Finite Mixture of C-vines for Complex Dependence

O. Evkaya *[1], **C. Yozgatlıgil**[2], **and A. S. Kestel**[2]

[1] *Atilim University*
[2] *Middle East Technical University*
[2] *Middle East Technical University*

Recently, there has been an increasing interest on the combination of copulas with a finite mixture model. Such a framework is useful to reveal the hidden dependence patterns observed for random variables flexibly in terms of statistical modeling. The combination of vine copulas incorporated into a finite mixture model is also beneficial for capturing hidden structures on a multivariate data set. In this respect, the main goal of this study is extending the study of Kim et al. (2013) with different scenarios. For this reason, finite mixture of C-vine is proposed for multivariate data with different dependence structures. The performance of the proposed model has been tested by different simulated data set including various tail dependence properties.

Keywords: copula; dependence; finite mixture; C-vine; tail dependence

1 Full Inference on C-vine Copula

This section is introduced to recall inference procedures of parameters in Vine copula, exemplified by C-vine copula. Generally, p-dimensional C-vine copula density can be written as in 1,

$$f(\boldsymbol{x};\boldsymbol{\phi_{cvine}}) = \prod_{k=1}^{p} f_k(x_k) \prod_{i=1}^{p-1}\prod_{j=1}^{p-i} c_{i,i+j|1:(j-1)} \left(F(x_i|x_1,...,x_{i-1}), F(x_{i+j}|x_1,...,x_{i-1}); \boldsymbol{\beta}_{i,i+j|(i+1):(i+j-1)}\right) \tag{1}$$

where $f_k(x_k)$ denotes the marginal densities, $c_{i,i+j|1:(j-1)}$ are the bivariate copula density functions with parameter(s) $\beta_{i,(i+j)|(i+1):(i+j-1)}$, and ϕ_{cvine} is the set of all parameters in p-dimensional C-vine density.

*Corresponding author: ozanevkaya@gmail.com

There exist one root node in the tree construction of C-vine model which results in following illustration in 4-dimension given by equation 2,

$$\begin{aligned} f(x_1, x_2, x_3, x_4; \boldsymbol{\phi}) = & c_{12}(F(x_1), F(x_2); \beta_{12}) c_{13}(F(x_1), F(x_3); \beta_{13}) \\ & c_{14}(F(x_1), F(x_4); \beta_{14}) \\ & c_{23|1}(F(x_2|x_1), F(x_3|x_1); \beta_{23|1}) c_{24|1}(F(x_2|x_1), F(x_4|x_1); \beta_{24|1}) \\ & c_{34|12}(F(x_3|x_1, x_2), F(x_4|x_1, x_2); \beta_{34|12}) \prod_{k=1}^{4} f_k(x_k) \end{aligned} \tag{2}$$

Under such multivariate framework, full inference on C-vine copula can be derived using the log-likelihood function presented in 3,

$$\begin{aligned} L(\boldsymbol{\phi}) = & \sum_{i=1}^{p-1} \sum_{j=1}^{p-i} \sum_{n=1}^{N} \log c_{i,i+j|(1):(j-1)} \\ & (F(x_{i,n}|x_{1,n}, ..., x_{i-1,n}), F(x_{i+j,n}|x_{1,n}, ..., x_{i-1,n}); \boldsymbol{\beta}_{i,i+j|(i+1):(i+j-1)}) \end{aligned} \tag{3}$$

and following three consecutive steps:

Step 1 Decide which variable is used as a root node in the first tree T_1 of a C-vine copula (i.e. joining the variables in which the root node variable is selected based on its significant relations with other variables)

Step 2 Specify the family and parametric shape of each pair-copula in an assumed C-vine copula

Step 3 Estimate all parameters of C-vine by maximizing the log-pseudo likelihood function given in 3

2 Finite Mixture of C-vines

The finite mixture model is introduced to connect m component C-vine densities to detect complex and hidden dependence structures in multivariate data with the related EM algorithm for estimating the parameters in the model.

Assume that a p-dimensional random vector X=(X_1, ..., X_p) is said to be generated from a mixture of M- component C-vine densities, where its density function is defined as in 4.

$$g(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m f(\boldsymbol{x}, \boldsymbol{\phi_m}) \tag{4}$$

where π_m is the mixing proportion of the m-th component s.t. $0 < \pi_m < 1$ and $\sum_{m=1}^{M} \pi_m = 1$. Besides, ϕ_m is the m-th component-specific parameter vector for the C-vine density described in 4. Note that θ is the set of all parameters with dimension p and denoted by Θ, full product space (the simplex of π_m and the cross product space of ϕ_m). Here p is the total number of free parameters to be estimated and $p = (M-1) + \sum_{m=1}^{M} dim(\phi_m)$. For the estimation of equation 4, both the number of components M and the parameters θ are required to estimate, using the following EM-algorithm setup, proposed previously by Dempster et al. (1977).

Assume that N observations randomly drawn from a M component C-vine density given in 4, denoted as x_k=($x_{k,1}$, ..., $x_{k,p}$) where $k = 1, ..., p$. Then, log-likelihood of θ is described as given in 5

$$L(\boldsymbol{\theta}) = \log(\prod_{n=1}^{N} g(\boldsymbol{x}_n, \boldsymbol{\theta})) = \log(\prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m f(\boldsymbol{x}_n, \boldsymbol{\phi}_m)) \tag{5}$$

Let z_n=($z_{n1}, ..., z_{nm}, ..., z_{nM}$) denotes latent variables, where $z_{nm} = 1$ if x_n drawn from the m-th component and $z_{nm} = 0$ otherwise. Here, z_n is i.i.d. from a multinomial distribution, i.e. z_n is Mult($M, \pi = (\pi_1, ..., \pi_m)$). Under this setup, the complete log-likelihood for the complete data set y_n=(x_n,z_n) is given by equation 6.

$$\begin{aligned} L(\boldsymbol{\theta})_c &= \log \prod_{n=1}^{N} \prod_{m=1}^{M} [\pi_m f(\boldsymbol{x}_n, \boldsymbol{\phi}_m)]^{z_{nm}} \\ &= \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log(\pi_m) + \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log(f(\boldsymbol{x}_n, \boldsymbol{\phi}_m)) \end{aligned} \tag{6}$$

Starting with initial values (initial guesses) for the parameters, θ^0, the repeated E-th and M-th step of the EM algorithm (to compute the successive estimates, θ^s) is described as follows:

E-step Calculates the conditional expectation of $L(\theta)_c$ given the observed data and current parameter estimates for θ. Such a computation is equivalent to the calculation of posterior probability that x_n belongs to the m-th

component, given the current values of the parameters formulated as in equation 7

$$\widehat{z_{nm}}^{(s)} = E[z_{nm}|\boldsymbol{x}, \theta^s] = P[z_{nm} = 1|\boldsymbol{x}, \theta^s] = \frac{\pi_m^{(s)} f(\boldsymbol{x}_n, \boldsymbol{\phi}_m^{(s)})}{\sum_{l=1}^{M} \pi_l^{(s)} f(\boldsymbol{x}_n, \boldsymbol{\phi}_l^{(s)})} \quad (7)$$

M-step Computes the parameter estimates for each component independently, ($\pi_1^{(s+1)}$, ..., $\pi_m^{(s+1)}$,..., $\pi_M^{(s+1)}$) and ($\phi_1^{(s+1)}$, ..., $\phi_m^{(s+1)}$, ..., $\phi_M^{(s+1)}$) by maximizing the expected complete-data log-likelihood from E-step. It is also possible to obtain closed form solution for $\pi_m^{(s+1)} = \frac{\sum_{n=1}^{N} \widehat{z_{nm}}^{(s)}}{N}$. Afterwords, the estimation of $\phi_m^{(s+1)}$ in the m-th component C-vine or D-vine density function is equivalent to deriving the parameter estimates weighted by $\widehat{z_{nm}}^{(s)}$ for the parameters in a C-vine density in equation 4.

The E-step and the M-step are iterated until $L(\theta^{s+1}) - L(\theta^s)$ is smaller than a pre-specified tolerance value (ie. 10^{-6} or 10^{-8}), as a result of a nice property of EM algorithm that the log-likelihood is not decreased during the iteration. Based on the above setup, the given algorithm is run with multiple starting values randomly drawn from the parameter space and the best values is chosen from multiple local maximizer having the highest log-likelihood value. To accomplish the full inference on mixture of C-vines, three well known model selection criteria values are used:

- Akaike's Information Criterian (AIC) as defined $AIC = -2\log(L(\widehat{\theta})) + 2p$
- Bayesian Information Criterian (BIC) as defined $BIC = -2\log(L(\widehat{\theta})) + p\log(n)$
- Consistent AIC (CAIC) as defined $CAIC = -2\log(L(\widehat{\theta})) + p(\log(n) + 1)$

where $\widehat{\theta}$ is the estimate of p-dimensional θ defined in 4. In this study, the main objective is identifying the full inference on C-vine mixture model based on different scenarios. For this reason, the whole procedure for the full inference of C-vine copula can be summarized as follows:

Step 1 Derive the normalized ranks of d-dimensional observed data

Step 2 Decide the root node of each C-vine density by calculating all pairwise correlations.

Step 3 Consider different candidates of copulas for all pairs in an assumed mixture model

Step 4 Given a copula family, fit a mixture vine copula with M component and estimate the parametersin each model by employing the EM- algorithm

Step 5 Select the best fitted model by finding the model with smallest values of model selection criterians such as AIC, BIC and CAIC.

For the last step, naturally, even if the selection criteria measures give meaningful conclusions, they are not enough to decide the best fitted model. For this reason, available GOF tests are also required to improve the model selection part. Especially, Clarke and Vuong tests are widely used GOF test for comparing two different vine copulas might be considered in model selection.

3 Simulation Results

To test the performance of the mixture model, the base mixture model is constructed using Clayton and Gumbel pairs with parameters ($\beta_{12}^C = 8, \beta_{13}^C = 7, \beta_{23|1}^C = 6$) and ($\beta_{12}^G = 9, \beta_{13}^G = 6, \beta_{23|1}^G = 5$), respectively. As the number of parameters described, as a simple case, the mixture of C-vines are considered in 3-dimension with positive strong tail dependencies.

Here, as a simulated data, 2 component equally weighted mixture of C-vines is considered under the proposed mixture model to investigate the data generating process performs well or not. In this setup, the number of observations has been increased from $N = 50$ (small data set) to $N = 1000$ (large data set) to see the differences in parameter estimation process. For now, parameter estimation results of 2-component C-vine mixture with Clayton pairs are presented for illustration. Here, the estimated parameters for each pair of both components are obtained using the average value and the median values of 1000 different run given in () and [] table, respectively.

In table 1, as it is expected, parameter estimations of the first component are very close to true value since the correct copula family at each step is predefined as Clayton at the beginning for the simulated data. Besides, the number of observations has positive impact on closing the gap between parameter estimates and true values. Generally, the most suitable model will be determined by comparing the model comparison values among different scenarios like Clayton-Clayton, Clayton-Gumbel, Frank-Gumbel, assumed pair copulas in mixture model. As we expected, the most plausible result will be obtained from the Clayton-Gumbel pair families selection for the first-second component, same as the original simulated data.

Table 1: Parameter Estimation and Model Comparison Values

	Number of Observations				
	50	100	250	500	1000
$\widehat{\beta^C_{12}}$	(7.19)[7.3]	(7.95)[8.05]	(8.77)[8.82]	(8)[8]	(8)[8]
$\widehat{\beta^C_{13}}$	(6.49)[6.6]	(7.08)[7.15]	(7.76)[7.71]	(7)[7]	(7)[7]
$\widehat{\beta^C_{23\|1}}$	(4.29)[4.24]	(4.29)[4.25]	(4.81)[4.81]	(6)[6]	(6)[6]
$\widehat{\beta^C_{12}}$	(7.6)[7.66]	(7.29)[7.27]	(7.12)[7.22]	(7.07)[6.98]	(7.65)[7.39]
$\widehat{\beta^C_{13}}$	(6.3)[6.32]	(5.8)[5.79]	(5.55)[5.66]	(5.5)[5.43]	(6.03)[5.87]
$\widehat{\beta^C_{23\|1}}$	(3.32)[3.11]	(2.85)[2.53]	(2.53)[2.42]	(2.56)[2.54]	(2.14)[2.17]
AIC	-194	-363	-783	-1536	-3001
BIC	-188	-356	-773	-1524	-2986
CAIC	-185	-353	-770	-1521	-2983

References

[1] R.B Nelsen. *An Introduction to Copulas* 2nd edn. Springer Series in Statistics. Springer, New York, NY, 2006.

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

[3] T. Bedford, R. Cooke. Vines a new graphical model for dependent random variables. *Ann. Stat.*. 30(4):1031–1068, 2002.

[4] D. Kim, J.M. Kim, S.M. Liao, Y.S. Jung. Mixture of D-vine Copulas for Modeling Dependence. *Computational Statistics and Data Analysis*. 64:1–19, 2013.

[5] E.C. Brechman, U. Schepsmeir. Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. *Journal of Statistical Software*. 52(3):1–27, 2013.