Vine copula mixture models and clustering for non-Gaussian data

Submitted to Econometrics and Statistics

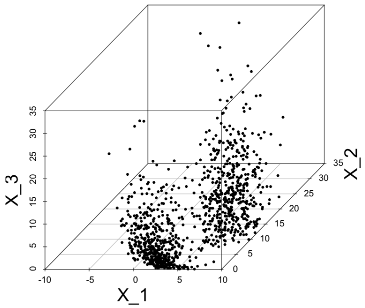
Prof. Claudia Czado
Özge Sahin <ozge.sahin@tum.de>

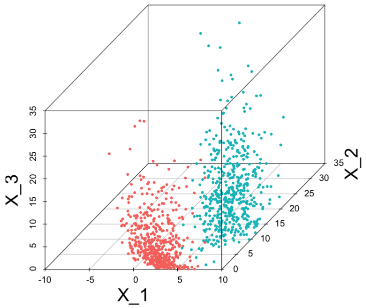
CMStatistics 2020

December 2020

How to find hidden groups in data in a probabilistic framework with vine copulas?

3-dimensional scatter plots of simulated data on x-scale with 2 groups and 500 observations per group





Outline

1. Introduction to mixture models and model-based clustering
2. Vine copula mixture models (vcmm)
3. Model selection and parameter estimation in vcmm
4. Model-based clustering with vcmm
5. Results

Mixture models and model-based clustering

- Formalize the notion of **clusters** (groups, components) through their probability distribution,
- An observation $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top \rightarrow$ realization of a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$,
- Data $\rightarrow$ d-dimensional n observations coming from k hidden components,
- $\pi_j \rightarrow$ mixture weight of the jth component (for $j = 1, \ldots, k$, $\pi_j \in (0,1), \sum_j^k \pi_j = 1$),
- $g_j(.\,; \boldsymbol{\psi}_j) \rightarrow$ density of the jth component for $j = 1, \ldots, k$,
- The **density of a finite mixture model** for $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ at $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$:

$$g(\boldsymbol{x}; \boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x}; \boldsymbol{\psi}_j). \tag{1}$$

Use vine copulas to have flexible component densities for continuous data

- **Bivariate copula**: Distribution on $[0,1]^2$ with univariate uniform margins.
- **Vine copula**: Distribution on $[0,1]^d$ with univariate uniform margins, where bivariate copulas and a nested set of trees determine dependence structure [Aas et al., 2009].

Sklar's Theorem [Sklar, 1959]

A d-dimensional density can be decomposed into products of marginal densities and bivariate copula densities assuming absolute continuity of random variables:

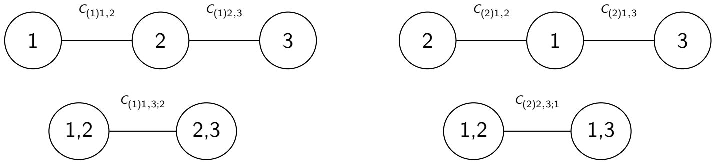
$$g(\boldsymbol{x}) = c\big(F_1(x_1), \ldots, F_d(x_d)\big) \cdot f_1(x_1) \cdots f_d(x_d), \quad \boldsymbol{x} \in \mathbb{R}^d. \qquad (2)$$

- $g_j(.\,;\psi_j)$ in Equation (3) $\rightarrow$ **simplified vine copula** with **parametric** marginal distributions and pair copulas.

$$g(\boldsymbol{x};\boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x};\psi_j). \qquad (3)$$

Selection and parameter estimation problems

Total number of component k is known



(a) First component (b) Second component

Selection problems for each component $j = 1, \ldots, k$:

1. The marginal distributions $\mathcal{F}_j = \{F_{1(j)}, \ldots, F_{d(j)}\}$,
2. The vine tree structure $\mathcal{V}_j$,
3. The pair copula families $\mathcal{B}_j(\mathcal{V}_j)$.

Accordingly, **parameter estimation** problems:

4. The marginal parameters $\gamma_j(\mathcal{F}_j)$,
5. The pair copula parameters $\boldsymbol{\theta}_j(\mathcal{B}_j(\mathcal{V}_j))$.

1. Marginal distribution selection via AIC

Assume a partition of d-dimensional n observations, $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ for $i = 1, \ldots, n$, into k components.
For $j = 1, \ldots k$:

- $n_j \rightarrow$ the total number of observations in the jth component,
- $\boldsymbol{x}_{(j)i_j} = (x_{(j)i_j,1}, \ldots, x_{(j)i_j,d})^\top \rightarrow$ the observations belonging to the jth component for $i_j = 1, \ldots n_j$ $\Big(\bigcup_{\forall(j,i_j)} \boldsymbol{x}_{(j)i_j} = \bigcup_{\forall i} \boldsymbol{x}_i, \quad \sum_{j=1}^{k} n_j = n\Big)$,
- $\boldsymbol{x}_{p(j)} = (x_{(j)1,p}, \ldots, x_{(j)n_j,p})^\top \rightarrow$ pth variable in the jth component for $p = 1, \ldots d$.

For each candidate for marginal distribution on the variable $\boldsymbol{x}_{p(j)}$, first find the parameters that maximize the (weighted) log-likelihood $\ell(\hat{\gamma}_{p(j)})$, then **marginal distribution with the lowest AIC**, $\hat{F}_{p(j)}$, given by $-2 \cdot \ell(\hat{\gamma}_{p(j)}) + 2 \cdot |\hat{\gamma}_{p(j)}|$ is selected.

2/3. Vine tree structure and pair copula families selection via a greedy algorithm

For $j = 1, \dots k$ and $p = 1, \dots d$, obtain the u-data of the jth component by applying probability integral transformation $\hat{\boldsymbol{u}}_{p(j)} = \hat{F}_{p(j)}(\boldsymbol{x}_{p(j)}; \hat{\boldsymbol{\gamma}}_{p(j)})$.

Vine tree structure selection $\mathcal{V}_j$: proceed sequentially tree by tree, starting from the tree level 1 and find the maximum spanning tree at each tree level. Edge weight is the absolute Kendall's τ value between the pair of nodes forming the edge [Dißmann et al., 2013].

Pair copula family selection $\mathcal{B}_j(\mathcal{V}_j)$: After learning the vine tree structure, for a parametric pair copula associated with an edge e in $\mathcal{V}_j$, first estimate the parameters that maximize the (weighted) log-likelihood $\ell(\hat{\boldsymbol{\theta}}_{(j)e_a,e_b;D_e})$. Later choose the copula family with the lowest AIC [Dißmann et al., 2013].

4/5. Estimate the parameters with the ECM algorithm [Meng and Rubin, 1993]

- The log-likelihood $\ell(\boldsymbol{\eta})$ of the given data $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ for $i = 1, \ldots, n$:

$$\ell(\boldsymbol{\eta}) = \log \prod_{i=1}^{n} g(\boldsymbol{x}_i; \psi) = \log \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x}_i; \psi_j). \quad (4)$$

- The **unknown** true assignment of the observations to a component,
- Introduce latent variables $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,k})^\top$, where

$$z_{i,j} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \text{ belongs to the } j\text{th component}, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

and $\sum_{j=1}^{k} z_{i,j} = 1$ for $i = 1, \ldots n$.

The **complete data log-likelihood** $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ of the complete data $\boldsymbol{y}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top$:

$$\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x}) = \log \prod_{i=1}^{n} \prod_{j=1}^{k} [\pi_j \cdot g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j)]^{z_{i,j}} = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j). \quad (6)$$

Iterate over E- and CM-steps

The complete data log-likelihood

$$\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log g_j(\boldsymbol{x}_i; \psi_j)$$

The **E-step** $\rightarrow$ Given the observed data and current parameter estimates, calculate the conditional expectation of $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$.
The **CM-steps** $\rightarrow$ Maximize the expected complete data log-likelihood from the E-step over the set of parameters.

The complete data log-likelihood

$$\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log g_j(\boldsymbol{x}_i; \psi_j)$$

Our steps at the $(t+1)$th iteration:

1. ***E-step*** *(Posterior probabilities)*

$$r_{i,j}^{(t+1)} = \frac{\pi_j^{(t)} g_j(\boldsymbol{x}_i; \psi_j^{(t)})}{\sum_{j=1}^{k} \pi_j^{(t)} g_j(\boldsymbol{x}_i; \psi_j^{(t)})} \quad \text{for} \quad i = 1, \dots n \quad \text{and} \quad j = 1, \dots k. \tag{7}$$

2. ***CM-step 1*** *(Mixture weights)*

$$\pi_j^{(t+1)} = \arg\max_{\pi_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log \pi_j \quad \text{for} \quad j = 1, \dots k. \tag{8}$$

A *closed form solution* exists.

The complete data log-likelihood

$$\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log g_j(\boldsymbol{x}_i; \psi_j)$$

3. ***CM-step 2*** *(Marginal parameters)*:

$$\gamma_j^* = \arg\max_{\gamma_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log g_j(\boldsymbol{x}_i; \gamma_j, \boldsymbol{\theta}_j^{(t)}) \quad \text{for} \quad j = 1, \ldots k. \tag{9}$$

   A closed form solution does not exist → a numeric solution.

4. ***CM-step 3*** *(Pair copula parameters)*:

$$\boldsymbol{\theta}_j^* = \arg\max_{\boldsymbol{\theta}_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log g_j(\boldsymbol{x}_i; \gamma_j^{(t+1)}, \boldsymbol{\theta}_j) \quad \text{for} \quad j = 1, \ldots k \tag{10}$$

   A closed form solution does not exist → a numeric solution.

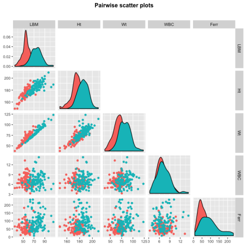
Model-based clustering with vcmm

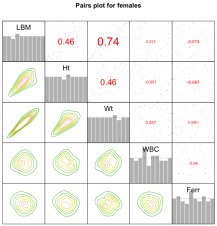
vcmm-1 and vcmm-2

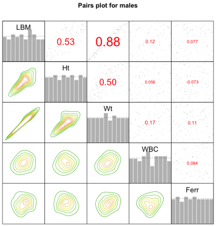
1: **Input**: d-dimensional n observations to cluster and total number of clusters k.
2: **Output**: A clustering partition of the observations $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$, estimated model components and parameters of the jth cluster $\hat{\mathcal{F}}_j$, $\hat{\gamma}_j$, $\hat{\mathcal{V}}_j$, $\hat{\mathcal{B}}_j(\hat{\mathcal{V}}_j)$, $\hat{\theta}_j\left(\hat{\mathcal{B}}_j(\hat{\mathcal{V}}_j)\right)$, $\hat{\pi}_j$ for $j = 1, \ldots, k$.
3: **Step 1**: Initial clustering assignment (via k-means)
4: **Step 2**: Initial marginal distribution and vine copula model (Markov tree) selection
5: A univariate margin $\rightarrow$ normal, student's t with d.o.f. 3, logistic, gamma, log-normal, log-logistic distribution
6: Pair copula families $\rightarrow$ parametric with a single parameter, BB1, BB6, BB8 copulas and their rotations
7: **Step 3**: Iterative parameter estimation with the ECM algorithm
8: **if** $\frac{\ell(\boldsymbol{\eta}^{(t+1)}) - \ell(\boldsymbol{\eta}^{(t)})}{\ell(\boldsymbol{\eta}^{(t)})} < 0.00001$ **then**
9: **break**
10: **end if**
11: **Step 4**: Temporary clustering assignment
12: **Step 5**: Temporary marginal distribution and vine copula model (full tree) selection
13: A univariate margin $\rightarrow$ Same as the line 5
14: Pair copula families $\rightarrow$ Same as the line 6
15: **Step 6**: Final clustering assignment

vcmm captures non-Gaussian components in AIS data better than other methods

Australian Institute of Sport (AIS) data with 13 measurements made on 102 male(green) and 100 female(red) athletes. Here a subset of 5 variables with non-Gaussian and asymmetric patterns:





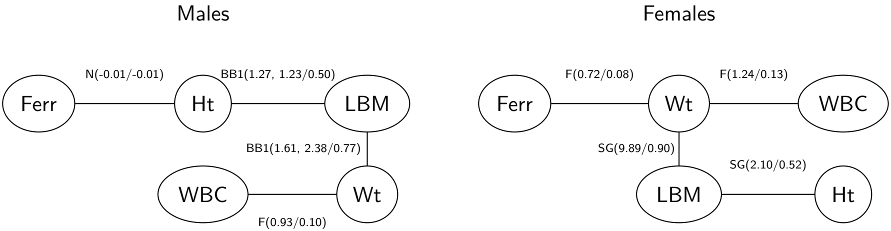


Comparison of clustering algorithm performances for a seed:

Model	vcmm-1	vcmm-2	GMM	skew normal	t	skew t	k-means
Misclassification rate	**0.03**	0.07	0.09	0.42	0.30	0.37	0.34
BIC	**6903**	6939	7062	7158	7100	7121	-
Number of free parameters	43	44	30	51	42	52	-

Interpretation of the dependence structure within the clusters of AIS data

The first tree level of the estimated vine copula model for females and males, where N: Gaussian, C: Clayton, SG: Survival Gumbel, and F: Frank copula (estimated parameter(s)/Kendall's τ):



Summary

- A vine copula mixture model, called vcmm, for continuous data allowing all types of vine tree structures, parametric pair copulas and margins,
- Assuming the number of components in the data is known, a data-driven approach for remaining selection problems and the ECM algorithm for parameter estimation,
- Due to its parametric nature, a nice interpretation of the structure of the data,
- A new model-based clustering algorithm that incorporates realistic interdependence structures of clusters and shows how the dependence structure varies within clusters of the data,
- Capture the non-Gaussian components hidden in the data better than the standard clustering methods,
- More in the paper.

Thank you for your attention!

References

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

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013).
Selecting and estimating regular vine copulae and application to financial returns.
Computational Statistics and Data Analysis, 59:52 – 69.

Meng, X.-L. and Rubin, D. B. (1993).
Maximum Likelihood Estimation via the ECM Algorithm: A General Framework.
Biometrika, 80(2):267–278.

Sklar, A. (1959).
Fonctions de Répartition à n Dimensions et Leurs Marges.
Publications de L'Institut de Statistique de L'Université de Paris, (8):229–231.