Astronomy & Astrophysics manuscript no. 38585corr
August 14, 2020

Modeling high-dimensional dependence in astronomical data

R. Vio[1], T.W. Nagler[2], and P. Andreani[3]

[1] Chip Computers Consulting s.r.l., Viale Don L. Sturzo 82, S.Liberale di Marcon, 30020 Venice, Italy
e-mail: robertovio@tin.it
[2] Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden
e-mail: t.w.nagler@math.leidenuniv.nl
[3] ESO, Karl Schwarzschild strasse 2, 85748 Garching, Germany
e-mail: pandrean@eso.org e-mail: pandrean@eso.org

Received....; accepted....

ABSTRACT

Fixing the relationship of a set of experimental quantities is a fundamental issue in many scientific disciplines. In the 2D case, the classical approach is to compute the linear correlation coefficient ρ from a scatterplot. This method, however, implicitly assumes a linear relationship between the variables. Such an assumption is not always correct. With the use of the partial correlation coefficients, an extension to the multidimensional case is possible. However, the problem of the assumed mutual linear relationship of the variables remains. A relatively recent approach that makes it possible to avoid this problem is the modeling of the joint probability density function (PDF) of the data with copulas. These are functions that contain all the information on the relationship between two random variables. Although in principle this approach also can work with multidimensional data, theoretical as well computational difficulties often limit its use to the 2D case. In this paper, we consider an approach based on so-called vine copulas, which overcomes this limitation and at the same time is amenable to a theoretical treatment and feasible from the computational point of view. We applied this method to published data on the near-IR and far-IR luminosities and atomic and molecular masses of the Herschel reference sample, a volume-limited sample in the nearby Universe. We determined the relationship of the luminosities and gas masses and show that the far-IR luminosity can be considered as the key parameter relating the other three quantities. Once removed from the 4D relation, the residual relation among the latter is negligible. This may be interpreted as the correlation between the gas masses and near-IR luminosity being driven by the far-IR luminosity, likely by the star formation activity of the galaxy.

Key words. Methods: data analysis – Methods: statistical

1. Introduction

Modeling the relationship of a set of experimental quantities is not straightforward. Often, no theoretical hints are available that would allow us to fix the dependence among the involved variables. Hence, the work has to be entirely based on the analysis of the data. In the 2D case, an example is represented by the scatterplots and the computation of the corresponding linear correlation coefficients ρ. Its extension to the multidimensional case is possible with the partial correlation coefficients. The main limitation of this approach is the implicit assumption of linear relationships among the variables under study. This is often an unrealistic condition. For this reason, a relatively recent alternative consists of modeling the joint probability distribution function (PDF) of the data. However, this task is not trivial even in the 2D case. Families of bivariate PDFs are available (Balakrishnan & Lai 2010), but are not very flexible and are difficult to use. Things worsen for the multidimensional case (Kotz et al. 2000). A relatively recent alternative is based on copulas. These are simply multivariate cumulative distribution functions (CDF) with standard uniform margins. They are used to describe the dependence between random variables, and their main role is to disentangle margins and the dependence structure (Nelsen 2006; Durante & Sempi 2016; Hofert et al. 2018). With copulas it is possible to decompose a joint probability distribution into their margins and a function that couples them. The copula is that coupling function.

In cosmology, 2D copulas have been used by Scherrer et al. (2010) for the determination of the PDF of the density field of the large-scale structure of the Universe, by Lin & Kilbinger (2015) and Lin et al. (2016) to predict weak-lensing peak counts, and by Sato et al. (2010, 2011) for the precise estimation of cosmological parameters. Other astronomical applications include the determination of the far-UV and far-IR bivariate luminosity function of galaxies (Takeuchi 2010; Takeuchi et al. 2011), the determination of the K-band and the submillimeter luminosity function (Andreani et al. 2014), and the bivariate luminosity versus the mass functions of the of the local HRS galaxy sample (Andreani et al. 2018).

In principle, the copula approach can work with multidimensional data, but theoretical as well computational difficulties often limit its use to the 2D case. Recently, however, vine copulas have been proposed in the statistical literature as an approach that overcomes this limitation and at the same time is amenable to a theoretical treatment and feasible from the computational point of view. The strength of vine copulas is that they allow, in addition to the separation of margins and dependence by the copula approach, tail asymmetries and separate multivariate component modeling. This is accommodated by constructing multivariate copulas using only bivariate building blocks, which can be selected independently. These building blocks are glued together to form valid multivariate copulas by appropriate conditioning (Joe 2015; Czado 2019). This makes vine copulas a very flex-

Article published by EDP Sciences, to be cited as https://doi.org/10.1051/0004-6361/202038585

ible and reliable tool even in the case of very high-dimensional data.

For this paper, we made use of multidimensional copulas, described in Sects. 2 and 3, and in particular of vine copulas, outlined in Sects. 4 and 5. We applied them to a data set related to a complete nearby sample of galaxies that has been observed at various wavelengths (Andreani et al. 2018, and references therein) and show its use to highlight the relation to the physical properties of the galaxies in Sect. 6.

2. What are copulas?

A d-dimensional copula $C_{1,\dots,d}(\boldsymbol{u})$, $\boldsymbol{u} = (u_1, \dots, u_d) \in [0,1]^d$ is simply a multivariate CDF with standard uniform univariate margins. Its importance is due to Sklar's theorem: for any d-dimensional CDF $F(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$, with univariate margins $F_1(x_1), \dots, F_d(x_d)$, a d-dimensional copula $C_{1,\dots,d}(\boldsymbol{u}) : [0,1]^d \to [0,1]$ exists, such that

$$F(\boldsymbol{x}) = C_{1,\dots,d}(F_1(x_1), \dots, F_d(x_d)) = C_{1,\dots,d}(u_1, \dots, u_d), \tag{1}$$

where $u_1 = F_1(x_1), \dots, u_d = F_d(x_d)$. The converse also holds, i.e. given a d-dimensional copula $C_{1,\dots,d}(\boldsymbol{u})$ and univariate CDFs $F_1(x_1), \dots, F_d(x_d)$, the CDF $F(\boldsymbol{x})$ defined by Eq. (1) is a d-dimensional CDF with margins $F_1(x_1), \dots, F_d(x_d)$. This means that copulas are those functions which combine the univariate margins $F_1(x_1), \dots, F_d(x_d)$ to form the d-dimensional CDF $F(\boldsymbol{x})$. In other words, copulas link multivariate CDFs to their univariate margins. The importance of copula is more evident if the PDFs $f(\boldsymbol{x})$ are considered. Indeed, it can be shown that

$$f(x_1, \dots, x_d) = c_{1,\dots,d}(F_1(x_1), \dots, F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d), \tag{2}$$

where

$$c_{1,\dots,d}(u_1, \dots, u_d) = \frac{\partial^d C_{1,\dots,d}(u_1, \dots, u_d)}{\partial u_1 \cdots \partial u_d}. \tag{3}$$

From Eq. (2), any joint PDF $f(\boldsymbol{x})$ can be factorized into the product of two terms. One is the product of the marginal PDFs $\{f_i(x_i)\}$ and the other is the copula density $c_{1,\dots,d}(\boldsymbol{u})$. The first term provides information on the statistical properties of the individual random variables $\{x_i\}$ whereas the second term provides information on their mutual dependence. Therefore, the importance of $c_{1,\dots,d}(\boldsymbol{u})$ lies in the fact that it describes the dependence structure among the random variables in separation of the associated marginal PDFs.

If a set of n d-dimensional random data $\{\boldsymbol{x}_k\}$, $k = 1, \dots, d$, with $\boldsymbol{x}_k = \{x_{p,k}\}$, $p = 1, \dots, n$, is available and the margins $\{F_k(x_k)\}$ and the corresponding PDFs $\{f_k(x_k)\}$ are known, the standard procedure to estimate $f(x_1, \dots, x_d)$ is as follows: first, compute the standard uniform variates $\boldsymbol{u}_k = F_k(\boldsymbol{x}_k)$, then fit their joint distribution by a copula $C_{1,\dots,d}(\boldsymbol{u}|\boldsymbol{\theta})$, which belongs to a continuous parametric family with characteristic parameters $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_{n_p}\}$. After that, Eq. (2) provides the joint PDF. A common method for fitting the copula is based on an estimate of the parameters $\boldsymbol{\theta}$ through a maximum likelihood method, but other techniques are also possible (Hofert et al. 2018). Often, however, the margins are not available. In this case, an alternative is to fit each set $\boldsymbol{x}_k$ with a PDF belonging to the Johnson, generalized Lambda, or any other family of parametric PDFs (Vio et al. 1994; Karian & Dudewicz 2011) (see also Appendix A), to compute the uniform random variates $\boldsymbol{u}_k = F_k(\boldsymbol{x}_k)$ and then, as before, to fit a copula. When the margins are not estimable with sufficient accuracy (e.g., because of little available data), a useful nonparametric variant is the computation of the random variates $\boldsymbol{u}_k$ by means of the so called pseudo-observations $u_{p,k} = R_{p,k}/(n+1)$ with $R_{p,k}$ the rank of $x_{p,k}$ among $(x_{1,k}, \dots x_{n,k})$. Since in general one has no indication of which kind of copula is suited for the data of interest, the typical solution is to fit a set of copulas and to choose that which provides the best result.

In principle, the above procedures can be applied to any d-dimensional data set. The point is that most of the parametric copula families available in literature are 2D (e.g., see Joe 2015), and the few available for a multidimensional analysis are not flexible enough. An alternative approach based on a nonparametric copula estimate has been also proposed (e.g., Nagler & Czado 2016; Nagler et al. 2017).

3. Preliminary considerations

Given that most of the available copula families are 2D, it is unclear how a d-dimensional PDF $f(x_1, \dots, x_d)$ can be computed. A possible solution is to express Eq. (2) in terms of 2D copulas. The starting point is that $f(x_1, \dots, x_d)$ can be factorized into the form

$$\begin{aligned} f(x_1, \dots, x_d) = f(x_d) \cdot f(x_{d-1}|x_d) \\ \cdot f(x_{d-2}|x_{d-1}, x_d) \cdots f(x_1|x_2, \dots, x_d), \end{aligned} \tag{4}$$

with $f(x_k|\boldsymbol{y})$ being the conditional PDF of the random variable x_k given the vector of random variables $\boldsymbol{y}$. Now, it can be proved (Czado 2019) that

$$f(x_k|\boldsymbol{y}) = c_{x_k y_j|\boldsymbol{y}_{-j}}(F(x_k|\boldsymbol{y}_{-j}), F(y_j|\boldsymbol{y}_{-j})|\boldsymbol{y}_{-j}) \cdot f(x_k|\boldsymbol{y}_{-j}), \tag{5}$$

where $c_{x_k y_j|\boldsymbol{y}_{-j}}(.,.)$ is the conditional copula density,

$$F(x_k|\boldsymbol{y}) = \frac{\partial C_{x_k, y_j|\boldsymbol{y}_{-j}}(F(x_k|\boldsymbol{y}_{-j}), F(y_j|\boldsymbol{y}_{-j})|\boldsymbol{y}_{-j})}{\partial F(y_j|\boldsymbol{y}_{-j})}, \tag{6}$$

$C_{x_k y_j|\boldsymbol{y}_{-j}}(.,.)$ is the conditional copula, y_j is one arbitrarily chosen component of $\boldsymbol{y}$, and $\boldsymbol{y}_{-j}$ denotes the y-vector, excluding this component. The key point is that these conditional PDFs are expressed in terms of 2D copula densities. The same holds for the PDF $f(x_1, \dots, x_d)$. For example, in the 3D case it is

$$\begin{aligned} f(x_1, x_2, x_3) = & f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \\ & \cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \\ & \cdot c_{13|2}(F(x_1|x_2), F(x_3|x_2)|x_2). \end{aligned} \tag{7}$$

In fact, the decomposition (4) is not unique since the indices of the variables $\{x_k\}$ can be permuted. For instance, a decomposition equivalent to (7) is

$$\begin{aligned} f(x_1, x_2, x_3) = & f_2(x_2) \cdot f_1(x_1) \cdot f_3(x_3) \\ & \cdot c_{21}(F_2(x_2), F_1(x_1)) \cdot c_{13}(F_1(x_1), F_3(x_3)) \\ & \cdot c_{23|1}(F(x_2|x_1), F(x_3|x_1)|x_1). \end{aligned} \tag{8}$$

Although the problem of estimating the PDF $f(x_1, \dots, x_d)$ has been simplified by means of Eqs. (4)-(6), it is still hard to deal with. The conditional copulas $C_{x_k y_j|\boldsymbol{y}_{-j}}$ and corresponding densities $c_{x_k y_j|\boldsymbol{y}_{-j}}$ are difficult to estimate. For this reason, usually the conditional copula densities are simplified into the form

$$c_{x_k y_j|\boldsymbol{y}_{-j}}(F(x_k|\boldsymbol{y}_{-j}), F(y_j|\boldsymbol{y}_{-j})|\boldsymbol{y}_{-j}) \approx c_{x_k y_j|\boldsymbol{y}_{-j}}(F(x|\boldsymbol{y}_{-j}), F(y_j|\boldsymbol{y}_{-j})). \tag{9}$$

Something similarly occurs to the corresponding conditional copulas. This simplification does not only make the problem easier to deal with, but it permits the use of the large set of available continuous parametric families of 2D copulas. This makes the method quite flexible. For instance, in the 3D case, the decomposition can be written in the form

$$\begin{aligned} f(x_1,x_2,x_3) =& f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3) \\ &\cdot c_{12}(F_1(x_1),F_2(x_2);\boldsymbol{\theta}_{12})\cdot c_{23}(F_2(x_2),F_3(x_3);\boldsymbol{\theta}_{23}) \\ &\cdot c_{13|2}(F(x_1|x_2),F(x_3|x_2);\boldsymbol{\theta}_{13|2}), \end{aligned} \tag{10}$$

where the 2D copula densities $c_{12}(.,.;\boldsymbol{\theta}_{12})$, $c_{23}(.,.;\boldsymbol{\theta}_{23})$ and $c_{13|2}(.,.|.;\boldsymbol{\theta}_{13|2})$ can be chosen of different types.

4. Vine copulas: The theory

For high-dimensional distributions, there is a huge number of possibilities for decompositions into 2D copulas, named pair-copulas, like Eqs. (7) and (8). All these possibilities can be organized according to graphical models called "regular vines". Two special cases, called D-vine and C-vine (Aas et al. 2009), have been introduced as a simplification. Each model gives a specific way of decomposing a density.

Figure 1 shows the graphical structure of a D-vine for a four-dimensional problem. This structure is formed by three levels or trees. Each circle or ellipsis constitutes a node, and each pair of nodes is joined by an edge. The label of a node in a given tree is given by the label of the edges of the tree at its immediate left. The label of an edge is given by the indices contained in the joined nodes with the conditional index given by the common one. For example, in the central tree node $(1,2)$ is connected to node $(2,3)$. The common index is 2, hence the label of the joining edge is $(1,3|2)$. Each edge represents a pair-copula density, and the edge label corresponds to the subscript of the pair-copula density. The indices of the CDFs that appear as the argument of a specific pair-copula density are given by the labels of the nodes connected by the corresponding edge. According to this rule, the first tree produces the terms $c_{12}(F_1(x_1),F_2(x_2))$, $c_{23}(F_2(x_2),F_3(x_3))$ and $c_{34}(F_3(x_3),F_4(x_4))$. The second tree produces the terms $c_{13|2}(F(x_1|x_2),F(x_3|x_2))$ and $c_{24|3}(F(x_2|x_3),F(x_4|x_3))$. Finally, the last tree produces the term $c_{14|23}(F(x_1|x_2,x_3),F(x_4|x_2,x_3))$. The decomposition of $f(x_1,x_2,x_3,x_4)$ is given by the product of these terms:

$$\begin{aligned} f(x_1,x_2,x_3,x_4) =& f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3)\cdot f_4(x_4) \\ &\cdot c_{12}(F_1(x_1),F_2(x_2))\cdot c_{23}(F_2(x_2),F_3(x_3)) \\ &\cdot c_{34}(F_3(x_3),F_4(x_4)) \\ &\cdot c_{13|2}(F(x_1|x_2),F(x_3|x_2))\cdot c_{24|3}(F(x_2|x_3),F(x_4|x_3)) \\ &\cdot c_{14|23}(F(x_1|x_2,x_3),F(x_4|x_2,x_3)). \end{aligned} \tag{11}$$

For a d-dimensional density $f(x_1,\ldots,x_d)$, this procedure provides the decomposition formula

$$\begin{aligned} f(x_1,\ldots,x_d) =& \prod_{k=1}^{d} f(x_k) \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} \\ & c_{i,i+j|i+1,\ldots,i+j-1}(F(x_i|x_{i+1},\ldots,x_{i+j-1}),F(x_{i+j}|x_{i+1},\ldots,x_{i+j-1})), \end{aligned} \tag{12}$$

where index j identifies the trees, while index i runs over the edges in each tree.

Figure 2 shows the graphical structure of a C-vine again for a 4D problem. While in a D-vine no node in any tree is connected to more than two edges, in a C-vine each tree has a unique node, known as the root node, which is connected to all the other nodes. The rules for labeling the nodes and the edges are identical to those of the D-vine. For a C-vine, the decomposition formula is

$$\begin{aligned} f(x_1,\ldots,x_d) =& \prod_{k=1}^{d} f(x_k) \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{j,i+j|1,\ldots,j-1} \\ & (F(x_j|x_1,\ldots,x_{j-1}),F(x_{j+i}|x_1,\ldots,x_{j-1})). \end{aligned} \tag{13}$$

Although in principle the decompositions provided by the C-vines and the D-vines should be equivalent, things are actually different because of the simplification (9). In general, D-vines are often useful when there is a natural ordering of the variables (e.g., by time), whereas C-vines might be advantageous when a particular variable is known to drive the interactions of the other variables. In such a situation, this variable can be located at the root node of the leftmost tree. In many practical applications, however, no a priori information is available allowing us to decide which kind of vine to use. As a consequence, the decision has to be based on which model better fits the data.

5. Vine copulas: Computational issues

The flexibility of vine copulas complicates the parameter estimation and the model selection. One needs to select the appropriate parametric families for each pair-copula, estimate the parameters, and find a good structure for the vine trees. Thankfully, these problems can mostly be solved in separation per pair-copula and per tree level.

We remind the reader that a copula $C_{1,\ldots,d}(\boldsymbol{u})$ is the distribution function of a random variate $\boldsymbol{u} = (u_1,\ldots,u_d)$. In what follows, we assume that for all variables $i = 1,\ldots,d$, n uniform variates $\boldsymbol{u}_k = \{u_{p,k}\}$, $p = 1,\ldots,n$, are available. As mentioned in Sect. 2, these are commonly obtained by transforming the original data $\{\boldsymbol{x}_k\}$ by means of $\boldsymbol{u}_k = F_k(\boldsymbol{x}_k)$.

5.1. Model fitting in the 2D case

We first consider the simpler 2D case. Let us suppose that we have available the variates $\{(u_{p,1},u_{p,2})\}$, $p = 1,\ldots,n$, from a parametric copula model $c_{12}(u_1,u_2;\boldsymbol{\theta})$. Then, the parameters $\boldsymbol{\theta}$ can be estimated by maximum-likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{p=1}^{n} \ln c_{12}(u_{p,1},u_{p,2};\boldsymbol{\theta}). \tag{14}$$

In practice, since the true copula is unknown, it is necessary to choose a parametric copula density $c_{12}^{\mathcal{F}_\kappa}(.,.)$ from a set of families $\{\mathcal{F}_\kappa\}$, $\kappa = 1,\ldots,m$, with n_{p_k} parameters each. This is commonly done by estimating the parameters $\hat{\boldsymbol{\theta}}_\kappa$ for each copula density and then by choosing the one with either the lowest Aikaike information criterion (AIC) or the lowest Bayesian information criterion (BIC) (see Appendix B) where (Czado 2019)

$$\mathrm{AIC}_\kappa = -2\sum_{p=1}^{n} \ln c_{12}^{\mathcal{F}_\kappa}(u_{p,1},u_{p,2};\hat{\boldsymbol{\theta}}_\kappa) + 2n_{p_\kappa}, \tag{15}$$

$$\mathrm{BIC}_\kappa = -2\sum_{p=1}^{n} \ln c_{12}^{\mathcal{F}_\kappa}(u_{p,1},u_{p,2};\hat{\boldsymbol{\theta}}_\kappa) + \ln(n)n_{p_\kappa}. \tag{16}$$

5.2. Iterating through tree levels

The methods above work for a single pair-copula. It is straightforward to apply them to all pair-copulas in the first tree level, but the same is not true in later tree levels. The reason is that, as is shown by Eq. (5), the estimate of a d-dimensional PDF requires the conditional uniform variates $u_{k|-j} = F(x_k|\boldsymbol{y}_{-j})$ and $u_{j|-j} = F(y_j|\boldsymbol{y}_{-j})$, which, however, are not directly available.

To solve this issue, for the moment we suppose that the tree structure is known and the pair-copulas up to the $(\ell-1)$th tree level have been fit. In the ℓth tree, pair-copulas have the form $c_{i,j|D}$, where D is a set of $\ell-1$ variable indices called "conditioning set". Then there are always edges with indices $(i, r|D\backslash k)$ and $(j, s|D\backslash s)$ in the $(\ell-1)$th tree.[1] With the help of the so called h-functions,

$$h_{i|r;D}(u_i|u_r) = \int_0^{u_i} c_{i,r|D\backslash r}(t, u_r)dt, \tag{17}$$

and

$$h_{j|s;D}(u_j|u_s) = \int_0^{u_j} c_{j,s|D\backslash s}(t, u_s)dt, \tag{18}$$

it can be shown that $c_{i,j|D}(.,.)$ is the joint copula density of the random variables

$$u_{i|D} = h_{i|r;D}(u_{i|D\backslash r}|u_{r|D\backslash r}), \tag{19}$$

and

$$u_{j|D} = h_{j|s;D}(u_{j|D\backslash s}|u_{s|D\backslash s}). \tag{20}$$

Here, the point is that the arguments of the h-functions have the same form of the corresponding conditional uniform variables, but the conditioning set has one index fewer. Therefore, it is possible to iterate the above equation until $D = \emptyset$, which corresponds to the first tree, where data are available. Because the pair-copulas $c_{i,r|D\backslash r}$ and $c_{j,s|D\backslash s}$ have already been estimated, we can substitute the estimated models in the expressions above. In this way, we can transform data from pair-copulas in one tree into data required for the estimation in the next tree. For example, we can express $u_{1|23} = F(x_1|x_2, x_3)$ required in Eq. (11) as

$$u_{1|23} = h_{1|2;3}(u_{1|2}|u_{3|2}) = \int_0^{u_{1|2}} c_{13|2}(t, u_{3|2})dt, \tag{21}$$

where

$$u_{1|2} = h_{1|2}(u_1|u_2) = \int_0^{u_1} c_{12}(t, u_2)dt, \tag{22}$$

$$u_{3|2} = h_{3|2}(u_3|u_2) = \int_0^{u_3} c_{23}(u_2, t)dt, \tag{23}$$

and $u_1 = F_1(x_1), u_2 = F_2(x_2), u_3 = F_3(x_3)$. The analytical form of the h-functions is available for the most common copulas (Joe 1997; Schepsmeier & Stöber 2013).

5.3. Finding the tree structure

The remaining issue is how to select the right tree structure. For a C-vine, we need to specify which variable serves as the root node in every tree. For a D-vine, it is sufficient to specify the order of variables in the first tree. If $d > 4$, there are also structures other than D- and C-vines.

[1] Symbol $H\backslash r$ means the set H minus its element r.

Algorithm 1 Iterative fitting of vine copula models

Input: Observations $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$.

for tree levels $\ell = 1, \ldots, d-1$:

1. Calculate empirical Kendall's τ values $\tau_{i,j|D_e}$ for all possible edges $e = (i, j \mid D_e)$.
2. Select the spanning tree E_m maximizing $\sum_{e\in E_m} |\tau_e|$.
3. **for all** $e \in E_m$:
   (i) Based on data $\boldsymbol{u}_{i_e|D_e}, \boldsymbol{u}_{j_e|D_e}$, fit a copula model $c_{i_e,j_e|D_e}$ as in Section 5.1.
   (ii) Compute corresponding h-functions $h_{i_e|j_e;D_e}, h_{j_e|i_e;D_e}$ using formulas (17) and (18).
   (iii) Set
   $$\boldsymbol{u}_{i_e|D_e\cup j_e} = h_{i_e|j_e;D_e}(\boldsymbol{u}_{i_e|D_e}|\boldsymbol{u}_{j_e|D_e}),$$
   $$\boldsymbol{u}_{j_e|D_e\cup i_e} = h_{j_e|i_e;D_e}(\boldsymbol{u}_{j_e|D_e}|\boldsymbol{u}_{i_e|D_e}).$$
   end for

end for

To select an appropriate structure, the heuristic proposed by Dissman et al. (2013) can be used. Their idea is to capture the strongest dependencies as early as possible in the tree structure. Here, "strength" is defined as the absolute value of Kendall's τ (for the definition of this quantity, see Appendix C). We start in the first tree and compute the (empirical) pair-wise Kendall's τ for all variable pairs. Then, we choose the tree that maximizes the sum of absolute pair-wise Kendall's τ. We fit pair-copula models for the edges and compute data for the next tree. On these data, we again compute the Kendall's τ for all possible pairs and select the maximum spanning tree.[2] We continue this way, iterating between structure selection, model fitting, and transforming the data until the whole model is fit. A summary of the whole procedure is given in Algorithm 1 and implemented in the VineCopula R-package (Nagler et al. 2019).

6. Application to an experimental set of data

6.1. Data set

We made use of the data published in Andreani et al. (2018) and complemented the molecular mass values with additional CO(1-0) line data taken at the NRO 45m antenna at Nobeyama (Andreani et al. 2020a,b). The data set consists of the K-band luminosity, L_K, the infrared luminosity, L_{FIR}, the atomic hydrogen mass, M_{HI}, and the molecular mass, M_{H_2}, derived from the CO(1-0) line luminosity toward the volume-limited local galaxy sample, the Herschel reference survey (HRS) (Boselli et al. 2010). The data set is extensively described in Andreani et al. (2018) and references therein. Being volume limited, the sample contains all the galaxies above a given threshold of K-band luminosity, and the analysis would not be largely affected by a flux selection effect.

These variables were chosen because they are related to the main overall physical properties of the sample and their relation to the star formation activity in the galaxies. We aimed to inves-

[2] A spanning tree is a subset of a graph, which has all the vertices covered with the minimum possible number of edges.

tigate the relationship of those properties and derive insights into the driving physical mechanism in their interstellar medium.

6.2. Data analysis and interpretation

As the first step of the analysis, the PDF of each of the quantities logLK=$\log_{10}L_K$, logLIR=$\log_{10}L_{FIR}$, logMHI=$\log_{10}M_{HI}$ and logMH2v=$\log_{10}M_{H_2}$ were modeled by means of the generalized lambda distribution (GLD) family (Karian & Dudewicz 2011, and references therein). The members of this family are four-parameter PDFs, which are known for their high flexibility and the large range of shapes that they can reproduce. The starship method has been adopted to fix the parameters. The reason is that this method finds the parameters that transform the data closest to the uniform distribution, which is an attractive characteristic when working with copulas. The results of the fit are shown in Fig. 3. After this step, the procedure presented in the previous section was applied with the random variates $\boldsymbol{u}$, computed by means of the estimated margins. The results are shown in Tab. 1 and Fig. 4. The original Kendall's τ coefficients in Tab. 1 are related to the strengths of the relation between the quantities without being dependent on the derived margins. This shows that the strongest correlations occur between the far-IR luminosity L_{FIR} and the gas masses, first molecular M_{H_2} and then atomic M_{HI}, while L_{FIR} is weakly correlated with the near-IR K-band luminosity, L_K. On the other side, Fig. 4 indicates the type of vine structure selected, specifically a C-vine, and provides the list of pair-copulas singled out for each edge. For each pair-copula, the values of the corresponding coefficients and of the lower and upper tail dependence coefficients are also shown (for the meaning of last two quantities see Appendix D). The Kendall's τ (for Tree 1) and partial Kendall's τ (for Trees 2 and 3) associated to each edge are also shown. This last quantity measures the dependence between two variables after the effect of other variables (the common indices of two nodes) has been removed.[3] In order to check the reliability of the obtained results, the procedure was repeated with the random variates $\boldsymbol{u}$ given by the pseudo-observations. As the comparison of Fig. 5 with Fig. 4 shows, the differences are not substantial. The fact that for the highest trees the copulas selected by the two methods are different is not significant. Indeed, one has to take present that when the Kendall's τ between the random variates coming from two PDFs or two conditional PDFs is close to zero (i.e., they are almost uncorrelated), there are various kinds of copulas that can provide similar reconstructions of the corresponding bivariate data distribution. In other words, in the reconstruction of a multivariate distribution, the specific types of copula are only meaningful for values of the Kendall's τ significantly different from zero.

These results can be more clearly interpreted by looking at Fig. 6. As explained in Sect. 5, the structure selection algorithm tries to capture the strongest dependencies first. Figure 6 shows a plot of the tree structure labeled with the Kendall's τ (for Tree 1) and partial Kendall's τ (for Trees 2 and 3). Here, the logLIR quantity as been selected as root node. This means that it is strongly correlated to all other variables and that it drives part of the dependence between the other variables. In the second tree, the effect of logLIR on the dependence between the others has been removed. There is only some weak negative dependence left.

All this can be interpreted with the fact that although from Tab. 1 the quantities M_{HI} and M_{H_2} appear positively dependent, such dependence appears to be driven entirely by their dependence on the quantity L_{FIR}. Once the dependence of L_{FIR} is removed from the relation with the other quantities the residual relations M_{HI} with M_{H_2} and L_K with M_{HI} are negatively dependent (albeit this dependence is quite weak). This means that the dependence shown in Tab. 1 is driven entirely by their dependence on L_{FIR}.

Since L_{FIR} is dominated by the thermal dust emission heated by FUV photons by massive stars and residing in molecular clouds, which are cocoons of star formation processes, this result confirms that the physical properties of the galaxies are driven by their star formation.

For completeness, in Fig. 7 we show the original data versus the data simulated from the the estimated 4D joint PDF of which the 2D slices are shown in Fig. 8. Figure 7 shows a good agreement between original and simulated data, while Fig. 8 demonstrates the one-to-one relation between the couple of variables.

7. Conclusions

In this work, a flexible and effective approach to modeling the relationship of a set of experimental multidimensional quantities is presented. This approach consists of modeling the joint PDF of the data by means of a special type of copula called a vine copula. Classical copulas are functions that contain all of the information on the relationship between two random quantities. Their major limitation is that they are unable to model multidimensional data. Vine copulas overcome this limitation by expressing the joint PDFs as the product of a set of 2D copula densities and the 1D PDFs corresponding to each quantity. In particular, two types of vine copulas have been considered: the C-vine and the D-vine. This approach makes the estimation of the joint PDFs amenable to a theoretical treatment and feasible from the computational point of view.

We applied this method to published data on the near-IR and far-IR luminosities and atomic and molecular masses of the HRS. We find that the far-IR luminosity, L_{FIR}, is the key player in driving the galaxy properties in this sample. Despite its original selection in the K-band, the HRS sample shows that it is L_{FIR} that plays a fundamental role. Removing its dependence from the other variables, the K-band luminosity, and the atomic and molecular masses, makes it clear that the established relation among these quantities does not show up any more.

The L_{FIR} in this sample is dominated by the thermal dust emission heated by FUV photons produced by massive stars in molecular clouds. Our analysis therefore highlightsthat the star formation activity of these galaxies is the key parameter driving the galaxy evolution.

Acknowledgements.

References

Aas, K., Czado, C., Frigessi, A., & Bakken, H. 2009, Insurance: Mathematics and Economics, 44, 182
Andreani, P., Spinoglio, L., Boselli, A. et al. 2014, A&A 566, A70
Andreani, P., Boselli, A., Ciesla, L. et al. 2018, A&A 617, A33
Andreani, P., Miyamoto Y., Kaneko H., Boselli, A., Tatematsu K., Sorai K. Vio R., submitted
Andreani, P., Miyamoto Y., Kaneko H., Boselli, A., Tatematsu K., Sorai K., in preparation
Balakrishnan, N., & Lai, C.D. 2010, Continuous Bivariate Distributions (New York: Springer)

[3] The "partial Kendall's τ" is computed by means of the Kendall's τ between the variates $u_{i|D}$and $u_{j|D}$ in Eqs. (19) and (20). It provides a measure of the relationship between u_i and u_j when the influence of the variates corresponding to the set D is removed.

Table 1. Sample Kendall's τ.

	logLK	logLIR	logMHI	logMH2v
logLK	1.00	0.25	0.03	0.25
logLIR	0.25	1.00	0.49	0.61
logMHI	0.03	0.49	1.00	0.33
logMH2v	0.25	0.61	0.33	1.00

Boselli, A., Eales, S., Cortese, L., et al. 2010, PASP, 122, 261
Burnham, K.P., & Anderson, D.R. 2002, Model Selection and Multimodel Inference (New York: Springer)
Czado, C. 2019, Analyzing Dependent Data with Vine Copulas (New York: Springer)
Durante, F., & Sempi, C. 2016, Principles of Copula Theory (New York: CRC Press)
Dissmann, J., Brechmann, E.C., Czado, C. and Kurowicka, D., 2013. Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics & Data Analysis, 59, pp.52-69.
Kotz, S., Balakrishnan, N., & Johnson, N.L. 2000, Continuous Multivariate Distributions Vol. 1 (New York: John Wiley & Sons, Inc.)
Nagler, T. & Czado, C. 2016, Journal of Multivariate Analysis, 151, 69
Nagler, T., Schellhase, C., & Czado, C. 2017, Dependence Modeling, 5, 99
Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E.C., Graeler, B, & Erhardt, T. 2019, VineCopula: Statistical Inference of Vine Copulas. R package version 2.3.0. https://CRAN.R-project.org/package=VineCopula
Nelsen, R.B. 2006, An Introduction to Copulas (New York: Springer Science + BusinessMedia, Inc.)
Hofert, M., Kojadinovic, I., Mächler, M., & Yan, J. 2018, Elements of Copula Modeling with R (New York: Springer)
Joe, H. 1997, Multivariate Models and Dependence Concepts (Dordrecht: Springer-Science+Business Media)
Joe, H. 2015, Dependence Modeling with Copulas (New York: CRC Press)
Karian, Z.A., & Dudewicz, E.J. 2011, Handbook of Fitting Statistical Distributions with R (New York: CRC Press)
Lin, C.A., & Kilbinger, M. 2015, A&A, 583, A70
Lin, C.A., Kilbinger, M., & Pires, S. 2016, A&A, 593, A88
Sato, M., Ichiki, K., & Takeuchi, T. 2010, Phys. Rev. Lett., 105, 251301
Sato, M., Ichiki, K., & Takeuchi, T. 2011, Phys. Rev. D, 83, 023501
Schepsmeier, U, & Stöber, J. 2013, Statistical Papers, 55, 525
Scherrer, R.J., Berlind, A.A., Mao, Q., & McBride, C.K. 2010, AJ, 708, L9
Takeuchi, T.T. 2010, MNRAS, 406, 1830
Takeuchi, T.T., Sakurai, A., Yuan, F.T., & Burgarella, D. 2011, Earth Planet Space, 65, 281
Vio, R., Fasano, G., Lazzarin, M., & Lessi, O. 1994, A&A, 289, 640

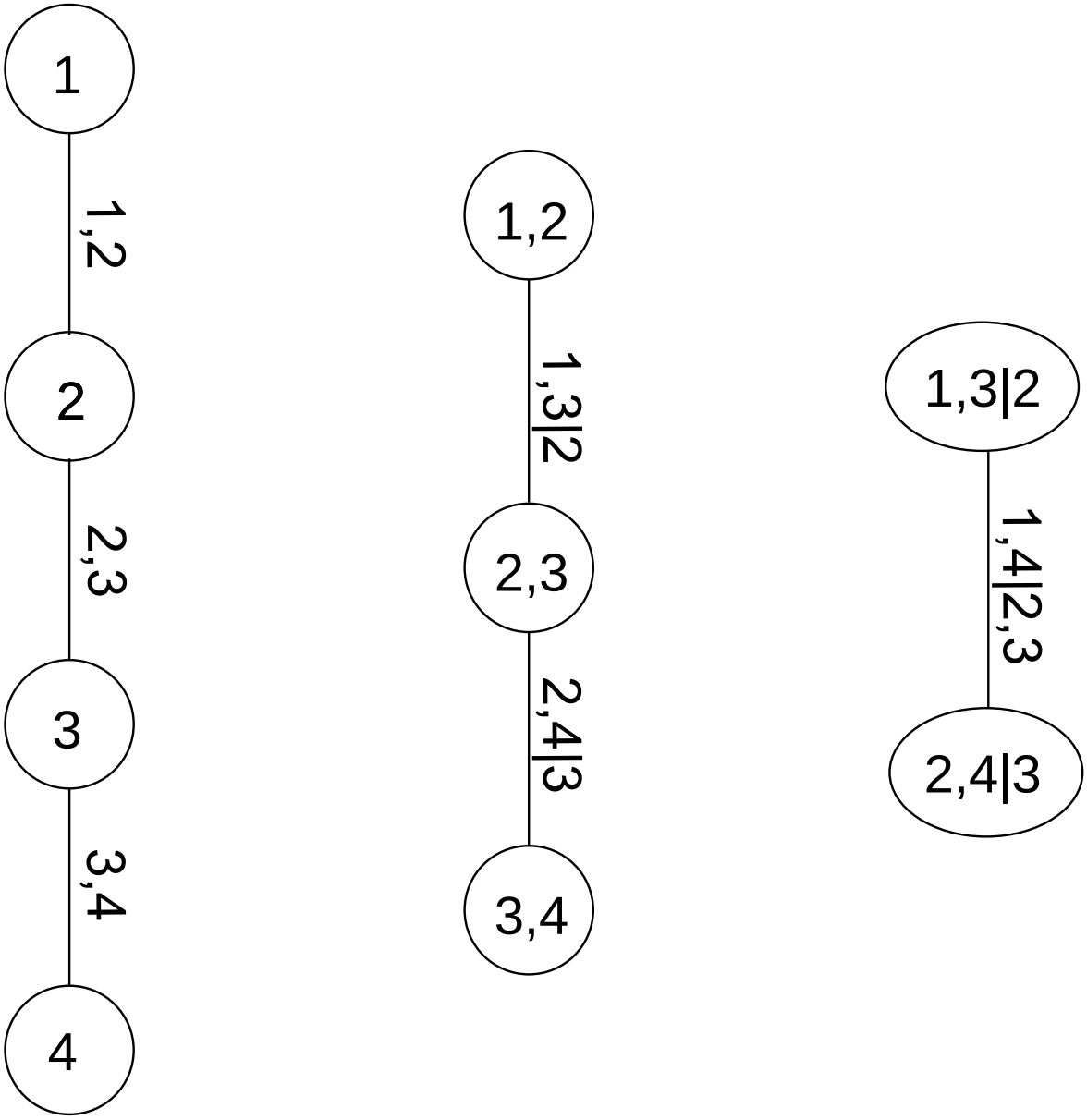

Fig. 1. Example of tree structure of a 4D D-vine copula (see text).

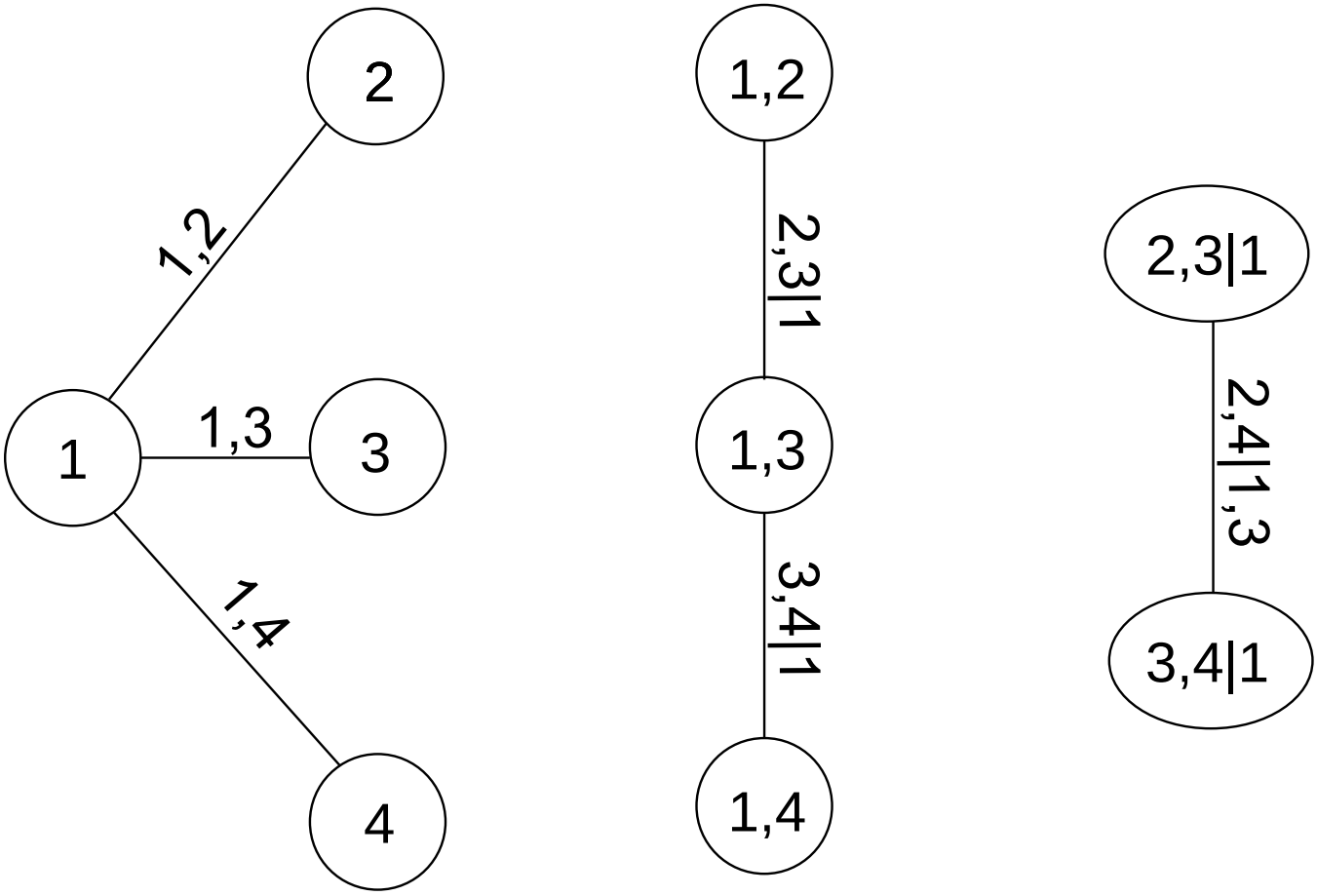

Fig. 2. Example of tree structure of a 4D C-vine copula (see text).

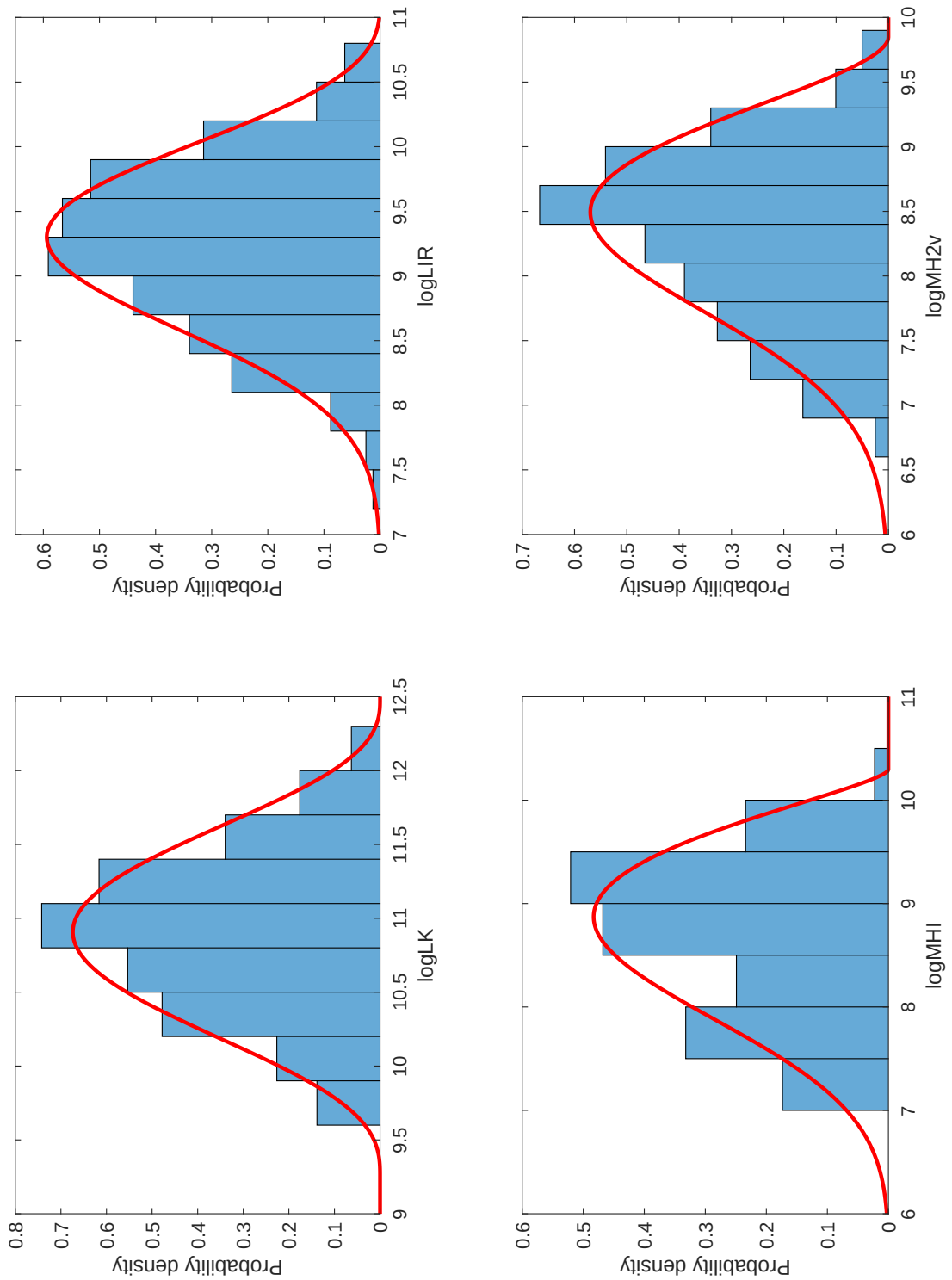

Fig. 3. Histograms of logLK, logLIR, logMHI, and logMH2v data. The red lines provide the PDF obtained by the fit of these data with the generalized lambda distribution family.

tree	edge	copula	par1	par2	tau	utd	ltd
1	2,1	Tawn2	3.31	0.34	0.29	0.33	0
	2,3	Gauss	0.69	0.00	0.49	0	0
	4,2	BB8	4.88	0.92	0.61	0	0
2	3,1|2	BB8_90	-1.66	-0.90	-0.18	0	0
	4,3|2	Frank	-1.02	0.00	-0.11	0	0
3	4,1|3,2	Clayton	0.26	0.00	0.12	0	0.07

type: C-vine

1 → logLK; 2 → logLIR; 3 → logMHI; 4 → logMH2v

Fig. 4. Results concerning the application of the procedure described in Sects. 5 and 6.2 to the data corresponding to Fig. 3 with the random variates $\boldsymbol{u}$ computed by means of the estimated generalized Lambda PDFs (see text). Here, a copula is associated to each edge and Kendall's τ for Tree 1, a partial Kendall's τ for Trees 2 and 3 (column "tau") and upper (column "utd"), respectively lower (column "ltd") tail-dependence coefficients. These are theoretical quantities corresponding to the selected copulas of which the estimated parameters are given in the columns "par1" and "par2". A description of these copulas can be found in Czado (2019). BB8_90 refers to copula BB8 rotated 90°.

```
 tree    edge |     copula   par1   par2 |  tau   utd    ltd
-----------------------------------------------------------------
   1      2,1 |      Tawn2   3.66   0.35 |  0.30  0.34     0
          2,3 |      Gauss   0.69   0.00 |  0.49     0     0
          4,2 |        BB8   4.46   0.93 |  0.60     0     0
   2    3,1|2 | Clayton_90  -0.33   0.00 | -0.14     0     0
        4,1|2 |    BB8_180   1.57   0.89 |  0.16     0     0
   3  4,3|1,2 |      Frank  -0.88   0.00 | -0.10     0     0
-----------------------------------------------------------------
type: C-vine
1 → logLK; 2 → logLIR; 3 → logMHI; 4 → logMH2v
-----------------------------------------------------------------
```

Fig. 5. As in Fig. 4 but with random variates ***u*** computed by means of the pseudo-observations (see text). Clayton_90 and BB8_180 mean copula Clayton rotated 90° and copula BB8 rotated 180°, respectively.

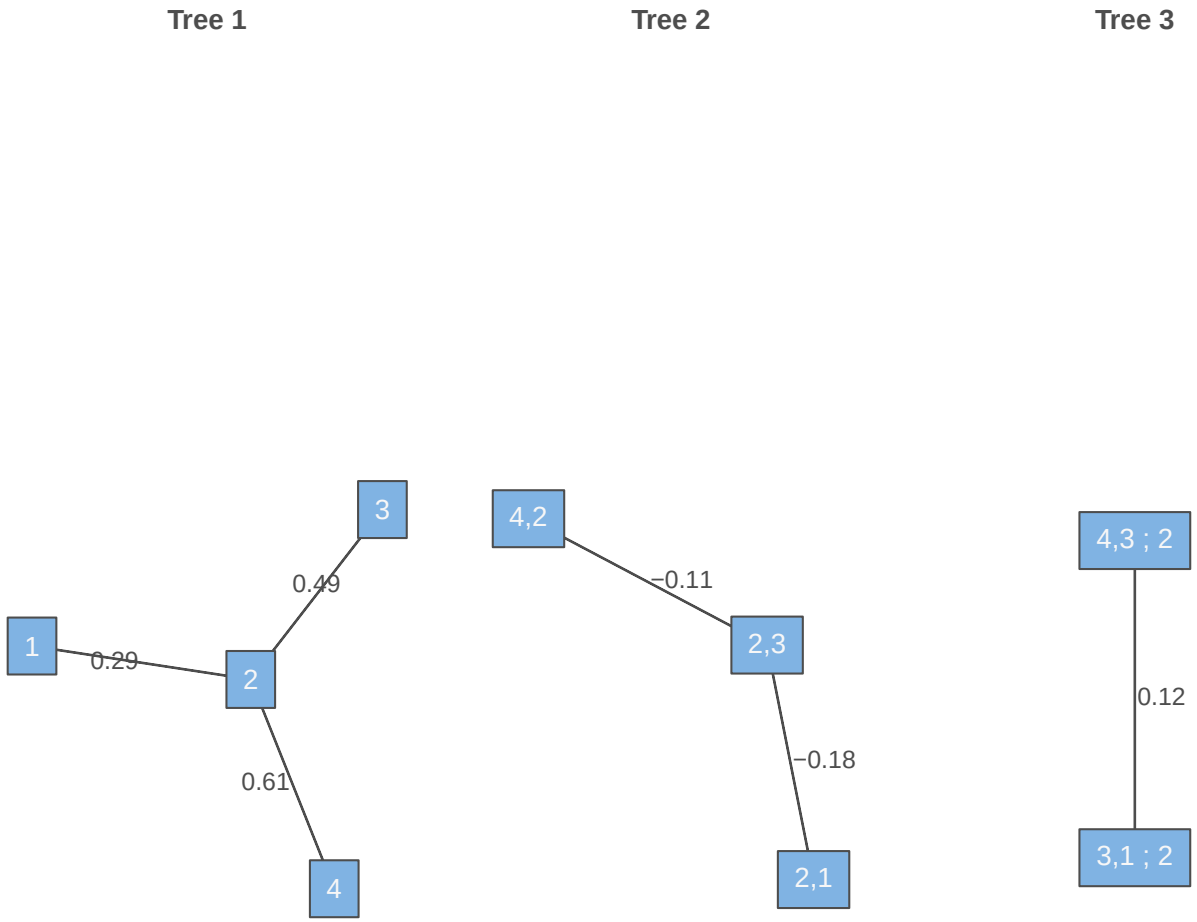

Fig. 6. C-vine structure selected by the procedure described in Sects. 5 and 6.2 to the data corresponding to Fig. 3. Each edge in Tree 1 is associated to a Kendall's τ, whereas for Trees 2 and 3 they are associated to a partial Kendall's τ. Here, 1 → logLK, 2 → logLIR, 3 → logMHI, 4 → logMH2v.

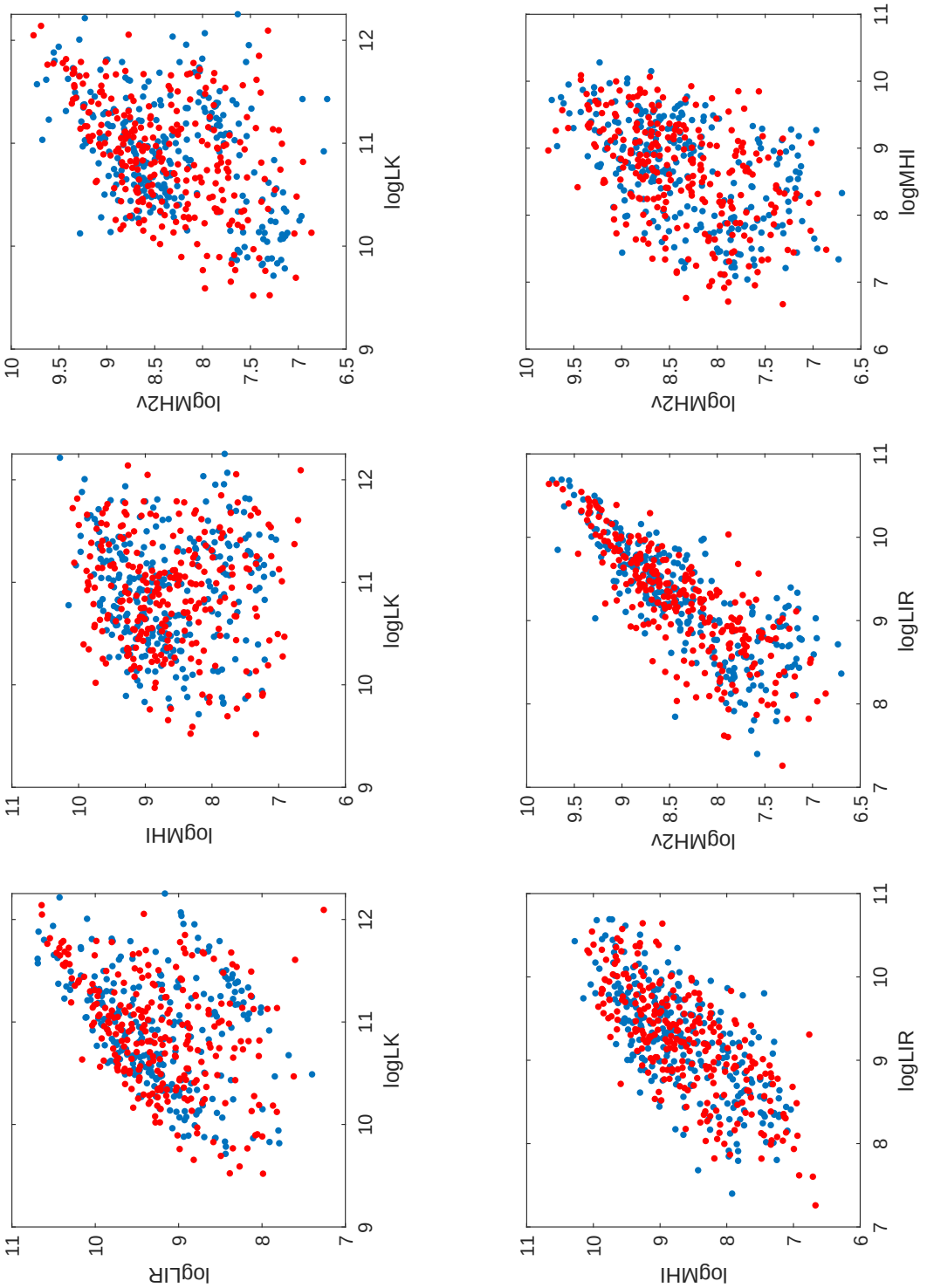

Fig. 7. Original logLK, logLIR, logMHI, and logMH2v data (blue circles) versus the corresponding simulated data obtained from the estimated 4D joint PDF (red circles).

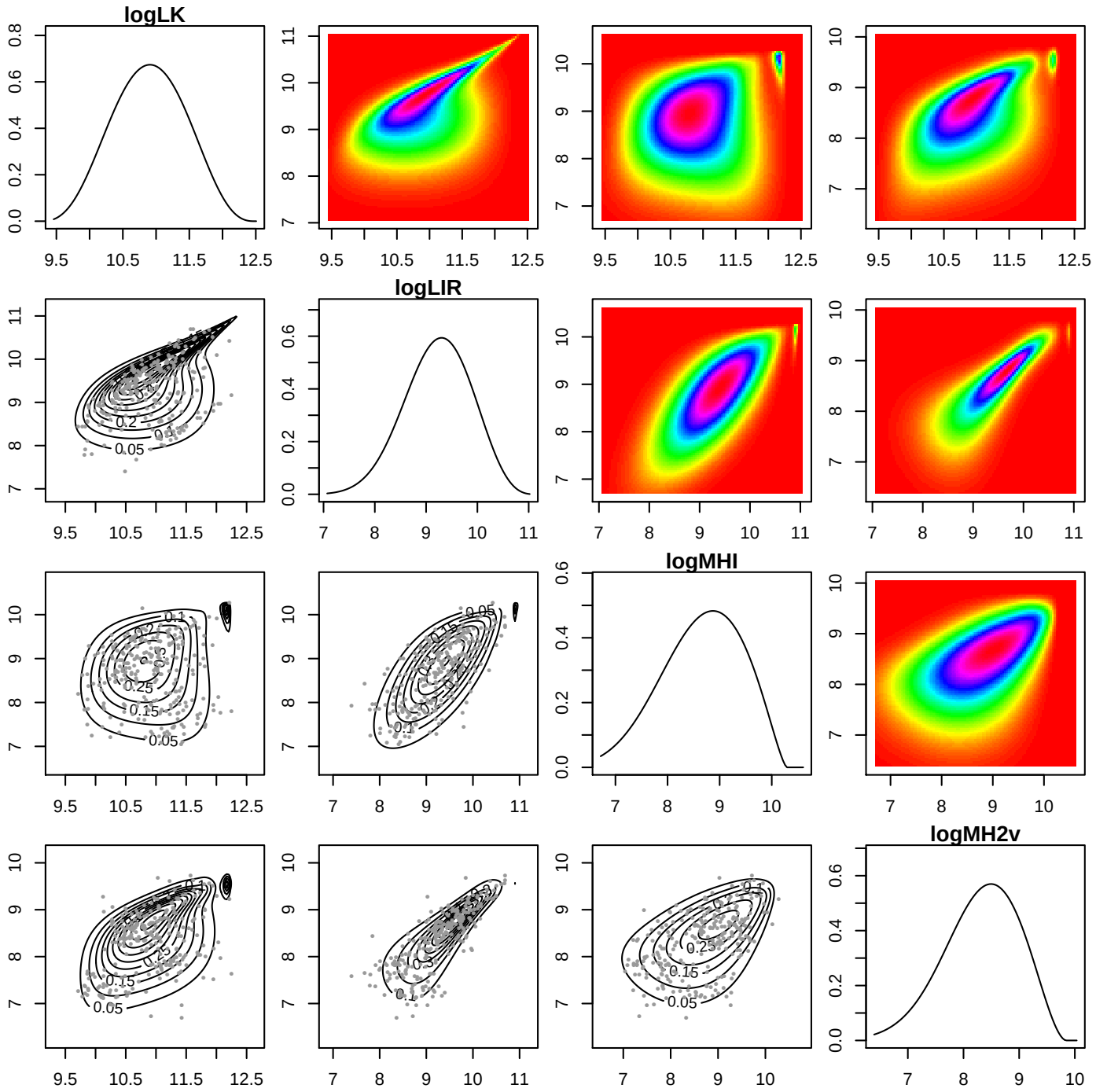

Fig. 8. Two-dimensional slices of the estimated 4D joint PDF for the data corresponding to Fig. 3. Along the diagonal are the fit PDFs of Figure 3. The right panels show the slices in which colors correspond to the intensity of the relation, while the left panels report, on the same slices, the data points and the iso-contours.

Appendix A: The Johnson's distribution and the generalized Lambda distribution families

The Johnson's distributions and the generalized Lambda distributions (GLD) are both four-parameter families that are used for fitting distributions to a wide variety of data sets. In particular, the Johnson's system is based on three different PDFs, $f_U(x)$, $f_B(x)$, and $f_L(x)$ according to the fact that the random variable x is unbounded, bounded both above and below, and bounded only below:

$$f_U(x) = \frac{\eta}{\sqrt{2\pi[(x-\epsilon)^2+\lambda^2]}} \times \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{(x-\epsilon)}{\lambda} + \sqrt{\left(\frac{x-\epsilon}{\lambda}\right)^2 + 1}\right)\right]^2\right\}, \quad \text{(A.1)}$$

for $-\infty < x < \infty$,

$$f_B(x) = \frac{\eta\lambda}{\sqrt{2\pi}(x-\epsilon)(\lambda - x + \epsilon)} \times \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{x-\epsilon}{\lambda - x + \epsilon}\right)\right]^2\right\}, \quad \text{(A.2)}$$

for $\epsilon \leq x \leq \lambda + \epsilon$, and

$$f_L(x) = \frac{\eta}{\sqrt{2\pi}(x-\epsilon)} \times \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{x-\epsilon}{\lambda}\right)\right]^2\right\}, \quad \text{(A.3)}$$

for $x \geq \epsilon$. In literature, various methods are available for the selection of the appropriate type of PDF as well for the estimate of the parameter (e.g., Vio et al. 1994; Karian & Dudewicz 2011).

The PDFs corresponding to the GLD family are given by

$$f_\lambda(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}}, \quad \text{(A.4)}$$

where $x = Q(y; \lambda_1, \lambda_2, \lambda_3, \lambda_4)$ with

$$Q(y; \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2}, \quad \text{(A.5)}$$

and $0 \leq y \leq 1$. Concerning this family, various methods are also available for the estimate of the parameters (e.g., Karian & Dudewicz 2011).

Appendix B: AIC and BIC criteria

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are two criteria for model selection from a finite set of models (Burnham & Anderson 2002). They are based on the maximum value $\hat{L}$ of the likelihood function for the model as well on the number n_p of free parameters it contains. The idea is that, when fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both the BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model. In particular,

$$\mathrm{AIC} = 2n_p - 2\ln(\hat{L}), \quad \text{(B.1)}$$

whereas

$$\mathrm{BIC} = n_p \ln(n) - 2\ln(\hat{L}), \quad \text{(B.2)}$$

with n being the number of data. In practical applications, a set of models is chosen, the corresponding quantity $\hat{L}$ evaluated, Eq. (B.1) or Eq. (B.2) used, and finally the model with the lowest AIC or BIC selected. The difference between AIC and BIC is how much model complexity (i.e., the number of parameters) is penalized. For $n \geq 8$, the BIC penalty is stronger. Both criteria give a mathematical guarantee to find the "best" model as the sample size increases. The BIC assumes that the true model is among the set of candidates, but the AIC does not. These criteria are useful in the context of the vine copulas since the different types of bivariate copulas considered for their construction contain a different number of free parameters.

Appendix C: The Kendall's τ

When working with copulas the relationship between two random quantities is typically measured by means of the Kendall's τ. The reason can be understood by looking at Fig. C.1, which shows the the realization of 1000 independent copies of a bivariate random vector (x_1, x_2) from the Gaussian, exponential and Cauchy PDFs and of the same number of a bivariate random vector (u_1, u_2) from the uniform PDF. These realizations appear quite different from one another, as well as the corresponding linear correlation coefficients ρ. Here, the point is that the first three sets of random numbers $\{(x_{1,i}, x_{2,i})\}$ were obtained from the set of uniform random pairs $\{(u_{1,i}, u_{2,i})\}$ by means of the transformations:

$$(x_1, x_2) = (F^{-1}(u_1), F^{-1}(u_2)), \quad \text{(C.1)}$$

where $F^{-1}(u)$ is the inverse CDF corresponding to the various PDFs. This is a common method to simulate random numbers from a given PDF. What this figure indicates is that the different appearance of the realizations is not due to the intrinsic relationship between the random quantities, but rather to their margins. Since with copulas one wants to disentangle margins from the dependence structure, the latter should be measured in a way that does not depend on the marginal distributions. This is what the Kendall's τ does.

If (x_1', x_2') is and independent copy of (x_1, x_2), τ is defined as

$$\tau = \mathbb{P}\left[(x_1 - x_1')(x_2 - x_2') > 0\right] - \mathbb{P}\left[(x_1 - x_1')(x_2 - x_2') < 0\right], \quad \text{(C.2)}$$

that is, it is the probability of concordance minus the probability of discordance of the random pairs (x_1, x_2) and (x_1', x_2'). The rationale behind this definition is that if there is positive dependence between the variable x_1 and x_2, then when x_1 increases or decreases, a similar behavior has to be expected for x_2. It can be demonstrated (Hofert et al. 2018) that

$$\tau = 4\int_0^1 \int_0^1 c(u_1, u_2) C(u_1, u_2) du_1 du_2 - 1, \quad \text{(C.3)}$$

meaning that τ effectively depends only on the underlying copula.

The sample version $\hat{\tau}$ of τ is given by

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{sign}[(x_{i1} - x_{j1})(x_{i2} - x_{j2})], \quad \text{(C.4)}$$

where n is the number of observations and $\mathrm{sign}[x] = 1$ if $x > 0$, $\mathrm{sign}[x] = 0$ if $x = 0$ and $\mathrm{sign}[x] = -1$ if $x < 0$. As expected, $\hat{\tau}$ is the same for all the realizations in Fig. C.1.

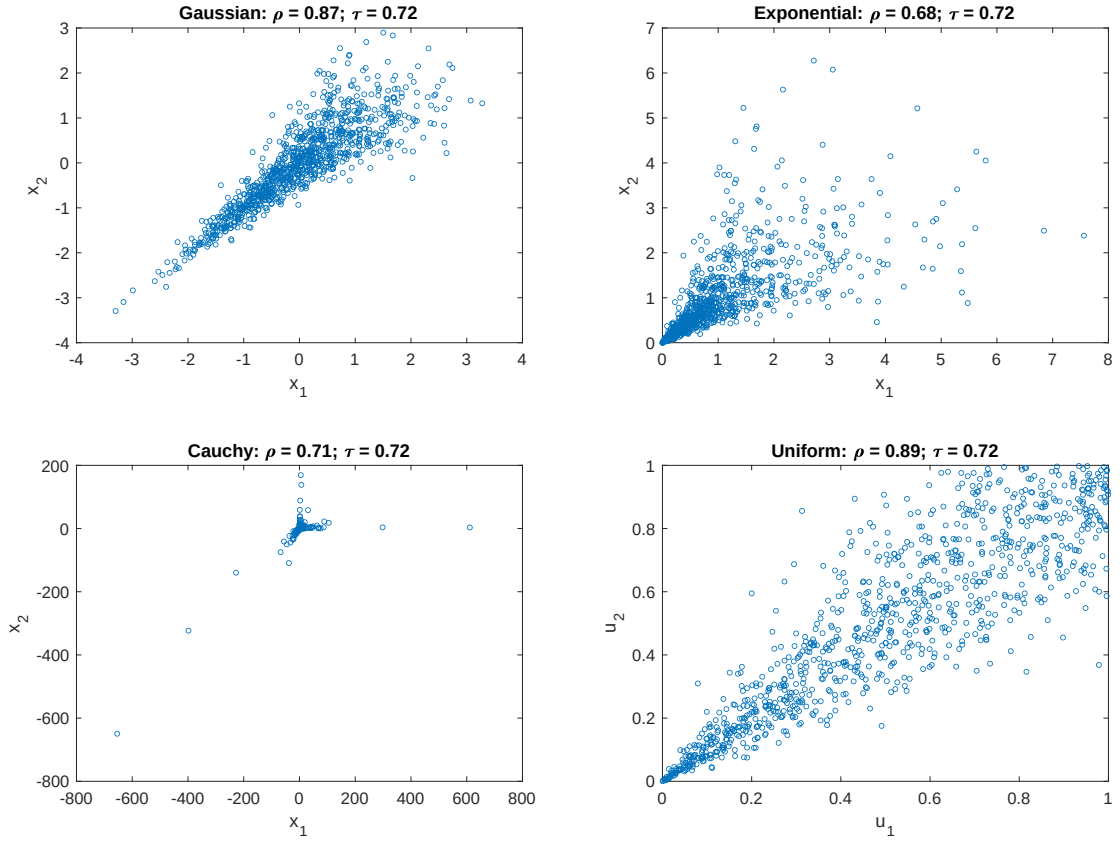

Fig. C.1. Numerical realization of 1000 independent copies of a bivariate random vector (x_1, x_2) from the Gaussian, exponential, and Cauchy PDFs obtained from the set of uniform random pairs $\{(u_{1,i}, u_{2,i})\}$, shown in the bottom-right panel, by means of the transformations $(x_1, x_2) = (F^{-1}(u_1), F^{-1}(u_2))$ where $F^{-1}(u)$ is the inverse CDF corresponding to the various PDFs.

Appendix D: Tail-dependence coefficients

There are situations where in the 2D scatterplot of a set of data, the points appear concentrated in one or both the tails of their joint distribution. For instance, this is the case for the scatterplots in Fig. C.1, where a concentration of points in the lower-left tail of the joint distribution is evident. Joint distributions characterized by well-developed tails indicate a high probability of joint occurrence of extremely small and/or large values. In some practical applications, it is useful to have an estimate of this probability. Given the margins $F_1(x_1)$ and $F_2(x_2)$ and the copula $C(u_1, u_2)$, the coefficients of lower and upper tail dependence provide such an estimate and are defined as

$$\lambda_l = \lim_{t\to 0^+} \mathbb{P}(x_2 \le F_2^{-1}(t) | x_1 \le F_1^{-1}(t)); \\ = \lim_{t\to 0^+} \frac{C(t,t)}{t}, \tag{D.1}$$

respectively,

$$\lambda_u = \lim_{t\to 1^-} \mathbb{P}(x_2 > F_2^{-1}(t) | x_1 > F_1^{-1}(t)); \\ = \lim_{t\to 1^-} \frac{1-2t+C(t,t)}{1-t}. \tag{D.2}$$

These coefficients are conditional probabilities that measure the tendency of the random variable x_2 to behave as the random variable x_1. When their value is close to one, it means tail dependence (i.e., high probability of joint extreme values), when close to zero it means tail independence (i.e., low probability of joint extreme values). The analytical expression of λ_l and λ_u is available for various parametric copulas. For instance, the random points in the bottom-left panel of Fig. C.1 has been generated through a Clayton copula with coefficient $\theta = 5$, for which $\lambda_l = 0.87$ and $\lambda_u = 0$. These values also hold for the other distributions in the same figure.