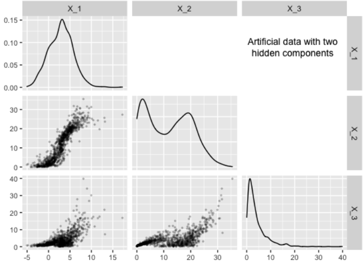
Vine copula mixture models and clustering for non-Gaussian data

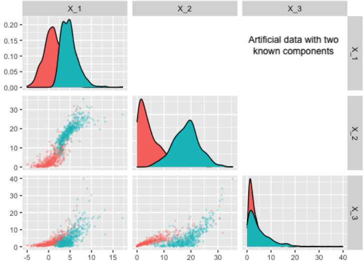
Econometrics and Statistics, 22, 136-158.

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

ETH-UCPH-TUM Workshop on Graphical Models

How to find hidden groups in data in a probabilistic framework?





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

Outline

1. Mixture models

2. Vine copulas

3. Vine copula mixture models (VCMM)

4. Model-based clustering with VCMM

Mixture models

- 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 $j = 1, \ldots, k$, $\pi_j \in (0,1), \sum_j^k \pi_j = 1$),
- $g_j(.\,; \psi_j) \rightarrow$ 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 $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$:

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

How to select densities of each component?

Previous works

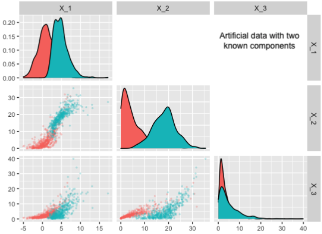
The density of a finite mixture model for $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ at $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$:

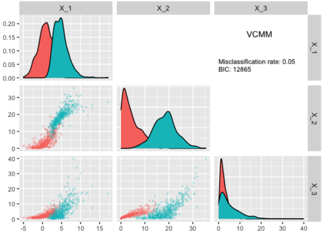
$$g(\boldsymbol{x}; \boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x}; \boldsymbol{\psi}_j). \tag{2}$$

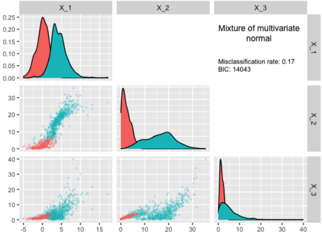
$g_j(.\,; \boldsymbol{\psi}_j) \rightarrow$ multivariate Gaussian distribution, multivariate t distribution, their skewed formulations, copulas.

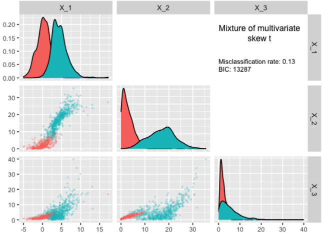
not flexible enough in representing different asymmetric or/and tail dependencies for different pairs of variables

Need a flexible framework to represent different asymmetric and tail dependencies for pairs of variables: vine copulas









Sklar's Theorem:the density of a d-dimensional distribution can be decomposed into the product of its univariate marginal densities and the associated copula density

- $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$,
- the joint cumulative distribution function (cdf) F,
- the univariate marginal distributions $F_1, \ldots F_d$ (absolutely continuous) and densities $f_1, \ldots, f_d$,
- a copula density c of the random vector $\boldsymbol{F} = \left(F_1(X_1), \ldots F_d(X_D)\right)^\top \in [0, 1]^d$,

Thanks to Sklar's theorem [Sklar, 1959], the d dimensional joint density g can be written as:

$$g(\boldsymbol{x}) = c\left(F_1(x_1), \ldots, F_d(x_d)\right) \cdot f_1(x_1) \cdots f_d(x_d), \quad \boldsymbol{x} \in \mathbb{R}^d. \quad (3)$$

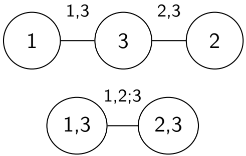
Vine copulas can be considered as a generalization of multivariate Gaussian distributions parametrized in terms of $d-1$ correlations and $\frac{(d-1)(d-2)}{2}$ partial correlations

Avoid the constraint of positive definiteness with an alternative parametrization of the correlation matrix by sequences of correlations and partial correlations being algebraically independent [Joe, 2014], e.g., for $d = 3$,

- $(\rho_{12}, \rho_{13}, \rho_{23;1}) \in (-1, 1)^3$,
- $(\rho_{12}, \rho_{23}, \rho_{13;2}) \in (-1, 1)^3$,
- $(\rho_{13}, \rho_{23}, \rho_{12;3}) \in (-1, 1)^3$.

Vine copulas' building plan is given by a vine tree structure and uses bivariate copulas that are algebraically independent glued together by conditioning

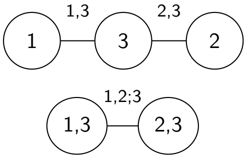
Example of a 3-dimensional vine tree structure that can represent the correlation matrix of a 3-dimensional Gaussian distribution with $(\rho_{13}, \rho_{23}, \rho_{12;3}) \in (-1, 1)^3$



Vine copulas can approximate many multivariate distributions

Thanks to Sklar's theorem [Sklar, 1959], the d dimensional joint density g can be written as:

$$g(\boldsymbol{x}) = c\big(F_1(x_1), \ldots, F_d(x_d)\big) \cdot f_1(x_1) \cdots f_d(x_d), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{4}$$



$$\begin{aligned} g(x_1, x_2, x_3; \psi) =& c_{1,3}\Big(F_1(x_1; \gamma_1), F_3(x_3; \gamma_3); \boldsymbol{\theta}_{1,3}\Big) \cdot c_{2,3}\Big(F_2(x_2; \gamma_2), F_3(x_3; \gamma_3); \boldsymbol{\theta}_{2,3}\Big) \\ & \cdot c_{1,2;3}\Big(F_{1|3}(x_1|x_3; \gamma_1, \gamma_3, \boldsymbol{\theta}_{1,3}), F_{2|3}(x_2|x_3; \gamma_2, \gamma_3, \boldsymbol{\theta}_{2,3}); \boldsymbol{\theta}_{1,2;3}, x_3\Big) \\ & \cdot f_1(x_1; \gamma_1) \cdot f_2(x_2; \gamma_2) \cdot f_3(x_3; \gamma_3). \end{aligned} \tag{5}$$

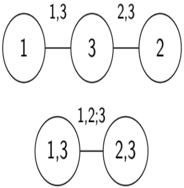
Use vine copulas to have flexible component densities for continuous data

The density of a finite mixture model for $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ at $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$:

$$g(\boldsymbol{x}; \boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x}; \boldsymbol{\psi}_j). \tag{6}$$

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

Many selection problems exist in vine copula mixture models



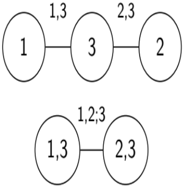
$$
\begin{aligned}
g(x_1, x_2, x_3; \psi) = & c_{1,3}\left(F_1(x_1; \gamma_1), F_3(x_3; \gamma_3); \boldsymbol{\theta}_{1,3}\right) \\
& \cdot c_{2,3}\left(F_2(x_2; \gamma_2), F_3(x_3; \gamma_3); \boldsymbol{\theta}_{2,3}\right) \\
& \cdot c_{1,2;3}\left(F_{1|3}(x_1|x_3; \gamma_1, \gamma_3, \boldsymbol{\theta}_{1,3}), F_{2|3}(x_2|x_3; \gamma_2, \gamma_3, \boldsymbol{\theta}_{2,3}); \boldsymbol{\theta}_{1,2;3}\right) \\
& \cdot f_1(x_1; \gamma_1) \cdot f_2(x_2; \gamma_2) \cdot f_3(x_3; \gamma_3).
\end{aligned}
$$

The total number of components k hidden in the data $\rightarrow$ **known**

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 $\mathcal{B}_j(\mathcal{V}_j)$.

Many parameter estimation problems exist in vine copula mixture models



$$
\begin{aligned}
g(x_1, x_2, x_3; \psi) = & c_{1,3}\left(F_1(x_1; \gamma_1), F_3(x_3; \gamma_3); \boldsymbol{\theta}_{1,3}\right) \\
& \cdot c_{2,3}\left(F_2(x_2; \gamma_2), F_3(x_3; \gamma_3); \boldsymbol{\theta}_{2,3}\right) \\
& \cdot c_{1,2;3}\left(F_{1|3}(x_1|x_3; \gamma_1, \gamma_3, \boldsymbol{\theta}_{1,3}), F_{2|3}(x_2|x_3; \gamma_2, \gamma_3, \boldsymbol{\theta}_{2,3}); \boldsymbol{\theta}_{1,2;3}\right) \\
& \cdot f_1(x_1; \gamma_1) \cdot f_2(x_2; \gamma_2) \cdot f_3(x_3; \gamma_3).
\end{aligned}
$$

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

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

Follow a data-driven approach – more in the paper: [Sahin and Czado, 2022]

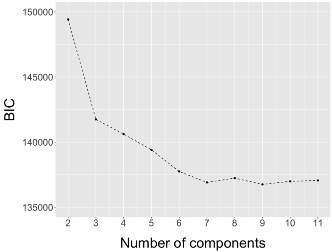
1. Marginal distribution selection via BIC
2. Vine tree structure and pair copula families selection via a greedy algorithm
3. Estimate the parameters with ECM algorithm [Meng and Rubin, 1993]

Different DGPs and real data sets are used for clustering benchmarking

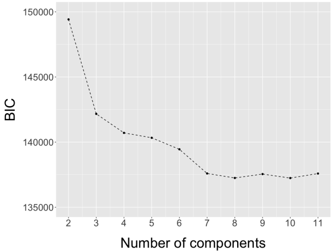
- Assign the observations to the clusters (components) with the final posterior probabilities:
  $\boldsymbol{x}_i \in \mathcal{C}_{j^*} \iff j^* = \underset{j=1,\ldots,k}{\arg\max}\, r_{i,j}^{(s+1)}$ for $i = 1, \ldots, n$.
- Clustering quality comparison based on the **BIC** and **misclassification rate**,
- the VCMM's **sensitivity, stability**, and **computational cost**,
- **4 DGPs**: three variables, two clusters with known labels, 100 or 500 observations in each cluster, replicate 100 times,
- **3 real data sets**,
- the R package `vineclust` [Sahin, 2021].

Real data set 3: the VCMM's optimal number of component selection is not stable based on the BIC

- Sachs Protein data analyzed by [Sachs et al., 2005],
- Continuous logarithmized levels of 11 phosphorylated proteins and phospholipids in 6161 individual cells, subjected to general and specific molecular interventions,
- [Zhang and Shi, 2017] work with the two-component Gaussian mixture copula Bayesian network.



(a) VCMM with k-means



(b) VCMM with hierarchical clustering

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 a modification of the ECM algorithm for parameter estimation.
- 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.
- Clustering benchmarking analyses with the VCMM.
- The `R` package `vineclust` to run simulations and the model-based clustering algorithm,.
- Future research for the number of component selection, variable selection, parsimonious VCMM, stable initial partition, and the inclusion of discrete variables.

Thank you for your attention!

References

Joe, H. (2014).
Dependence modeling with copulas.
CRC press.

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

Prates, M. O., Cabral, C. R. B., and Lachos, V. H. (2013).
mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions.
Journal of Statistical Software, 54(12):1–20.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005).
Causal protein-signaling networks derived from multiparameter single-cell data.
Science, 308(5721):523–529.

Sahin, Ö. (2021).
vineclust.

Sahin, Ö. and Czado, C. (2022).
Vine copula mixture models and clustering for non-Gaussian data.
Econometrics and Statistics, 22:136–158.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016).
Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models.
R Journal, 8(1):289–317.

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.

Zhang, Q. and Shi, X. (2017).
A mixture copula bayesian network model for multimodal genomic data.
Cancer Informatics, 16:1176935117702389.