree T_{j+1}.
- **Proximity condition:** Two nodes in tree T_{j+1} can be joined by an edge only if the corresponding edges in tree T_j share a node.

Example in five dimensions



Density

$$f = f_1 \cdot f_2 \cdot f_3 \cdot f_4$$

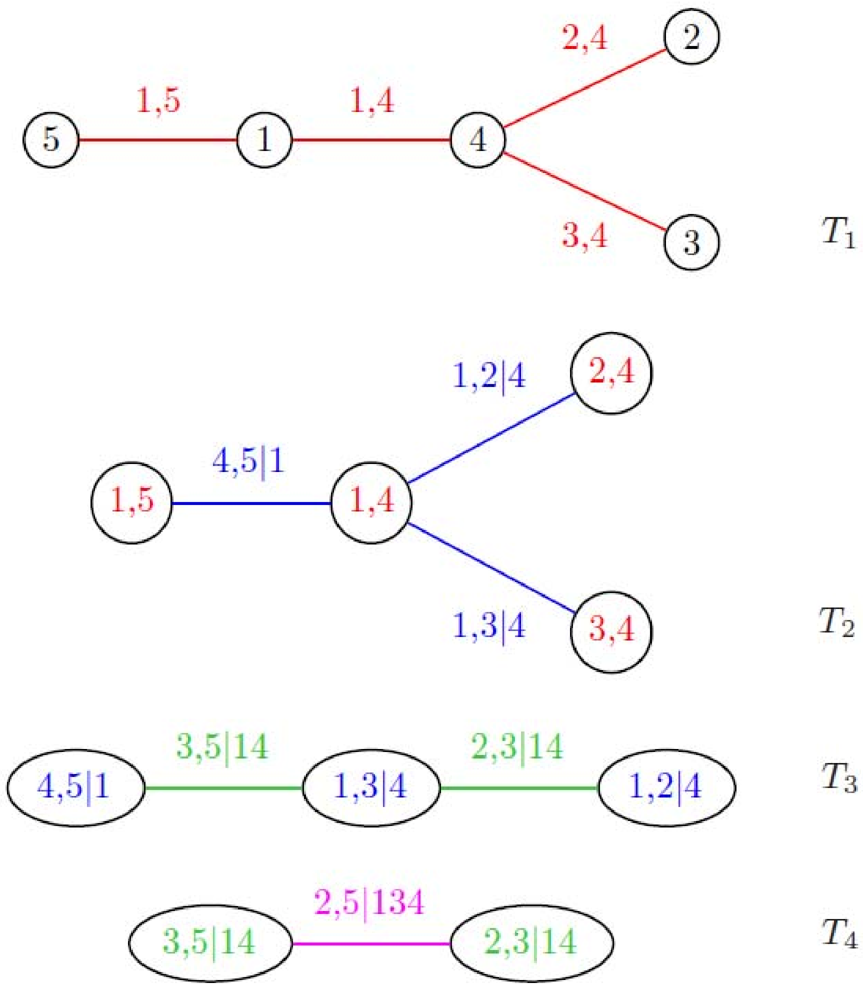
$\cdot\; c_{14} \cdot c_{15} \cdot c_{24} \cdot c_{34}$

$\cdot\; c_{12;4} \cdot c_{13;4} \cdot c_{45;1}$

$\cdot\; c_{23;14} \cdot c_{35;14}$

$\cdot\; c_{25;134}$

Matrix representation

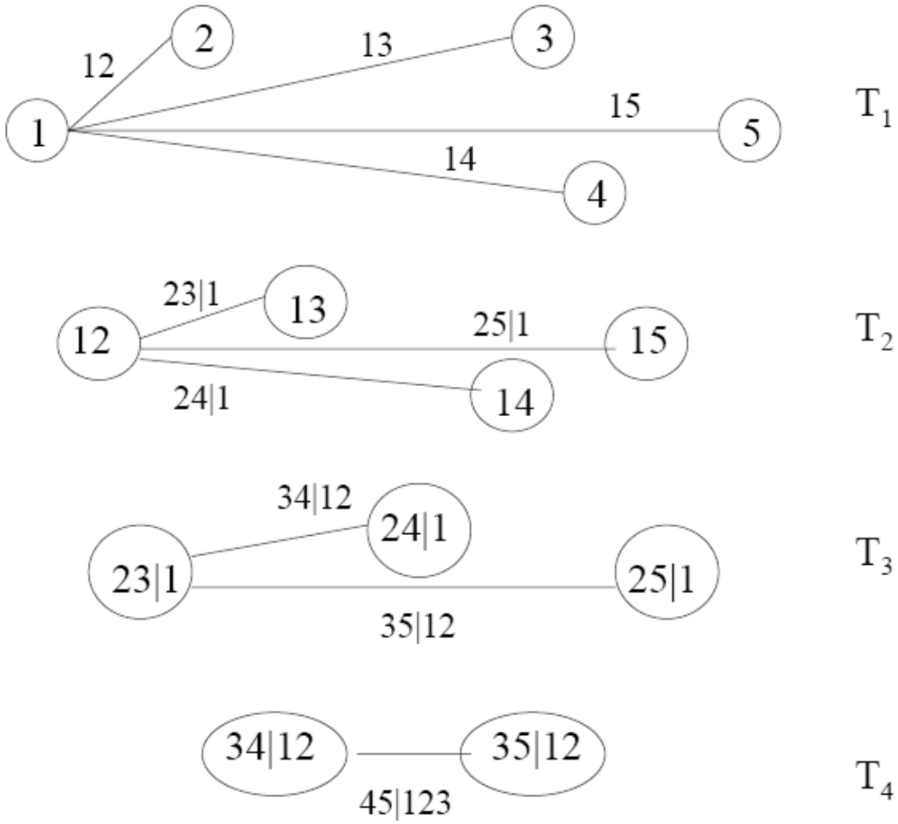


Matrix

Morales-Napoles (2008) shows how a lower triangular matrix may be used to store a regular vine.

$$M = \begin{pmatrix} 5 & & & & \\ 2 & 2 & & & \\ 3 & 3 & 3 & & \\ 4 & 1 & 1 & 1 & \\ 1 & 4 & 4 & 4 & 3 \end{pmatrix}$$

Special case: C-vine

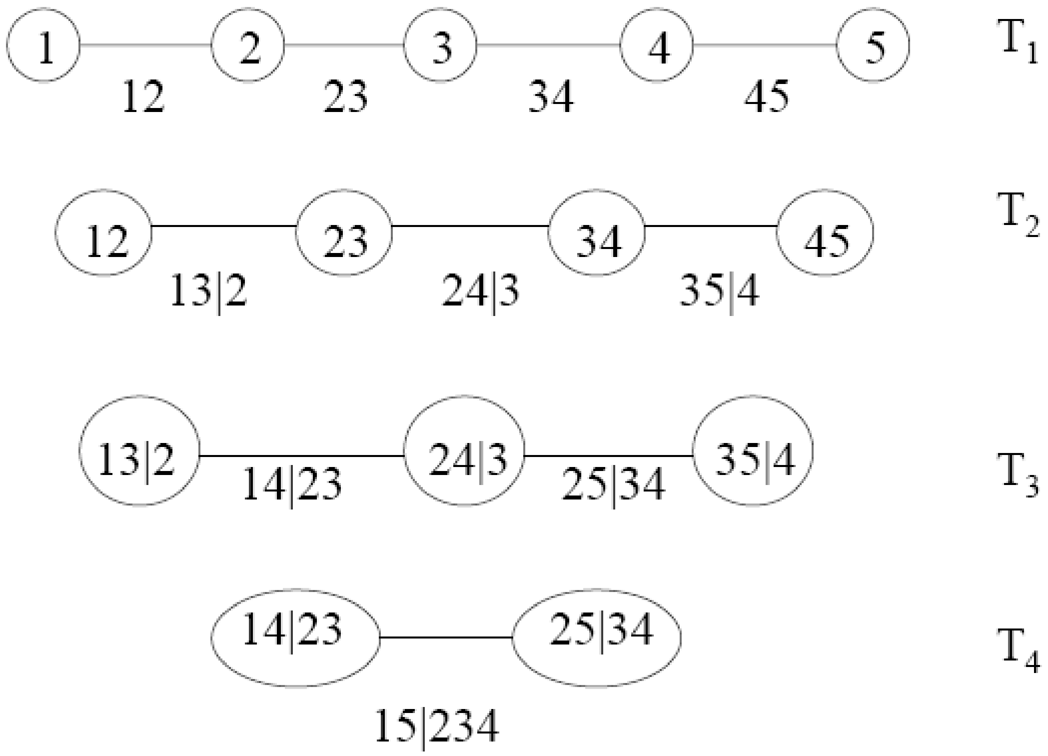


Each tree has a unique node that is connected to n-j edges.

$$
\begin{aligned}
f_{12345} &= f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \\
&\cdot\ c_{12} \cdot c_{13} \cdot c_{14} \cdot c_{15} \\
&\cdot\ c_{23;1} \cdot c_{24;1} \cdot c_{25;1} \\
&\cdot\ c_{34;12} \cdot c_{35;12} \\
&\cdot\ c_{45;123}
\end{aligned}
$$

Useful for ordering of importance

Special case: D-vine



No node in any tree is connected to more than two edges.

$$
\begin{aligned}
f_{1234} &= f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \\
&\cdot \; c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{45} \\
&\cdot \; c_{13;2} \cdot c_{24;3} \cdot c_{35;4} \\
&\cdot \; c_{14;23} \cdot c_{25;34} \\
&\cdot \; c_{15;234}
\end{aligned}
$$

Useful for temporal ordering.

General density expressions

- C-vine (Aas et al. 2009)

$$f(x_1, \dots x_d) = \left[\prod_{k=1}^{d} f(x_k)\right] \times \left[\prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{j,j+i;1,\dots,j-1}\right]$$

- D-vine (Aas et al. 2009)

$$f(x_1, \dots x_d) = \left[\prod_{k=1}^{d} f(x_k)\right] \times \left[\prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i,i+j;i+1,\dots,i+j-1}\right]$$

- Regular vine (Dißmann et al. 2013)

$$f(x_1, \dots, x_d) = \left[\prod_{k=1}^{d} f_k(x_k)\right] \times \left[\prod_{j=d-1}^{1}\prod_{i=d}^{j+1} c_{m_{j,j},m_{i,j};m_{i+1,j},\dots,m_{n,j}}\right]$$

Here, $m_{i,j}$ refers to element (i,j) in the matrix representation of the R-vine.

Conditional distribution functions

- The conditional distributions needed as copula arguments at level *j* are obtained as partial derivatives of the copulae at level *j-1*
- This is due to the following result of Joe (1996) stating that under regularity conditions we have:

$$F(x|\boldsymbol{v}) = \frac{\partial\, C_{x,v_j|\boldsymbol{v}_{-j}}(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))}{\partial F(v_j|\boldsymbol{v}_{-j})}$$

The terms **tree** and **level** are used as synonyms in this talk

The h-function

- It turns out that we only need the special case of *F(x|**v**)* when ***v*** is univariate and *x* and *v* are uniformly distributed on [0,1], i.e.

$$F(x|v) = \frac{\partial\, C_{x,v}(x, v, \Theta)}{\partial v}$$

where Θ is the set of copula parameters.

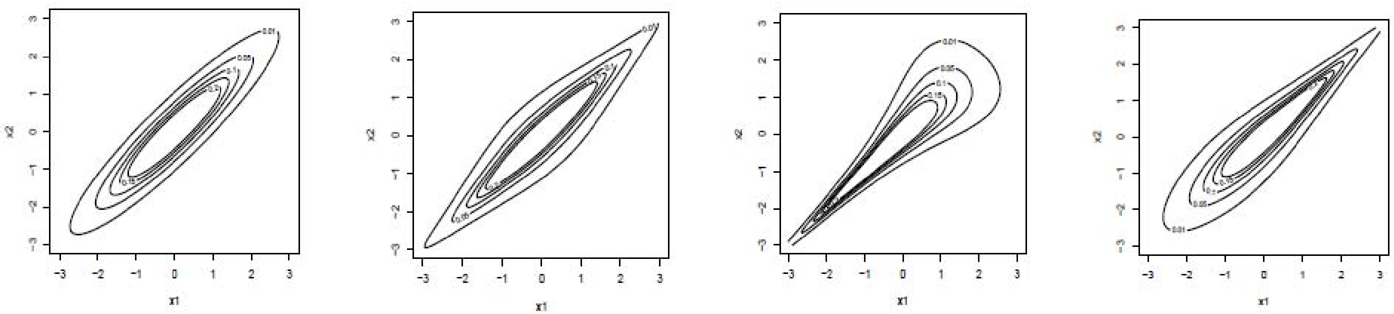
- From now on *F(x|v)* is denoted the **h-function**:

$$h(x, v, \Theta) = F(x|v) = \frac{\partial\, C_{x,v}(x, v, \Theta)}{\partial v}.$$

Building blocs

- The resulting multivariate distribution will be valid even if the bivariate copulae involved in the pair-copula construction are of different type.

- One may for instance combine the following types of pair-copulae
  - Gaussian (no tail dependence)
  - Clayton (lower tail dependence)
  - Gumbel (upper tail dependence)
  - Student (upper and lower tail dependence)



www.nr.no

Parameter estimation

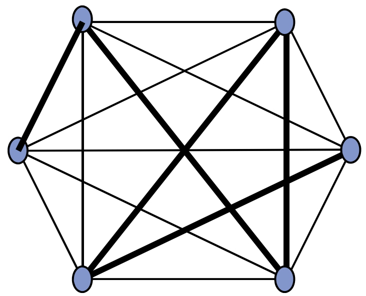
Three elements

- Full inference for a pair-copula decomposition should consider the following three tasks:

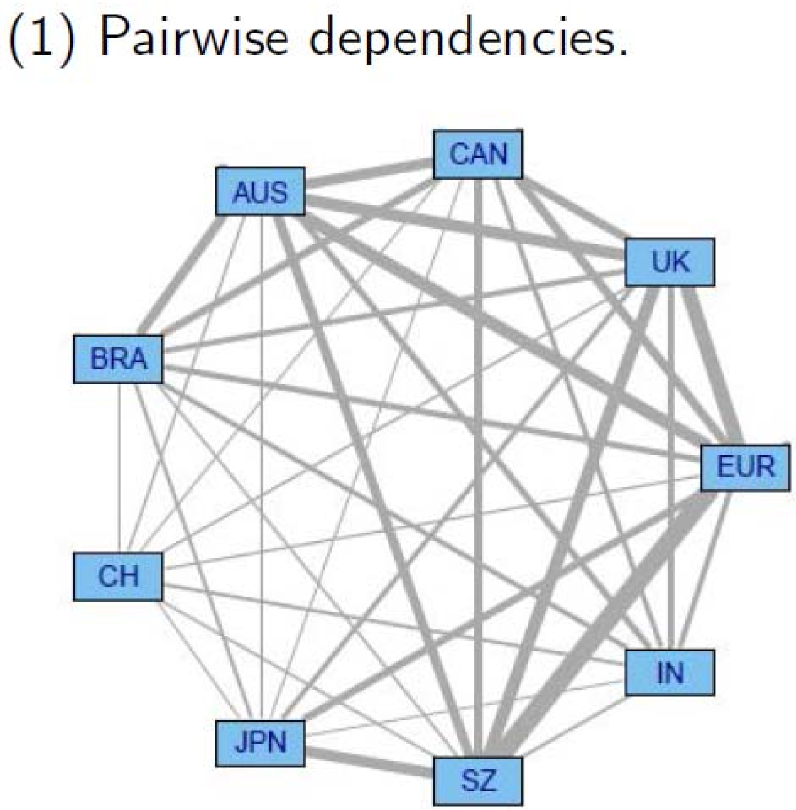
1. The selection of a specific factorisation.
2. The choice of pair-copula types.
3. The estimation of the parameters of the chosen pair-copulae.

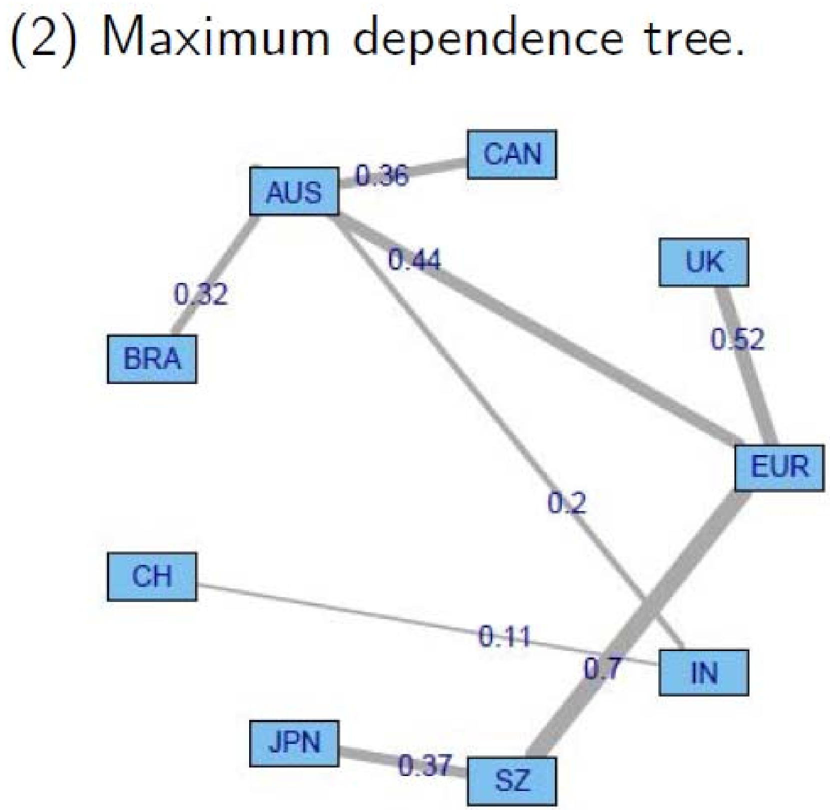
Which factorisation?

- The current idea is to capture the strongest pairwise dependencies in the first levels.
- Hence, for each tree we first calculate an empirical dependence measure (e.g. Kendall's tau) for all variable pairs, and then we select the tree on all nodes that maximizes the sum of absolute empirical dependencies using the spanning tree algorithm of Prim.



How does this look like for Tree 1?





Choice of copula-types

- The following procedure may be used to select copula types:

1. Determine which parametric classes to use at level 1 by plotting the original data, and/or by applying a Goodness-of-Fit (GoF) test.

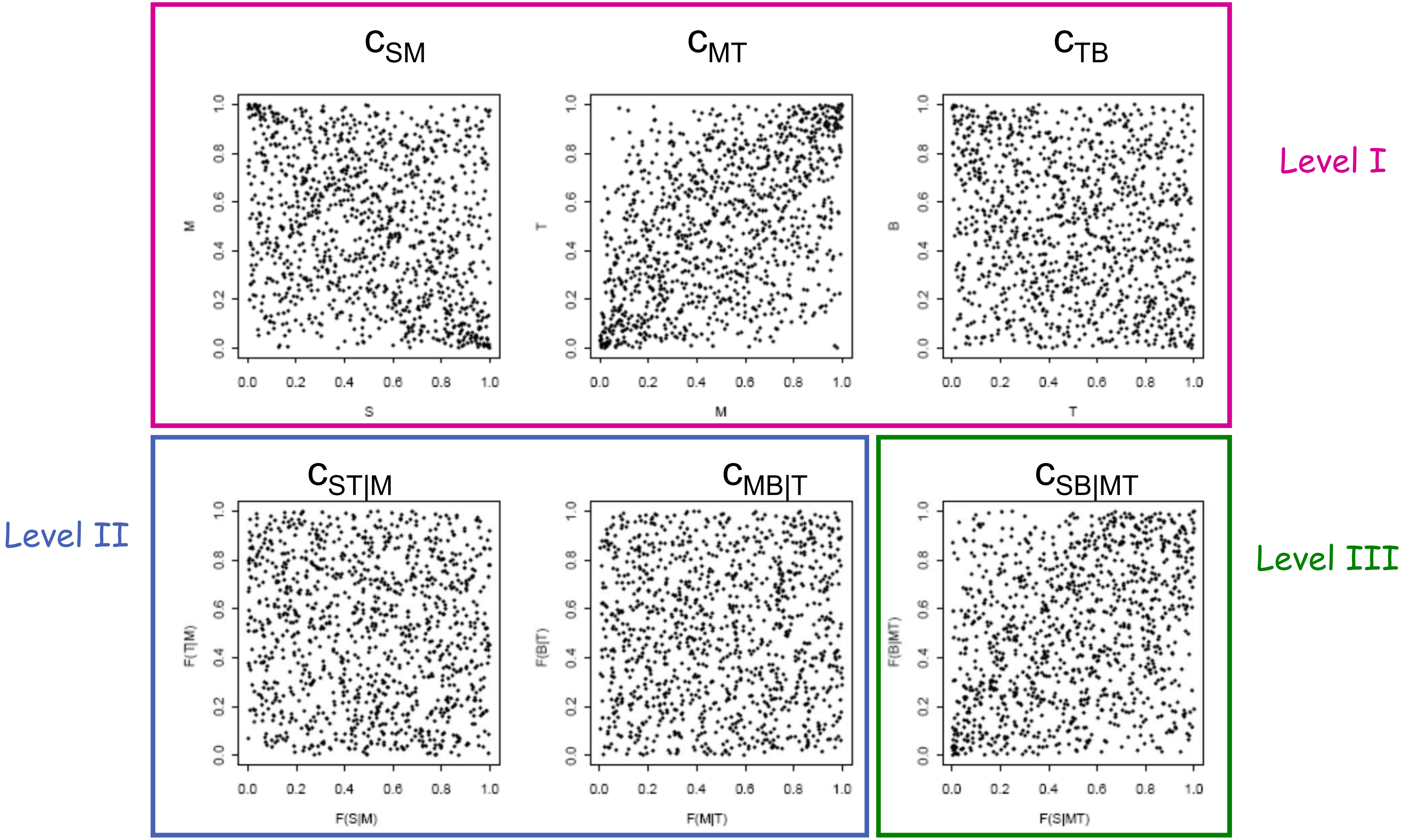
2. Estimate the parameters of the selected copulae using the original data.

3. Transform observations as required for level 2, using the parameters from level 1 and the $h(\cdot)$ functions for the selected copulas.

4. Repeat 1-3 for all levels 2,3...

This procedure is also denoted **sequential** or **stepwise semi-parametric** estimation.

Example



C_{SM}
C_{MT}
C_{TB}
Level I
$C_{ST|M}$
$C_{MB|T}$
Level II
$C_{SB|MT}$
Level III
S
M
T
B
F(S|M)
F(T|M)
F(M|T)
F(B|T)
F(S|MT)
F(B|MT)
0.0 0.2 0.4 0.6 0.8 1.0

Likelihood evaluation

Algorithm 3 Likelihood evaluation for canonical vine

log-likelihood = 0
for $i \leftarrow 1, 2, \ldots, n$
  $\boldsymbol{v}_{0,i} = \boldsymbol{x}_i$.
end for
for $j \leftarrow 1, 2, \ldots, n-1$
  for $i \leftarrow 1, 2, \ldots, n-j$
    log-likelihood = log-likelihood + $L(\boldsymbol{v}_{j-1,1}, \boldsymbol{v}_{j-1,i+1}, \Theta_{j,i})$
  end for
  for $i \leftarrow 1, 2, \ldots, n-j$
    $\boldsymbol{v}_{j,i} = h(\boldsymbol{v}_{j-1,i+1}, \boldsymbol{v}_{j-1,1}, \Theta_{j,i})$
  end for
end for

$$v_{j,i,t} = F(x_{i+j,t}|x_{1,t}, \ldots x_{j,t})$$

Θ_{ji} are the parameters of copula density $c_{j,j+i|1,\ldots,j-1}(\cdot)$

$$L(\boldsymbol{x}, \boldsymbol{v}, \Theta) = \sum_{t=1}^{T} \log\left(c(x_t, v_t, \Theta)\right)$$

The SSP-estimator

- Full semi-parametric maximum likelihood estimation (SP) has shown to be consistent and asymptotically normal (Genest, 1995, Tsukahara, 2005).
- However, it is computationally too heavy in high dim.
- Hence, people tend to use the stepwise semi-parametric (SSP-) approach (Aas et. al., 2009) instead.
- In the SSP approach, the parameters of the vine are sequentially estimated starting from the top tree.
- The performance of SSP and SP is quite similar, but SSP is computationally much faster than SP.

Properties of the SSP-estimator

- Hobæk Haff (2011a) have shown that
  - The SSP-estimator is less efficient than the SP-estimator in general.
  - This loss of efficiency may however be rather low.
  - The SSP-estimator is semiparametrically efficient for the Gaussian copula.
- Hobæk Haff (2011b) have shown that
  - The finite sample bias and MSE of SSP are higher than those of SP (the difference increases with increasing dependency).
  - With a small sample size or misspecification of the model, the difference between SP and SSP however becomes smaller.

Simplifying assumption

- Generally, the parameters of the conditional density $c_{13|2}(F(x_1|x_2), F(x_3|x_2))$ depends on the value of $x_{2.}$
- Inference requires however the simplifying assumption that all pair copulae depend on the conditioning variables through the two conditional distribution functions that constitute their arguments only, and not directly.
- As shown in Hobæk Haff et. al. (2010), this seems not to be a severe restriction.

Application:

Market risk model for the largest Norwegian bank, DNB

Data set

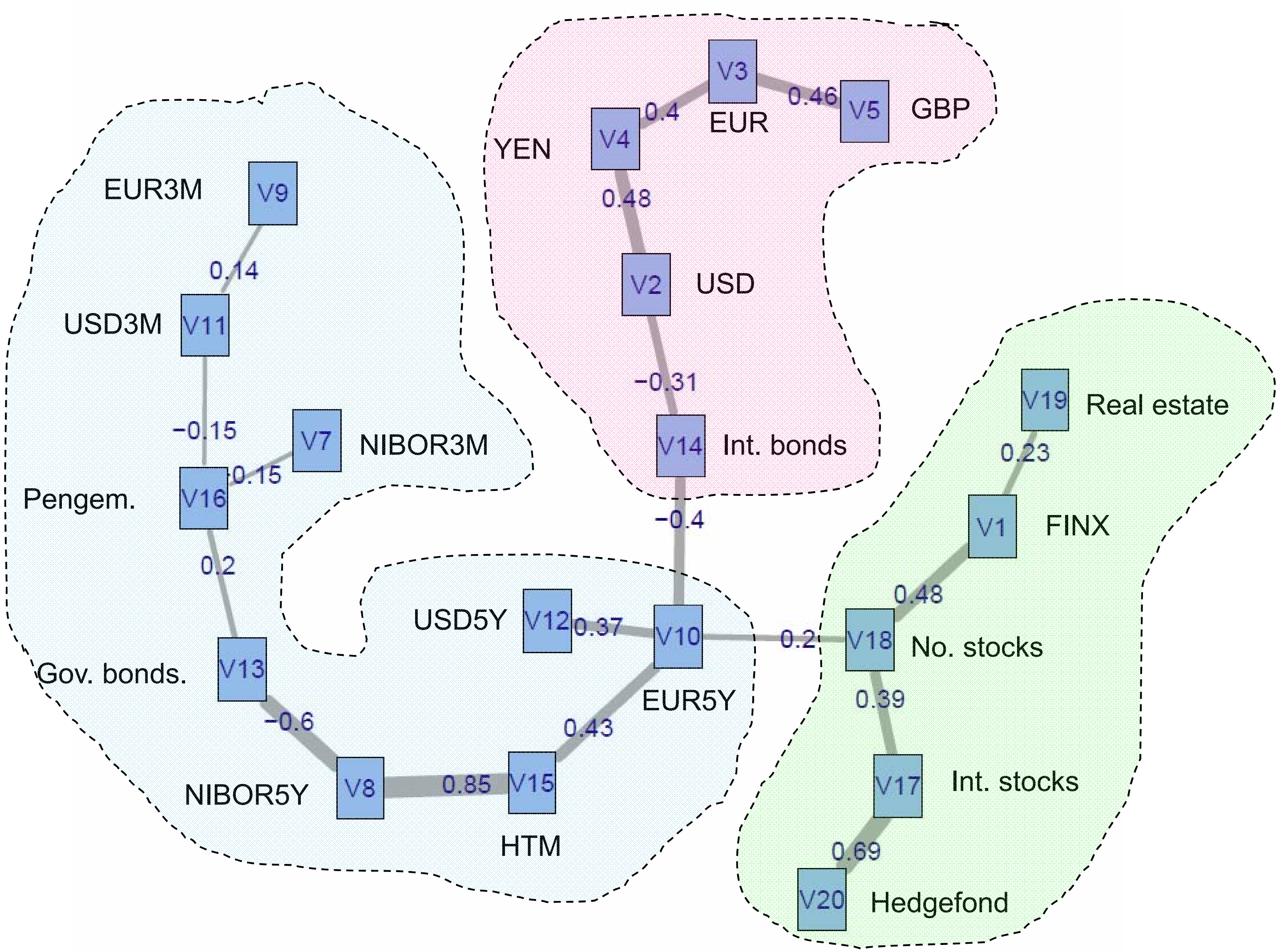
- 19 financial variables that constitute the market portfolio of DNB.
- Daily log returns from March 2003 to March 2008 (1107 obs.) are used.

ID	description	ID	description
V1	Norwegian Financial Index	V12	5-year US Government Rate
V2	USD-NOK exchange rate	V13	Norwegian bond index (BRIX)
V3	EURO-NOK exchange rate	V14	Citigroup World Government Bond Index (WGBI)
V4	YEN-NOK exchange rate		
V5	GBP-NOK exchange rate	V15	Norwegian 6-year Swap Rate
V7	3-month Norwegian Inter Bank Offered Rate	V16	ST2X - Government Bond Index (fix modified duration of 0.5 years)
V8	Norwegian 5-year Swap Rate	V17	Morgan Stanley World Index (MSCI)
V9	3-month Euro Interbank Offered Rate	V18	OSEBX - Oslo Stock Exchange main index
V10	5-year German Government Rate	V19	Oslo Stock Exchange Real Estate Index
V11	3-month US Libor Rate	V20	S&P Hedge Fund Index

Modelling procedure

- Fit appropriate ARMA-GARCH models for log-return time series.
- Fit an R-vine as well as a multivariate Student-t copula (for comparison) to standardized residuals
- Pair-copulas are selected from a range of 11 bivariate families using AIC:
  - Independence copula, Gaussian, t, Clayton, rotated Clayton (90°), Gumbel, rotated Gumbel (90°), Frank, Joe, Clayton-Gumbel (BB1), Joe-Clayton (BB7).

First tree of R-vine



Results

Copula	Log likelihood	No. of param.	AIC
R-vine	6390.75	92	-12597.50
Student-t	6324.98	172	-12305.96

Number of parameters:

Note that the number of parameters to be estimated for a 19-dimensional R-vine usually is at least d(d-1)/2. The reason why the number in the table is 92 and not 171 is that a large amount of the pair-copulae in this R-vine are identified as the independence copula, using the bivariate independence test based on Kendall's tau as described in Genest and Favre (2007) .

Truncation (I)

- The number of parameters in an R-vine grows quadratically with the dimension.
- Hence, it would be useful to be able to reduce the model complexity.
- In Brechmann et. al (2012) we have studied the problem of determining whether an R-vine may be **truncated**.
- By a truncated R-vine at level K, we mean an R-vine with all pair-copulae with conditioning set larger than or equal to K set to independence copulae.

Truncation (II)

- We fit one tree at a time and use the likelihood ratio test of Vuong (1989) to determine whether an additional tree provides a significant gain in the model fit.

$$
\begin{array}{lllllllll}
T_1: & c_{12} & & c_{23} & & c_{34} & & c_{45} \\
T_2: & & c_{13|2} & & c_{24|3} & & c_{35|4} & \\
\hline
T_3: & & & c^{ind}_{14|23} & & c^{ind}_{25|34} & & \\
T_4: & & & & c^{ind}_{15|234} & & &
\end{array}
$$

Truncation (III)

Copula	Log likelihood	No. of param.	AIC
R-vine	6390.75	92	-12597.50
Student-t	6324.98	172	-12305.96
6-level R-vine	6274.47	77	-12394.94
4-level R-vine	6234.05	68	-12332.10

Conclusion:

We conclude from this that the most important dependencies in this data set are actually captured in the first four to six trees, meaning that the corresponding R-vine copula may be truncated at level 6, or even at level 4, depending on the desired level of parsimony (and of course at the expense of accuracy).

Recent advances connected to PCC

Applications

- Finance
- Insurance
- Genetics
- Marketing
- Health
- Hydrology
- Infrastructure modeling
- Image analysis




PCC types

- Non-simplified PCC (Acar et. al., 2012).
- PCC with time-varying parameters (Almeida et. al., 2012, So & Yeung, 2013).
- Regime-switching PCC (Chollete et. al., 2008, Stöber & Czado, 2013).
- Non-parametric PCC (Haff & Segers, 2013, Kauermann & Schellhase, 2013).
- Spatial PCC (Gräler & Pebesma, 2011).
- PCC with discrete margins (Panagiotelis et. al., 2012).
- PCC for longitudinal data (Smith et. al., 2010).
- PCC with Lévy copulas (Grothe & Nicklas, 2013).

www.nr.no

Summary

Summary

- Pair-copula decomposed models represent a very flexible and intuitive way of constructing higher-dimensional copulae.
- Simulation and inference are straight-forward (but time-consuming in higher dimensions).
- Sequential and MLE parameter estimation of C-, D- and R-vines are available in R packages CDVine and VineCopula.