Copula-based directional dependence networks for multivariate data

Subho Majumdar

School of Statistics, University of Minnesota

Table of contents

Outline

Background

- Copulas model joint dependency structure in financial and survival data.
  - **Pros**: Distribution-free: invariant under monotone transformations on data, can model different kinds of tail-dependence
  - **Cons**: Multivariate parametric models insufficient
- **Copula selection** done by
  - AIC/BIC [5]
  - minimizing distance to empirical copula function [4]
- **Multivariate dependency**
  - Vine copulas provide a conditional pairwise dependency tree, but decomposition of pdf is not unique.
  - [1] did pairwise copula on multivariate genetic data, but no copula selection.

Overview of work

- Derived depth-based (robust?) estimators of copula parameters
- Choose parametric copula families to model different tail dependencies
- Calculate ML or depth-based estimate, choose best-fitting copula
- Repeat for all pairs of variables in multivariate data: gives dependency structure
- Application on two datasets
- Future work

Outline

Defnition

- $C : [0,1]^2 \to [0,1]$, uniformly distributed marginals
- (**Sklar's theorem**) Any bivariate cdf H on random variables X, Y, with marginals F, G respectively, there always exists a copula function C s.t.

$$H(x, y) = C(F(x), G(y))$$

i.e. $C(u, v) = H(F^{-1}(u), G^{-1}(v))$ for $(u, v) \in [0,1]^2$.

Two types of Copula:

- **Inplicit copula**- cdf given as an integral, no closed form
- **Explicit copula**- cdf has closed form

Example

- **Gaussian copula**: Copula parameter is $\rho \in [-1, 1]$, the correlation coefficient. **Models low tail-dependency**.

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2xy + y^2}{2(1-\rho^2)}\right\} dxdy$$

- *t* **copula**: Parameters $\rho \in [-1, 1], \nu =$ degree of freedom. **Jointly heavy-tailed**.

$$C_{\rho,\nu}(u, v) = \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left\{-\frac{x^2 - 2xy + y^2}{2(1-\rho^2)}\right\}^{-(\nu+2)/2} dxdy$$

$\nu \geq 30 \equiv$ Gaussian copula.

Example

- **Clayton copula**: Parameter $\delta \in (0, \infty)$. $\delta \to 0$ means independence, $\delta \to \infty$ means perfect dependence. **Models heavy dependency on left tail**.

$$C_\delta(u, v) = (u^{-\delta} + v^{-\delta} + 1)^{-1/\delta}$$

- **Gumbel copula**: Parameter $\delta \in [1, \infty)$. $\delta = 1$ means independence, $\delta \to \infty$ means perfect dependence. **Models heavy dependency on right tail**.

$$C_\delta(u, v) = \exp\left[-\left\{(-\log u)^\delta + (-\log v)^\delta\right\}^{1/\delta}\right]$$

Note: Both Clayton and Gumbel copulae are used for positive dependence. For negative dependence, 90° (or 270°) rotated versions are used, which have the same expressions for $C_\delta(u, v)$, but $\delta \in (-\infty, 0)$ for Clayton and $\in (-\infty, 1]$ for Gumbel copula, respectively.

Likelihood depth

- Due to [2][3]. Consider n iid observations from $Z \sim f_\theta$ with $\theta \in \Theta \subset \mathbb{R}^p$, likelihood function $L(\theta, z) = f(\theta, z)$.
- **Nonfit** A parameter $\theta \in \Theta$ is nonfit wrt data $(z_1, ..., z_n)$ when $\exists \quad \theta' \neq \theta \in \Theta$ s.t. $L(\theta', z_i) > L(\theta, z_i)$, or equivalently for log-likelihood: $l_i(\theta') > l_i(\theta)$ for $i = 1, , , n$.
- **Likelihood depth** at θ is the minimum proportion of observations that need to be deleted from the data to make θ a nonfit.
- **Tangent Likelihood depth** Same as likelihood depth under regularity conditions.

$$d_T(\theta, \mathbf{z}) = \frac{1}{n} \inf_{u \neq \mathbf{0_p}} \#\{i : u^T \nabla l_i(\theta) \leq 0\}$$

Maximum Likelihood Depth Estimator (MLDE)

- A parameter value (not necessarily unique) that maximizes the likelihood depth over parameter space Θ:

$$\hat{\theta}_d \in \arg\max_{\theta \in \Theta} d_T(\theta, \mathbf{z})$$

- Under regularity conditions, depth at each $\theta \in \Theta$ i.e. $d_T(\theta, \mathbf{z})$ converges uniformly to its population analogue $d_T(\theta, P)$, where P is a valid probability measure. Same holds for the point with maximum depth, given it is unique in the population.

- May give biased estimate of true parameter.

Outline

Copula parameter estimation

1. Based on Kendall's τ
2. **Maximum Likelihood estimates**
3. **Unbiased estimate based on Max. likelihood depth**

Provides biased estimates for true parameters. For 4 types of copulas relation between MLDEs and true parameters determined numerically:

- **Cubic equation** for Gaussian ($c_3 = 1.2222, c_2 = 3.6434, c_1 = -1.4215, c_0 = 0.0004, \rho_0 = 0.461$) and t-copula ($c_0, c_1, c_2, c_3, \rho_0$ depend on df ν):

$$\rho = \begin{cases} \text{sign}(\hat{\rho}_d)\left[c_3|\hat{\rho}_d|^3 + c_2|\hat{\rho}_d|^2 + c_1|\hat{\rho}_d| + c_0\right] & \text{if } |\hat{\rho}_d| > \rho_0 \\ 0 & \text{if } |\hat{\rho}_d| \leq \rho_0 \end{cases}$$

- **Linear equation** for Clayton ($a_0 = -0.5302, a_1 = 0.7163$) and Gumbel copula ($a_0 = 0.02, a_1 = 0.706$):

$$\delta = a_0 + a_1\hat{\delta}_d$$

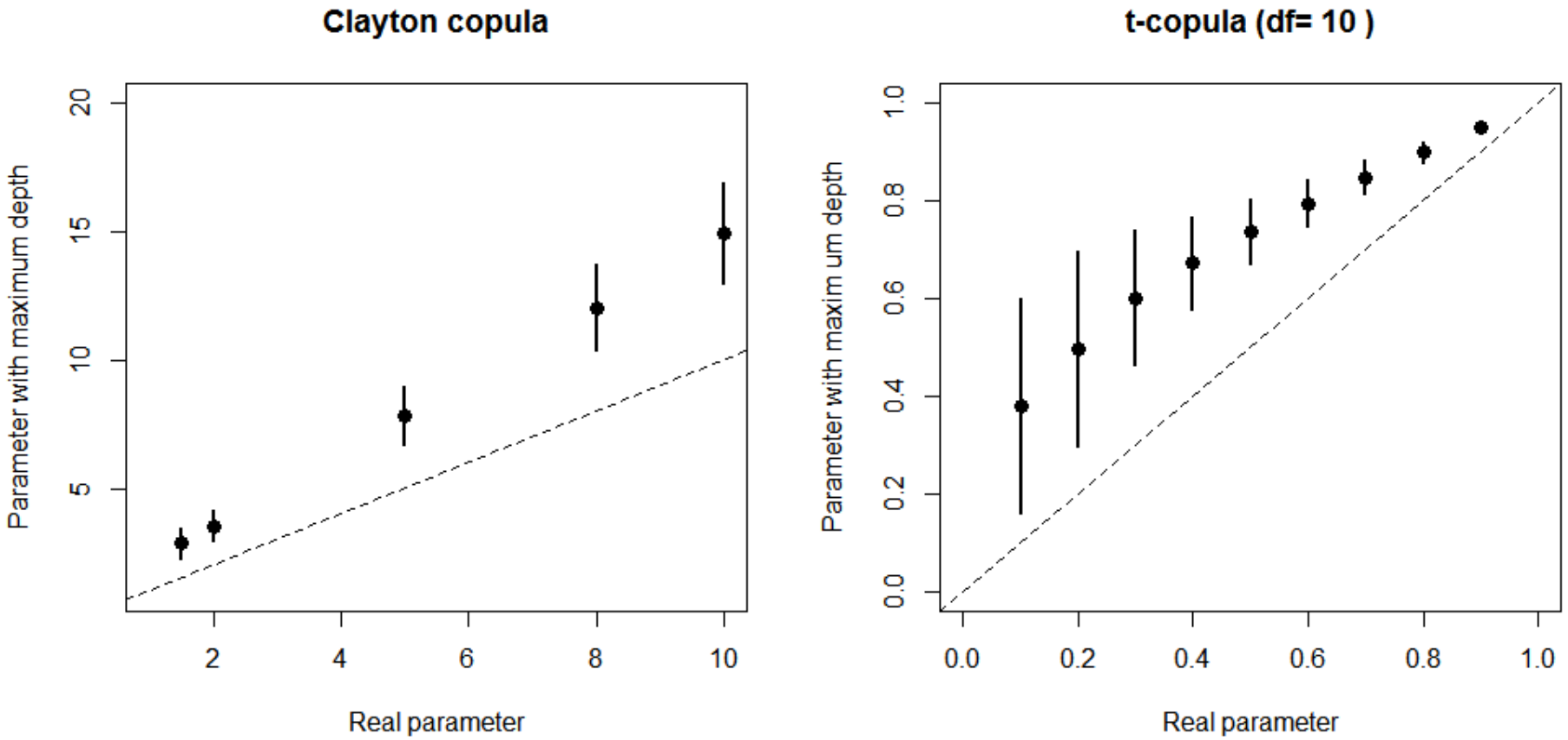

Figure : Mean and standard deviations of simulated parameters with max depth

- **Empirical copula** For data $\mathbf{Z}_1, ..., \mathbf{Z}_n$ with $\mathbf{Z}_i = (X_i, Y_i)$

$$C_e(u, v) = \frac{1}{n} \sum_{i=1}^{n} I(U_i \leq u, V_i \leq v)$$

  where $\mathbf{U} = F_n(\mathbf{X}), \mathbf{V} = G_n(\mathbf{Y})$ are pseudosamples obtained from the data using marginal empirical distributions F_n, G_n respectively.

- Euclidean distance of a copula C from empirical copula:

$$d(C, C_e) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left[C\left(\frac{i}{n}, \frac{j}{n}\right) - C_e\left(\frac{i}{n}, \frac{j}{n}\right) \right]^2}$$

- For the four of copula families, estimate parameters by ML or depth-based method, then choose the copula that minimizes the above distance.

Constructing the graph

1. Test for independence between a variable pair using test based on asymptotic normality of Kendall's τ (n = sample size):

$$\hat{\tau} \sim AN\left(0, \frac{2(2n+5)}{9n(n-1)}\right)$$

2. Apply algorithm on next page.

3. At the end of the algorithm we end up with two graphs, $\mathbf{C}_M$ and $\mathbf{C}_D$, giving best-fitting copulae, obtained by the two respective methods, for each pair of variables.

Algorithm 1 Algorithm to obtain pairwise copula dependence network

procedure CopNetwork(data matrix $\mathbf{D} \in \mathbb{R}^{n \times p}$)
  Set $i = 1, j = 1$.
  top:
  Set $\mathbf{X} = i^{th}$ column of $\mathbf{D}$, $\mathbf{Y} = j^{th}$ column of $\mathbf{D}$.
  Check for independence of $\mathbf{X}$ and $\mathbf{Y}$ using test above.
  if Independent **then**
    goto *update*
  else
    if Kendall's $\tau > 0$ **then**
      $S =$ {Gaussian, t, Clayton, Gumbel}
    else
      $S =$ {Gaussian, t, rotated Clayton, rotated Gumbel}
    Select $C_{ij,M} \in S$ as the best fitting copula, parameter estimated by ML method.
    Select $C_{ij,D} \in S$ as the best fitting copula, parameter estimated by depth-based method.
  update:
  if $j = p$ **then**
    if $i = p$ **then Stop**
    else
      Set $i \leftarrow i + 1, j \leftarrow i$, **goto** *top*
  else
    Set $j \leftarrow j + 1$, **goto** *top*

Outline

Data example: Breast Cancer data

- From UCI Machine Learning Reopsitory: `https://archive.ics.uci.edu/ml/datasets/Breast+Tissue`
- Impedance measurements from 106 breast tissue samples.
- 9 measurement variables (I0, PA500, HFS, DA, Area, A/DA, Max IP, DR, P) and a class variable specifying the class of Breast Cancer the patient has (6 or 4 classes).

Variable name	Description
I0	Impedivity (ohm) at zero frequency
PA500	Phase angle at 500 KHz
HFS	High-frequency slope of phase angle
DA	Impedance distance between spectral ends
AREA	Area under spectrum
A/DA	Area normalized by DA
MAX IP	Maximum of the spectrum
DR	Distance between I0 and real part of the maximum frequency point
P	Length of the spectral curve

Table : Description of measurement variables in Breast Cancer data

Breast Cancer data: analysis

- We ignore the class variable due to small sample sizes in each class and obtain the networks from the measurement variables only.
- 28 significantly dependent variable pairs among 36 possible ones.
- Variables except the two phase angle related variables: HFS and PA500, are all dependent on one another.
- Most of these dependencies are symmetric, and heavy-tailed (t-copula) as per the ML method, but light-tailed as per the depth-based method.

Breast Cancer data: networks

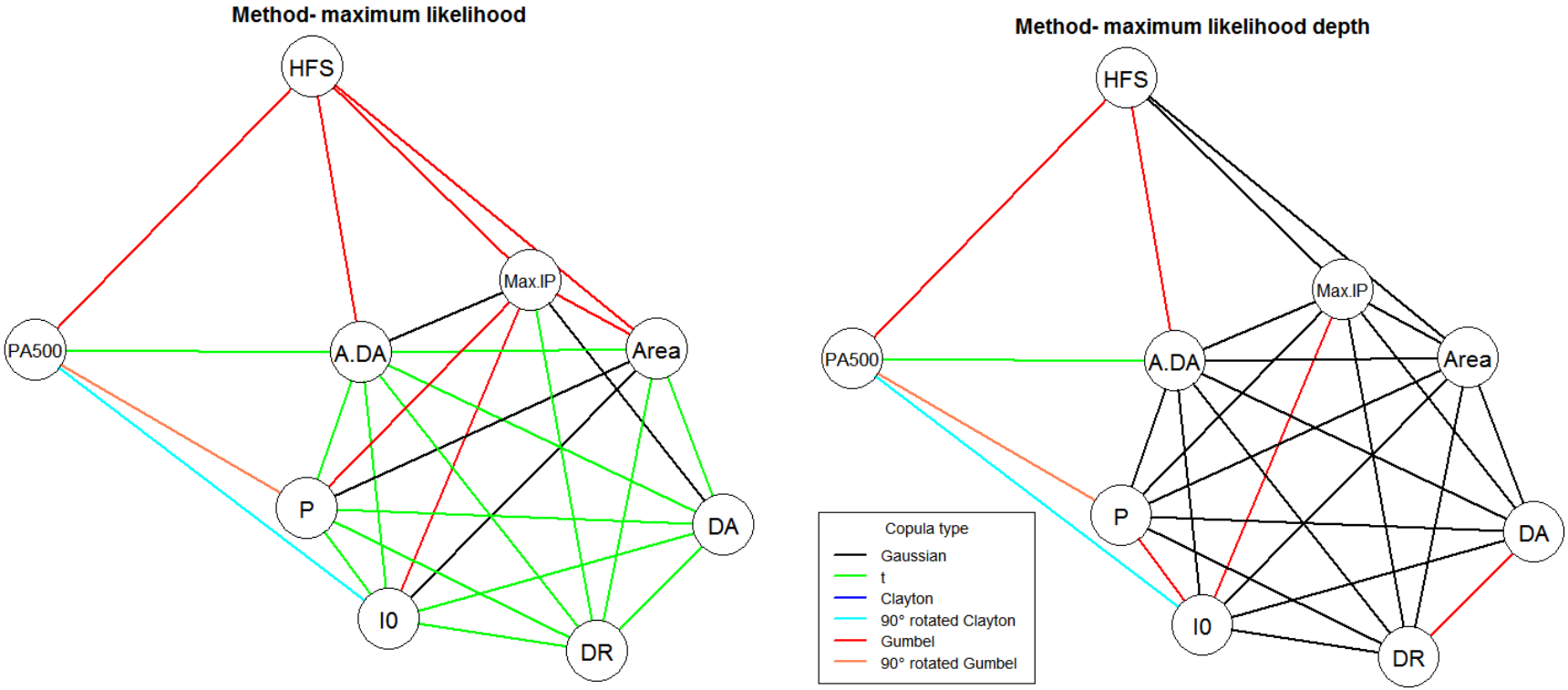

Figure : Graph of Breast Cancer variables obtained by best-fitting (top) ML copula, (Bottom) MLD copula

Data example: Cardiotocography (CTG) data

- Measurements relating to 2126 fetal cardiotocograms: `https://archive.ics.uci.edu/ml/datasets/Cardiotocography`
- 21 predictor variables: Variables 1-7 give numerical measurements (grp A), 8-11 are about variability of cardiograms with respect to time (grp B), Variables 12-21 give measurements of heart-rate histograms (grp C).
- DS, the sixth variable has most of its values set at 0, so we exclude it from our analysis.

CTG data: Variable group-specific analysis

- Over 90% of connections between the 20 variables (176 among 190 possible) found significant in the initial screening for dependence.
- Instead of plotting the graph we analyze the dependence structures within and between the 3 variable groups.
- 3 within-group (AA, BB, CC) and 3 between-group (AB, AC, BC) interactions.

Interaction	Indep.	Gaussian	t	Clayton	rot. Clayton	Gumbel	rot. Gumbel
AA	2	0	0	7	1	0	5
BB	1	2	1	0	1	0	1
CC	2	11	9	2	7	5	9
AB	1	4	3	2	11	1	2
AC	3	16	2	16	18	2	3
BC	5	7	3	7	11	2	5

Table : Summary of best-fitting copulae for within (Top 3) and between (bottom 3) group dependencies in CTG data

CTG data: Dependence structure for different fetal classes

- Within-group dependencies for the 3 fetal classes are also compared.
- Networks are plotted for groups A and B in 3 sample classes.
- For variable group C, the summary of copulae fit between the 45 variable-pairs is summarized in table below. Highlights include a high amount of independence in suspect class and high asymmetric dependencies in pathologic class.

Sample class	Indep.	Gaussian	t	Clayton	rot. Clayton	Gumbel	rot. Gumbel
Normal	4	13	8	4	5	7	4
Suspect	11	8	8	3	3	10	2
Pathologic	3	9	2	2	11	7	11

Table : Best-fitting copulae by ML method for group C variables in CTG data

CTG data: networks

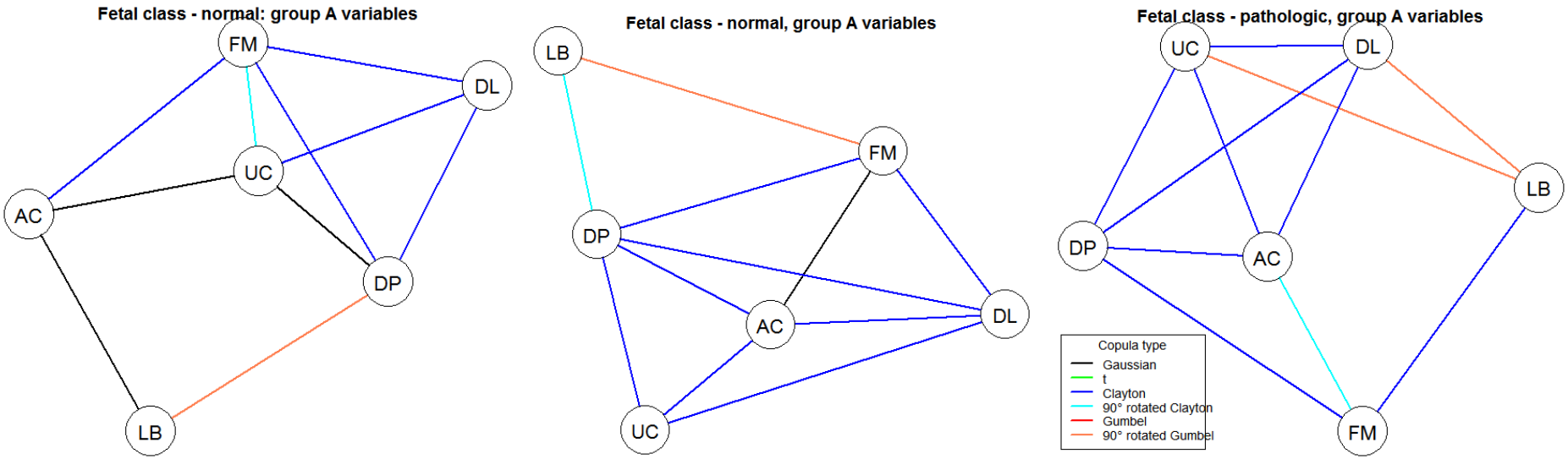

Figure : Graph of group A variables for 3 classes

CTG data: networks

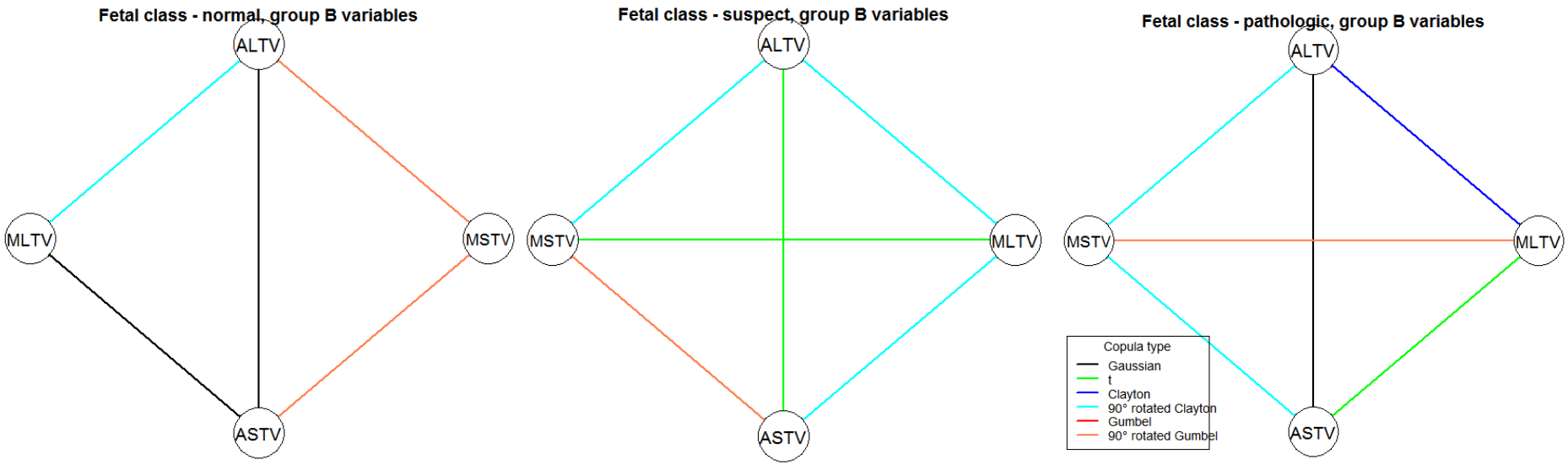

Figure : Graph of group B variables for 3 classes

Summary and future work

- Formulation of a distribution-free method to analyze the nature of dependencies between pairs of variables in a multivariate dataset.
- Application on two real datasets

Future works include:

- **Using conditional bivariate copula** for each variable-pair to eliminate effect of other variables
- **Detailed simulation studies** to compare between the two methods of copula parameter estimation
- **Analyze genetic data** and compare the methodology with other known methods

Acknowledgement

I thank my adviser Prof. Snigdhansu Chatterjee for his guidance and valuable inputs throughout the project.

References

KIM, J.-M., JUNG, Y.-S., SUNGUR, E., HAAN, K.-H., PARK, C., AND SOHN, I.
A copula method for modeling directional dependence of genes.
BMC Bioinformatics 9 (2008).

MIZERA, I.
On depth and deep points: A calculus.
The Annals of Statistics 30 (2002), 1681–1736.

MIZERA, I., AND MÜLLER, C.
Location-scale depth.
Journal of the American Statistical Association 99 (2004), 949–989.

ROMANO, C.
Calibrating and simulating copula functions: An application to the Italian stock market.
Tech. rep., Capitalia, November 2002.

SMITH, M.
Modelling sample selection using Archimedean copulas.
Econometrics Journal 6 (2003), 99–123.

THANK YOU!