Journal de la Société Française de Statistique
Vol. 154 No. 1 (2013)

Selection strategies for regular vine copulae

Titre: Stratégies de sélection pour les grappes régulières de copules

Claudia Czado[1], Stephan Jeske[2] and Mathias Hofmann[2]

Abstract: Regular vine (R-vine) copulae are a very flexible class of multivariate copulae, which have received increasing interest in finance and insurance. We will introduce these copulae, discuss their scope and parameter estimation. Since the class of R-vines is huge, model class selection is vital. Recently a top down and a bottom up approach for model selection have been developed. We will discuss these approaches and introduce some useful extensions based on using p-values of goodness-of-fit tests as selection weights. The use of R-vine copulae will be illustrated for a data set involving log concentrations of chemicals in water samples. The performance of these selection procedures are investigated through simulation.

Résumé : Les grappes régulières de copules ("R-vines" en anglais) forment une famille flexible de copules de plus en plus fréquemment utilisée dans le domaine de la finance et de l'assurance. Dans un premier temps, nous présentons ces copules et nous discutons leur estimation. La classe des grappes régulières de copules étant très riche, il est crucial de disposer d'outils de sélection de modèles. Dans un deuxième temps, nous nous intéressons ainsi à un algorithme descendant de sélection de modèles récemment suggéré et nous en proposons une extension fondée sur des tests d'adéquation. L'utilisation de grappes régulières de copules et des algorithmes de sélection étudiés est enfin illustrée sur des données de concentrations chimiques.

Keywords: copulae, regular vines, model selection
Mots-clés : grappes régulières de copules, regular vines, sélection de modèles
AMS 2000 subject classifications: 62F07, 62H20, 62G32

1. Introduction

For modeling high dimensional data which exhibit non Gaussian dependency pattern often a copula approach (e.g. [Joe, 1997, Nelson, 2006]) is taken. This allows to model the marginal distributions and the dependence structure separately. The dependence structure in d dimensions is described by the copula, which is a distribution function on the d-dimensional hypercube with uniform margins. Common choices for the copula are elliptical (e.g. [Frahm et al., 2003]), and Archemedean copulae. This however assumes a similar dependence pattern among all pairs of variables. For many data applications this restriction might not be satisfied, therefore it was first suggested in [Joe, 1996] to build multivariate copulae using only bivariate copula building blocks. These building blocks are called pair copulae. Later a set of linked trees was used in [Bedford and Cooke, 2001, Bedford and Cooke, 2002], which was called a regular (R) vine to identify the pair copulae. The term *vine* was used since the induced dependence structure can be visualized resembling a grape vine. In particular each edge of a tree correspond to a pair copula. The pair

[1] Zentrum Mathematik, Technische Universität München.
E-mail: cczado@ma.tum.de
[2] Zentrum Mathematik, Technische Universität München.
E-mail: jeske@ma.tum.de and E-mail: hofmann@ma.tum.de

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

copulae in the first tree characterize pairwise unconditional dependencies, while the pair copulae in higher trees model the conditional dependency between two variables given a set of variables. In addition the number of conditioning variables grows with the tree number. This process was called a pair copula construction (PCC) in [Aas et al., 2009].

When all trees have a path like structure, the subclass of D-vines results, while the case of star like structures is called a C-vine. In general parametric bivariate copula families are chosen as building blocks and they can be chosen arbitrarily to yield a valid multivariate copula. Therefore a vine copula has three components: first the tree structure, then the copula family for each edge in the tree structure and finally the corresponding dependence parameter(s) for each pair copula.

First estimation procedures assuming the tree structure and the pair copula families to be known were developed in [Kurowicka and Cooke, 2008], while standard maximum likelihood (ML) inference for D-vines was studied in [Aas et al., 2009]. Under regularity conditions the standard asymptotic theory remains valid. The Hessian matrices based on analytic derivatives are derived and calculated numerically to provide asymptotic standard error estimates for the ML estimates in [Stöber et al., 2012]. To facilitate ML estimation a sequential estimation procedure was developed in [Aas et al., 2009], which requires only separate parameter optimization for each pair copula. These sequential estimates can be used as starting values for the ML optimization. The asymptotic theory for these sequential estimates was derived in [Hobak Haff, 2012].

Since the number of regular vines is very large ([Morales-Nápoles et al., 2010]) it is vital to develop efficient selection strategies for all components of a vine specification. Early approaches restricted themselves to C- and D-vines. Since higher trees capture conditional dependencies, the order of the nodes were chosen in such a way that a large number of strongest pairwise dependencies are captured in the first tree. In applications one often sees a decrease in the magnitude of the pairwise conditional dependencies as the number of conditioning variables increased. This observation and the fact that sequential estimates are becoming less precise as the number of conditioning variables increases led to the development of two different sequential model selection algorithms for R-vine copulae ([Dißmann et al., 2013] and [Kurowicka, 2011]) .

In particular the selection procedure developed in [Dißmann et al., 2013] is a top down strategy. For the first tree all pairwise Kendall's τ estimates are calculated and their absolute value is used as edge weight to find a tree, which maximizes the sum of edge weights among all possible trees. This can be accomplished by using a maximal spanning tree (MST) algorithm. In [Dißmann et al., 2013] the Algorithm of Prim (see for example [Cormen et al., 2009, Section 23.2]) was utilized. In the next step the copula families and their parameters for all pair copulae in the top tree are selected using the smallest AIC as suggested by [Brechmann, 2010, Brechmann et al., 2012]. These choices are then used to estimate all pairwise Kendall's τ values for edges which maintain the proximity condition necessary for the R-vine tree structure, i.e. they are eligible to be an edge in Tree 2. This allows then to apply again the MST algorithm to select Tree 2. The corresponding copula families and their parameters are chosen again by AIC. This way of proceeding results in estimating the complete R-vine specification by selecting the strongest pairwise conditional dependencies first.

In contrast the selection procedure in [Kurowicka, 2011] is a bottom up strategy by selecting the weakest conditional dependencies for the highest trees. Here partial correlations as pairwise dependency measure are used. Since partial correlations and conditional correlations agree for elliptical distributions, this is a valid procedure in the case of elliptical copulae. We will later see

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

that the vine distribution is defined by an appropriate set of bivariate conditional distributions.

In this paper we compare and illustrate these two selection procedures. In addition we extend the selection procedure of [Dißmann et al., 2013] to allow for different weight measures. The choice of the absolute value of Kendall's τ as weight ignores the fit of the data to the chosen copula family. Therefore we consider weights based on the goodness-of-fit (GOF) of the chosen copula family such as given by the corresponding p-value of a GOF test.

2. Regular vine copulae

The class of regular vine copulae is an extremely flexible class of multivariate copulae, which has been first introduced in [Bedford and Cooke, 2001, Bedford and Cooke, 2002]. Likelihood based inference was considered in [Dißmann et al., 2013] including an efficient storage of the R-vine specification and its joint density function. For the convenience of the reader we recall the definition of the R-vine.

Definition 2.1 (R-vine)**.**

1. **R-vine tree:** *$\mathcal{V} = (T_1, \ldots, T_{d-1})$ is an R-vine tree on d elements if*
   (a) *T_1 is a tree with nodes $N_1 = \{1, \ldots, d\}$ and a set of edges denoted E_1.*
   (b) *For $i = 2, \ldots, d-1$, T_i is a tree with nodes $N_i = E_{i-1}$ and edge set E_i.*
   (c) *For $i = 2, \ldots, d-1$ and $\{a, b\} \in E_i$ with $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$ it must hold that $\#(a \cap b) = 1$, where $\#$ denotes the cardinality of a set.* (proximity condition).
2. **R-vine copula specification:** *$(\mathbf{F}, \mathcal{V}, \mathcal{C})$ is an R-vine copula specification if $\mathbf{F} = (F_1, \ldots, F_d)$ is a vector of continuous invertible univariate distribution functions, $\mathcal{V}$ is an d-dimensional R-vine tree structure and $\mathcal{C} = \{C_e | e \in E_i, i = 1, \ldots, d-1\}$ is a set of copulae with C_e being a bivariate copula, a so-called pair-copula.*

Figure 1 gives an example of a vine tree structure on 7 nodes. The special case where each node in tree T_1 has degree 2 is called a D-vine, while a C-vine is characterized by the property that each tree T_i has a unique node of degree $i-1$. This node is called the root node for the corresponding tree.

The proximity condition ensures that there can only be an edge between the nodes e_1 and e_2 in T_i if the edges e_1 and e_2 in T_{i-1} share a node in T_{i-1}. This allows to identify each edge e as $e = (a_e, b_e | D_e)$, where $\{a_e, b_e\}$ are the conditioned set and D_e the conditioning set of e. Here a_e and b_e are single elements of $\{1, \ldots, d\}$ and D_e is a subset of $\{1, \ldots, d\}$. See [Bedford and Cooke, 2002, Part 4] and [Kurowicka and Cooke, 2008, Chapter 4.4] for details. For example in Figure 1 the nodes $1,7|5$ and $1,4|5$ in T_3 share as edges in T_2 the node $1,5$, therefore the edge between the nodes $1,7|5$ and $1,4|5$ in T_3 is denoted by $4,7|1,5$ and can be found as node in T_4.

In the following we assume that all distributions are absolutely continuous. We denote the corresponding density of the marginal distribution function F_i by f_i for $i = 1, \ldots, d$ and the density of the bivariate copula C_e for edge e by c_e, respectively. Given an R-vine copula specification the following theorem was shown in [Bedford and Cooke, 2001, Bedford and Cooke, 2002].

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

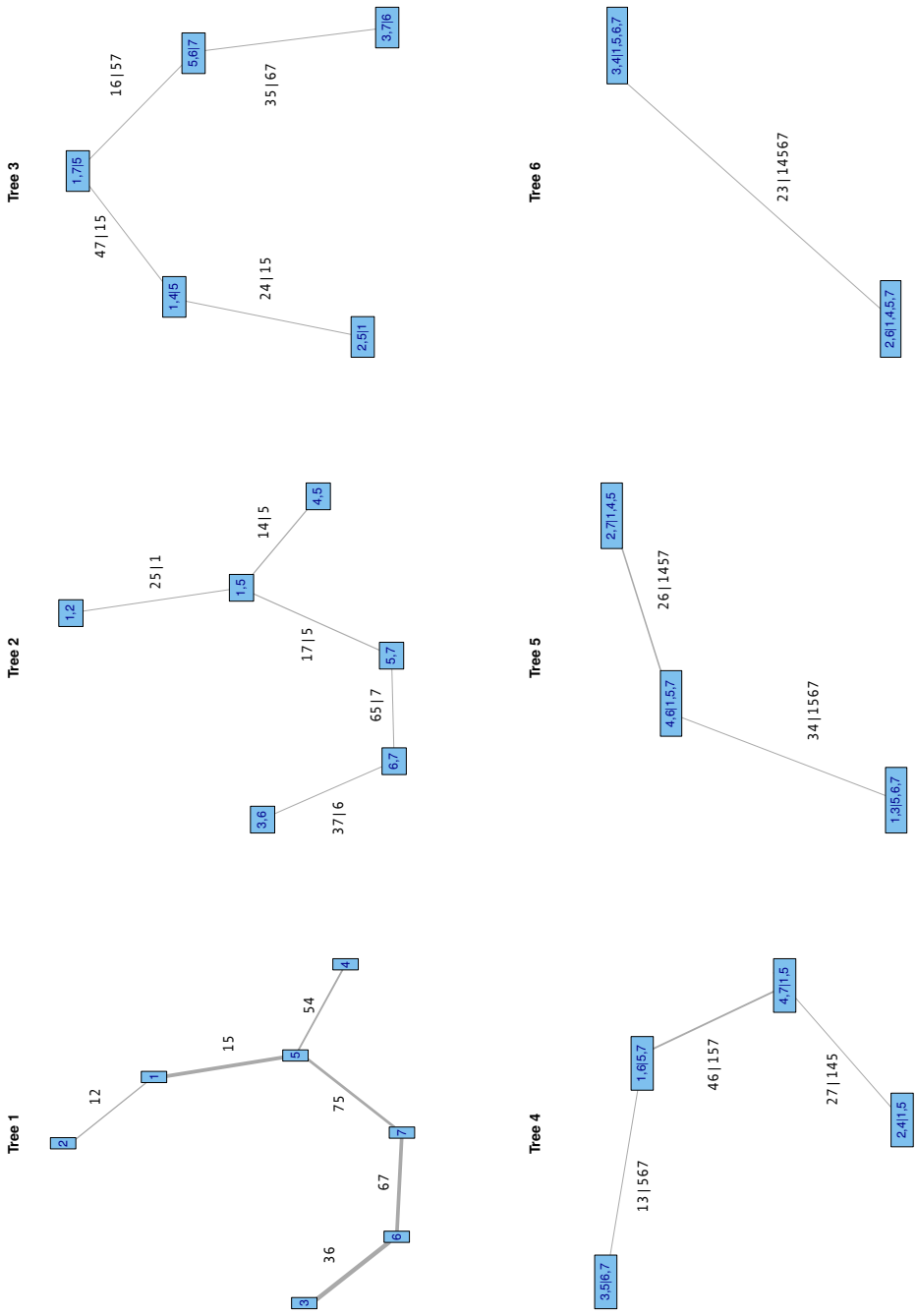

FIGURE 1. *Tree structure with edge and node labels for a seven dimensional R-vine*

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

Theorem 2.2. *Let* $(\mathbf{F}, \mathcal{V}, B)$ *be an R-vine copula specification on d elements. There is a unique d-dimensional distribution F that realizes this R-vine copula specification with density*

$$f_{1\ldots d}(\mathbf{x}) = \prod_{k=1}^{d} f_k(x_k) \prod_{i=1}^{d-1} \prod_{e \in E_i} c_{a_e,b_e|D_e}\left(F_{a_e|D_e}(x_{a_e}|\mathbf{x}_{D_e}), F_{b_e|D_e}(x_{b_e}|\mathbf{x}_{D_e})\right), \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_d)'$, $e = (a_e, b_e|D_e)$ *and* $\mathbf{x}_{D_e}$ *contains the variables in* D_e, *i.e.,* $\mathbf{x}_{D_e} = \{x_i | i \in D_e\}$. *Here* $F_{a_e|D_e}$ *denotes the conditional distribution function of* X_{a_e} *given* $\mathbf{X}_{D_e} = \mathbf{x}_{D_e}$ *when* $\mathbf{X} = (X_1, \ldots, X_d)'$ *has distribution F.*

Since in (1) the pair copula $c_{a_e,b_e|D_e}(\cdot,\cdot)$ does not depend on the conditioning value $\mathbf{x}_{D_e}$, not all multivariate distributions can be represented as a vine distribution. This assumption is called the simplifying assumption and will be assumed throughout the paper.

To evaluate the joint density in (1) conditional distribution functions which occur as arguments of the pair copulae have to be evaluated. It was shown in [Joe, 1996] that these can be found by partial differentiation of a pair copula. In particular if we need to calculate $F_{a|D}(x_a|\mathbf{x}_D)$, then

$$F_{a|D}(x_a|\mathbf{x}_D) = \frac{\partial C_{a,v|D_{-v}}(F_{a|D_{-v}}(x_a|\mathbf{x}_{D_{-v}}), F_{v|D_{-v}}(x_v|\mathbf{x}_{D_{-v}}))}{\partial F_{v|D_{-v}}(x_v|\mathbf{x}_{D_{-v}})} \tag{2}$$

holds. Here $v \in D$ and D_{-v} is D with v removed. In [Dißmann et al., 2013] it was shown that (2) can be used to compute all required conditional distribution required in (1) recursively. This fully specifies the vine density (1). There are $\frac{d(d-1)}{2}$ pair copula terms in (2). For each term a parametric pair copula family together with their parameter value(s) have to be selected. In our analysis we allow as bivariate copula families the independence, Gaussian (N), Student t , Gumbel (G) and their rotated versions by 90, 180 and 270 degree. The rotation by 180 degree corresponds to the survival Gumbel (SG). This gives a nice range of dependence structures, such as no, symmetric and asymetric tail dependence. However other bivariate copula families can be added easily.

We would like to mention that the multivariate Gaussian copula and the Student t copula with a common degree of freedom ν can be represented as a vine distribution under the simplifying assumption. Here all pair copulae are bivariate Gaussian or Student t copulae, respectively. The parameters of the pair copulae correspond to conditional correlations, which are also partial correlations for elliptical copulae. For the Student t copula with degree of freedom ν the pair copulae in the first tree have as degree of freedom ν, while in Tree $i > 1$ the degree of freedom is $\nu + i - 1$. The only Archemedean copula which can be represented as a vine with density (1) is the Clayton copula. See [Stöber et al., 2012] for further classification results with regard to vine distributions. The results of [Hobak Haff et al., 2010] give examples where vines provide a good approximation to multivariate distributions, while in [Acar et al., 2012] a three dimensional example is provided, where this is not the case. Tail dependence properties of vine distributions are studied in [Nikoloulopoulos et al., 2012].

3. Parameter estimation for R-vine distribution with specified R-vine tree structure and bivariate copula families for pair copula terms

To estimate the parameters of an R-vine distribution for a given R-vine tree structure and bivariate copula families for each pair copula term, we require an i.i.d. sample from the R-vine distribution.

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

Given this multivariate sample we need to estimate marginal and copula parameters. Since joint parameter estimation requires high dimensional optimization often a two step approach is followed, where first the marginal parameters are either estimated parametrically or non-parametrically separately for each component. The first approach is called the inference functions for margin (IFM) approach discussed in [Joe and Xu, 1996], while the second approach is called maximum pseudo likelihood (MPL) introduced in [Genest et al., 1995]. After this is accomplished the appropriate (either parametric or non parametric) probability integral transform is formed to obtain data in the copula domain $[0,1]^d$. This so called copula data are then utilized to estimate the copula parameters. There has been considerable effort to adjust for the estimation error made by possibly mis-specifying the marginal models (see [Chen and Fan, 2006] and references therein) however it is our experience that parameter estimates are only affected if the margins are severely misspecified (see [Kim et al., 2007] where exponential margins were fitted using normal margins). This can be avoided by a careful statistical analysis of the marginal fit. Therefore we prefer to use the IFM approach after the marginal models are chosen in such a way that there is no statistical evidence against the marginal model choice. This means we use the derived copula data to fit the copula parameters. For R-vine copulae with known tree structure there are several copula parameter estimation methods available. The first one is the sequential estimation method first suggested in [Aas et al., 2009] and later studied in detail by [Hobak Haff, 2012], maximum likelihood also discussed in [Aas et al., 2009] and Bayesian estimation for D-vine copulae in [Min and Czado, 2010]. For general R-vine copulae [Gruber and Czado, 2013] give a Bayesian estimation approach which estimates the tree structure, pair copulae and their parameters starting with Tree 1 to Tree $d-1$.

The sequential estimation method provides fast estimates of the copula parameters for a given the R-vine tree structure since each pair copula term is optimized separately. To illustrate sequential estimation consider an R-vine copula in three dimensions with density

$$c(u_1,u_2,u_3) = c_{12}(u_1,u_2;\theta_{12})c_{23}(u_2,u_3;\theta_{23})c_{13|2}(F_{1|2}(u_1|u_2;\theta_{12}),F_{3|2}(u_3|u_2;\theta_{23});\theta_{13|2}).$$

Here the copula parameters θ_{12} and θ_{23} are estimated using the copula sample $\{(u_{i1},u_{i2}),i=1,\ldots,n\}$ and $\{(u_{i2},u_{i3}),i=1,\ldots,n\}$, respectively. This gives estimates $\hat{\theta}_{12}$ and $\hat{\theta}_{23}$ for θ_{12} and θ_{23}, respectively. We form now pseudo conditional copula observations $u_{i1|2} := F_{1|2}(u_{i1}|u_{i2};\hat{\theta}_{12})$ and $u_{i3|2} := F_{3|2}(u_{i3}|u_{i2};\hat{\theta}_{32})$ for $i=1,\ldots,n$ to be used to estimate $\theta_{13|2}$. This approach can be generalized to arbitrary R-vine copulae and precise recursions are given for example for C-vines in [Czado et al., 2012]. These estimates are also often used as starting values for the optimization required for joint maximum likelihood estimation.

4. Top down strategies using maximal spanning tree algorithms with weights

4.1. Basic algorithm

We describe now the basic algorithm underlying the top down selection strategies for R-vines. We assume that we have copula data $\{(u_{\ell 1},...,u_{\ell d}),\ell=1,...N\}$ available. For this we either know the marginal distributions or have them estimated parametrically or non-parametrically using ranks. The backbone is the availability of appropriate weights, which summarize some chosen characteristics of the bivariate conditional distribution of (U_j,U_k) given a set of variables $\mathbf{U}_D$.

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

Algorithm 4.1 Sequential method to select an R-vine model based on weights

Input: Data $(u_{\ell 1},\dots u_{\ell d})$, $\ell = 1,...,N$ **(realizations of i.i.d. random vectors on** $[0,1]^d$**).**
Output: R-vine copula specification, i.e., $\mathscr{V}$, $\mathscr{C}$.

1: Calculate the weight $w_{j,k}$ for all possible variable pairs $\{j,k\}, 1 \le j < k \le d$.
2: Select the spanning tree that maximizes the sum of weights, i.e.,

$$\max \sum_{e=\{j,k\} \text{ in spanning tree}} w_{j,k}.$$

3: For each edge $\{j,k\}$ in the selected spanning tree, select a bivariate copula family and estimate the corresponding parameter(s).
4: **for** $i = 2,\dots,d-1$ **do** {Iteration over the trees}
5: Based on the selected copula families and their parameters in Tree $i-1$ calculate the weights $w_{j,k|D}$ for all conditional variable pairs $\{j,k|D\}$ that can be part of Tree T_i, i.e., all edges fulfilling the proximity condition (see Definition 2.1).
6: Among these edges, select the spanning tree that maximizes the sum of weights, i.e.,

$$\max \sum_{e=\{j,k|D\} \text{ in spanning tree}} w_{j,k|D}.$$

7: For each edge $\{j,k|D\}$ in the selected spanning tree, select a bivariate copula family and estimate the corresponding parameter(s).
8: **end for**

Here D does not contain the indices j and k. We will restrict to positive weights which measure the strength of dependence between two variables. It is assumed that a higher weight induces a better fit to the chosen characteristic. Using the sequential estimation approach we have pseudo observations for these conditional distributions available, which allow us to estimate weights sequentially. More precisely we assume that we can calculate a positive weight for each edge $e = (j,k|D)$ allowed by the proximity condition of Definition 2.1 in Tree i denoted by $w_{jk|D}$. For Tree 1 the weights will depend on the copula data only, while the weights for Tree $i > 1$ will depend on the pseudo conditional copula observations using the tree structure of Trees $k \le i-1$, its copula families and corresponding parameters. In Tree 1 all edges connecting pairs of nodes are allowed by the proximity condition and so we calculate weights $w_{j,k}$ for all pairs (j,k). Given weights we can apply for example the Algorithm of Prim ([Cormen et al., 2009, Section 23.2]) to select the tree structure that maximizes the sum of weights as Tree 1. The pair copula families are chosen by the lowest AIC attained by the corresponding bivariate pseudo conditional observations. Here their corresponding parameters are estimated in the same way as in the sequential estimation procedure. For Trees $i > 1$ all edges are considered which are allowed by the proximity condition applied to Tree $i-1$ and their weights determined. Among these edges again the Algorithm of Prim is used to select the tree structure. Copula families are selected and their parameters are estimated as discussed above. This basic top down selection procedure is summarized in Algorithm 4.1. Since a selected edge for a tree cannot be dropped once it is selected in this algorithm, we do not expect to identify the best or the true R-vine tree structure in general, but rather we hope for reasonable candidate models.

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

4.2. *Choice of weights*

To apply the Algorithm 4.1 we need to choose weights. We discuss now four different choices representing different characteristics of the bivariate conditional distributions which build up the regular vine distribution. The first choice discussed will be a measure of dependence, while the next two choices choose edges where the pseudo data are fitted well by the class of pair copula families considered. The final choice is a combination between dependence and goodness of fit of the pseudo data.

4.2.1. *Absolute Kendall's τ (TAU)*

The data collector is often interested in capturing the strongest pairwise dependencies in the data. This corresponds in a regular vine model to the task of selecting the $d-1$ strongest pairwise dependencies for Tree 1. The most common dependency measure is Kendall's τ, which captures non linear dependencies and is invariant under monotone transformations of the margins. Since we require that these variables form a tree we choose those $d-1$ pairs of variables which maximize the sum of the absolute value of Kendall's τ among all pairs which form a tree. Since the true Kendall's τ values are unknown, we use empirical estimates. For edges in Trees $i > 1$ we determine the appropriate pseudo observations to estimate the corresponding pairwise Kendall's τ.

4.2.2. *Akaike information criterion (AIC)*

Since we are interested in finding a good fitting R-vine to the data, we are interested in choosing the pair copula families in such a way that they fit the corresponding (pseudo) observations well. The most prominent goodness-of-fit measure is the Akaike ([Akaike, 1973]) information measure, however this does not allow to assign statistical significance such as the p-value corresponding to a statistical goodness-of-fit test. We select the pair copulae from the set of bivariate copula families considered and their parameters to each pair of variables separately and calculate the corresponding AIC based on bivariate conditional pseudo copula data. In the next step we choose the copula family with the lowest AIC and assign this value as edge weight for each pair. Using the sequential R-vine selection of Algorithm 4.1 this gives us Tree 1. For Trees $i > 1$ we utilize the pseudo observations to fit all pair copula terms which are allowed by the proximity condition and determine the lowest AIC among the copula families and their parameters for each pair copula term. This determines the R-vine in a sequential way. Note that for Tree 1 this is the smallest AIC among all regular vines allowed by the copula families where all conditional copulae are set to the independence copulae. For Tree 2 this corresponds to the smallest AIC among all regular vines with a fixed Tree 1, arbitrary Tree 2 and independence copulae for conditional copulae with more than one conditioning variable. A similar interpretation can be made for Trees $i > 2$.

4.2.3. *Copula goodness-of-fit p-value (GOF)*

As already mentioned the drawback of AIC based weights are that they do not allow a quantitative assessment of the goodness-of-fit. Therefore we consider the following approach. The performance of the sequential estimation procedure relies on the selection of a pair copula term for the

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

corresponding pair of pseudo data values. Thus it is natural to consider a copula goodness-of-fit measure for this selection. There is a large literature on copula goodness-of-fit tests including large scale simulations (see for example [Genest et al., 2009] and [Berg, 2009]). Here one is interested in testing whether the unknown pair copula C belongs to a chosen parametric copula family $\mathbf{C} = \{C_\theta, \theta \in \Theta\}$ or not, i.e.

$$H_0 : C \in \mathbf{C} \text{ versus } H_1 : C \notin \mathbf{C}.$$

Since we are interested in applying the GOF test to the bivariate (pseudo conditional) copula data we restrict to this case. Based on a data sample $\{(u_{1,i}, u_{2,i}), i = 1, \ldots, n\}$ a difference measure between the empirical bivariate copula C_n and the fitted copula $C_{\hat{\theta}_n}$, where $\hat{\theta}_n$ is an estimate of θ, is considered. In [Genest et al., 2009] and [Berg, 2009] it was shown that a very powerful difference measure among several alternatives is the Cramér von Mises statistic defined as

$$S_n = n \int_{-\infty}^{\infty} \left[C_n(u_1, u_2) - C_{\hat{\theta}_n}(u_1, u_2) \right]^2 dC_n(u_1, u_2) = \sum_{i=1}^{n} \left[C_n(u_{1,i}, u_{2,i}) - C_{\hat{\theta}_n}(u_{1,i}, u_{2,i}) \right]^2.$$

While [Genest and Remillard, 2008] provide a computationally expensive parametric bootstrap procedure for the calculation of the corresponding p-value, [Kojadinovic and Yan, 2011] and [Kojadinovic et al., 2011] developed a multiplier approach to obtain approximate p-values much faster. The corresponding implementation in the statistical software R is described in [Kojadinovic and Yan, 2010]. Since we have to select among several bivariate copula families for each pair copula term in the joint density we proceed with a pre-selection using the AIC criterion to select the parametric bivariate copula family with the smallest AIC. For this choice for $\mathbf{C}$ we apply the above copula goodness-of-fit test using the corresponding pseudo observations to calculate the approximate p-value using the fast multiplier approach of [Kojadinovic and Yan, 2011]. A small p-value indicates that the bivariate copula family does not fit the (pseudo) observations adequately. Finally these p-values are chosen as edge weights for the sequential R-vine selection procedure. It should be emphasized that since the pairs of pseudo observations considered are not necessarily disjoint, the p-values only give a rough guide of the goodness-of-fit of the complete data. We like to add that this way of proceeding ignores the uncertainty in the parameter estimation in the previous trees when pseudo observations are used and thus can only be regarded as approximately theoretically justified.

4.2.4. Copula goodness of fit p-value times absolute Kendall's τ value (GOF-TAU)

As a final weight we study a combination of dependency strength and goodness-of-fit measure as weight. This allows us to mitigate the effect of parameter estimation errors when forming the pseudo observations, while allowing for copula families which fit the data best among the families considered. Therefore we use the product between the absolute Kendall's τ value and the copula goodness-of-fit p-value as weight.

5. Bottom up strategy using partial correlations (PARTIAL)

This selection strategy was developed in [Kurowicka, 2011]. It is a bottom up strategy based on selecting Tree $d-1$ first and then selecting Trees $d-i$ for $i \geq 2$ sequentially. It is also motivated

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

by choosing an R-vine with decreasing dependency strength as Tree i increases. For a bottom up strategy this means that one is interested in finding Tree $d-1$ with the lowest possible dependence strength as measured by the lowest absolute partial correlation. In [Kurowicka, 2011] an algorithm is developed which allows to build up a valid R-vine tree structure from Tree $d-1$ until Tree 1 by giving appropriate conditions for the selection of edges in Tree i given that Tree $i+1$ is already specified. In addition if there are several choices for an edge in Tree i, the one with the lowest absolute partial correlation is chosen. For further details consult [Kurowicka, 2011]. Since partial correlations are equal to conditional correlations for elliptical distributions (see [Baba et al., 2004]) we expect this strategy to perform well, when all pair copula families are elliptical. This strategy chooses only the tree structure of the vine distribution. The pair copula families and their parameters are then chosen by AIC using the pseudo data for Trees $i > 1$.

6. Selection strategies for C- and D-vines

In some data applications there might be a natural order such as a time order in the variables. Such a case can be handled by using a D-vine copula. In other situations an order of the importance of the variables is known and this can be handled by a C-vine copula model. Therefore it is also interesting to restrict the model class to C- and D-vines respectively. For these restricted vine copula classes we can adapt the discussed selection strategies for R-vines.

In particular a D-vine tree structure is completely determined by the order of the components. Therefore it is enough to select the first tree. For this compute for each pair as above an appropriate weight based on the sample. The problem of finding the permutation of the nodes that maximizes the sum of weights is a Hamiltonian path problem, for which there are many approximate algorithms of this NP hard problem.

For C-vines we need to determine an order of the root nodes for each tree. We follow the approach taken in [Czado et al., 2012]. For the first tree we again compute weights for each pair of variables. For each variable $i = 1, \ldots, d$ we then determine $s_i := \sum_{j \neq i}^{d} w_{ij}$ and select i_1 which maximizes s_i. This gives the first root node. For the next root node we compute $s_{i|i_1} := \sum_{j \neq i, i_1}^{d} w_{ij|i_1}$ for $i \neq i_1$. The node i_2 which maximizes $s_{i|i_1}$ is then selected as the second root node for the C-vine. We proceed in a similar fashion to determine the complete order of the root nodes.

7. Illustration and simulation

As an illustration we consider the classical hydro-geochemical stream and sediment reconnaissance data discussed in [Cook and Johnson, 1981], where the log concentrations of seven chemicals in 655 water samples were recorded. This data are contained in the R package `copula` as data set `uranium`. The seven chemicals are uranium (U), lithium (L), colbalt (Co), potassium (K), cesium (Cs), scandum (Sc) and titanium (Ti). The triplet of (Co, Ti, Sc) was analyzed in [Ben Ghorbal et al., 2009] and [Cook and Johnson, 1981], who argue that neither an elliptical nor an extreme-value copula can fit this triplet. In Figure 2 pairwise scatter plots of the rank transformed copula data are given in the upper triangular part of the figure. The rank transformations are applied to each margin separately. Empirical contour plots of normalized copula data can be used to identify reasonable pairwise dependency models. More precisely for the normalized copula data we use the inverse cdf of the standard normal distribution for each component to transform to

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

TABLE 1. *Estimated log likelihood, number of parameters, AIC and BIC values corresponding to the chosen copula models*

Class	Strategy	Log Likelihood	# Parameters	AIC	BIC
Gauss	MLE	751.31	21	-1460.6	-1366.4
t-copula	MLE	824.22	22	-1604.5	-1505.8
C-vine	TAU	847.08	31	-1632.2	-1493.1
D-vine	TAU	854.16	31	-1646.3	-1507.3
R-vine	TAU	856.99	32	-1650.0	-1506.5
	GOF	862.24	28	-1668.5	-1542.9
	PARTIAL	864.00	29	-1670.0	-1539.9
	AIC	858.44	31	-1654.9	-1515.8
	GOF-TAU	859.00	30	-1658.0	-1523.5

dependent data with standard normal margins. For example a Gaussian copula gives elliptical contours, while a Student t copula with low degree of freedom produces a diamond shape and a Gumbel/Clayton copula produce pear like shapes. Figure 2 shows that there are non symmetric pairwise dependencies present in the data, which will be confirmed by our analysis.

R-vine copulae using the five selection strategies were now fitted to the copula data using the sequential selection and estimation approach discussed above. For the strategy TAU the resulting trees and their pair copula families together with derived Kendall's τ values are given in Figure 3, which has the same tree structure as the R-vine used for illustration in Section 2. For the other strategies the resulting first trees are given in Figure 4. This shows that the selection procedures discussed choose different models. However there are some common elements. For example the selection strategy TAU and AIC differ only by one edge. The triplets (Ti, Sc, Co) and (K, Cs, U) are common to selection strategies TAU and PARTIAL. The GOF strategy has only two pairs (Sc, Co) and (Ti, Cs) in common with the strategy TAU, while the structure for (Co, Sc, Ti, Cs, U) from the GOF-TAU strategy can be found also in the first tree of the TAU strategy.

We also fit a multivariate Gaussian copula and a multivariate t copula with common degree of freedom using maximum likelihood as baseline models. To assess which of the five chosen R-vine copulae fits the data best, we computed the corresponding log likelihood, the number of parameters, the overall AIC and BIC value as given in Table 1, showing that the GOF and the PARTIAL selection strategy give the best fit according to BIC, which prevents overfitting in large data sets. These two strategies also perform best with regard to the AIC criterion. For comparison we also provide the results for C- and D-vine models using the adapted TAU strategy for these model classes as discussed in Section 6. This shows that the fits of these subclasses of regular vine models can be improved using the full class of regular vine copulae.

As noted by a referee the `uranium` data set contains a large number of ties in the margins. Therefore the assumption of absolutely continuously distributed margins is questionable. To see if the ties are affecting the performance of the model selection the referee suggested to proceed as for the `loss-alae` data analysis of [Kojadinovic and Yan, 2010]. For brevity and since we are using this data only as an illustration of the model selection procedures we do not include such an extended analysis.

Additionally it is our experience that it depends on the true R-vine copula model which selection strategy performs the best. To illustrate this, we used the five selected R-vine copula models for the `uranium` data set as possible regular vine senarios to be investigated. We simulated from

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

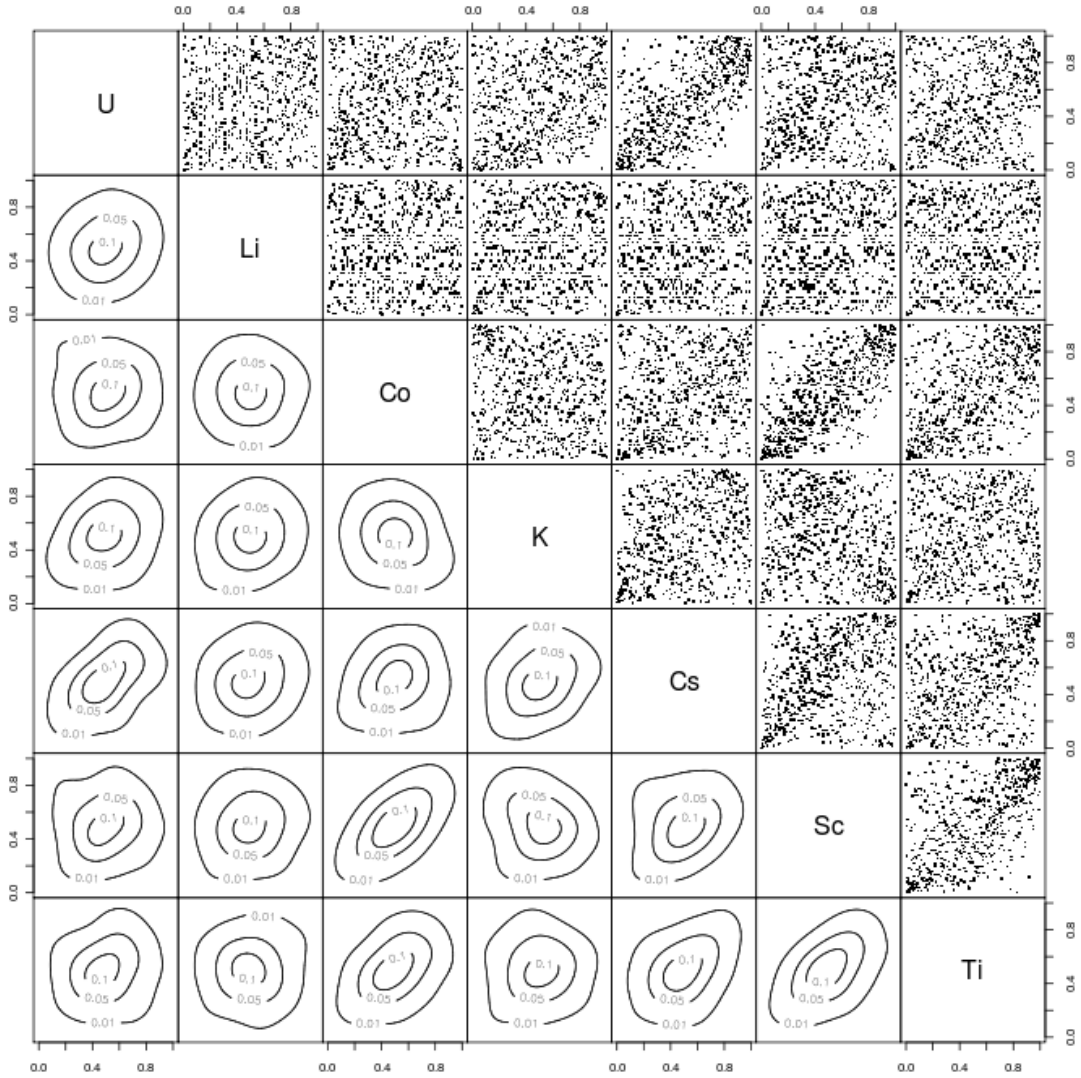

FIGURE 2. *Pairwise scatter plots of the copula data (transformed ranks) together with empirical contour plots of the corresponding normalized copula data (transformed to have standard normal margins) for the uranium data set*

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

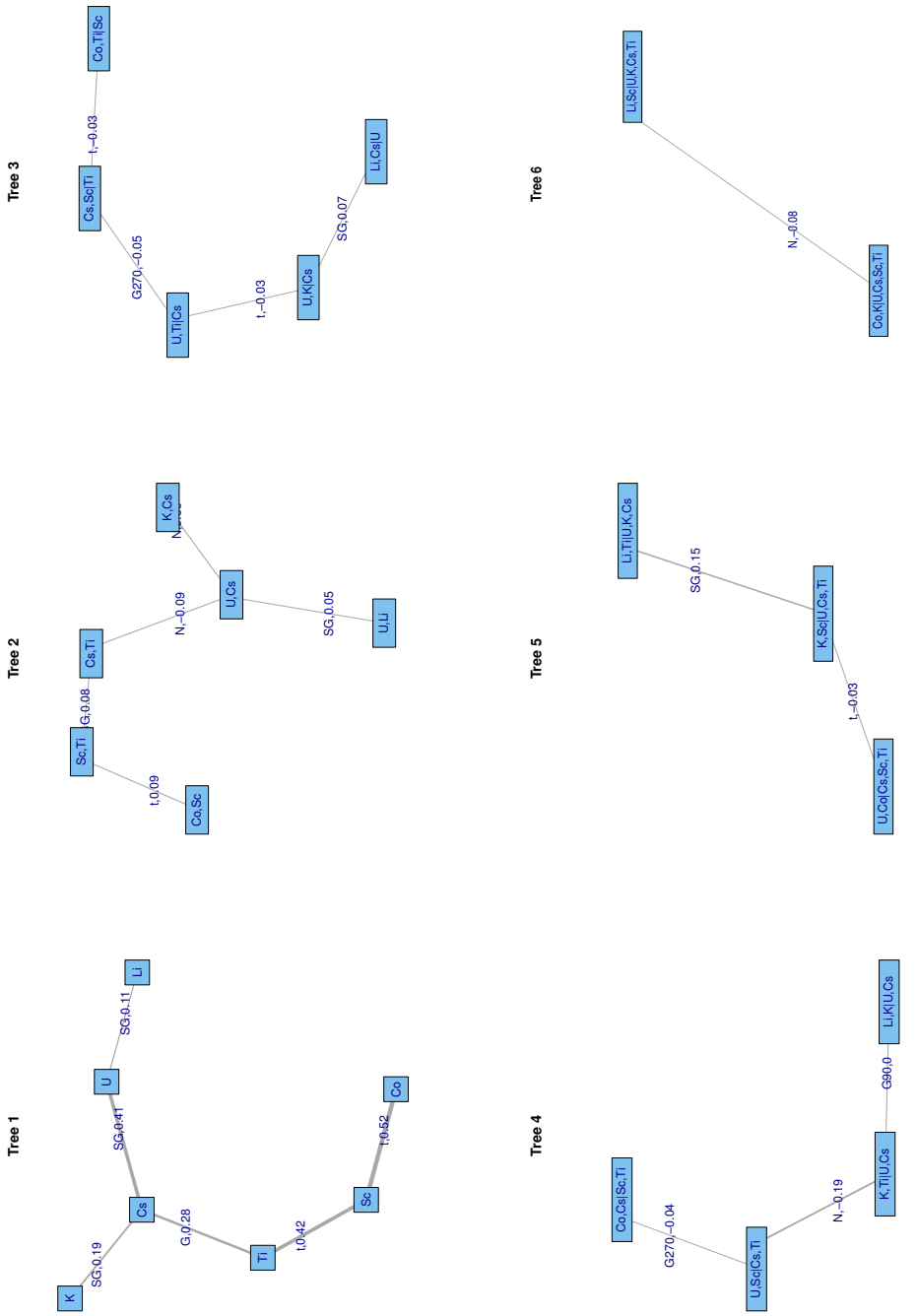

FIGURE 3. *Fitted R-vine using the TAU selection strategy applied to the uranium data (Gaussian (N), Student t (t), Gumbel (G) and their rotations by 90, 180 and 270 degree)*

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

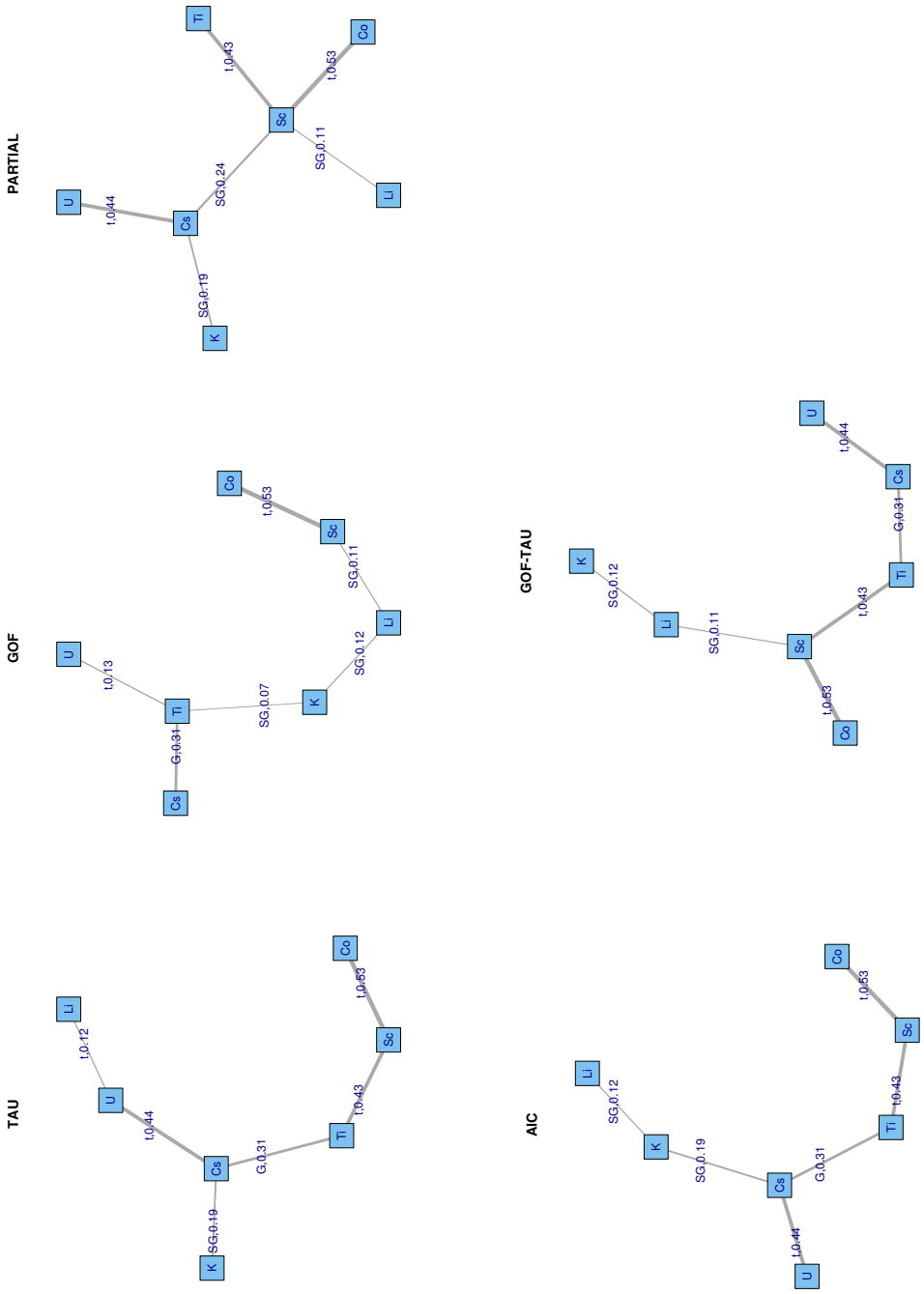

FIGURE 4. *Selected first tree together with copula families and estimated Kendall's τ for the five selection strategies applied to the uranium data (Gaussian (N), Student t (t), Gumbel (G) and their rotations by 90, 180 and 270 degree)*

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

TABLE 2. *Estimated median and 5% and 95% quantiles of the BIC based on 100 data sets simulated from the five selected R-vine models for the* `uranium` *data and fitted with the five selection strategies*

	Selection strategy				
	TAU	GOF	PARTIAL	AIC	GOF-TAU
MODEL TAU					
5%	-1762.6	-1727.7	-1741.9	-1762.3	-1738.6
50%	**-1626.5**	-1590.3	-1608.3	**-1627.9**	-1609.0
95%	-1492.4	-1450.3	-1464.7	-1500.1	-1490.2
MODEL GOF					
5%	-1728.7	-1735.1	-1729.7	-1726.1	-1727.0
50%	-1583.7	-1577.2	**-1591.5**	-1586.0	-1585.8
95%	-1456.5	-1449.4	-1454.7	-1456.8	-1458.9
MODEL PARTIAL					
5%	-1773.2	-1783.3	-1779.3	-1786.7	-1779.6
50%	-1616.1	-1608.4	-1617.8	-1623.4	**-1634.4**
95%	-1454.6	-1456.3	-1451.7	-1463.8	-1469.2
MODEL AIC					
5%	-1800.8	-1758.4	-1768.1	-1800.9	-1781.0
50%	-1627.3	-1591.8	-1609.5	**-1633.9**	-1609.2
95%	-1474.1	-1438.4	-1453.2	-1481.2	-1478.1
MODEL GOF-TAU					
5%	-1767.8	-1752.6	-1769.2	-1764.2	-1768.9
50%	-1588.7	-1563.2	-1581.0	**-1595.5**	-1575.0
95%	-1435.9	-1421.4	-1430.4	-1440.4	-1425.0

each of the five models 100 data sets of size 655. Subsequently we used the 5 strategies to fit each simulated data set. The resulting median values and 5% and 95% empirical quantiles of the BIC statistics are given in Table 2. The smallest median BIC values per model are bolded. All strategies but the GOF strategy is chosen in these simulation settings, however the GOF strategy performed strongly when applied to the `uranium` data.

As a final comment note that a more detailed analysis of the triplet (Co,Ti,Sc) in [Acar et al., 2012] revealed that this data set might be best fitted by a nonparametric conditional copula depending on the exact value of the conditioning variable. However the method presented in [Acar et al., 2012] is currently restricted to the three dimensional case and extension to higher dimension would require efficient nonparametric smoothing methods in dimensions higher than 2.

8. Discussion and outlook

In this paper we introduced R-vine copulae and distributions, which can be used to model different dependency patterns for different pairs. These models extend considerably the class of eliptical and Archemedean copulae. For these models different parameter estimation methods such as the sequential, maximum likelihood and Bayesian for a specified tree structure and pair copula families. The class of R-vine tree structures is very large and further flexibility is introduced by choosing from a set of parametric bivariate copula families. The potential of this flexibility was illustrated in an application involving concentrations of chemicals in water samples.

Since the model is formulated in a sequential way using a set of sequentially linked trees, sequential selection procedures were developed. There are two basic sequential selection proce-

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

dures, one starting from the top tree to the bottom and one from the bottom to the top tree. For the top down procedure of [Dißmann et al., 2013] we provided some extensions by considering different characteristics chosen as weights in a maximal spanning tree algorithm. The application and the simulation study illustrated that there is so far no winner for all data situations. This is to be expected because of the sequential nature of the selection. For comparison forward selection procedures for selection covariates in linear models are known to be not optimal as well.

Since the model class is very large it is not feasible to always find a single best model, however we are able to select reasonable candidate models and discuss their common dependency structures. These common structures indicate parts of the model, where uncertainty is lower than other parts. Selection of the best performing models were based on popular information criteria such as AIC and BIC. However selected models are nonnested and therefore we could also base our decisions on the tests suggested in [Vuong, 1989]. These tests have been applied in the context of truncating R-vine based models in [Brechmann et al., 2012]. There also strategies for finding more parsimonious models are suggested. In addition if the model selection is based on pseudo (conditional) copula data the use of AIC and BIC is only approximately justified.

Vine based models have been applied successfully in higher dimensions ([Chollete et al., 2008, Brechmann and Czado, 2012b]) and current development both in theory and applications are contained in [Kurowicka and Joe, 2011]. They have been extended to include time varying copula parameters ([Almeida et al., 2012]), regime switching copulae ([Chollete et al., 2008, Stöber and Czado, 2013]), factor copula models ([Krupskii and Joe, 2012]), cross serial dependence ([Brechmann and Czado, 2012a]) and multivariate option pricing ([Bernard and Czado, 2012]). Similar PCC decompositions are available for discrete data ([Panagiotelis et al., 2012]) and are used to construct Non Gaussian distributions on directed acyclic graphs ([Bauer et al., 2012]). In addition there are the R libraries `CDVine` ([Schepsmeier and Brechmann, 2012]) for C- and D-vines and `VineCopula` ([Schepsmeier et al., 2012]) for R-vines available. This shows that these copula models provide an open and wide field for complex dependency models both for applications and theory.

Acknowledgment

The authors would like to thank the referees for helpful suggestions which improved the manuscript considerably.

References

[Aas et al., 2009] Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula construction of multiple dependence. *Insurance Mathematics and Economics*, 44:182–198.

[Acar et al., 2012] Acar, E., Genest, C., and Neštlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, pages 267–281.

[Almeida et al., 2012] Almeida, C., Czado, C., and Manner, H. (2012). Modeling high dimensional time-varying dependence using D-vine scar models. Preprint.

[Baba et al., 2004] Baba, K., Shibata, R., and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, 46(4):657–664.

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

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

[Bedford and Cooke, 2001] Bedford, T. and Cooke, R. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268.

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

[Ben Ghorbal et al., 2009] Ben Ghorbal, N., Genest, C., and Neštlehová, J. (2009). On the Ghoudi, Khoudraji and Rivest test for extreme-value dependence. *Canadian Journal of Statistics*, 37:534–552.

[Berg, 2009] Berg, D. (2009). Copula goodness-of-fit testing: an overview and power comparison. *The European Journal of Finance*, 15(7-8):675–701.

[Bernard and Czado, 2012] Bernard, C. and Czado, C. (2012). Multivariate option pricing using copulae. *Applied Stochastic Models in Business and Industry*, pages n/a–n/a.

[Brechmann, 2010] Brechmann, E. (2010). Truncated and simplified regular vines and their applications. Master's thesis, Technische Universität München.

[Brechmann and Czado, 2012a] Brechmann, E. and Czado, C. (2012a). COPAR - Multivariate time series modeling using the COPula AutoRegressive model. Preprint.

[Brechmann and Czado, 2012b] Brechmann, E. and Czado, C. (2012b). Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50. Preprint.

[Brechmann et al., 2012] Brechmann, E., Czado, C., and Aas, K. (2012). Truncated regular vines in high dimensions with applications to financial data. *Canadian Journal of Statistics*, 40(1):68–85.

[Chen and Fan, 2006] Chen, X. and Fan, Y. (2006). Estimation and model selection of semi-parametric copula based multivariate dynamic models under copula misspecification. *Canadian Journal of Statistics*.

[Chollete et al., 2008] Chollete, L., Heinen, A., and Valdesogo, A. (2008). Modeling international financial returns with a multivariate regime switching copula. Preprint.

[Cook and Johnson, 1981] Cook, R. and Johnson, M. (1981). A family of distributions for modeling nonelliptically symmetric multivariate data. *Journal of the Royal Statistical Society, Series B*, 43:210–218.

[Cormen et al., 2009] Cormen, T., Leiserson, E., Rivest, R., and Stein, C. (2009). *Introduction to Algorithms*. The MIT Press, 3rd edition.

[Czado et al., 2012] Czado, C., Schepsmeier, U., and Min, A. (2012). Maximum likelihood estimation of mixed C-vine pair copula with application to exchange rates. *Statistical Modelling*, 12:229–255.

[Dißmann et al., 2013] Dißmann, J., Brechmann, E., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69.

[Frahm et al., 2003] Frahm, G., Junker, M., and Szimayer, A. (2003). Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, 63(3):275–286.

[Genest et al., 1995] Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

[Genest and Remillard, 2008] Genest, C. and Remillard, B. (2008). Validity of the parametric bootstrap for goodness of fit testing in semiparametric models. *Annales de l'Institut Henri Pointcare: Probabilities et Statistiques*, 44:1096–1127.

[Genest et al., 2009] Genest, C., Remillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213.

[Gruber and Czado, 2013] Gruber, L. and Czado, C. (2013). Sequential bayesian model selection of regular vine copulas. *Preprint*.

[Hobak Haff, 2012] Hobak Haff, I. (2012). Parameter estimation for pair-copula constructions. *Bernoulli (in Press)*.

[Hobak Haff et al., 2010] Hobak Haff, I., Aas, K., and Frigessi, A. (2010). On the simplified pair-copula construction - simply useful or too simplistic? *Journal of Multivariate Analysis*, 101:1296–1310.

[Joe, 1996] 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, editor, *Distributions with Fixed Marginals and Related Topics*.

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

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238

[Joe and Xu, 1996] Joe, H. and Xu, J. (1996). The estimation method of inference functions for margins of multivariate models. *Technical Report 166, Department of Statistics, University of British Columbia*.

[Kim et al., 2007] Kim, G., Silvapulle, M., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, 51(6):2836–2850.

[Kojadinovic and Yan, 2010] Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20.

[Kojadinovic and Yan, 2011] Kojadinovic, I. and Yan, J. (2011). A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems. *Statistics and Computing*, 21:17–30.

[Kojadinovic et al., 2011] Kojadinovic, I., Yan, J., and Holmes, M. (2011). Fast large-sample goodness-of-fit test for copulas. *Statistica Sinica*, 21:841–871.

[Krupskii and Joe, 2012] Krupskii, P. and Joe, H. (2012). Factor copula models for multivariate data, with applications to financial data. Preprint.

[Kurowicka, 2011] Kurowicka, D. (2011). Optimal truncation of vines. In *Dependence Modelling: Vine Copula Handbook*. World Scientific Publishing Co.

[Kurowicka and Cooke, 2008] Kurowicka, D. and Cooke, R. (2008). *Uncertainty Analysis with High Dimensional Dependence Modelling*. John Wiley & Sons, Ltd.

[Kurowicka and Joe, 2011] Kurowicka, D. and Joe, H. (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific Publishing Co., Singapore.

[Min and Czado, 2010] Min, A. and Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8(4):511–546.

[Morales-Nápoles et al., 2010] Morales-Nápoles, O., Cooke, R., and Kurowicka, D. (2010). About the number of vines and regular vines on n nodes. *Preprint*.

[Nelson, 2006] Nelson, R. (2006). *An Introduction to Copulas*. Springer Science+Business Media, Inc.

[Nikoloulopoulos et al., 2012] Nikoloulopoulos, A., Joe, H., and Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics & Data Analysis*, 56(11):3659–3673.

[Panagiotelis et al., 2012] Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

[Schepsmeier and Brechmann, 2012] Schepsmeier, U. and Brechmann, E. (2012). *CDVine: Statistical inference of C- and D-vine copulas*. R package version 1.1-9.

[Schepsmeier et al., 2012] Schepsmeier, U., Stoeber, J., and Brechmann, E. (2012). *VineCopula: Statistical inference of vine copulas*. R package version 1.0.

[Stöber and Czado, 2013] Stöber, J. and Czado, C. (2013). Detecting regime switches in the dependence stucture of high dimensional financial data. To appear in Computational Statistics & Data Analysis.

[Stöber et al., 2012] Stöber, J., Joe, H., and Czado, C. (2012). Simplified pair copula constructions - limits and extensions. Preprint.

[Vuong, 1989] Vuong, Q. (1989). Ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333.

© Société Française de Statistique et Société Mathématique de France (2013) ISSN: 2102-6238