Nested Archimedean copulas: Structure estimation and goodness of fit

by Nathan Uyttendaele (ISBA*; UCL**)

nathan.uyttendaele@uclouvain.be

Workshop “Copulas and Extremes”

Grenoble, France

November 19-20, 2013

*Institut de Statistique, Biostatistique et Sciences Actuarielles;
**Université Catholique de Louvain.

First things first

Structure estimation: a first approach based on trivariate pieces

Structure estimation: a second approach based on binary structures

Structure estimation: a third approach inspired by phylogenetics

Is a NAC a good idea for my data?
(work in progress)

The future of research about NACs

Copulas

Let $(X_1, \ldots, X_d)$ be a vector of continuous random variables. Then the unique copula of this vector is defined as

$$C(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d)$$

with

$$(U_1, \ldots, U_d) = (F_{X_1}(X_1), \ldots, F_{X_d}(X_d))$$

and where $F_{X_1}, \ldots, F_{X_d}$ are the marginal CDFs of $(X_1, \ldots, X_d)$.

By studying the copula function, you study how the variables depend on each other, how they interact.

Archimedean Copulas

Archimedean copulas are a class of copulas that admit the (simple) representation

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d))$$

where ψ is called the generator and must be d-monotone on $[0, \infty)$, see McNeil and Nešlehová (2009) for more details.

Example of generator (Clayton generator):

$$\psi(x) = (1 + x)^{-1/\theta}$$

$$\theta \in (0, \infty)$$

Archimedean Copulas

An Archimedean copula is defined through its generator, ψ.

Estimation is usually performed either by assuming ψ is known up to some Euclidean parameter(s) or by not assuming anything about ψ, i.e., you have to estimate the whole ψ function, see Genest et al. (2011).

An important drawback of Archimedean copulas

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d))$$

Since Archimedean copulas are highly symmetric functions, all margins of the same dimension are equal.

For modeling purposes, this becomes an increasingly strong assumption as the dimension d grows.

Introducing asymmetries: nested Archimedean copulas

Asymmetries, that is, more realistic dependencies, can be modeled by a hierarchical structure of Archimedean copulas, obtained by plugging in Archimedean copulas into each other (Joe, 1997).

Nested Archimedean copulas

Start from an Archimedean copula (not necessarily a bivariate one):

$$C_0(u_i, \bullet\,) = \psi_0(\psi_0^{-1}(u_i) + \psi_0^{-1}(\,\bullet\,))$$

where the argument • is replaced by another Archimedean copula (again, not necessarily a bivariate one), such as

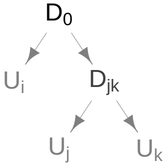
$$C_{jk}(u_j, u_k) = \psi_{jk}(\psi_{jk}^{-1}(u_j) + \psi_{jk}^{-1}(u_k))$$

in order to get a nested Archimedean copula of the form

$$C_0(u_i, C_{jk}(u_j, u_k)) = \psi_0\big(\psi_0^{-1}(u_i) + \psi_0^{-1}(\psi_{jk}(\psi_{jk}^{-1}(u_j) + \psi_{jk}^{-1}(u_k)))\big)$$

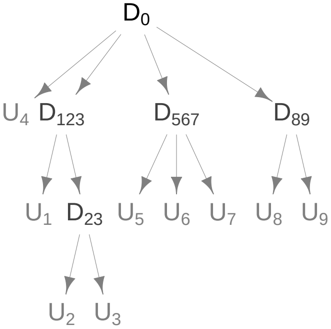
Nested Archimedean copulas

The way the two previous Archimedean copulas were plugged in corresponds to the following structure, which we will refer to as λ_{jk} later on:



Nested Archimedean copulas

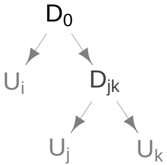
The hierarchical structure inherent to any nested Archimedean copula is actually a roadmap of dependencies:



Nested Archimedean copulas

Nested Archimedean copulas, also called hierarchical Archimedean copulas (HAC), are made up of two things: a tree structure and a collection of generators, one for each internal node of the structure.

$$C_0(u_i, C_{jk}(u_j, u_k)) = \psi_0\big(\psi_0^{-1}(u_i) + \psi_0^{-1}(\psi_{jk}(\psi_{jk}^{-1}(u_j) + \psi_{jk}^{-1}(u_k)))\big)$$



Notice that Archimedean copulas are a special case of Nested Archimedean copulas. It is implied throughout this presentation that the class of NACs encompass the class of Archimedean copulas.

Nested Archimedean copulas: estimation of the structure

Based on an iid sample of size n from $(X_1, ..., X_d)$ and admitting the joint distribution of $(U_1, ..., U_d)$ is a nested Archimedean copula,

how to estimate the structure by making the less and weakest assumptions possible about that NAC?

Even beter: how to estimate the structure by making NO assumption at all about that NAC?

If you assume all generators accross the (unknown) structure are known up to one Euclidean parameter and that the parameter's values are strictly increasing as you go down in the structure, see Okhrin et al. (2013).

First things first

Structure estimation: a first approach based on trivariate pieces

Structure estimation: a second approach based on binary structures

Structure estimation: a third approach inspired by phylogenetics

Is a NAC a good idea for my data?
(work in progress)

The future of research about NACs

Recovering a target structure from trivariate structures

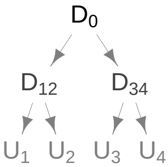
Key point for the first approach: if the structure of (U_i, U_j, U_k) is known for any distinct $i, j, k \in \{1, ..., d\}$, then the structure of $(U_1, ..., U_d)$ can be retrieved. That is, it is sufficient to know the marginal structure of all possible sets of three variables to retrieve the target structure.

Proof: see Segers and Uyttendaele (2013)*

*Paper to appear in CSDA soon, manuscript already available on the CSDA website.

Recovering a target structure from trivariate structures: an example

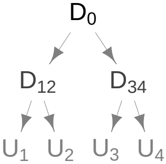
Suppose $d = 4$, that is we have (U_1, U_2, U_3, U_4). Say the structure is:



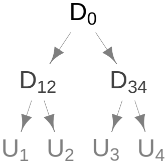
There are $\binom{4}{3} = 4$ marginal trivariate structures for this vector.

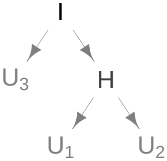
What are they?

Recovering a target structure from trivariate structures: an example

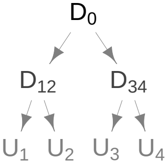


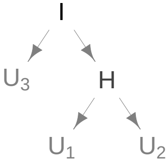
Recovering a target structure from trivariate structures: an example

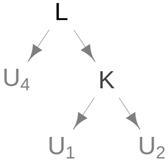




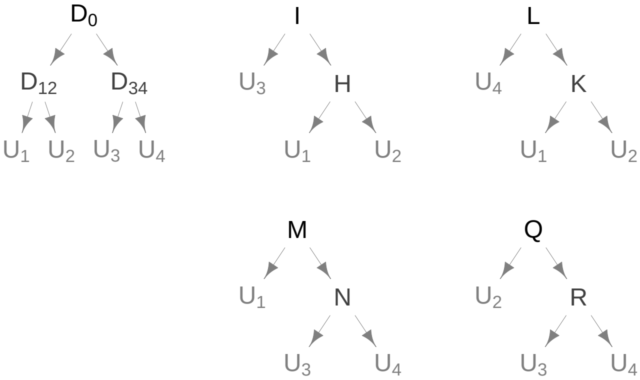
Recovering a target structure from trivariate structures: an example



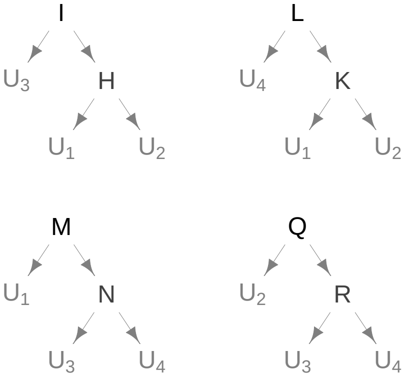




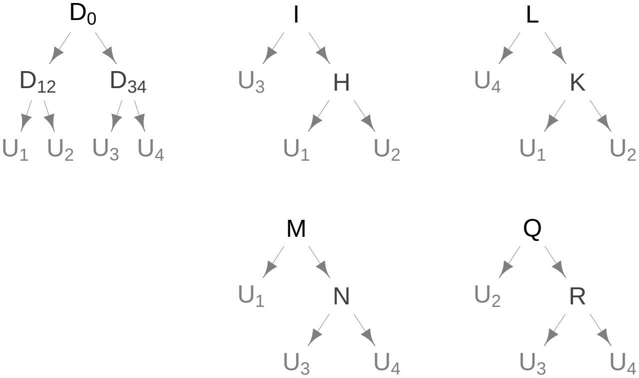
Recovering a target structure from trivariate structures: an example



Recovering a target structure from trivariate structures: an example

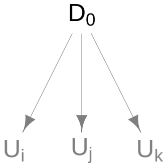


Recovering a target structure from trivariate structures: an example

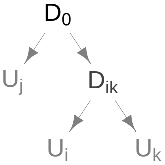


Estimation of a trivariate structure

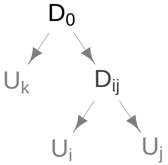
Another key point: for any (U_i, U_j, U_k), there are only four possible structures:



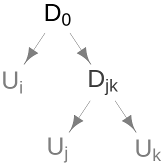
Λ_{ijk}



λ_{ik}



λ_{ij}



λ_{jk}

Estimation of a trivariate structure

We need to perform comparisons between empirical bivariate distributions in order to choose the correct trivariate structure. Not an easy problem. But....

...it is known from Genest and Rivest (1993) that the Kendall distribution of a pair of variables (X_j, X_k) fully determines the copula of that pair if it is an Archimedean copula. And, in a NAC, any bivariate distribution is actually an Archimedean copula. Kendall distributions are univariate.

Kendall distribution function

Define a random variable W_{jk} as

$$W_{jk} = C_{jk}(U_j, U_k)$$

Then the map defined, for all $w \in [0, 1]$, by

$$K_{jk}(w) = P(W_{jk} \leq w)$$

is the Kendall distribution function (Barbe et al. 1996; Nelsen et al. 2003; Genest and Rivest 2001).

Estimation of a trivariate structure: the procedure

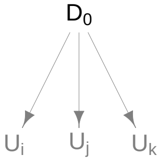
First, estimate the three empirical Kendall distributions (Genest, Nešlehová, and Ziegel, 2011), that is, get $\widehat{K}_{ij}, \widehat{K}_{ik}, \widehat{K}_{jk}$.

Calculate a distance between any two empirical Kendall distributions. Get $\delta_{ij,ik}, \delta_{ij,jk}, \delta_{ik,jk}$.

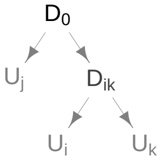
Pick the smallest of the three distances. Is it significantly smaller than the two others?

Say $\delta_{ij,jk}$ is the minimum distance. Is it significantly smaller than $\delta_{ij,ik}$ or $\delta_{ik,jk}$?

If no, then the true structure must be Λ_{ijk}, also called the trivial trivariate structure:



If yes, then the true structure must be λ_{ik}:



This last problem can be formalized as

$$\begin{array}{ll} H_0: & \text{the true structure is } \Lambda_{ijk}\text{, the trivial structure.} \\ H_1: & \text{the true structure is not } \Lambda_{ijk}. \end{array}$$

and if H_0 is rejected, being able to identify the smallest distance among $\delta_{ij,ik}, \delta_{ij,jk}$ and $\delta_{ik,jk}$ also enables to easily pick a structure among the three remaining structures λ_{jk}, λ_{ik} and λ_{ij}.

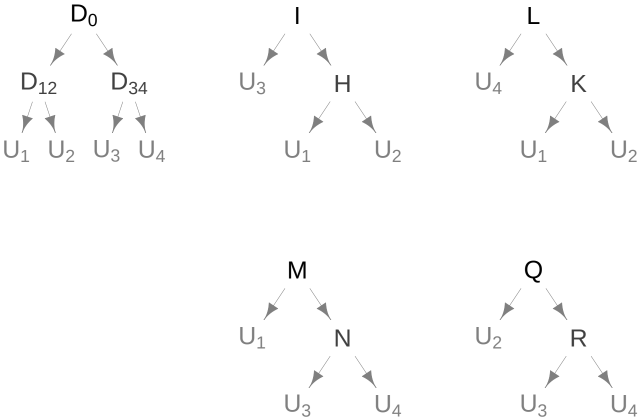
The test statistic for this test is equal to the difference of the minimum distance and the average of the two remaining distances. The H_0 distribution can be found thanks to some bootstrap.

The biggest difficulty of this first approach

A given target structure can always be broken down into a set of trivariate structures. This set of trivariate structures can then be used to recover the target structure.

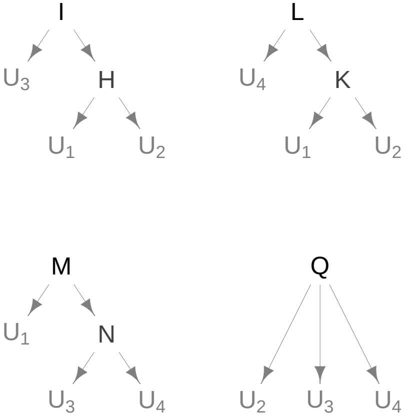
BUT, given a set of trivariate structures, it is not always possible to retrieve a target structure. Such sets of trivariate structures are called *broken* or *faulty* sets.

The biggest difficulty of this first approach: an example



The biggest difficulty of this first approach: an example

?



The biggest difficulty of this first approach

A quick fix is however suggested in Segers and Uyttendaele (2013) to ensure we always end up with a d-variate estimated structure at the end of this first approach.

Unfortunately, if the true distribution of $(U_1, \ldots, U_d)$ is not a NAC, the end result of this quick fix is usually that the final estimated structure is the one of a d-variate Archimedean copula, that is, a trivial structure of dimension d.

The biggest difficulty of this first approach

In summary, this first approach combined with the quick fix as suggested in Segers and Uyttendaele (2013) lacks robustness with respect to the NAC assumption of the data. If the true copula of the data is not a NAC, we usually end up with a d-variate trivial structure.

There is hope however that a better fix than the one suggested in Segers and Uyttendaele (2013) could come and solve this issue.

First things first

Structure estimation: a first approach based on trivariate pieces

Structure estimation: a second approach based on binary structures

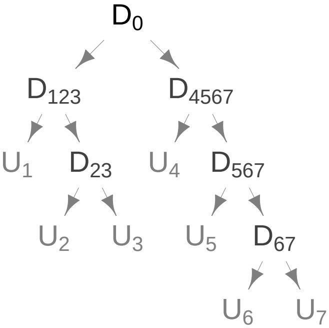
Structure estimation: a third approach inspired by phylogenetics

Is a NAC a good idea for my data?
(work in progress)

The future of research about NACs

Preliminary: definition of a binary tree

A binary tree is a tree such that each internal node of the tree has two and only two children.



A two-step procedure

The second approach to estimate a NAC structure is a two-step procedure and requires a weak assumption on the NAC for the first step.

First, estimate a binary tree on $U_1, \ldots, U_d$.

Second, check which parts of the binary tree can be collapsed.

First step: estimation of a binary structure

Based on an iid sample of size n from $(X_1, ..., X_d)$ such that the distribution of $(U_1, \dots, U_d)$ is a NAC, estimate a distance for every couple (X_i, X_j) with distinct $i, j \in \{1, \dots, d\}$.

Cluster the variables one at the time according to the estimated distances to get a binary tree on $(X_1, ..., X_d)$.

This approach makes sense only if the estimated distances are measures of dependence (a large dependence being translated by a small distance) and you assume that the NAC structure is such that the dependence increases as we go down the structure.

First step: estimation of a binary structure

Suggestion of measures of dependence between two variables we could use:

Kendall's τ,

A distance between the (theoretical) Kendall distribution of two independent variables and the empirical Kendall distribution of the two variables,

Hoeffding's D statistic.

Remark: these three distances are all such that $\widehat{\text{dist}}(X_i, X_j) = \widehat{\text{dist}}(U_i, U_j)$. So the binary tree on $(X_1, ..., X_d)$ is actually also the binary tree on $(U_1, \ldots, U_d)$.

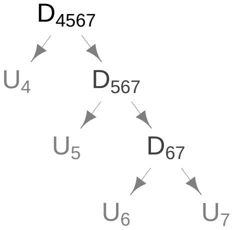
Step 2: collapsing of the binary tree

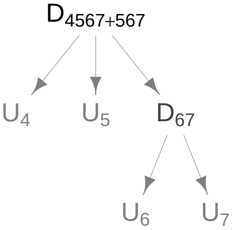
If for some reason you know that the NAC structure is a binary tree, then you should obviously skip this second step.

If you have no clue about what the true NAC structure is, inspect all internal edges of the estimated binary structure and collapse two linked nodes into one if a criterion is not met.

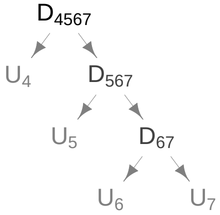
Step 2: collapsing of the binary tree

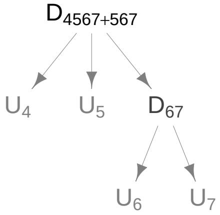
Suppose you end up with the left structure as binary structure. You check if the nodes D_{4567} and D_{567} can be collapsed into one, and if so, you end up with the structure on the right.





Step 2: a criterion based on Kendall's τ for collapsing

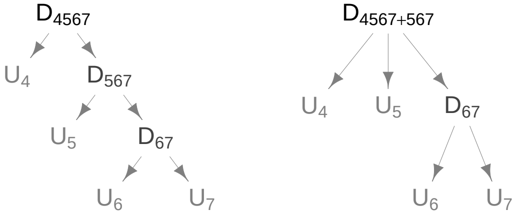




For each pair of variables such that the two variables of the pair are related through the *parent* node, estimate Kendall's τ between the two variables. Average all Kendall's τ you get this way.
For each pair of variables such that the two variables of the pair are related through the *child* node, estimate Kendall's τ between the two variables. Average all Kendall's τ you get this way.

If the absolute difference between the average Kendall's τ of the parent node and the average Kendall's τ of the child node is lower than a threshold τ_c, collapse the two nodes into one.

Step 2: a criterion based on the comparison of trivariate pieces for collapsing



If we break down the two structures into trivariate pieces, we see that the left structure is only made of trivariate pieces where we always have two variables that are more related than a third one.

In the structure on the right, the trivariate structures of (U_4, U_5, U_6) and of (U_4, U_5, U_7) are trivial trivariate structures.

Step 2: a criterion based on the comparison of trivariate pieces for collapsing

Test 1:
H_0: the structure of (U_4, U_5, U_6) is the trivial structure, denoted Λ_{456}.
H_1: H_0 is wrong.

Test 2:
H_0: the structure of (U_4, U_5, U_7) is the trivial structure, denoted Λ_{457}.
H_1: H_0 is wrong.

Use as test statistic for both tests either the bootstrap test statistic or the Friedman test statistic, both discussed in the first approach to estimate a NAC structure.

If the average p-value of the two tests is lower or equal to a threshold α, do not collapse the nodes D_{4567} and D_{567} into one.

Advantages and disadvantages of this two-step approach to estimate a NAC structure

The main advantage of this second, two-step approach to estimate a NAC structure over the first approach is that we do not lack robustness with respect to the assumption that the data have a NAC as true copula.

The main disadvantage is that this second approach assumes that the NAC of $(U_1, \ldots, U_d)$ is such that the dependence (measured in terms of one of the three distances defined earlier) between the random variables strictly increases as we go down the NAC structure. It remains however a weak assumption about the NAC.

First things first

Structure estimation: a first approach based on trivariate pieces

Structure estimation: a second approach based on binary structures

Structure estimation: a third approach inspired by phylogenetics

Is a NAC a good idea for my data?
(work in progress)

The future of research about NACs

Phylogenetics

Phylogeneticians encountered the following problem a long time ago: how to retrieve a target structure that will represent as well as possible an input set of trees, this set including trees of various sizes, conflicting trees and also missing trees (that is, some information to build the target structure is actually lacking).

Phylogenetics

Methods solving this problem are called *supertree methods* by phylogeneticians.

Some interesting references to get started are Bininda-Emonds (2004), Wilkinson et al. (2005) or Swenson et al. (2012).

Supertree methods in R

There are only two supertree methods in R, both in the phytools package.

As input, both methods only accept a set of binary trees.

The output tree is also a binary tree, and is unrooted.

Structure estimation: a third approach

Suggestion: for every set of distinct $i, j, k \in \{1, ..., d\}$, pick the best non-trivial trivariate structure for (U_i, U_j, U_k), that is, pick the best structure among λ_{ij}, λ_{ik} and λ_{jk}. We already know it is an easy problem.

Use then the resulting set of estimated trivariate structures (only binary structures) as input set for the two supertree methods available in the phytools package.

Get an unrooted binary tree as output, root it*, and try to collapse some of the internal nodes using one of the two criterions introduced in the second approach to estimate a NAC structure.

*See the blog of Liam Revell, author of the phytools package, for more details about the rooting.

Advantages of the third approach to estimate a NAC structure

This third approach to estimate a NAC structure has the advantages of the two previous approaches:

It is not required to assume anything about the NAC from which we want to estimate a structure (first approach),

AND the structure estimation is robust with respect to the NAC assumption (second approach).

First things first

Structure estimation: a first approach based on trivariate pieces

Structure estimation: a second approach based on binary structures

Structure estimation: a third approach inspired by phylogenetics

Is a NAC a good idea for my data?
(work in progress)

The future of research about NACs

Is a NAC a good idea for my data?

Proposition 1 from Okhrin et al. (2009):

"Let F be an arbitrary multivariate distribution function based on a NAC. Then F can be uniquely recovered from the marginal distribution functions and all bivariate copula functions."

Since all bivariate copula functions in a NAC are actually Archimedean distributions, Proposition 1 can be conveniently rewritten as:

"Let C be a NAC on $(U_1, \ldots, U_d)$. Then C can be uniquely recovered from all Kendall distributions, one for each pair (U_i, U_j) of random variables."

Is a NAC a good idea for my data?

By definition of a NAC, if two pairs of random variables are such that the two variables (U_i, U_j) of the first pair are related through a node A and that the two variables (U_k, U_l) of the second pair are also related through the same node A, then both pairs (U_i, U_j) and (U_k, U_l) have the same Kendall distributions.

Such pairs are said to belong to the same equivalence class (Segers and Uyttendaele, 2013).

Is a NAC a good idea for my data?

Thus to check if the data are compatible with a NAC, one could:

For each pair of variables, get the (unconstrained) empirical Kendall distribution.

Estimate a NAC structure on the data, $\hat{\lambda}$.

For each pair of variables, get the (constrained) empirical Kendall distribution under the assumption that the pair comes from a NAC with structure equal to $\hat{\lambda}$. *All pairs belonging to the same equivalence class are assigned the same empirical Kendall distribution: just average all (unconstrained) empirical Kendall distributions within the equivalence class. The equivalence classes are defined by* $\hat{\lambda}$*.*

Finally, check for discrepancies between the (unconstrained) empirical Kendall distributions and the (constrained) empirical Kendall distributions pair by pair of random variables.

Is a NAC a good idea for my data?

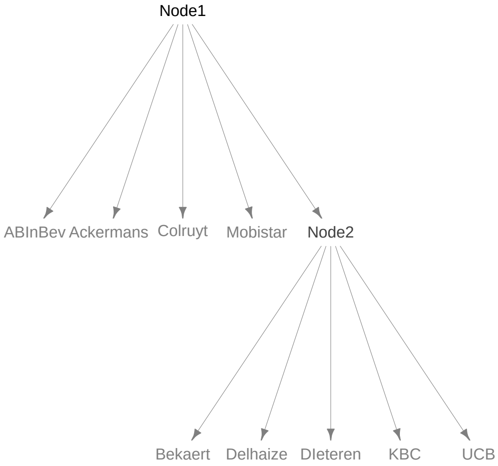
If there are too many large discrepancies, you can exclude a NAC as the true copula for your data.

However, if you fail to spot serious discrepancies, do NOT conclude a NAC will fit the data well! Failure to spot serious discrepancies is not enough to conclude a NAC will fit the data well.

The diagnostic tool developed here only allows to maybe exclude a NAC as true copula for your data, nothing more.

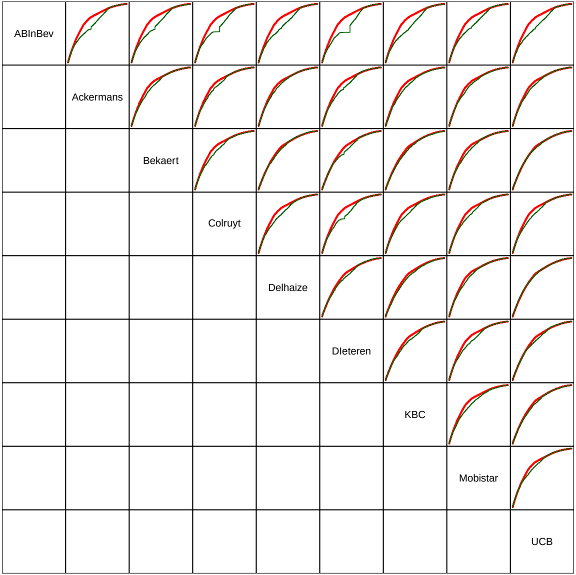
Is a NAC a good idea for my data? Examples

Dataset = daily logreturns from 2000 to 2006 of d=9 companies based in Belgium.



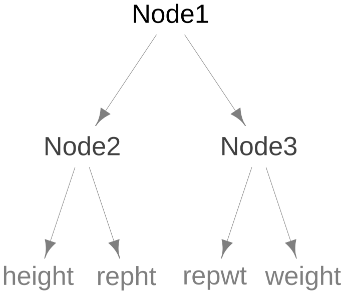
Is a NAC a good idea for my data? Examples

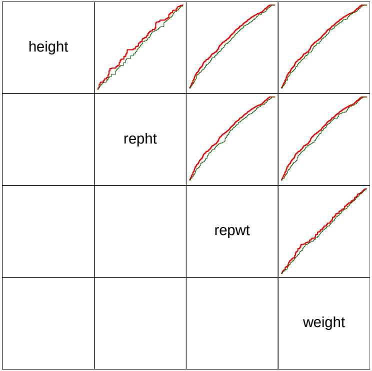
Dataset = daily logreturns from 2000 to 2006 of d=9 companies based in Belgium.



Is a NAC a good idea for my data? Examples

The Davis dataset (library “car”) has 200 rows and 4 columns. The subjects were men and women engaged in regular exercise.





Is a NAC a good idea for my data? Examples

Dataset “States”, 51 rows and 6 columns. The observations are the U.S. states and Washington, D.C.

pop - Population: in 1,000s.

SATV - Average score of graduating high-school students in the state on the verbal component of the Scholastic Aptitude Test (a standard university admission exam).

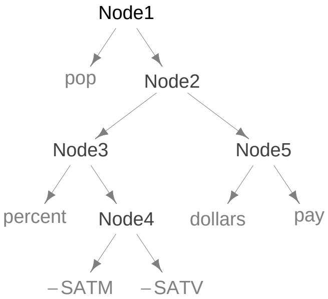
SATM - Average score of graduating high-school students in the state on the math component of the Scholastic Aptitude Test.

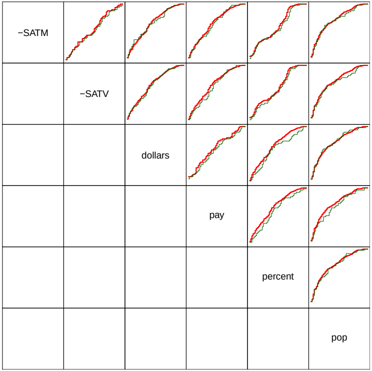
percent - Percentage of graduating high-school students in the state who took the SAT exam.

dollars - State spending on public education, in dollars per student.

pay - Average teacher’s salary in the state, in dollars.

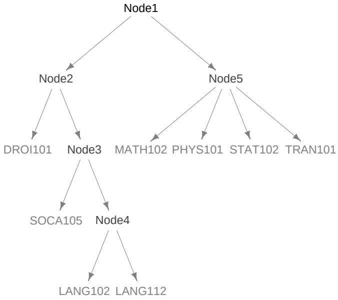
Is a NAC a good idea for my data? Examples

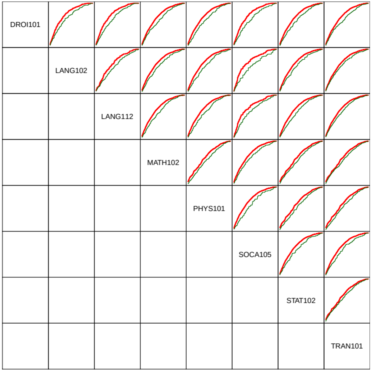




Is a NAC a good idea for my data? Examples

Examination results of more than 243 students at ULB (Université Libre de Bruxelles).





First things first

Structure estimation: a first approach based on trivariate pieces

Structure estimation: a second approach based on binary structures

Structure estimation: a third approach inspired by phylogenetics

Is a NAC a good idea for my data?
(work in progress)

The future of research about NACs

The future of research about NACs

Sufficient and necessary condition(s) on the generators to ensure we have a copula?

Estimation of the generators under these condition(s)?

Data generation from an estimated (structure + generators) NAC?

Better goodness-of-fit tests or diagnostic tools?

NACs are a class of copulas with the potential to become even more popular than the Archimedean class of copulas. However this will not happen untill the above issues are properly solved.

References I

Barbe, P., C. Genest, K. Ghoudi, and B. Rémillard (1996), "On Kendall's process." *Journal of Multivariate Analysis*, 58, 197–229.

Bininda-Emonds, Olaf RP (2004), "The evolution of supertrees." *Trends in Ecology & Evolution*, 19, 315–322.

Genest, C., J. Nešlehová, and J. Ziegel (2011), "Inference in multivariate Archimedean copula models." *Test*, 20, 223–256.

Genest, C. and L-P Rivest (1993), "Statistical inference procedures for bivariate Archimedean copulas." *Journal of the American Statistical Association*, 88, 1034–1043.

Genest, C. and L.P. Rivest (2001), "On the multivariate probability integral transformation." *Statistics & Probability Letters*, 53, 391–399.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

McNeil, A. J. and J. Nešlehová (2009), "Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions." *The Annals of Statistics*, 37, 3059–3097.

References II

Nelsen, R.B., J.J. Quesada-Molina, J.A. Rodríguez-Lallena, and M. Úbeda-Flores (2003), "Kendall distribution functions." *Statistics & Probability Letters*, 65, 263–268.

Okhrin, O., Y. Okhrin, and W. Schmid (2009), "Properties of hierarchical Archimedean copulas." SFB 649 discussion paper 2009,014, Berlin.

Okhrin, Ostap, Yarema Okhrin, and Wolfgang Schmid (2013), "On the structure and estimation of hierarchical Archimedean copulas." *Journal of Econometrics*, 173, 189–204.

Segers, J. and N. Uyttendaele (2013), "Nonparametric estimation of the tree structure of a nested Archimedean copula." *arXiv:1304.1384, to appear in CSDA*.

Swenson, M Shel, Rahul Suri, C Randal Linder, and Tandy Warnow (2012), "Superfine: fast and accurate supertree estimation." *Systematic biology*, 61, 214–227.

Wilkinson, Mark, James A Cotton, Chris Creevey, Oliver Eulenstein, Simon R Harris, Francois-Joseph Lapointe, Claudine Levasseur, James O Mcinerney, Davide Pisani, and Joseph L Thorley (2005), "The shape of supertrees to come: tree shape related properties of fourteen supertree methods." *Systematic biology*, 54, 419–431.