Copulas and Vines (stat08012)

Luciana Dalla Valle, University of Plymouth, UK *

Abstract

Copulas and vines allow us to model the distribution of multivariate random variables in a flexible way. This article introduces copulas via Sklar's theorem, explains how pair copula constructions are built by decomposing multivariate copula densities and illustrates vine graphical representations.

Keywords: *Copulas, Conditional Distributions, Dependence, Pair Copula Constructions, Sklar's Representation, Vines.*

*email: luciana.dallavalle@plymouth.ac.uk

1 Copulas

A copula is a multivariate distribution function with uniform marginals on the interval $[0, 1]$. According to Sklar's theorem [25], given d continuous random variables $X_1, \ldots, X_d$, any joint multivariate distribution $F(x_1, \ldots, x_d)$ of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ can be uniquely determined as a copula C of its univariate marginal distributions $F_1(x_1), \ldots, F_d(x_d)$, via the expression

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d); \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ denotes the set of parameters of the copula, which we will omit in the remainder, to simplify the notation (see **stat00943**).

Once applied to the univariate marginals, the copula returns the multivariate joint distribution, enclosing all the information about the dependence structure of the variables [21], [24].

The joint density function of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is therefore,

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \times f_1(x_1) \times \ldots \times f_d(x_d) \tag{1}$$

where $c(F_1(x_1), \ldots, F_d(x_d))$ is the d-variate copula density.

For example, in the 4-dimensional case and the (1) becomes

$$f(x_1, \ldots, x_4) = c(F_1(x_1), \ldots, F_4(x_4)) \times f_1(x_1) \times \ldots \times f_4(x_4).$$

Several families of copulas are available to capture different types of symmetric and asymmetric dependencies among the marginals. The most popular families are the elliptical <stat01214>, such as the Gaussian <stat01215> and Student's t, and the archimedean cop-

ulas, such as the Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7 and BB8 (see **stat07523**, **stat04393** and [5]).

Due to their flexibility in modelling the distribution of multivariate random variables, copulas have become very popular and have been applied to a wide variety of fields, ranging from finance [7], [8], [10], [11], [15], to social sciences [13] and engineering [19].

2 Pair Copula Constructions

In recent years, the literature on bivariate copulas (called *pair copulas*) has thrived, with regular developments of new contributions. In contrast, the use to multivariate copulas has been more limited, due to analytical and computational complexity. In order to overcome these limitations, Bedford and Cooke [2], [3] introduced a flexible class of multivariate copulas that is constructed using a set of bivariate copulas as building blocks. The decomposition of a multivariate copula into bivariate copulas is called *pair copula construction* (PCC) and it allows to express the multivariate density of a random vector as a product of pair copula densities.

Assuming that $f(x_1, \ldots, x_d)$ is the joint density of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$, then it factorizes (uniquely up to re-labeling of the variables) into a product of conditional densities

$$f(x_1, \ldots, x_d) = f_d(x_d) \times f_{d-1|d}(x_{d-1}|x_d) \times \ldots \times f_{1|2\cdots d}(x_1|x_2, \ldots, x_d). \tag{2}$$

For a 4-dimensional density, equation (2) takes the following form

$$f(x_1, \ldots, x_4) = f_4(x_4) \times f_{3|4}(x_3|x_4) \times f_{2|3,4}(x_2|x_3, x_4) \times f_{1|2,3,4}(x_1|x_2, x_3, x_4). \tag{3}$$

By Sklar's theorem the joint density of the subvector (X_d, X_{d-1}) can be expressed in terms

of a copula density, as follows

$$f(x_{d-1}, x_d) = c_{d-1,d}(F_{d-1}(x_{d-1}), F_d(x_d)) \times f_{d-1}(x_{d-1}) \times f_d(x_d),$$

where $c_{d-1,d}(\cdot,\cdot)$ is an arbitrary pair copula density. Hence, the conditional density of $X_{d-1}|X_d$ can be easily written as

$$f_{d-1|d}(x_{d-1}|x_d) = c_{d-1,d}(F_{d-1}(x_{d-1}), F_d(x_d)) \times f_{d-1}(x_{d-1}).$$

Then, the conditional densities of (2) can be decomposed into the appropriate pair copula times a conditional marginal density. More precisely, for a generic element $X_\jmath$ of the vector $\mathbf{X}$ we obtain

$$f_{x_\jmath|\mathbf{v}}(x_\jmath|\mathbf{v}) = c_{x_\jmath,v_\ell|\mathbf{v}_{-\ell}}(F_{x_\jmath|\mathbf{v}_{-\ell}}(x_\jmath|\mathbf{v}_{-\ell}), F_{v_\ell|\mathbf{v}_{-\ell}}(v_\ell|\mathbf{v}_{-\ell})) \times f_{x_\jmath|\mathbf{v}_{-\ell}}(x_\jmath|\mathbf{v}_{-\ell}), \tag{4}$$

where $\mathbf{v}$ is the conditioning vector, v_ℓ is a generic component of $\mathbf{v}$, $\mathbf{v}_{-\ell}$ is the vector $\mathbf{v}$ without the component v_ℓ, $F_{x_\jmath|\mathbf{v}_{-\ell}}(\cdot|\cdot)$ is the conditional distribution of $x_\jmath$ given $\mathbf{v}_{-\ell}$, and $c_{x_\jmath,v_\ell|\mathbf{v}_{-\ell}}(\cdot,\cdot)$ is the conditional pair copula density. For example, the second factor, $f_{3|4}(x_3|x_4)$, in the right-hand side of (3) can be easily decomposed into the pair-copula $c_{3,4}(F_3(x_3), F_4(x_4))$ and a marginal density $f_3(x_3)$:

$$f_{3|4}(x_3|x_4) = c_{3,4}(F_3(x_3), F_4(x_4)) \times f_3(x_3).$$

One of the possible decompositions of the third factor in the right-hand side of (3), using the (4), is

$$f_{2|3,4}(x_2|x_3, x_4) = c_{2,3|4}(F_{2|4}(x_2|x_4), F_{3|4}(x_3|x_4)) \times f_{2|4}(x_2|x_4),$$

for the appropriate pair copula $c_{2,3|4}$, applied to the transformed variables $F_{2|4}(x_2|x_4)$ and $F_{3|4}(x_3|x_4)$.

Any d-dimensional joint multivariate distribution function can thus be expressed as a product of pair copulas by recursively plugging equation (4) in equation (2). Since in the (4) the conditional distributions of the form $F_{x|\mathbf{v}}(\cdot|\cdot)$ are not directly observable, they are calculated using Joe's result [20]

$$F_{x|\mathbf{v}}(x|\mathbf{v}) = \frac{\partial C_{x,v_\ell|\mathbf{v}_{-\ell}}(F(x|\mathbf{v}_{-\ell}), F(v_\ell|\mathbf{v}_{-\ell}))}{\partial F(v_\ell|\mathbf{v}_{-\ell})}. \tag{5}$$

If the conditioning set $\mathbf{v}$ is univariate, $\mathbf{v} = v$ and expression (5) can be written as

$$F(x|v) = \frac{\partial C_{x,v}(x, v, \boldsymbol{\theta})}{\partial v} = h(x, v, \boldsymbol{\theta}), \tag{6}$$

where $\boldsymbol{\theta}$ denotes the set of parameters of the copula, and $F(x|v)$ is named the *h function*. The forms of the *h functions* for the main classes of copulas are given in [1] and in [4]. For example, $F_{3|4}(x_3|x_4)$ can be determined using expression (6) as follows

$$F_{3|4}(x_3|x_4) = \frac{\partial C_{3,4}(F_3(x_3), F_4(x_4))}{\partial F_4(x_4)}.$$

3 Vines

PCCs can be represented through a graphical model called *regular vine* (R-vine). An R-vine $\mathcal{V}(d)$ on d variables is a nested set of trees $T_1, \ldots, T_{d-1}$, where the variables are represented by nodes linked by edges, each associated with a certain pair copula in the corresponding PCC. The edges of tree T_ι are the nodes of tree $T_{\iota+1}$, $\iota = 1, \ldots, d-1$. In an R-vine, if two edges of tree T_ι share a common node, they are represented in tree $T_{\iota+1}$ by nodes joined

by an edge. Note that there are many different orderings of the variables yielding different R-vines. We can distinguish two special subclasses of regular vines, *canonical* or *C-vines* and *drawable* or *D-vines*, each of them giving a specific way of decomposing the density. A C-vine is an R-vine where each tree T_ι has a unique node that is connected to $d-\iota$ edges [6]. Conversely, a D-vine is an R-vine where all nodes in tree T_ι are adjacent to at most two other nodes [9]. A pair copula density is associated to any edge, with the edge label indicating the subscript of the pair copula density [12]. An example of a 4-dimensional D-vine is provided in Figure 1. The vine consists of three trees T_ι, $\iota = 1, \ldots, 3$, where each edge corresponds to a pair copula density. Each edge may belong to a different copula family and the edge label corresponds to the subscript of the pair copula density, e.g. edge $14|23$ corresponds to the copula density $c_{14|23}(\cdot)$.

<Figure 1 near here>

The joint density of the D-vine represented in Figure 1 is given by

$$f(x_1, \ldots, x_4) = \prod_{j=1}^{4} f_j(x_j) \times c_{12} \times c_{23} \times c_{34} \times c_{13|2} \times c_{24|3} \times c_{14|23},$$

where $c_{ab} = c_{ab}(F(x_a), F(x_b))$.

More generally, the joint density of a D-vine of dimension d takes the form

$$\begin{aligned} f(x_1, \ldots, x_d) \;=\; & \prod_{j=1}^{d} f_j(x_j) \times \\ & \prod_{\iota=1}^{d-1} \prod_{i=1}^{d-\iota} c_{i,i+\iota|i+1,\ldots,i+\iota-1}(F(x_i|x_{i+1}, \ldots, x_{i+\iota-1}), F(x_{i+\iota}|x_{i+1}, \ldots, x_{i+\iota-1})) \end{aligned}$$

which is the product of d marginal densities f_j and $d(d-1)/2$ bivariate copulas $c_{i,i+\iota|i+1,\ldots,i+\iota-1}(\cdot,\cdot)$ evaluated at the conditional distribution functions $F(\cdot|\cdot)$.

4 Vines Inference

Inference on R-vines involves the specification of the vine structure, the choice of the copula family for each pair copula and the estimation of their parameters.

In order to select a suitable R-vine decomposition, a sequential "top-down" approach is commonly adopted, specifying the structure of the first tree and then proceeding similarly for higher-order trees [1]. This approach is based on the definition of a tree on all nodes (named spanning tree), which maximizes the sum of absolute pairwise dependencies, measured, for example, by Kendall's tau [14]. The subsequent trees are are built in a similar way, under the additional restriction that the proximity condition must be fulfilled. This specification allows to capture the strongest dependencies in the first tree, thus obtaining a parsimonious model. An alternative "bottom-up" strategy starts by selecting the weakest conditional dependencies in higher-order trees and then specifying lower-order trees analogously [23].

Given the selected tree structure, copula families for each pair of variables are generally selected one by one, using the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) or the Copula Information Criterion (CIC) [17]. This choice is made amongst a large set of families, comprising elliptical copulas as well as archimedean copulas and their rotated versions, to cover a large range of possible dependence structures.

Note that conditional independence between variables may reduce the number of levels of the pair copula decomposition, and hence simplify the construction. From a graphical point of view, conditional independence removes edges in the R-vine, performing the so-called "pruning". Pruning may be implemented using a copula goodness-of-fit-test [16] or, more generally, the Cramér-von Mises test [18].

After the specification of the vine structure and the pair copula families, the copula parameters $\boldsymbol{\theta}$ are then generally estimated using the maximum likelihood method [1]. The most

computationally efficient approach is the sequential method, for which the R-vine estimation procedure is performed level by level for all trees, until the R-vine is completely specified. Alternatively, Bayesian inference can be adopted to estimate the copula parameters [22].

5 Related Articles

stat00943

stat01214

stat01215

stat07523

stat03684

stat04393

stat07457

References

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

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

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

[4] Czado, C., Schepsmeier, U. & Min, A. (2012). Maximum Likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, **12**, 229–255.

[5] Dalla Valle, L. (2016) The Use of Official Statistics in Self-Selection Bias Modeling. *Journal of Official Statistics*, **32**, 887–905.

[6] Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas. *Quality Technology and Quantitative Management*, **11**, 111–131.

[7] Dalla Valle, L. (2010) Measuring Operational Risk in a Bayesian Framework, in *Rethinking Risk Measurement and Reporting. Uncertainty, Bayesian Analysis and Expert Judgement*, K. Böcker, ed, Risk Books, London, chap. 14, pp. 395–422.

[8] Dalla Valle, L. (2009) Bayesian Copulae Distributions, with Application to Operational Risk Management, *Methodology and Computing in Applied Probability*, **11**, 95–115.

[9] Dalla Valle, L., De Giuli, M.E., Tarantola, C. & Manelli, C. (2016) Default Probability Estimation via Pair Copula Constructions. *European Journal of Operational Research*, **249**, 298–311.

[10] Dalla Valle, L., Fantazzini, D. & Giudici, P. (2007) Empirical Studies with Operational Loss Data: Dalla Valle, Fantazzini and Giudici Study, in *Operational Risk : a Guide to Basel II Capital Requirements, Models, and Analysis*, A.S. Chernobai, S.T. Rachev & F.J. Fabozzi, eds., John Wiley & Sons, pp. 274–277.

[11] Dalla Valle, L. & Giudici, P. (2008) A Bayesian Approach to Estimate the Marginal Loss Distributions in Operational Risk Management, *Computational Statistics and Data Analysis*, **52**, 3107–3127.

[12] Dalla Valle, L. & Kenett, R. (2015) Official Statistics Data Integration for Enhanced Information Quality. *Quality and Reliability Engineering International*, **31**, 1281–1300.

[13] Dalla Valle, L., Leisen, F. & Rossini, L. (2016) Bayesian Nonparametric Conditional Copula Estimation of Twin Data. `arXiv:1603.03484 [stat.ME]`.

[14] Dißmann, J., Brechmann, E.C., Czado, C. & Kurowicka, D. (2013) Selecting and Estimating Regular Vine Copulae and Application to Financial Returns. *Computational Statistics and Data Analysis*, **59**, 52–69.

[15] Fantazzini, D., Dalla Valle, L. & Giudici, P. (2008) Copulae and Operational Risks, *International Journal of Risk Assessment and Management*, **9**, 238–257.

[16] Genest, C. & Favre, A.C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **12**, 347–368.

[17] Grønneberg, S. & Hjort, N.L. (2014) The Copula Information Criteria. *Scandinavian Journal of Statistics*, **41**, 436–459.

[18] Hobæk Haff, I. & Segers, J. (2015) Nonparametric estimation of pair-copula constructions with the empirical pair-copula. *Computational Statistics and Data Analysis*, **84**, 1–13.

[19] Jane, R., Dalla Valle, L., Simmonds, D. & Raby, A. (2016) A Copula Based Approach for the Estimation of Wave Records Through Spatial Correlation. *Coastal Engineering*, **117**, 1–18.

[20] Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *IMS lecture notes*, **28**, 120–141.

[21] Joe, H. (1997). *Multivariate model and dependence concepts*, Monographs on Statistics an Applied Probability, **73**, Chapman & Hall, London.

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

[23] Kurowicka, D. (2011) Optimal truncation of vines, in *Dependence Modeling: Vine Copula Handbook*, D.Kurowicka & H. Joe, eds., World Scientific Publishing Co.

[24] Nelsen, R. B. (1999). *An introduction to copulas*, Springer-Verlag, New York.

[25] Sklar, M. (1959). Fonctions de répartition á $n$dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

6 Figures

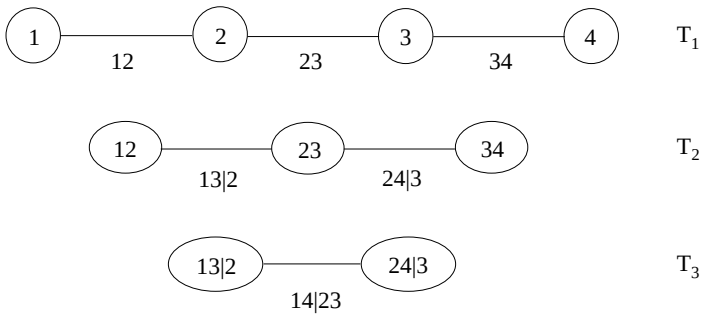

Figure 1: 4-dimensional D-vine graphical representation