Computational Statistics and Data Analysis 117 (2018) 109–127

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Vine copula based likelihood estimation of dependence patterns in multivariate event time data

Nicole Barthel [a,*], Candida Geerdens [b], Matthias Killiches [a], Paul Janssen [b], Claudia Czado [a]

[a] Department of Mathematics, Technische Universität München, Boltzmannstraße 3, 85748 Garching, Germany
[b] Center for Statistics, I-BioStat, Universiteit Hasselt, Agoralaan 1, 3590 Diepenbeek, Belgium

ARTICLE INFO

Article history:
Received 6 April 2017
Received in revised form 22 July 2017
Accepted 27 July 2017
Available online 21 August 2017

Keywords:
Dependence modeling
Multivariate event time data
Maximum likelihood estimation
Right-censoring
Survival analysis
Vine copulas

ABSTRACT

In many studies multivariate event time data are generated from clusters having a possibly complex association pattern. Flexible models are needed to capture this dependence. Vine copulas serve this purpose. Inference methods for vine copulas are available for complete data. Event time data, however, are often subject to right-censoring. As a consequence, the existing inferential tools, e.g. likelihood estimation, need to be adapted. A two-stage estimation approach is proposed. First, the marginal distributions are modeled. Second, the dependence structure modeled by a vine copula is estimated via likelihood maximization. Due to the right-censoring single and double integrals show up in the copula likelihood expression such that numerical integration is needed for its evaluation. For the dependence modeling a sequential estimation approach that facilitates the computational challenges of the likelihood optimization is provided. A three-dimensional simulation study provides evidence for the good finite sample performance of the proposed method. Using four-dimensional mastitis data, it is shown how an appropriate vine copula model can be selected for data at hand.

1. Introduction

In many studies, primary interest lies in the time until a prespecified event occurs. Often, the data appear in clusters. For example, in Laevens et al. (1997) time to mastitis infection in udder quarters of primiparous cows is observed. The cow is the cluster and the infection times of the four udder quarters are the clustered data. For an accurate analysis of clustered data flexible models are needed to describe the underlying dependence pattern. Copulas provide the right tools for this goal. A d-dimensional copula $\mathbb{C}$ is a distribution function on $[0, 1]^d$ with uniformly distributed margins. According to Sklar (1959), a copula is a dependence function that interconnects the marginal survival functions S_j, $j = 1, \ldots, d$, and thereby models the joint survival function S, i.e. with $t_j \geq 0$

$$S(t_1, \ldots, t_d) = \mathbb{C}\{S_1(t_1), \ldots, S_d(t_d)\}.$$

For clusters of size two, a large catalog of bivariate copula families exists. For clusters of size more than two, popular multivariate copulas such as exchangeable (EAC) and nested Archimedean copulas (NAC) (Embrechts et al., 2003; Hofert, 2008; Joe, 1993; Nelsen, 2006) only induce restrictive association patterns. For instance, in EAC models all marginal copulas

* Corresponding author.
E-mail addresses: nicole.barthel@tum.de (N. Barthel), candida.geerdens@uhasselt.be (C. Geerdens), matthias.killiches@tum.de (M. Killiches), paul.janssen@uhasselt.be (P. Janssen), cczado@ma.tum.de (C. Czado).

http://dx.doi.org/10.1016/j.csda.2017.07.010
0167-9473/

show exactly the same type (and even strength) of tail-dependence. In NAC models, the nesting condition limits all building blocks to stem from the same copula family leading again to the same type (but not strength) of tail-dependence. More flexible models are thus needed to capture complex association patterns present in clustered data. This is a difficult but at the same time a challenging exercise. Flexible alternatives for EAC and NAC include Joe–Hu copulas (Joe and Hu, 1996) and vine copulas (Aas et al., 2009; Bedford and Cooke, 2002; Czado, 2010; Kurowicka and Joe, 2010; Kurowicka and Cooke, 2006). A Joe–Hu copula corresponds to a mixture of positive powers of max-infinitely divisible bivariate copulas. The induced association pattern is completely determined by the mixture and by the choice of bivariate copulas. The idea of a vine copula is to decompose the joint density of the clustered event times into a cascade of bivariate copula densities via conditioning. So, in both approaches bivariate copulas or bivariate copula densities are the building blocks. Given the variety of well-studied bivariate copulas, it is clear that Joe–Hu copulas and vine copulas allow a flexible modeling of the within-cluster association in clustered event time data.

For the above mentioned copula models the focus is usually on complete, i.e. non-censored, data. However, event time data are often subject to right-censoring. This means that for some observations the true event time is not observed but instead a lower (censored) time is registered. For example, in the mastitis study cows may be lost to follow-up (e.g. due to death) or may experience the event after the end of the study (censored at study end). The presence of right-censoring in clustered event time data complicates the statistical analysis substantially, but its incorporation is indispensable to arrive at a sound statistical analysis. Since this is not straightforward, the application of copulas to right-censored clustered data has been less explored. Recently, Geerdens et al. (2016) studied, for right-censored data, the model flexibility of Joe–Hu copulas as compared to less elaborate EAC and NAC models. Vine copulas have not yet been studied for right-censored clustered event times. Therefore, our main objective is to develop a likelihood based estimation approach using the flexible class of vine copulas. Using the theorem of Sklar (1959) and following the ideas in Shih and Louis (1995), we proceed in two steps. In step one, the survival margins are modeled. Here, any estimation technique for univariate right-censored event time data can be used, e.g. maximum likelihood estimation or the nonparametric Kaplan–Meier estimator. Focus, however, lies in detecting the inherent dependence pattern using vine copula based likelihood estimation in the second step. Due to right-censoring, numerical integration is needed, making the global likelihood optimization computationally challenging. We introduce a sequential estimation approach to find a fair trade-off between the numerical demand caused by data complexity and the accuracy of the estimates.

In Section 2, we describe the construction of vine copulas; we consider trivariate and quadruple data. The mastitis data are described in detail in Section 3. Following the ideas in Shih and Louis (1995), Section 4 contains the likelihood function for right-censored quadruple event time data. In particular, we provide the likelihood expression in terms of vine copula components and therewith extend existing vine copula concepts to the setting of right-censored clustered time-to-event data. In this section, we also discuss how to deal with numerical aspects of the presented optimization method. A simulation study is performed in Section 5 to demonstrate the good finite sample performance of our approach. In Section 6, we revisit the mastitis data. Conclusions and remarks are collected in Section 7.

2. Vine copulas

First, we recall the definition of vine copulas (Aas et al., 2009; Bedford and Cooke, 2002; Czado, 2010; Kurowicka and Cooke, 2006; Kurowicka and Joe, 2010) and explain how vine copulas are constructed following the approach taken in Czado (2010).

The basic idea is to decompose a d-dimensional copula density $\mathbb{c}$ into a product of $d\,(d-1)\,/2$ so-called pair-copulas via conditioning. The latter are copulas associated to bivariate conditional distributions. It is essential to note that the representation of $\mathbb{c}$ in terms of pair-copulas is not unique. Depending on the conditioning strategy, there is a variety of possible decompositions. To organize the structure of a d-dimensional vine-copula, Bedford and Cooke (2002) propose a sequence of linked trees. More precisely, a set of connected trees $\mathcal{V} := (\mathcal{T}_1, \ldots, \mathcal{T}_{d-1})$ is called a regular vine (R-vine) on d elements with the set of edges $E(\mathcal{V}) := E_1 \cup \cdots \cup E_{d-1}$ and the set of nodes $N(\mathcal{V}) := N_1 \cup \cdots \cup N_{d-1}$ if the following holds:

1. $\mathcal{T}_1$ is a tree with nodes $N_1 = \{1, \ldots, d\}$ and edges E_1.
2. For $j = 2, \ldots, d-1$, $\mathcal{T}_j$ is a tree with nodes $N_j = E_{j-1}$ and edges E_j.
3. *(Proximity condition)* For $j = 2, \ldots, d-1$, whenever two nodes of $\mathcal{T}_j$ are connected by an edge, the associated edges of $\mathcal{T}_{j-1}$ share a node.

Two four-dimensional examples of possible tree sequences, also called vine structures, are visualized in Fig. 1. We see that except for the labeling of the nodes the two examples illustrate the only possible ways to arrange the four nodes in $\mathcal{T}_1$. In the vine structure on the right, there exists a unique node in $\mathcal{T}_j$, $j = 1, 2, 3$, that is connected to $d-j$ edges. This vine structure suggests an ordering by importance and is referred to as a C-vine. The vine structure on the left is called a D-vine. Here, no node is connected to more than two edges implying a serial ordering. In particular, in dimension three C-vines and D-vines are equivalent. In this paper, we concentrate on D-vines. The derived concepts, however, can easily be applied to C-vines (Barthel, 2015).