Multivariate Copula Expressed by Lower Dimensional Copulas

Edith Kovács[a], Tamás Szántai[b]

[a]*Department of Mathematics, ÁVF College of Management of Budapest, Villányi út 11-13, H-1114 Budapest, Hungary*
Corresponding author,
E-mail address: kovacs.edith@avf.hu
[b]*Institute of Mathematics, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary*
E-mail address: szantai@math.bme.hu

Abstract

Modeling of high order multivariate probability distribution is a difficult problem which occurs in many fields. Copula approach is a good choice for this purpose, but the curse of dimensionality still remains a problem. In this paper we give a theorem which expresses a multivariate copula by using only some lower dimensional ones based on the conditional independences between the variables. In general the construction of a multivariate copula using this theorem is quite difficult, due the consistency properties which have to be fulfilled. For this purpose we introduce the sample derivated copula, and prove that the dependence between the random variables involved depends just on this copula and on the partition. By using the sample derivated copula the theorem can be successfully applied, in order to to construct a multivariate discrete copula by using some of its marginals.

Keywords:
Multivariate copula, Junction tree, Conditional independence, Sample derivated copula.

1. Introduction

First we motivate why should we model the multivariate distribution by copulas from an information theoretical point of view. The information content of a multivariate probability distribution depends only on its copula

density. In [12] and [1] one can see this result for the two-dimensional case and the same is true for more dimensions, too.

In this paper we prove a theorem which links the multivariate probability distribution assigned to a junction tree to the multivariate copula. It is known that the probability distribution assigned to a junction tree uses the conditional independence structure underlying the random variables so the copula introduced here will have this property, too.

In this introductory part we describe the main concepts and introduce the notations which we will use in the paper. In the second section we prove a theorem which links a multivariate copula to the junction tree probability distribution. In the third section we will introduce the concept of Sample Derivated Copula (SDC) which makes possible the exploitation of the conditional independences between the random variables. We prove that the information content of the probability distribution given by a partition set depends only on the SDC. In the fourth section we apply the junction tree approach to the SDC.

We finish the paper with conclusions and possible applications.

Let $V = \{1, \ldots, n\}$ be a set of vertices. A hypergraph is a set V of vertices together with a set Γ of subsets of V. A hypergraph is acyclic if no elements in Γ are subsets of other elements, and if the elements of Γ can be ordered $(K_1, \ldots, K_m)$ to have the *running intersection property*: for all $j \geq 2$, exists $i < j : K_i \supseteq K_j \cap (K_1 \cup \ldots \cup K_{j-1})$ [8].

It is convenient to introduce the so called separator sets $S_j = K_j \cap (K_1 \cup \ldots \cup K_{j-1})$, where $S_1 = \phi$.

We note here that if $R_j = K_j \backslash S_j$ then S_j separates (in graph terms) the vertices in R_j from the vertices in $(K_1 \cup \ldots \cup K_{j-1}) \backslash S_j$.

We mention here that a hypergraph (V, Γ) is acyclic if and only if Γ can be considered to be the set of cliques of a chordal (triangulated) graph [9],[16].

In the following we consider acyclic hypergraphs with the property that the union of all sets in Γ is V. We denote the separator set by $\mathcal{S}$ and refer to the acyclic hypergraph as $(V, \Gamma, \mathcal{S})$.

Let $V = \{1, 2, \ldots, n\}$ be the set of indices of the continuous random variables $X = \{X_1, \ldots, X_n\}$. We suppose that the probability density functions of $X_1, \ldots, X_n$ exist and denote them by $f_{X_1}, \ldots, f_{X_n}$.

We need the following notations:

- $F_{X_i}(x_i) = P(X_i < x_i; X_j = \infty \text{ for all } j \neq i)$ stands for the univariate marginal cumulative distribution function corresponding to the variable

X_i,

- The joint probability density function and the joint cumulative distribution function of $(X_1, \ldots, X_n)^T$ is denoted by $f_{\mathbf{X}}(\mathbf{x})$ and $F_{\mathbf{X}}(\mathbf{x})$, respectively,
- $D = \{i_1, \ldots, i_d\} \subset V$, $\mathbf{X}_D = (X_{i_1}, \ldots, X_{i_d})^T$, $\mathbf{x}_D = (x_{i_1}, \ldots, x_{i_d})^T$,
- The d-th order marginal probability density function and the d-the order marginal cumulative distribution function of $\mathbf{X}_D$ is denoted by $f_{\mathbf{X}_D}(\mathbf{x}_D)$ and $F_{\mathbf{X}_D}(\mathbf{x}_D)$,respectively.

Having these notations we give the concept of the junction tree. It is known that the junction tree encodes the conditional independences between the variables. Let us remark here that from now on the indices of the random variables are assigned to the nodes of a graph. In the graph a set of nodes B separates a set of nodes A from another set of nodes C, where A, B, C are disjoint subsets of V, if and only if X_A and X_C are conditionally independent with respect to X_B (see the definition of the Markov random field).

Definition 1. A junction tree over X is a cluster tree, which is assigned to an acyclic hypergraph $(V, \Gamma, \mathcal{S})$ as follows:

1) Each cluster of the cluster tree consists of a subset X_K of X , where $K \in \Gamma$. To each cluster is assigned the joint marginal density function $f_{\mathbf{X}_K}(\mathbf{x}_K)$;
2) Each edge connecting to clusters is called separator and consists of a subset X_S of X, where S is a separator set. To each separator there is assigned the marginal probability density function $f_{\mathbf{X}_S}(\mathbf{x}_S)$;
3) The union of all clusters is X.

Definition 2. A junction tree probability distribution is a probability distribution assigned to the junction tree in the following way:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\prod\limits_{K \in \Gamma} f_{\mathbf{X}_K}(\mathbf{x}_K)}{\prod\limits_{S \in \mathcal{S}} \left(f_{\mathbf{X}_S}(\mathbf{x}_S) \right)^{v_{s-1}}},$$

where v_S is the number of those clusters which contain all the variables of $\mathbf{X}_S$.

It is useful to note here that since in the hypergraph $(V, \Gamma, \mathcal{S})$ S_j separates (in graph terms) the vertices in $R_j = K_j - S_j$ from the vertices in $(K_1 \cup \ldots \cup K_{j-1}) - S_j$ the random variables with indices in $R_j = K_j - S_j$ and the variables with indices in $(K_1 \cup \ldots \cup K_{j-1}) - S_j$ are conditionally independent with respect to the variables with indices in S_j.

Remark 1. Since the junction tree is assigned to an acyclic hypergraph, the running intersection property stands for the junction tree, too. It can be reformulated as follows. If two clusters contain a random variable, then all clusters on the path between these clusters contain this random variable.

First we call back the concept of copula and formulate the Sklar's theorem (see [3] and [13]).

Definition 3. A function $C : [0;1]^d \to [0;1]$ is called a d-dimensional copula if it satisfies the following conditions:

1) $C(u_1, \ldots, u_d)$ is increasing in each component u_i,
2) $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$ for all $u_k \in [0;1]$, $k \neq i$, $i = 1, \ldots, n$,
3) $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for all $u_i \in [0;1]$, $i = 1, \ldots, d$,
4) C is d-increasing, i.e for all $(u_{1,1}, \ldots, u_{1,d})$ and $(u_{2,1}, \ldots, u_{2,d})$ in $[0;1]^d$ with $u_{1,i} < u_{2,i}$ for all i, we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{\sum_{j=1}^{d} i_j} C(u_{i_1,1}, \ldots, u_{i_d,d}) \geq 0.$$

Due to Sklar's theorem if $X_1, \ldots, X_d$ are continuous random variables defined on a common probability space, with the univariate marginal cdf's $F_{X_i}(x_i)$ and the joint cdf $F_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$ then there exists a unique copula function $C_{X_1,\ldots,X_d}(u_1, \ldots, u_d) : [0;1]^d \to [0;1]$ such that by the substitution $u_i = F_i(x_i)$, $i = 1, \ldots, d$ we get

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = C_{X_1,\ldots,X_d}(F_1(x_1), \ldots, F_d(x_d))$$

for all $(x_1, \ldots, x_d)^T \in R^d$.

In the following we will use the vectorial notation $F_{\mathbf{X}_D}(\mathbf{x}_D) = C_{X_D}(\mathbf{u}_D)$, where $\mathbf{u}_D = \left(F_{X_{i_1}}(x_{i_1}), \ldots, F_{X_{i_d}}(x_{i_d})\right)^T$.

We need the following assertion:

$$
\begin{aligned}
&f_{X_{i_1},\ldots X_{i_d}}(x_{i_1},\ldots,x_{i_d}) = \\
&= \frac{\partial^d F_{X_{i_1},\ldots X_{i_d}}(x_{i_1},\ldots,x_{i_d})}{\partial x_{i_1}\cdots\partial x_{i_d}} \\
&= \frac{\partial^d C_{X_{i_1},\ldots,X_{i_d}}\left(F_{X_{i_1}}(x_{i_1}),\ldots,F_{X_{i_d}}(x_{i_d})\right)}{\partial x_{i_1}\cdots\partial x_{i_d}} \\
&= \left.\frac{\partial^d C_{X_{i_1},\ldots X_{i_d}}(u_{i_1},\ldots,u_{i_d})}{\partial u_{i_1}\cdots\partial u_{i_d}}\right|_{\substack{u_{i_k}=F_{X_{i_k}}(x_{i_k}),\\ k=1,\ldots d}} \cdot \prod_{k=1}^{d}\frac{\partial F_{X_{i_k}}(x_{i_k})}{\partial x_{i_k}} \\
&= c_{X_{i_1},\ldots X_{i_d}}\left(F_{X_{i_1}}(x_{i_1}),\ldots,F_{X_{i_d}}(x_{i_d})\right)\cdot\prod_{k=1}^{d} f_{X_{i_k}}(x_{i_k})
\end{aligned}
$$

In vectorial notation this can be written as

$$
f_{\mathbf{X}_D}(\mathbf{x}_D) = c_{\mathbf{X}_D}(\mathbf{u}_D)\cdot\prod_{i_k\in D} f_{X_{i_k}}(x_{i_k}) \tag{1}
$$

and from (1) we get

$$
c_{\mathbf{X}_D}(\mathbf{u}_D) = \frac{f_{\mathbf{X}_D}(\mathbf{x}_D)}{\prod_{i_k\in D} f_{X_{i_k}}(x_{i_k})} \tag{2}
$$

2. The multivariate copula associated to a junction tree probability distribution.

Theorem 1. *The copula density function associated to a junction tree probability distribution*

$$
f_{\mathbf{X}}(\mathbf{x}) = \frac{\prod_{K\in\Gamma} f_{\mathbf{X}_K}(\mathbf{x}_K)}{\prod_{S\in\mathcal{S}}\left[f_{\mathbf{X}_S}(\mathbf{x}_S)\right]^{v_S-1}},
$$

is given by

$$
c_{\mathbf{X}}(\mathbf{u}_V) = \frac{\prod_{K\in\Gamma} c_{\mathbf{X}_K}(\mathbf{u}_K)}{\prod_{S\in\mathcal{S}}\left[c_{\mathbf{X}_S}(\mathbf{u}_S)\right]^{v_S-1}}. \tag{3}
$$

PROOF.

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\prod\limits_{K\in\Gamma} f_{\mathbf{X}_K}(\mathbf{x}_K)}{\prod\limits_{S\in\mathcal{S}} \left[f_{\mathbf{X}_S}(\mathbf{x}_S)\right]^{v_S-1}} = \frac{\prod\limits_{K\in\Gamma} c_{\mathbf{X}_K}(\mathbf{u}_K) \cdot \prod\limits_{i_k\in K} f_{X_{i_k}}(x_{i_k})}{\prod\limits_{S\in\mathcal{S}} \left[c_{\mathbf{X}_S}(\mathbf{u}_S) \cdot \prod\limits_{i_k\in S} f_{X_{i_k}}(x_{i_k})\right]^{v_S-1}}. \quad (4)$$

The question that we have to answer is how many times appears in the nominator respectively in the denominator the probability density function $f_{X_i}(x_i)$ of each X_i random variable.

Since $\bigcup\limits_{K\in\Gamma} \mathbf{X}_K = X$ for each random variable $X_i \,\epsilon X$, $f_{X_i}(x_i)$ appears at least once in the nominator.

Now we prove that in the junction tree over X the number of clusters which contain a variable X_i is greater with 1 than the number of separators which contain the same variable. This is true for all $i = 1, \ldots, n$. This means $\#\{K \in \Gamma | X_i \in X_K\} = \#\{S \in \mathcal{S} | X_i \in X_S\} + 1$.

For a variable X_i we denote $\#\{S \in \mathcal{S} | X_i \in X_S\}$ by t.

Case: $t = 0$.

The statement is a consequence of the definition of junction tree, that is the union of all clusters is X, so every variable have to appear at least in one cluster. X_i can not appear in two clusters, because in this case there should exist a separator which contain X_i too, and we supposed that there is not such a separator ($t = 0$)

Case: $t > 0$

If two clusters contain the variable X_i, then every cluster from the path between the two clusters contain X_i (running intersection property). From this results that the clusters containing X_i are the nodes of a connected graph, and this graph is a tree. If this tree contain t separator sets then it contains $t+1$ clusters. All of these separators contain X_i, and each separator connects two clusters. So there will be $t + 1$ clusters that contain X_i.

Applying this result in formula (4) after simplification we obtain

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\prod\limits_{K\in\Gamma} c_{\mathbf{X}_K}(\mathbf{u}_K) \prod\limits_{i=1}^{n} f_{X_i}(x_i)}{\prod\limits_{S\in\mathcal{S}} \left[c_{\mathbf{X}_S}(\mathbf{u}_S)\right]^{v_S-1}}.$$

Dividing both sides by $\prod\limits_{i=1}^{n} f_{X_i}(x_i)$ we obtain

$$\frac{f_{\mathbf{X}}(\mathbf{x})}{\prod_{i=1}^{n} f_{X_i}(x_i)} = \frac{\prod_{K\in\Gamma} c_{\mathbf{X}_K}(\mathbf{u}_K)}{\prod_{S\in\mathcal{S}} \left[c_{\mathbf{X}_S}(\mathbf{u}_S)\right]^{v_S-1}}. \tag{5}$$

Equations (2) and (5) prove the statement of the theorem.

We saw that if the conditional independence structure underlying the random variables makes possible the construction of a junction tree, then the multivariate copula density associated to the joint probability distribution can be expressed as a product and fraction of lower dimensional copula densities.

A logical question is the following. What conditions are necessary for (5) to be a copula density? It is easy to see that the product and fraction of copulas are positive. So

$$c(\mathbf{u}) = \frac{\prod_{K\in\Gamma} c_K(\mathbf{u}_K)}{\prod_{S\in\mathcal{S}} \left[c_S(\mathbf{u}_S)\right]^{v_S-1}}$$

will be a copula density if and only if

$$\int_{[0;1]^n} \frac{\prod_{K\in\Gamma} c_K(\mathbf{u}_K)}{\prod_{S\in\mathcal{S}} \left[c_S(\mathbf{u}_S)\right]^{v_S-1}} d\mathbf{u} = 1.$$

This happens if the following consistency conditions are fulfilled for all connected clique pairs K_i and K_j:

$$\int_{[0;1]^{\#\{K_i\setminus S_{ij}\}}} c_{Ki}(\mathbf{u}_{K_i})\, d\mathbf{u}_{K_i\setminus S_{ij}} = \int_{[0;1]^{\#\{K_j\setminus S_{ij}\}}} c_{Kj}\left(\mathbf{u}_{K_j}\right) d\mathbf{u}_{K_j\setminus S_{ij}},$$

where $S_{ij} = K_i \cap K_j$. We emphasize here that all cliques are subsets of the set Γ of the acyclic hypergraph $(V, \Gamma, \mathcal{S})$.

These conditions are fulfilled if $c_{S_{ij}}\left(\mathbf{u}_{S_{ij}}\right)$ are marginal probability densities of $c_{K_i}(\mathbf{u}_{K_i})$, whenever S_{ij} connects a cluster K_i. This can be expressed by terms of copula function as follows.

For $\{k_1, \ldots, k_m\} = K_i$ and $\{s_1, \ldots, s_l\} = S_{ij}$, $\{s_1, \ldots, s_l\} \subset \{k_1, \ldots, k_m\}$ stands $C_m(u_{k_1}, \ldots, u_{k_m}) = C_l(u_{s_1}, \ldots, u_{s_l})$ for $u_{k_i} = 1$, when $k_i \notin S_{ij}$. Usually this condition is not fulfilled by copulas.

Finding multivariate copulas which fulfill the consistency conditions is not a trivial task.

A special type of conditional independence, when the graph underlying the random variables is starlike, is treated in [19]. Another type of special multivariate copula where the underlying conditional independence graph is a tree can be found in [6].

For discrete random variables, the conditional independences are exploited by the Markov random fields. In physics for two-valued random variables it is known the Ising model. In these cases the random variables take on a few values only. However many times the problem is hard. The great advantage of using the discrete approach is that the marginal probability distributions involved fulfill the consistency conditions (see [17] and [18]).

If we have an i.i.d. sample of size N from a continuous joint probability distribution then for each random variable we have N different values. For this case, the empirical copulas were introduced and first studied by P. Deheuvels in [5] who called them empirical dependence functions. Later in [10] and [11] there were introduced the so called discrete copulas. About the two-dimensional empirical copulas one can read in Nelsen's introductory book (see [13]). In the case when we are dealing with a sample drawn from a continuous joint probability distribution the size of these random variables would be too large, so we will apply a uniform partition and define the so called sample derivated copula.

3. The sample derivated copula.

Let $X_1, \ldots, X_n$ be continuous random variables in the same probability field. Let

$$\begin{array}{c} x_1^1, \ldots, x_n^1 \\ x_1^2, \ldots, x_n^2 \\ \vdots \\ x_1^N, \ldots, x_n^N \end{array} \tag{6}$$

be an i.i.d. sample of size N taken from the joint probability distribution of the random vector $(X_1, \ldots, X_n)^T$.

As any sample element occurs two times in the sample with probability zero, we can suppose that the sample elements are different.

We denote the set of the values of X_i in the sample by Λ_i. This set contains N values, for each random variable. The theoretical range of the continuous random variable X_i will be denoted by $\overline{\Lambda_i}$. For every i we denote by $\lambda_i^m = \min \overline{\Lambda_i} \in R$ and by $\lambda_i^M = \max \overline{\Lambda_i} \in R$. We suppose for simplicity that $\min \overline{\Lambda_i} \neq \min \Lambda_i$ and $\max \overline{\Lambda_i} \neq \max \Lambda_i$ For each random variable X_i we define a partition of Λ_i by $\mathcal{P}_i = \left\{x_0^{p_i} = \lambda_i^m, x_1^{p_i}, \ldots, x_{m_i-1}^{p_i}, x_{m_i}^{p_i} = \lambda_i^M\right\}$ with the following properties:

- For each random variable X_i, each interval $\left(x_{j-1}^{p_i}; x_j^{p_i}\right], j = 1, \ldots, m_i$ contains a given $n_i = \dfrac{N}{m_i} \in N$ number of values from the set Λ_i.
- Each $x_j^{p_i} \in \Lambda_i,\ j = 1, \ldots, m_i - 1$.

The partition with the above properties will be called uniform partition. We denote by $\mathcal{P}$ the set of partitions $\{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$.

Let be $\widetilde{X_i}$ the categorical random variable associated to the random variable X_i:

$$P(\widetilde{X_i} \in \left(x_{j-1}^{p_i}; x_j^{p_i}\right]) = \frac{1}{m_i}, j = 1, \ldots, m_i.$$

We assign to each $x^i \in \left(x_{j-1}^{p_i}; x_j^{p_i}\right]$ the number $u_j^i = \dfrac{j}{m_i}, j = 0, \ldots, m_i$. Obviously $u_0^i = 0$ and $u_{m_i}^i = 1$. Let $\widetilde{\Lambda}_i = \left\{u_j^i | j = 0, \ldots, m_i\right\}$. So we can define the following discrete uniform random variables:

$$\widetilde{U}_i = \begin{pmatrix} u_0^i & u_1^i & \ldots & u_{m_i-1}^i & u_{m_i}^i \\ 0 & \dfrac{1}{m_i} & \ldots & \dfrac{1}{m_i} & \dfrac{1}{m_i} \end{pmatrix}, i = 1, \ldots, n.$$

Now we transform the sample (6) using the above assignment. We denote the transformed sample by $\mathcal{T}$.

Definition 4. The function $\widetilde{c} : \prod_{i=1}^{n} \widetilde{\Lambda}_i \to R$ defined by

$$\left(u_{k_1}^1, \ldots, u_{k_n}^n\right) \mapsto \widetilde{c}\left(u_{k_1}^1, \ldots, u_{k_n}^n\right) = \frac{\#\left\{(u_{k_1}^1, \ldots, u_{k_n}^n) \in \mathcal{T}\right\}}{N}, k_i = 0, \ldots, m_i$$

will be called sample derivated copula distribution.

Remark 2. The maximum number of different vectors what the above defined sample derivated copula can take on equals to $\prod_{i=1}^{n} m_i$.

Definition 5. The function $\widetilde{C}_n^{\mathcal{P}} : \prod_{i=1}^{n} \widetilde{\Lambda}_i \subset [0;1]^n \to [0;1]$ defined by

$$\left(u_{k_1}^1, \ldots, u_{k_n}^n\right) \mapsto \widetilde{C}_n^{\mathcal{P}}\left(u_{k_1}^1, \ldots, u_{k_n}^n\right) =$$
$$= \frac{\#\left\{(u_{k_1}^1, \ldots, u_{k_n}^n) \in \mathcal{T} \mid u_1 \leq u_{k_1}^1, \ldots, u_n \leq u_{k_n}^n\right\}}{N}$$

will be called sample derivated copula.

Throughout the paper we use the notation $\widetilde{C}_n$ instead of $\widetilde{C}_n^{\mathcal{P}}$.

Theorem 2. *The sample derivated copula is a copula.*

Proof. 1) It is evident that $\widetilde{C}_n$ is increasing in its each component.

2) If exists s such that $u_{k_s}^s = 0$ then $\widetilde{C}_n\left(u_{k_1}^1, \ldots, u_{k_{s-1}}^{s-1}, 0, u_{k_{s+1}}^{s+1} \ldots, u_{k_n}^n\right) = 0$. This follows directly from the definition. The sample do not contain any vector with a negative coordinate.

3) If for all $s \neq l$ we have $u_{k_s}^s = 1$ then

$$\widetilde{C}_n\left(1, \ldots 1, u_{k_l}^l, 1 \ldots, 1\right) =$$
$$= \frac{\#\left\{(u_1, \ldots, u_s, \ldots u_n) \in \mathcal{T} \mid u_s \leq 1, \forall s \neq l \text{ and } u_l \leq u_{k_l}^l\right\}}{N} =$$
$$= \frac{1}{N} \sum_{0}^{k_l - 1} n_l = \frac{n_l \cdot k_l}{N} = \frac{k_l}{m_l} = u_{k_l}^l$$

4) $\widetilde{C}_n$ is n-increasing as it is a cumulative probability distribution function.

Remark 3. The sample derivated copula differs, from the empirical copula [5] and the discrete copula [10], [11]. One of the differences is that the cardinal of $\widetilde{\Lambda}_i$ is not necessary the same for all $i = 1, \ldots, n$. Another difference is that a marginal variable can take the same value in more than one vector (since $m_i < N$).

Theorem 3. *The sample derivated copula has the following consistency property. If all variables $u^s_{k_s} = 1$ for $s \in V \setminus \{l_1, \ldots, l_q\}$ then*

$$\widetilde{C}_n\left(u^1_{k_1}, \ldots, u^n_{k_n}\right) = \widetilde{C}_q\left(u^1_{l_1}, \ldots, u^q_{l_q}\right).$$

Proof.

$$\widetilde{C}_n\left(u^1_{k_1}, \ldots, u^n_{k_n}\right) =$$

$$= \frac{\#\left\{(u_1, \ldots, u_n) \in \mathcal{T} \mid u_s \leq 1, s \in V \setminus \{l_1, \ldots, l_q\}, u_{l_i} \leq u^i_{l_i}, i = 1 \ldots q\right\}}{N} =$$

$$= \frac{\#\left\{(u_1, \ldots, u_n) \in \mathcal{T} \mid u_{l_i} \leq u^i_{l_i}, i = 1 \ldots q\right\}}{N} =$$

$$= \widetilde{C}_q\left(u^1_{l_1}, \ldots, u^q_{l_q}\right).$$

Remark 4. In general copulas do not fulfill the consistency property..

Remark 5. This theorem assures the consistency statements that we need when constructing junction tree like copulas.

At the end of this part we convince the reader from an information theoretical point of view why should one use the uniform partition and the sample derivated copula.

In the following we suppose that each $\Lambda_i, i = 1, \ldots, n$ is partitioned in the same number of $m_i, i = 1, \ldots, n$ intervals as in the previous case. We denote now the partitioning points by $y^{p_i}_j,\ j = 0, 1, \ldots, m_i;\ i = 1, \ldots, n$. This partition is arbitrary (for example equidistant) which has not the property that each interval contains the same number of sample elements. The partition of Λ_i is given by $\left\{y^{p_i}_j \mid j \in \{0, 1, \ldots, m_i\}\right\}$ and is denoted by $\mathcal{P}'_i$. We denote by $\mathcal{P}'$ the set of partitions $\mathcal{P}'_1, \ldots, \mathcal{P}'_n$.

We denote the number of values of the variable X_i belonging to $\left(y^{p_i}_j; y^{p_i}_{j+1}\right] \cap \Lambda_i$ by k^i_j.

Let $\widetilde{Y}_i$ be the categorical random variable associated to X_i:

$$P\left(\widetilde{Y}_i \in \left(y^{p_i}_j; y^{p_i}_{j+1}\right]\right) = \frac{k^i_j}{N}, j = 0, 1, \ldots, m_i,$$

where

$$\sum_{j=1}^{m_i} \frac{k^i_j}{N} = 1.$$

The entropy of X_i determined by the partition $\mathcal{P}_i'$ is:

$$H_{\mathcal{P}_i'}(X_i) = H\left(\widetilde{Y}_i\right) = -\sum_{j=1}^{m_i} \frac{k_j^i}{N} \log \frac{k_j^i}{N},\ i = 1 \ldots n.$$

It can be seen that the entropy $H_{\mathcal{P}_i'}(X_i)$ depends on the number of intervals m_i and on k_j^i.

We introduce the following notation:

$$q_{j_1,\ldots,j_n}^{X_1,\ldots,X_n} = P\left(X_1 \in \left(y_{j_1-1}^{p_1}; y_{j_1}^{p_1}\right], X_2 \in \left(y_{j_2-1}^{p_2}; y_{j_2}^{p_2}\right], \ldots, X_n \in \left(y_{j_n-1}^{p_n}; y_{j_n}^{p_n}\right]\right),$$

where $j_i = 1, \ldots, m_i,\ i = 1, \ldots, n$.

The joint probability distribution determined by the partition $\mathcal{P}'$ has the joint entropy:

$$H_{\mathcal{P}'}(X_1, \ldots, X_n) = -\sum_{j_1=1}^{m_1} \cdots \sum_{j_n=1}^{m_n} q_{j_1,\ldots j_n}^{X_1,\ldots,X_n} \log_2 q_{j_1,\ldots j_n}^{X_1,\ldots,X_n}.$$

The information content of the joint probability distribution determined by the partition $\mathcal{P}'$ is:

$$\begin{aligned} I_{\mathcal{P}'}(X_1, \ldots, X_n) &= \sum_{i=1}^{n} H_{\mathcal{P}_i'}(X_i) - H_{\mathcal{P}'}(X_1, \ldots, X_n) = \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{k_j^i}{N} \log \frac{k_j^i}{m_i} + \sum_{j_1=1}^{m_1} \cdots \sum_{j_n=1}^{m_n} q_{j_1,\ldots j_n}^{X_1,\ldots,X_n} \log_2 q_{j_1,\ldots j_n}^{X_1,\ldots,X_n}. \end{aligned} \tag{7}$$

Remark 6. In this case the information content depends on the number of intervals m_i, the number of values in each one dimensional interval k_j^i, and the probabilities of belonging to the n-dimensional intervals.

If we regard the uniform partition $\mathcal{P}$ then the entropy of X_i is:

$$H_{\mathcal{P}_i}(X_i) = H\left(\widetilde{X}_i\right) = -\sum_{j=1}^{m_i} \frac{1}{m_i} \log \frac{1}{m_i} = \log m_i,\ i = 1, \ldots, n.$$

The entropy $H_{\mathcal{P}_i}(X_i)$ depends just on the number of intervals m_i.

We express now the probability

$$\begin{aligned} &p_{j_1,\ldots j_n}^{X_1,\ldots,X_n} = P\left(X_1 \in \left(x_{j_1-1}^{p_1}; x_{j_1}^{p_1}\right], \ldots, X_n \in \left(x_{j_n-1}^{p_n}; x_{j_n}^{p_n}\right]\right) = \\ &P\left(\widetilde{U}_1 = u_{j_1}, \ldots, \widetilde{U}_n = u_{j_n}\right) = \widetilde{c}(u_{j_1}, \ldots, u_{jn}) \end{aligned}$$

The joint probability entropy associated to the partition is:

$$H_{\mathcal{P}}(X_1,\ldots,X_n) = -\sum_{j_1=1}^{m_1}\cdots\sum_{j_n=1}^{m_n} p_{j_1,\ldots j_n}^{X_1,\ldots,X_n}\log_2 p_{j_1,\ldots j_n}^{X_1,\ldots,X_n} =$$

$$= -\sum_{j_1=1}^{m_1}\cdots\sum_{j_n=1}^{m_n}\widetilde{c}(u_{j_1},\ldots,u_{jn})\log_2\widetilde{c}(u_{j_1},\ldots,u_{jn}).$$

The information content determined by the partition $\mathcal{P}$ is:

$$\begin{aligned} I_{\mathcal{P}}(X_1,\ldots,X_n) &= \sum_{i=1}^{n} H_{\mathcal{P}_i}(X_i) - H_{\mathcal{P}}(X_1,\ldots,X_n) = \\ &= \sum_{i=1}^{n}\log m_i + \sum_{j_1=1}^{m_1}\cdots\sum_{j_n=1}^{m_n}\widetilde{c}(u_{j_1},\ldots,u_{jn})\log_2\widetilde{c}(u_{j_1},\ldots,u_{jn}) \end{aligned} \tag{8}$$

Remark 7. If we suppose that for all $i = 1,\ldots,n$ the number of intervals m_i is the same for the two discussed cases then comparing formulas (7) and (8) we can see that in the case of partition $\mathcal{P}$ the information content does not depend on the first sum of formula (8) but only on the sample derivated copula.

4. The junction tree approach applied to the sample derivated copula.

We introduced the sample derivated copula as a discrete probability distribution with uniform marginals. We proved for this special copula in Theorem 3 that the consistency properties are fulfilled.

Let $V = \{1,\ldots,n\}$ be again a set of vertices. Let be defined an acyclic hypergraph over V. We denote by Γ and $\mathcal{S}$ the set of clusters and separators of the hypergraph which determine a junction tree $\mathcal{J}$. The marginal probability distributions associated to the clusters $K = \{i_1,\ldots,i_t\} \in \Gamma$ are denoted by $\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right) = \widetilde{c}_{i_1,\ldots,i_t}\left(\widetilde{U}_{i_1},\ldots,\widetilde{U}_{i_t}\right)$. The marginal probability distributions associated to the separators are denoted in the same way by $\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)$. The joint discrete copula is shortly denoted by $\widetilde{c}\left(\widetilde{\mathbf{U}}\right)$ and the univariate marginals by $\widetilde{c}_i\left(\widetilde{U}_i\right), i = 1,\ldots,n$.

In this section we are going to use the following popular notation:

$$\sum_{\mathbf{u}} f\left(\widetilde{\mathbf{U}}\right) = \sum_{i_1=1}^{m_1}\ldots\sum_{i_n=1}^{m_n} f\left(\widetilde{U}_1 = u_{i_1}^1,\ldots,\widetilde{U}_n = u_{i_n}^n\right),$$

where $u_{i_k}^k, i_k = 1, \ldots, m_k$ are the possible values of the random variable $\widetilde{U}_k, k = 1, \ldots, n$ and f is an arbitrary n-dimensional function. This simplified notation is used for the products, too.

Definition 6. The junction tree distribution given by

$$\widetilde{c}_{\mathcal{J}}\left(\widetilde{\mathbf{U}}\right) = \frac{\prod\limits_{K\in\Gamma} \widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)}{\prod\limits_{S\in\mathcal{S}} \left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}}, \tag{9}$$

where v_S is the number of clusters connected by the separator S, is called copula junction tree distribution, or shortly junction tree copula.

The problem is finding the junction tree copula which fits to the sample derivated copula. The goodness of fitting will be quantified by the Kullback-Leibler divergence [4].

Theorem 4. *The Kullback-Leibler divergence between the approximation (9) and the sample derivated copula* $\widetilde{c}\left(\widetilde{\mathbf{U}}\right)$ *is given by the formula:*

$$\begin{aligned} &KL\left(\widetilde{c}_{\mathcal{J}}\left(\widetilde{\mathbf{U}}\right), \widetilde{c}\left(\widetilde{\mathbf{U}}\right)\right) = \\ &= -H\left(\widetilde{\mathbf{U}}\right) - \left[\sum_{K\in\Gamma} I\left(\widetilde{\mathbf{U}}_K\right) - \sum_{S\in\mathcal{S}} (v_S - 1)\, I\left(\widetilde{\mathbf{U}}_S\right)\right] + \sum_{i=1}^{n} \log_2 m_i. \end{aligned}$$

PROOF.

$$KL\left(\widetilde{c}_{\mathcal{J}}\left(\widetilde{\mathbf{U}}\right),\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\right)=$$

$$=\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\frac{\widetilde{c}\left(\widetilde{\mathbf{U}}\right)}{\widetilde{c}_{\Gamma}\left(\widetilde{\mathbf{U}}\right)}=\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\widetilde{c}\left(\widetilde{\mathbf{U}}\right)-\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\widetilde{c}_{\Gamma}\left(\widetilde{\mathbf{U}}\right)=$$

$$=-H\left(\widetilde{\mathbf{U}}\right)-\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\frac{\prod\limits_{K\in\Gamma}\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)}{\prod\limits_{S\in\mathcal{S}}\left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}}=-H\left(\widetilde{\mathbf{U}}\right)-$$

$$-\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\left[\log_2\prod_{K\in\Gamma}\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)-\log_2\prod_{S\in\mathcal{S}}\left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}\right]=$$

$$=-H\left(\widetilde{\mathbf{U}}\right)-\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\prod_{K\in\Gamma}\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)+$$

$$+\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\prod_{S\in\mathcal{S}}\left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}.$$

We add and substract the sum:

$$\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\prod_{K\in\Gamma}\prod_{i\in K}\widetilde{c}_i\left(\widetilde{U}_i\right). \tag{10}$$

It follows from the definition of the junction tree that $\bigcup\limits_{K\in\Gamma}K=V$, and each variable belongs once more in the clusters as in the separators. So by adding and substracting (10) we obtain the following:

$$KL\left(\widetilde{c}_{\mathcal{J}}\left(\widetilde{\mathbf{U}}\right),\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\right)=-H\left(\widetilde{\mathbf{U}}\right)-\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\frac{\prod\limits_{K\in\Gamma}\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)}{\prod\limits_{K\in\Gamma}\prod\limits_{i\in K}\widetilde{c}_i\left(\widetilde{U}_i\right)}+$$

$$+\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\frac{\prod\limits_{S\in\mathcal{S}}\left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}}{\prod\limits_{S\in\mathcal{S}}\left[\prod\limits_{i\in S}\widetilde{c}_i\left(\widetilde{U}_i\right)\right]^{v_S-1}}-\sum_{\mathbf{u}}\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\log_2\prod_{i=1}^{n}\widetilde{c}_i\left(\widetilde{U}_i\right)=$$

$$= -H\left(\widetilde{\mathbf{U}}\right) - \sum_{\mathbf{u}} \widetilde{c}\left(\widetilde{\mathbf{U}}\right) \sum_{K\in\Gamma} \log_2 \frac{\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)}{\prod\limits_{i\in K} \widetilde{c}_i\left(\widetilde{U}_i\right)} +$$

$$+ \sum_{\mathbf{u}} \widetilde{c}\left(\widetilde{\mathbf{U}}\right) \sum_{S\in\mathcal{S}} \log_2 \frac{\left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}}{\left[\prod\limits_{i\in S} \widetilde{c}_i\left(\widetilde{U}_i\right)\right]^{v_S-1}} - \sum_{\mathbf{u}} \widetilde{c}\left(\widetilde{\mathbf{U}}\right) \sum_{i=1}^{n} \log_2 \widetilde{c}_i\left(\widetilde{U}_i\right).$$

Since the sample derivated copula has the property that all $\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)$, $\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)$, $\widetilde{c}_i\left(\widetilde{U}_i\right)$ are consistent marginals of $\widetilde{c}\left(\widetilde{\mathbf{U}}\right)$ (see Theorem 3) we have the following relations:

- $\sum\limits_{\mathbf{u}} \widetilde{c}\left(\widetilde{\mathbf{U}}\right) \sum\limits_{K\in\Gamma} \log_2 \dfrac{\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)}{\prod\limits_{i\in K} \widetilde{c}_i\left(\widetilde{U}_i\right)} =$

$$= \sum_{K\in\Gamma} \sum_{\mathbf{u}_k} \widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right) \log_2 \frac{\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)}{\prod\limits_{i\in K} \widetilde{c}_i\left(\widetilde{U}_i\right)} = \sum_{K\in\Gamma} I\left(\widetilde{\mathbf{U}}_K\right);$$

- $\sum\limits_{\mathbf{u}} \widetilde{c}\left(\widetilde{\mathbf{U}}\right) \sum\limits_{S\in\mathcal{S}} \log_2 \dfrac{\left[\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)\right]^{v_S-1}}{\left[\prod\limits_{i\in S} \widetilde{c}_i\left(\widetilde{U}_i\right)\right]^{v_S-1}} =$

$$= \sum_{S\in\mathcal{S}} \sum_{\mathbf{u}_S} (v_S - 1)\, \widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right) \log_2 \frac{\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)}{\prod\limits_{i\in S} \widetilde{c}_i\left(\widetilde{U}_i\right)} = \sum_{S\in\mathcal{S}} (v_S - 1)\, I\left(\widetilde{\mathbf{U}}_S\right);$$

- $-\sum\limits_{\mathbf{u}} \widetilde{c}\left(\widetilde{\mathbf{U}}\right) \sum\limits_{i=1}^{n} \log_2 \widetilde{c}_i\left(\widetilde{U}_i\right) = \sum\limits_{i=1}^{n} H\left(\widetilde{U}_i\right) = \sum\limits_{i=1}^{n} \log_2 m_i;$

Here $I\left(\widetilde{\mathbf{U}}_K\right), I\left(\widetilde{\mathbf{U}}_S\right)$ are the information content of the probability distribution of the marginals $\widetilde{c}_K\left(\widetilde{\mathbf{U}}_K\right)$ and $\widetilde{c}_S\left(\widetilde{\mathbf{U}}_S\right)$ (see [4]).

By the substitution of these assertions we obtain:

$$KL\left(\widetilde{c}_{\mathcal{J}}\left(\widetilde{\mathbf{U}}\right),\widetilde{c}\left(\widetilde{\mathbf{U}}\right)\right)=$$
$$=-H\left(\widetilde{\mathbf{U}}\right)-\left[\sum_{K\in\Gamma}I\left(\widetilde{\mathbf{U}}_K\right)-\sum_{S\in\mathcal{S}}\left(v_S-1\right)I\left(\widetilde{\mathbf{U}}_S\right)\right]+\sum_{i=1}^{n}\log_2 m_i.$$

Remark 8. The difference $\sum_{i=1}^{n}\log_2 m_i - H\left(\widetilde{\mathbf{U}}\right)$ does not depend on the junction tree structure.

Definition 7. The difference

$$\sum_{K\in\Gamma}I\left(\widetilde{\mathbf{U}}_K\right)-\sum_{S\in\mathcal{S}}\left(v_S-1\right)I\left(\widetilde{\mathbf{U}}_S\right)$$

is called the weight of the junction tree copula.

It is easy to see that in order to find a better approximation using junction trees, wee have to find the junction tree having the largest weight.

Finding the best fitting k-width junction tree, (the largest cluster contains k elements) for $k>2$ is an NP-hard problem. For $k=2$ the problem is similar to the Chow-Liu approximation [2]. In this case it is possible to find the best fitting second order junction tree by Kruskal' or Prim' algorithm.

For $k\geq 3$ it can be successfully used a heuristic approach introduced by the authors in [7] and [14]. The idea is the fitting of a special kind of junction tree, called t-cherry junction tree.

5. Conclusions and possible applications

One of the advantages of the junction tree copula is that it reveals some of the conditional independences between the variables involved. This kind of dependence structure is not exploited by the copula function. Another advantage of the method is that a multivariate copula can be decomposed into some lower dimensional sample derivated copulas.

The sample derivated copula approach is useful in cases when nothing else is known about the probability distribution but an iid sample. If the uniform partition is applied the whole information content depends on the sample derivated copula.

The copula junction tree can be used in feature selection which is a key-question in many fields as finance, medicine and biostatistics.

We got very good numerical results in pattern recognition (see [15]). First we applied the uniform partition to discretize continuous random variables then constructed the t-cherry junction tree approximation. In this way we found the informative features and so reduced the dimension of the classifier.

References

[1] R.S. Calsaverini, R. Vicente, An information theoretic approach to statistical dependence: Copula information, arXiv:0911.4207v1, 2009.

[2] C.K. Chow, C.N. Liu, Approximating discrete probability distribution with dependence tree, IEEE Transactions on Informational Theory 14 (1968) 462–467.

[3] Gh. Constantin, I. Istratescu, Elements of Probabilistic Analysis, Kluwer Academic Publisher, Dordrecht, 1989.

[4] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Interscience, New York, 1991.

[5] P. Deheuvels, La fonction de dépendance empirique et ses propriétés – Un test non paramétrique d' independance, Académie Royale de Belgique – Bulletin de la Classe des Sciences – 5e Série 65 (1979) 274–292.

[6] S. Kirshner, Learning with tree-average densities and distributions, Advances in Neural Information Processing Systems (NIPS), 2007.

[7] E. Kovács, T. Szántai, On the approximation of discrete multivariate probability distribution using the new concept of t-cherry junction tree, Lecture Notes in Economics and Mathematical Systems, 633, Proceedings of the IFIP/IIASA/GAMM Workshop on Coping with Uncertainty, Robust Solutions, 2008, IIASA, Laxenburg, 39–56.

[8] S.L. Lauritzen, T. Speed, K. Vijayan, Decomposable graphs and hypergraphs, J. Aust. Math. Soc. A 36 (1984) 12–29.

[9] S.L. Lauritzen, D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society, Ser. B 50 (1988) No. 2 157–224.

[10] G. Mayor, J. Suner, J. Torrens, Copula-like operations on finite settings, IEEE Transactions on fuzzy systems 13 (2005) No. 4 468–477.

[11] R. Mesiar, Discrete copulas – what they are, In: Joint EUSFLAT-LFA 2005, Conference Proceedings (E. Montseny and P. Sobrevilla, eds.) Universitat Politecnica de Catalunya, Barcelona, 2005, 927-930.

[12] J. Ma, Z. Sun, Mutual information is copula entropy, arXiv: 0808.0845v1, 2008.

[13] R. Nelsen, An Introduction to Copulas, Springer, New York, 1999.

[14] Szántai, T. and E. Kovács, Hypergraphs as a mean of discovering the dependence structure of a discrete multivariate probability distribution, *Proc. Conference APplied mathematical programming and MODelling (APMOD), 2008*, Bratislava, 27-31 May 2008, *Annals of Operations Research*, to appear.

[15] T. Szántai, E. Kovács, Application of t-cherry junction trees in pattern recognition, Broad Research in Artificial Intelligence and Neuroscience (BRAIN), Special Issue on Complexity in Sciences and Artificial Intelligence, Eds. B. Iantovics, D. Radoiu, M. Marusteri and M. Dehmer, 2010, 40–45.

[16] R.E. Tarjan, M. Yannakakis, Simple linear time algorithms to test cordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs, SIAM J. Comput. 13 (1984) 566–579.

[17] N.N. Vorob'ev, Consistent families of measures and their extensions, Theory of Probability and Applications 7 (1962) 147–163.

[18] N.N. Vorob'ev, Markov measures and Markov extensions, Theory of Probability and Applications 8 (1963) 420–429.

[19] J. Yang, Y. Qi, Q.R.Wang, A class of multivariate copulas with bivariate Frechet marginal copulas, Insurance: Mathematics and Economics 45 (2009) 139–147.