International Journal of Automation and Computing 9(2), April 2012, 113-121
DOI: 10.1007/s11633-012-0624-6

Dependence Tree Structure Estimation via Copula

Jian Ma[1] Zeng-Qi Sun[1] Sheng Chen[2, 3] Hong-Hai Liu[4]

[1]Department of Computer Science, Tsinghua University, Beijing 100084, PRC
[2]Electronics and Computer Science, Faculty of Physical and Applied Sciences, University of Southampton, Southampton SO17 1BJ, UK
[3]Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[4]Intelligent Systems and Robotics Research Group, School of Creative Technologies, University of Portsmouth, Portsmouth PO1 2DJ, UK

Abstract: We propose an approach for dependence tree structure learning via copula. A nonparametric algorithm for copula estimation is presented. Then a Chow-Liu like method based on dependence measure via copula is proposed to estimate maximum spanning bivariate copula associated with bivariate dependence relations. The main advantage of the approach is that learning with empirical copula focuses on dependence relations among random variables, without the need to know the properties of individual variables as well as without the requirement to specify parametric family of entire underlying distribution for individual variables. Experiments on two real-application data sets show the effectiveness of the proposed method.

Keywords: Copula, empirical copula, dependence, tree structure learning, probability distribution.

1 Introduction

Dependence between random variables is of fundamental importance because it may imply essential statistical relations within real-world social, physical, or biological systems. A large amount of data sets have been collected from different fields, such as engineering, biology, social networks, finance and world-wide web. However, analysis of these data remains a challenge. Hence, dependence structure learning is one of the fundamental problems that is frequently encountered in all fields of science and engineering. The best established statistical methodology for dependence representation is graphical models or Bayesian networks[1−4]. A generic graphical model is difficult and expensive to obtain in many practical cases. A tree-structured graphical model, which approximates the true underlying dependence structure, considers only pairwise dependence. However, its representational simplicity, through bivariate dependence decomposition, reduces computational complexity and makes large-scale problem modelling and inferring tractable. Traditional methods on inferring graphical models involve maximum likelihood (ML), where parametric family of entire underlying distribution is specified, including marginal distributions of individual variables. Hypothesis selection on marginal distributions is central to the performance of structure learning to a large extent. But there is often a shortage of prior knowledge needed for such selection. Thus, it is of practical interest to find a method which can separate structure learning from parametric marginal specification.

Copula theory unifies the representation of multivariate dependence[5, 6], and it has found wide-ranging applications in finance, statistics, and machine learning[7−12]. The term "copula", coming from Latin, refers to the way that random variables relate to each other. According to Sklar theorem[13, 14], a multivariate distribution can be represented by its marginal distributions and a copula function, which represents dependence structure among random variables. Using copula, one can separate the marginal distributions from their joint density distribution and, therefore, study only statistical interrelations without knowing the properties of each variable. Based on empirical copula estimation, many dependence models can be further adopted and approximately inferred. An advantage of empirical copula is that it is a model-free non-parametric estimation of the underlying true copula. In particular, dependence tree structures can be identified as a special case of copula, concerning only pairwise dependence, known as bivariate copula. In this contribution, we propose inferring a dependence tree structure by Chow-Liu like algorithm[15] based on empirical copula estimation. The advantage of the proposed approach is that structure learning is free from the requirement to specify parametric family of entire underlying distribution for individual variables, and there is no need to know the properties of individual variables. Copula estimation is also believed to be robust to outliers.

Manuscript received July 14, 2010; revised May 30, 2011

2 Copula and copula space

A copula function is a distribution function on the cumulative distribution function (CDF) transforms of marginal distributions[5, 6]. The relation between joint CDF, marginal CDFs, and copula is stated in the following theorem.

Sklar's Theorem[13, 14]**.** Given a random vector $X = [X_1\ X_2 \cdots X_N]^{\mathrm{T}}$, where T denotes the transpose operator, its CDF $F(x)$ can be represented as

$$F(x) = C(u_1, u_2, \cdots, u_N) \tag{1}$$

where $x = [x_1\ x_2 \cdots x_N]^{\mathrm{T}} \in \mathbf{R}^N$, $\{u_i = F_i(x_i), 1 \leqslant i \leqslant N\}$ with $u_i \in I \triangleq [0,\ 1]$ are marginal CDFs of X, and C is a copula function. If $\{F_i\}$ are continuous, then C is unique.

Intuitively, an N-dimensional copula $C : I^N \rightarrow I$ can be viewed as a new CDF stretched onto $u = [u_1\ u_2 \cdots u_N]^{\mathrm{T}} \in I^N$ from the CDF of X. By applying derivative on (1), we can also represent probability density function (PDF) via copula, leading to the following definition of copula density. An N-dimensional copula density c corresponding to an N-

copula C is defined as

$$c(u) = \frac{\partial^N}{\partial u_1\, \partial u_2 \cdots \partial u_N} C(u) \tag{2}$$

where $u \in I^N$. With this definition of copula density, we have the following corollary of Sklar′s Theorem.

Corollary 1. The PDF $p(x)$ of X can be represented as

$$p(\mathbf{x}) = c(u) \prod_{i=1}^{N} p_i(x_i) \tag{3}$$

where $\{p_i(x_i), 1 \leqslant i \leqslant N\}$ are the marginal PDFs of X and c is a copula density.

The significance of Sklar′s Theorem is as follows. Learning a multivariate distribution is a highly complicated task but modelling univariate marginals is often straightforward. Once the univariate marginals are learnt, the required multivariate distribution is readily modelled using copulas. How to construct a multivariate copula is of importance in applications. In many cases we cannot write down an analytic copula. However, the set of all copula functions is a convex set enclosed by some minimal copula and maximum copula[6]. Thus, the convex combination of copulas (copula densities) is also a copula (copula density). For example, let $\{c_k(u)\}_{k=1}^K$ be a set of K copula densities. Then the mixture of $\{c_k(u)\}_{k=1}^K$

$$c(u) = \sum_{k=1}^{K} w_k c_k(u) \tag{4}$$

where $w_i \geqslant 0$, $1 \leqslant i \leqslant K$, and $\sum_{i=1}^K w_i = 1$, is also a copula density. Thus, mixture of copulas provides a flexible way of constructing multivariate copula representations. In this contribution, we are concerned with dependence tree structures. Under the tree-structured dependence, a copula density can be decomposed as the product of bivariate copulas[11], i.e.,

$$c(u) = \prod_{i,j \in \{1,2,\cdots,N\}} c(u_i, u_j) \tag{5}$$

where $c(u_i, u_j)$ denotes a bivariate copula density related to (X_i, X_j).

3 Empirical copula estimation

Many parametric inference methods for copula can be summarised as follows: starting with a parametric family of copula, either implicitly implied by PDF or explicitly specified, optimise the parameters under the ML framework. Using a nonparametric method will help to avoid the difficulty of choosing parametric model family when no a priori knowledge is available. Some works on estimation of copulas can be found in [16–18]. Empirical copula (copula density) was introduced in [19, 20], which approximates the copula (copula density) from samples based on order statistics[21].

Consider an independently identically distributed (i.i.d.) sample set $\{x^t = [x_1^t\ x_2^t \cdots x_N^t]^{\mathrm{T}} \in \mathbf{R}^N, t \in \{1, 2, \cdots, T\}\}$. Let $\{x_n^{(t)}\}$ be the order statistics of $\{x_n^t\}$ with the corresponding ranks[1] $1 \leqslant r_n^t \leqslant T$ so that $x_n^{\left(r_n^t\right)} = x_n^t$. An empirical copula $\hat{C}$ of the samples $\{x^t, 1 \leqslant t \leqslant T\}$ is defined on a $(T+1)$ lattice

$$L = \left\{ \left(\frac{t_1}{T}, \cdots, \frac{t_N}{T} \right) : t_n \in \{0, 1, \cdots, T\}, 1 \leqslant n \leqslant N \right\} \tag{6}$$

as follows

$$\hat{C}\left(\frac{t_1}{T}, \cdots, \frac{t_N}{T} \right) = \frac{1}{T} \sum_{t=1}^{T} \prod_{n=1}^{N} \mathcal{I}_{[r_n^t \leqslant t_n]} \tag{7}$$

where the indicator function

$$\mathcal{I}_{[r_n^t \leqslant t_n]} = \begin{cases} 1, & \text{if } r_n^t \leqslant t_n, \\ 0, & \text{otherwise.} \end{cases}$$

Using forward difference on lattice, an empirical copula density is derived in a same way as

$$\hat{c}\left(\frac{t_1}{T}, \cdots, \frac{t_N}{T} \right) = \sum_{i_1=1}^{2} \cdots \sum_{i_N=1}^{2} (-1)^{\sum_{n=1}^{N} i_n} \times \hat{C}\left(\frac{t_1 - i_1 + 1}{T}, \cdots, \frac{t_N - i_N + 1}{T} \right). \tag{8}$$

Based on (7), we have the estimation algorithm.

Algorithm 1, for empirical copula given a set of samples $\{\boldsymbol{x}^t = [x_1^t\ x_1^t \cdots x_N^t]^{\mathrm{T}}, t \in \{1, 2, \cdots, T\}\}$. According to (8), we have the estimation algorithm, Algorithm 2, for empirical copula density, which is just an accumulative process based on Algorithm 1. Algorithm 1 has a linear complexity $\mathrm{O}(TN)$ while Algorithm 2 has an exponential complexity $\mathrm{O}(TN \times 2^N)$. Using empirical copula when estimating dependence structure has many advantages. Firstly, with nonparametric empirical copula algorithm, we can estimate different dependence relations from data in a model-free way. Secondly, copulas are invariant under monotonically increasing transformation and, therefore, we do not have to normalise data during analysis. Thirdly, copulas are insensitive to outliers.

Algorithm 1. Empirical copula function $\hat{C}$.

Input: data $\{x_n^t\}$ of dimension N and size T; $u \in I^N$
for $n = 1$ to N do
  $r_n^t = \mathrm{rank}(x_n^t)$
end for
$m = 0$, $u_n = u_n \times T$
for $t = 1$ to T do
  Initialise $n = 1$
  while $r_n^t \leqslant u_n$ do
    $n = n + 1$
  end while
  if $n = N + 1$ then
    $m = m + 1$
  end if
end for
Output: $\hat{C}(u) = m/T$.

[1] Order statistics and rank statistics are basic and standard tools in nonparametric statistics and inference[21]. For example, consider $T = 4$ and $\{x_n^1 = 6.1, x_n^2 = 9.3, x_n^3 = 3.4, x_n^4 = 8.6\}$. The order statistics are $\{x_n^{(1)} = 3.4, x_n^{(2)} = 6.1, x_n^{(3)} = 8.6, x_n^{(4)} = 9.3\}$, and the rank statistics or rankings are $\{r_n^1 = 2, r_n^2 = 4, r_n^3 = 1, r_n^4 = 3\}$.

Algorithm 2. Empirical copula density function $\hat{c}$.

Input: data $\{x_n^t\}$ of dimension N and size T; $u \in I^N$
use Algorithm 1 to produce $\hat{C}(u)$
$\hat{c}(u) = 0$
for all $t = [t_1\ t_2 \cdots t_N]^{\mathrm{T}} \in \{1, 2\}^N$ do

$$\hat{c}(u) = \hat{c}(u) + (-1)^{\sum_{i=1}^{N} t_i} \hat{C}(u - \frac{t-1}{T})$$

end for
Output: $\hat{c}(u)$.

4 Learning dependence tree structures

In a dependence tree structure, the dependence relation represented by edges in the graph is equivalent to a product of a group of bivariate copulas. We propose inferring such a product of bivariate copulas from data by Chow-Liu type algorithm[15] based on empirical copula estimation.

4.1 Maximum spanning bivariate copula problem

Suppose that we want to approximate dependence relations with a structure $\mathcal{T}(a)$, where a is the parameter vector that specifies $\mathcal{T}$. Given a set of i.i.d. samples X generated from an N-dimensional random vector $\boldsymbol{X}$ having the PDF $p(x)$, a cost function $\mathcal{F}$ can be defined on X, which can be minimised with respect to a to infer $\mathcal{T}$

$$\min_a \mathcal{F}(a; X). \tag{9}$$

In many works, the objective function $\mathcal{F}$ is defined through the ML principle, which requires parametric assumptions on the multivariate density function $p(x)$. Now consider an N-dimensional copula density c of $\boldsymbol{X}$. We can obtain its empirical estimation $\hat{c}$ based on X, which contains all the dependence information in the data. Notice that in doing so we do not need to specify a parametric form for $p(x)$ or $c(u)$.

In particular, when $\mathcal{T}$ has a tree structure, the dependence relations of $\mathcal{T}$ can be covered by a product of bivariate copulas. An N-copula c decomposed in the product form of bivariate copulas consists of a product of the $N(N-1)$ bivariate copulas. We can further approximate this product of the $N(N-1)$ bivariate copulas with a maximum spanning tree (MST) consisting of a product of only $N-1$ bivariate copulas. We refer to such a representation as the maximum spanning bivariate copula (MSBC), which approximates the dependence tree structure among the N random variables. In such an MSBC problem, the objective function $\mathcal{F}$ can be defined as the sum of dependence measurements on the $N-1$ bivariate copulas. Thus we transform the structure learning into a fitting problem.

4.2 Dependence measures

Copula summarises all the dependence relations. Hence it naturally links with the existing dependence measures in statistics. It has been shown that the statistical measures of dependence, such as Kendall′s tau, Spearman′s rho or Gini′s gamma, can be calculated from copula function[6]. Using empirical copula (copula density) to approximate copula (copula density), we can calculate these measures approximately. For example, an estimation of Spearman′s rho or the correlation between two random variables based on empirical bivariate copula is given by

$$\rho(X_1, X_2) = \frac{12}{T^2 - 1} \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} \left(\hat{C}\left(\frac{t_1}{T}, \frac{t_2}{T}\right) - \frac{t_1 t_2}{T^2} \right) \tag{10}$$

where T is the order of the lattice (6).

Alternatively, mutual information (MI) is a natural dependence measure based on information theory[22]. The MI of the two random variables, X_1 and X_2, is given as

$$I(X_1, X_2) = \int_{x_1, x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p_1(x_1) p_2(x_2)} \mathrm{d}x_1 \mathrm{d}x_2 \tag{11}$$

where $p(x_1, x_2)$ is the joint PDF of X_1 and X_2, while $p_1(x_1)$ and $p_2(x_2)$ are the two marginal PDFs of X_1 and X_2, respectively. Note that copula density is actually a density on I^N. Let $c(u_1, u_2)$ be a copula density of X_1 and X_2, with $u_1 = F_1(x_1)$ and $u_2 = F_2(x_2)$ denoting the marginal CDFs of X_1 and X_2, respectively. According to Corollary 1 of (3), the MI (11) can be transformed into a copula density representation

$$I(X_1, X_2) = \int_{x_1, x_2} p_1(x_1) p_2(x_2) c(u_1, u_1) \log c(u_1, u_2) \mathrm{d}x_1 \mathrm{d}x_2. \tag{12}$$

Given a set of data samples $\{x_1^t, x_2^t\}_{t=1}^T$, we can estimate the MI (12) according to

$$\hat{I}(X_1, X_2) = \sum_{t=1}^{T} \hat{p}_1(x_1^t) \hat{p}_2(x_2^t) \hat{c}(\hat{u}_1^t, \hat{u}_2^t) \log \hat{c}(\hat{u}_1^t, \hat{u}_2^t) \tag{13}$$

where $\hat{c}(\hat{u}_1^t, \hat{u}_2^t)$ is an empirical copula density estimate of $c(\hat{u}_1^t, \hat{u}_2^t)$, $\hat{p}_i(x_i^t)$ for $1 \leqslant i \leqslant 2$ are the estimates of $p_i(x_i^t)$ for $1 \leqslant i \leqslant 2$, and $\hat{u}_i^t = \hat{F}_i(x_i^t)$ for $1 \leqslant i \leqslant 2$ are the estimates of $u_i^t = F_i(x_i^t)$ for $1 \leqslant i \leqslant 2$.

In (13), besides empirically estimated copula density, univariate marginal densities and CDFs are also estimated, for which there are many well-established methods. Univariate density estimation can be obtained, for example, using the naive Bayesian estimator, k-nearest neighbour estimator and kernel density estimator[23−25]. We adopt the univariate Gaussian kernel density estimator due to its simplicity in density estimation. The well-known univariate empirical distribution function

$$\hat{F}_i(x_i) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{I}_{[x_i^t \leqslant x_i]} \tag{14}$$

can be used to estimate the marginal CDF $F_i(x_i)$. According to Glivenko-Cantelli theorem[26], the empirical distribution function (14) converges to the true CDF almost surely as the number of observations $T \to \infty$, under the assumption of i.i.d. observations.

The correlation dependence measure (10) is based on second-order statistics and imposes a very low computational complexity, while the MI dependence measure (13) is based on higher-order statistics and imposes a much higher complexity than the measure (10). If the underlying probability distribution of the problem is Gaussian, then the correlation based dependence measure (10) is sufficient. In general, however, we believe that the MI based dependence

measure (13) may be better, simply because it can cope naturally with non-Gaussian distributions.

4.3 Construction algorithm for MSBC tree

Firstly, we approximate copula (copula density) based on samples in the form of product of bivariate copulas (copula densities). From the resulting dependence measure matrix $M_X = \{\rho_{i,j} = \rho(X_i, X_j), 1 \leqslant i < j \leqslant N\}$ or $M_X = \{\hat{I}_{i,j} = \hat{I}(X_i, X_j), 1 \leqslant i < j \leqslant N\}$, a complete graph $\mathcal{G}$ of $N(N-1)$ edges on the N random variables $\boldsymbol{X} = [X_1\ X_2 \cdots X_N]^{\mathrm{T}}$ is built where the weight of each edge indicates the degree of dependence between two variables. Constructing an MSBC tree is equal to finding an MST $\mathcal{T}$ of $\mathcal{G}$. According to graph theory, an MST can be constructed by some well-established algorithms in polynomial time, such as Kruskal′s algorithm[27] and Prim′s algorithm[28]. We adopt Prim′s algorithm in our approach.

Prim′s algorithm starts with a vertex set V containing the two vertices with the maximum weight edge. At each stage, a vertex $u \notin V$ is chosen to add into V, which connects with a vertex $v \in V$, does not contribute to looping in the resulting new V, and leads to the pair (u, v) having the maximum possible weight edge. The procedure is repeated until all the N vertices are included in V with the resulting MST $\mathcal{T}$ of $N-1$ edges. A similar problem has been studied by Chow and Liu[15], where Chow-Liu algorithm approximates the density with tree structure by constructing an MST with the MI as edge weight. We summarise our approach for constructing MSBC tree in Algorithm 3.

Algorithm 3. Estimating dependence tree structure via copula

Input: data $\{x_n^t\}$ of dimension N and size T; $u \in I^N$

Construct empirical copula $\hat{C}$ by Algorithm 1 or empirical copula density $\hat{c}$ by Algorithm 2.

Calculate the dependence measure matrix M_X according to (10) or (13).

Build the dependence tree $\mathcal{T}$ by Prim′s algorithm based on M_X.

Output: maximum spanning bivariate copula tree T.

5 Experimental results

We applied the proposed method to a simulated data set and the two real data sets, Abalone and Boston Housing[29], to study their inner dependency structures.

5.1 Simulated data

We generated a data set of 1000 samples from a 5-dimensional distribution of a random vector $[\mathrm{G1\ G2\ G3\ Gn\ Ge}]^{\mathrm{T}}$, of which the first three elements, G1, G2 and G3, were zero-mean Gaussian distributed and the other two, Gn and Ge, were governed by a Gaussian copula with the marginal Gaussian and exponential distributions, respectively. The data set was so designed such that G1 to G3 are related (correlated), while Gn and Ge are related. In other words, there exists a dependence relationship between G1, G2 and G3, while there exists a relationship between Gn and Ge.

Because the underlying distribution was governed by the Gaussian random variables and Gaussian copula, it was sufficient to use the correlation (10) as the dependence measure. The Algorithm 1 was first applied to the data set to estimate the empirical copula, and the scatter plot of the estimated empirical copula samples is given in Fig. 1. The scatter plot is obviously symmetric, as the scatter plot of G1 and Ge, for example, is the same as the scatter plot of Ge and G1. Of particular interest are the scatter plots in the last row of Fig. 1, which are the scatter plots of the marginal Gaussian distributed variables with the marginal exponential distributed variable. The "non-Gaussian" nature of the joint distributions is evident in the last row of Fig. 1. The Algorithm 3 was then run on the data set to derive an approximate dependence tree as illustrated in Fig. 2. As expected, a sub-graph exists between the three Gaussian variables, G1, G2 and G3, while the two copula variables, Gn and Ge, are grouped in another sub-graph. The link or dependence between the two sub-graphs is very week indeed.

5.2 Abalone

Abalone data set[29] was built to predict the age (rings) of abalone based on physical measurements of abalone body, such as weight and height. It consists of 4177 samples with 9 attributes, as listed in Table 1. The data is complete and has both continuous and discrete attributes. The problem is a regression task where some measurements are possibly intrinsically interrelated. Instead of predicting the age based on the other 8 attributes, we focus on the dependence relations among the 9 attributes, which may benefit the prediction task. There are a few outliers in the data set. In other moment-based dependence analysis, these outliers

Table 1 Attributes of Abalone data set

Abbreviation	Detailed description	Numerical type	Unit
S	Sex: M, F, and I (infant)	Nominal	–
L	Length: longest shell measurement	Continuous	mm
D	Diameter: perpendicular to length	Continuous	mm
H	Height: with meat in shell	Continuous	mm
ww	Whole weight: whole abalone	Continuous	g
sw	Shucked weight: weight of meat	Continuous	g
vw	Viscera weight: gut weight (after bleeding)	Continuous	g
shw	Shell weight: after being dried	Continuous	g
r	Rings: +1.5 gives the age in years	Integer	–

are usually eliminated by pre-processing step. Otherwise they may cause large deviation in dependence measure calculation. But this is unnecessary for our approach because copula estimation is less susceptible to outliers. When estimating empirical copula in the experiment, we set the order of lattice with different sizes empirically considering a trade-off between approximation accuracy and computational cost. We chose the correlation (10) as well as the MI (13) as dependency measures. During the MI estimation, the univariate Gaussian kernel density estimator with well-tuned width parameter and the univariate empirical CDF estimator were applied on different moderate sized subsets randomly sampled from the whole data set. The estimation values varied a little. With either the correlation or the MI as weights, many MSBC trees were built using Algorithm 3 in our experiments, and they varied a little depending on the experimental setups.

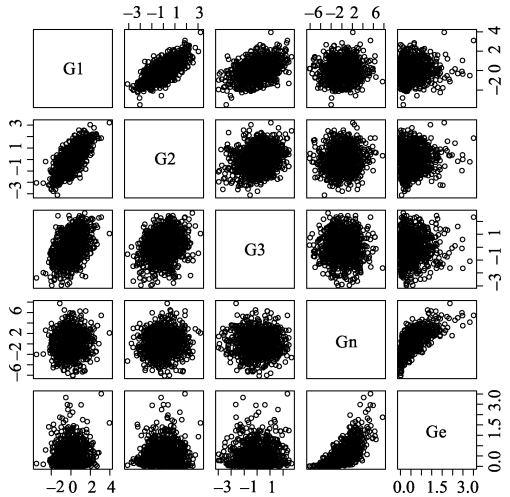


Fig. 1 Scatter plot of the estimated empirical copula samples for the simulated data set

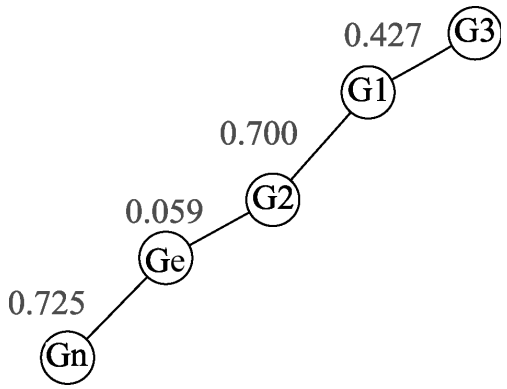


Fig. 2 A maximum spanning bivariate copula tree with correlation as edge weights generated from the simulated data set

For an illustration purpose, four attributes, "L", "H", "sw" and "vw", of the original data set are plotted in Fig. 3. As a comparison, the estimated empirical copulas of these four attributes are shown in Fig. 4. During empirical copula estimation, the effect of outliers diminishes, as can be seen clearly by comparing Fig. 3 with Fig. 4. In addition to robustness to outliers, we emphasise another fact that copula measures dependence relations and in doing so it does not need to consider individual properties of the variables. It can be observed from Fig. 3 that all the attributes possess very different individual properties as demonstrated in their very different pairwise non-Gaussian joint distributions. While in fact Fig. 4 shows that all the pairwise estimated copulas seem to have a very similar dependency structure.

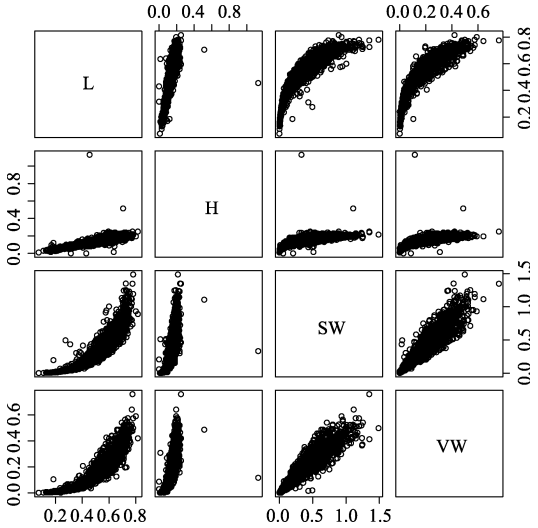


Fig. 3 Scatter plot of the original data samples of four attributes, "L", "H", "sw", and "vw", in Abalone data set

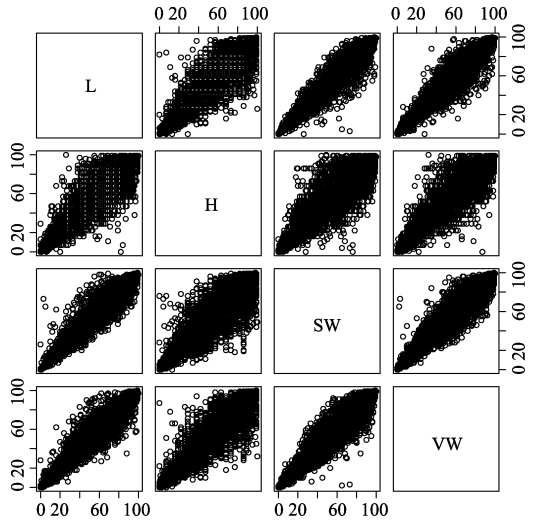


Fig. 4 Scatter plot of the estimated empirical copulas of four attributes, "L", "H", "sw", and "vw", in Abalone data set

Fig. 5 shows a typical MST constructed for Abalone data set, where edge weights were the corresponding correlation measures, while a typical MST constructed using the MI as edge weights is depicted in Fig. 6. Except for "sex" and "rings", the other seven attributes were linked with relatively strong weighted edges, as is seen from Figs. 5 and 6. It can in fact be learnt from all the MSBC trees constructed in our experiments that the edges linked these seven physical measurements form the backbone of all the estimated

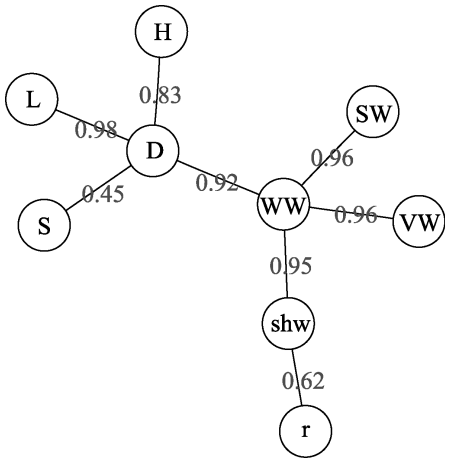


Fig. 5 A maximum spanning bivariate copula tree with correlation as edge weights generated from Abalone data set

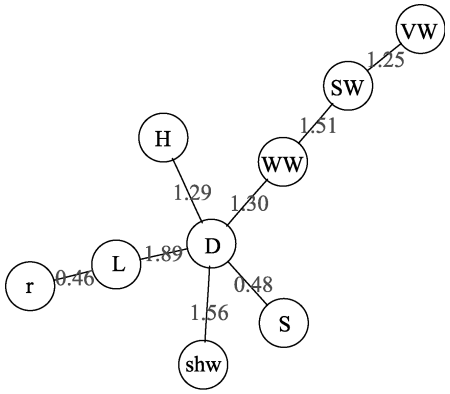


Fig. 6 A maximum spanning bivariate copula tree with mutual information as edge weights generated from Abalone data set

trees whilst the nodes for "sex" and "rings" are leaves randomly attached to the seven-node backbone. This can be interpreted as the reflection of the true abalone′s body growth process. Thus, "rings" and "sex" are not strongly related with the other seven physical attributes. Based on this observation, we argue that the original experimental design of predicting "rings" with the other attributes in abalone data set may not be a good one[2].

5.3 Housing

The Boston house price data set was from 1970 census, first published in [30]. It contains 506 samples with 14 mixed-type attributes, including 13 continuous attributes and 1 binary one, as listed in Table 2. Previous research mainly treated the problem as a regression task with the aim to predict "Medv" based on the other 13 attributes. In our experiment, we studied the dependence structure instead. Using copula to estimate dependence relations and to generate an MSBC tree, we hoped to find some previously unnoticed relations between the attributes. Such knowledge is extremely valuable in various social science policy studies, such as housing development and city planning.

In many previous studies, researchers proposed to transform the data into a suitably scaled one before further dependence analysis, by applying monotonically increasing functions, such as normalisation, nonlinear exponential or log functions. In our experiment, this was unnecessary due to copula′s invariance to such kinds of transformation. As with the previous abalone experiment, the MSBC algorithm, Algorithm 3, was run on the moderate data sub-sets randomly sampled from the whole Housing data set. Many MSBC trees were generated, and two of them are plotted in Figs. 7 and 8, respectively. Experimental results obtained indicate that only two links, "crim-rad" and "medv-lstat", appeared in all the estimated trees with relatively strong edge weights, indicating that these are the two strongest

Table 2 Attributes of Boston Housing data set

Abbreviation	Detailed description
crim	Per capita crime rate in town
zn	Proportion of residential land zoned for lots over 25 000 sq.ft.
indus	Proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	Nitric oxides concentration (parts per 10 million)
rm	Average number of rooms per dwelling
age	Proportion of owner-occupied units built prior to 1940
dis	Weighted distances to five Boston employment centres
rad	Index of accessibility to radial highways
tax	Full-value property-tax rate per USD 10 000
ptratio	Pupil-teacher ratio in town
b	$1000(B-0.63)^2$ where B is the proportion of blacks in town
lstat	Percentage lower status of the population
medv	Median value of owner-occupied homes in USD 1000′s

[2]From the viewpoint of experimental design, the dependent variable should be strongly linked or related to the input variables in order to build a meaningful or accurate prediction relationship or model. From our dependence analysis, we can see that the attribute "rings" has weak dependence relationships to the other 8 attributions. Therefore, it can be argued that the original design of predicting the age of abalone based on the 8 chosen physical measurements of abalone body was not well thought through, as it may not reflect accurately the true abalone′s body growth process.

interconnections. We also observed that there are two groups of attributes, one includes "nox, dis, indus, crim" and the other includes "medv, lstat, age, ptratio", which were interconnected and appeared in many constructed MSBC trees.

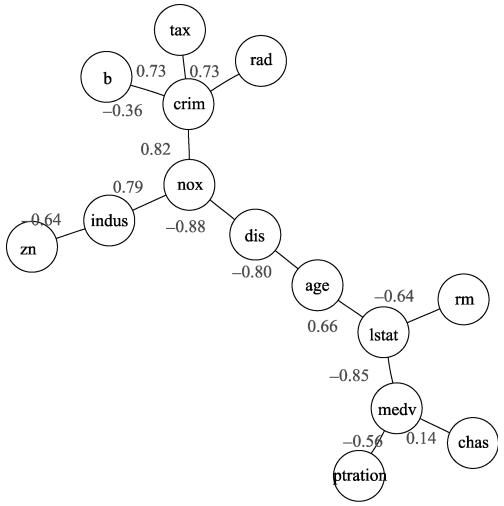


Fig. 7 A maximum spanning bivariate copula tree with mutual information as edge weights generated from Abalone data set

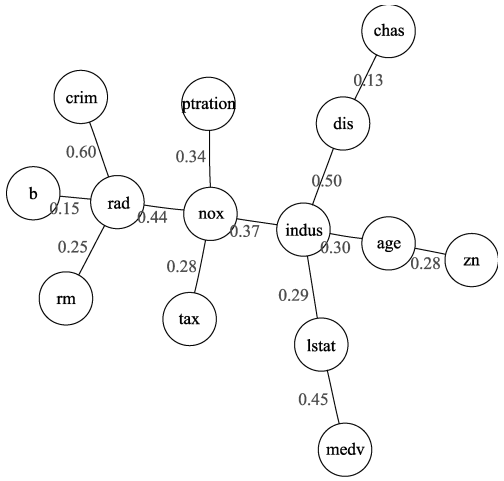


Fig. 8 A maximum spanning bivariate copula tree with mutual information as edge weights generated from Housing data set

5.4 Discussions

In our approach, the goal of structure learning based on copula is to construct an MST that spans the structure by maximising the total edge weights of the constructed tree based on certain dependence measure. In this dependence representation, only bivariate dependence relations are considered. Given the graph containing the $N(N-1)$ pairs of bivariate dependence relations for the N random variables, only $N-1$ relations are chosen to form an MST approximation. To examine the accuracy of this approximation, we can consider the ratio of the total edge weights of the MST to the sum of all the $N(N-1)$ weights. The values of this ratio over a number of experiments are plotted in Fig. 9 for both data sets. For Abalone data set, the MST contains 11 % of the total edges but it explains on average 20 % of the total edge weights. For Housing data set, the MST contains only 7 % of the total edges and yet it explains on average over 40 % of the total edge weights. This demonstrates the effectiveness of MSBC tree approximation.

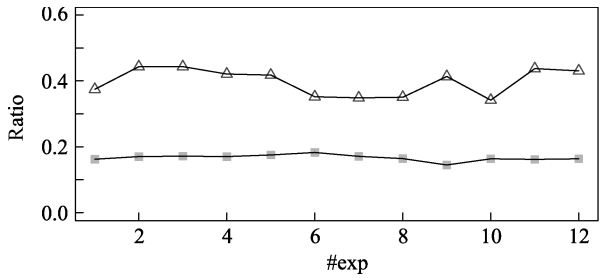


Fig. 9 Ratio of the total weights of estimated MSBC tree to the total weights of all the bivariate dependence relations over a number of experimental runs. Rectangles are for Abalone data set while triangles for Housing data set. #exp denotes number of experimental runs

6 Conclusions

We have proposed to estimate dependence tree structures using the copula method. Copula provides a general framework for representing dependence relations among random variables and it makes no assumption on the underlying distribution. A nonparametric estimation algorithm for empirical copula offers great flexibility in structure learning because the estimated empirical copula contains all the dependence information in the data. In particular, we have studied the learning process of the dependence structure that is represented by the bivariate dependence relations of the N random variables. Such a graph contains a total of $N(N-1)$ edges. A Chow-Liu like method based on empirical copula has been proposed to construct a maximum spanning tree with the strongest $N-1$ bivariate dependence relations. The effectiveness of this maximum spanning bivariate copula tree approximation has been demonstrated using one simulated data set as well as two real data sets.

References

[1] D. Heckerman, D. Geiger, D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[2] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 2, pp. 195–210, 1996.

[3] M. I. Jordan. *Learning in Graphical Models*, Holland: Kluwer Academic Publishers, 1998.

[4] C. M. Bishhop. A new framework for machine learning. *Computational Intelligence: Research Frontiers*, J. M. Zu-

rada, G. G. Yen, J. Wang, Eds., Heidelberg, Germany: Springer, pp. 1–24, 2008.

[5] H. Joe. *Multivariate Models and Dependence Concepts*, London, UK: Chapmann & Hall, 1997.

[6] R. B. Nelsen. *An Introduction to Copulas*. New York, USA: Springer, 1999.

[7] E. Bouyé, V. Durrleman, A. Nikeghbali, G. Riboulet, T. Roncalli. Copulas for finance — A reading guide and some applications, [Online], Available: http://ssrn.com/abstract=1032533, October 27, 2011.

[8] S. X. Chen, T. M. Huang. Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics*, vol. 35, no. 2, pp. 265–282, 2007.

[9] J. Ma, Z. Sun. Copula component analysis. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, ACM, London, UK, pp. 73–80, 2007.

[10] K. Abayomi, U. Lall, V. de la Pena. Copula based independent component analysis, [Online], Available: http://ssrn.com/abstract=1028822, October 27, 2011.

[11] S. Kirshner. Learning with tree-averaged densities and distributions. *Advances in Neural Information Processing Systems*, J. C. Platt, D. Koller, Y. Singer, S. Roweis, Eds., Cambridge, USA: MIT Press, pp. 761–768, 2000.

[12] X. H. Chen, W. B. Wu, Y. P. Yi. Efficient estimation of copula-based semiparametric Markov models. *The Annals of Statistics*, vol. 37, no. 6B, pp. 4214–4253, 2009.

[13] A. Sklar. Fonctions de repartition à n dimensions et leurs marges. *Publications de l′Institut de Statistique de l′Université de Paris*, vol. 8, pp. 229–231, 1959. (In French)

[14] L. Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, vol. 139, no. 11, pp. 3921–3927, 2009.

[15] C. K. Chow, C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[16] J. D. Fermanian. Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, vol. 95, no. 1, pp. 119–152, 2005.

[17] J. Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, [Online], Available: http://www.jstatsoft.org/v21/i04/paper, October 27, 2011.

[18] X. Shen, Y. Zhu, L. Song. Linear B-spline copulas with applications to nonparametric estimation of copulas. *Computational Statistics and Data Analysis*, vol. 52, no. 7, pp. 3806–3819, 2008.

[19] P. Deheuvels. La fonction de dépendance empirique et ses propriétés — Un test non paramétrique d′indépendance. *Académie Royale de Belgique — Bulletin de la Classe des Sciences — 5e Série*, vol. 65, pp. 274–292, 1979. (In French)

[20] P. Deheuvels. A non parametric test for independence. *Publications de l′Institut de Statistique de l′Université de Paris*, vol. 26, pp. 29–50, 1981.

[21] H. A. David, H. N. Nagaraja. *Order Statistics*, the 3rd Edition, New York, USA: John Wiley & Sons, 2003.

[22] T. M. Cover, J. A. Thomas. *Elements of Information Theory*, New York, USA: John Wiley & Sons, 1991.

[23] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London, UK: Chapmann & Hall, 1986.

[24] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[25] S. Chen, X. Hong, C. J. Harris. An orthogonal forward regression technique for sparse kernel density estimation. *Neurocomputing*, vol. 71, no. 4–6, pp. 931–943, 2008.

[26] F. Topsøe. On the Glivenko-Cantelli theorem. *Probability Theory and Related Fields*, vol. 14, no. 3, pp. 239–250, 1970.

[27] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. In *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[28] R. C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, vol. 36, pp. 1389–1401, 1957.

[29] A. Asuncion, D. J. Newman. UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, [Online], Available: http://archive.ics.uci.edu/ml/datasets.html, October 28, 2011.

[30] D. Harrison, D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.

Jian Ma received his B. Sc. and M. Sc. degrees in computer science from Hangzhou Dianzi University, Hangzhou, PRC in 2000 and 2003, respectively, and his Ph. D. degree in computer science and technology from Tsinghua University, Beijing, PRC in 2009. He is currently a post-doctoral researcher with the Department of Automation, Tsinghua University.

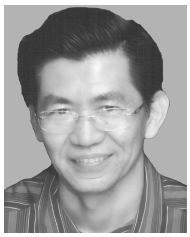
His research interests include machine learning, data analysis, and information theory.

E-mail: majian03@mails.tsinghua.edu.cn

Zeng-Qi Sun received his B. Sc. degree from the Department of Automatic Control, Tsinghua University, Beijing, PRC in 1966, and his Ph. D. degree in control engineering from Chalmers University of Technology, Gothenburg, Sweden in 1981. He is currently a full professor in the Department of Computer Science and Technology, Tsinghua University. He is the author and coauthor of over 200 research papers and eight books on computer control theory, intelligent control, and robotics.

His research interests include intelligent control, robotics, fuzzy systems, neural networks, and evolution computing.

E-mail: szq-dcs@tsinghua.edu.cn



Sheng Chen received his B. Eng. degree from Huadong Petroleum Institute, Dongying, PRC in January 1982, and his Ph. D. degree from the City University, London, UK in September 1986, both in control engineering. He was awarded the Doctor of Sciences (D. Sc.) degree by the University of Southampton, Southampton, UK in 2004. From October 1986 to August 1999, he held research and academic appointments at the University of Sheffield, the University of Edinburgh and the University of Portsmouth, all in UK. Since September 1999, he has been with the Electronics and Computer Science, the University of Southampton, UK, where he currently holds the post of professor of intelligent systems and signal processing. He is a Distinguished Adjunct Professor at King Abdulaziz University, Jeddah, Saudi Arabia. He has published over 450 research papers. He is a Chartered Engineer (CEng), a fellow of IET and a fellow of IEEE. In the database of the world′s most highly cited researchers, compiled by Institute for Scientific Information (ISI) of the USA, he is on the list of the highly cited researchers in the engineering category.

His research interests include wireless communications, adaptive signal processing for communications, machine learning, evolutionary computation methods, and intelligent control systems.

E-mail: sqc@ecs.soton.ac.uk (Corresponding author)

Hong-Hai Liu received his Ph. D. degree in robotics from Kings College, University of London, UK in 2003. He joined the University of Portsmouth, UK in September 2005, where he currently holds a post of professor of intelligent systems. He previously held research appointments at Universities of London and Aberdeen, UK, and project leader appointments in the industrial control and system integration industries. He has published over 200 research papers including three Best Paper Awards. He is a senior member of IEEE.

His research interests include computational intelligence methods and applications with a focus on those approaches which could make contributions to the intelligent connection of perception to action.

E-mail: honghai-liu@port.ac.uk