Copula Models for Dependent Data Analysis

Yihao Deng

Department of Mathematical Sciences

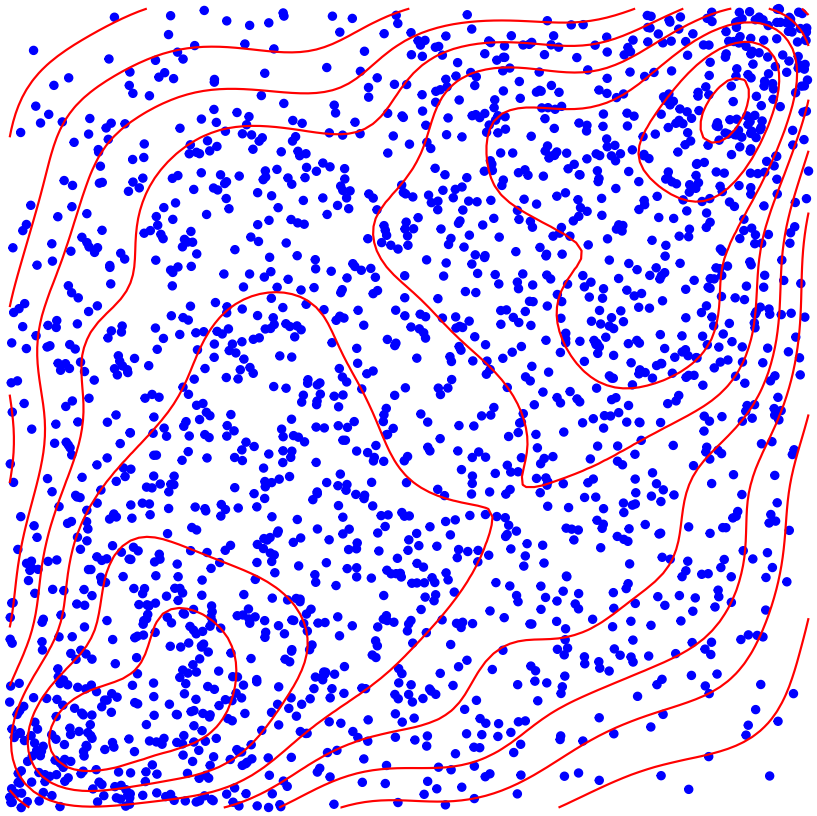
Purdue University Fort Wayne

December 5, 2019

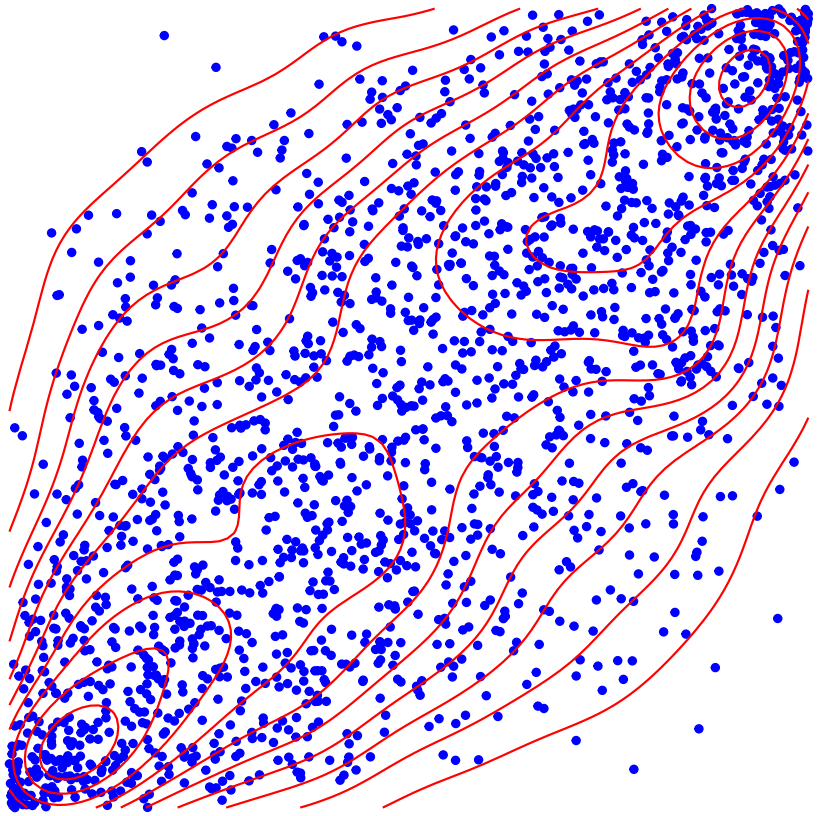
Dependent Data

- Data collected from family members (twins)
- Return of stocks from the same sector
- Health measures from the same person (height, weight, blood pressure, cholesterol levels, etc.)

Interest lies in the relation among the variables. The most popular measure is **correlation** coefficient, assuming variables are normally distributed.

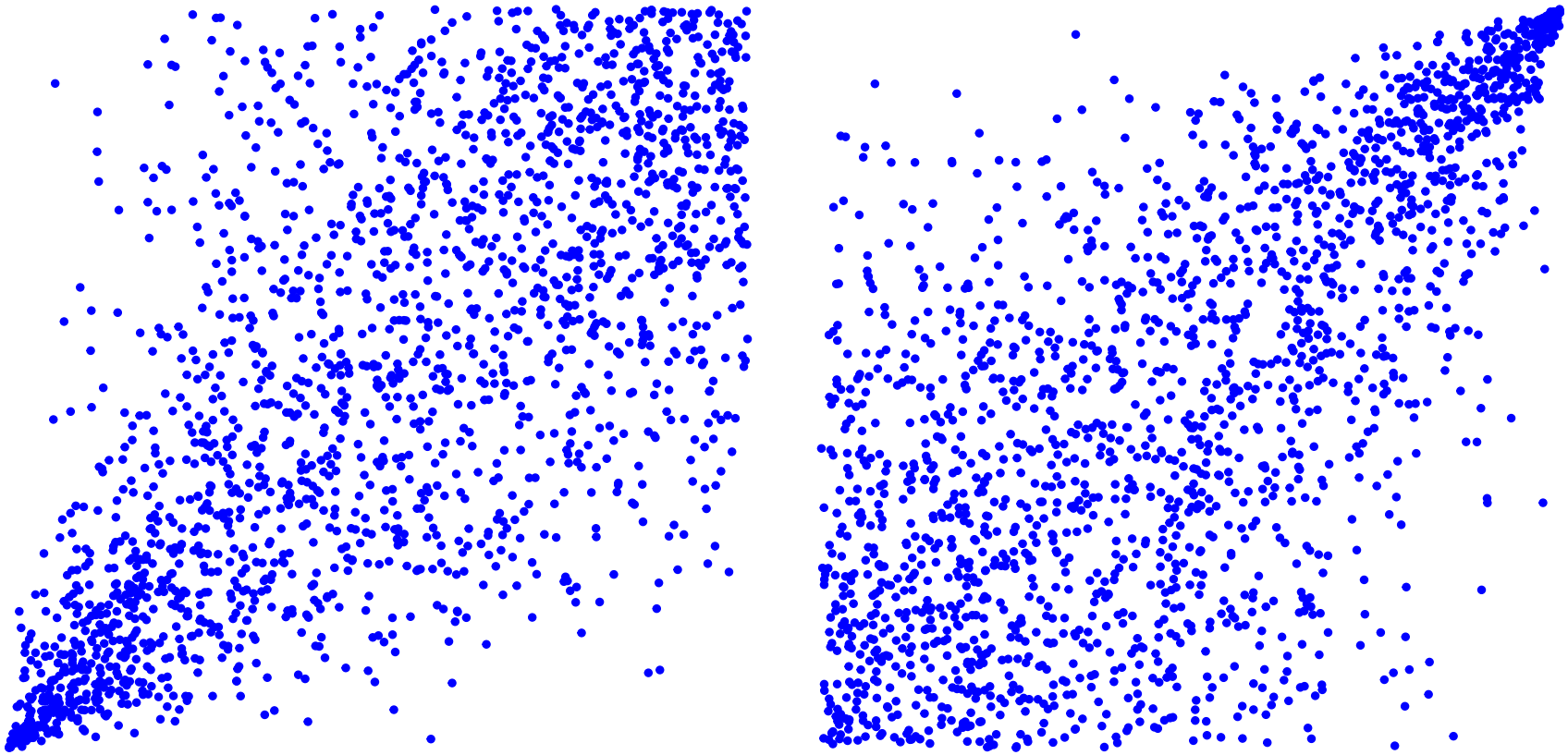


$\rho = 0.4$



$\rho = 0.7$

What If?



same dependence measure as in the previous normal case ($\rho = 0.7$).

Copula

A copula C is a joint cumulative distribution function (cdf) where all marginals are uniform on $(0, 1)$.

Suppose that $Y_i \sim F_i$ continuous, then $F_i(Y_i) \sim U(0, 1)$. The joint cdf H of $Y_1, \ldots, Y_k$ can be written as

$$H(y_1, \ldots, y_k) = C(F_1(y_1), \ldots, F_d(y_k))$$

Let $U_i = F_i(Y_i)$, then $Y_i = F_i^{-1}(U_i)$. The copula is given by

$$C(u_1, \ldots, u_k) = H(F_1^{-1}(u_1), \ldots, F_k^{-1}(u_k); \boldsymbol{\theta}) \tag{1}$$

Copula Examples

- Independence Copula:

$$C(u_1, u_2, \ldots, u_k) = u_1 \times u_2 \times \cdots \times u_k$$

- Gaussian Copula:

$$C(u_1, u_2, \ldots, u_k) = \Phi_k(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_k); \mathbf{R})$$

where

$$\Phi_k(z_1, \ldots, z_k) = \int_{-\infty}^{z_k} \cdots \int_{-\infty}^{z_1} \frac{1}{(2\pi)^{\frac{k}{2}} |\mathbf{R}|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{t}'\mathbf{R}^{-1}\mathbf{t}} dt_1 \ldots dt_k$$

and

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Copula Examples (continued)

- Archimedean Copula:

$$C(u_1, u_2, \ldots, u_k) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2) + \cdots + \psi^{-1}(u_k); \theta)$$

  - Clayton family: $\psi = (1+t)^{-1/\theta}$
  - Gumbel familty: $\psi = e^{-t^{1/\theta}}$
  - Frank family: $\psi = -\frac{1}{\theta}\ln(1 + e^{-t}(e^{-\theta} - 1))$
  - Joe family: $\psi = 1 - (1 - e^{-t})^{1/\theta}$

- Gaussian Copula:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1k} & \rho_{2k} & \rho_{3k} & \cdots & 1 \end{bmatrix}$$

  which should be positive definite.

- Archimedean Copula: Exchangeable dependence structure. Or the depenence among all pairs of variables are assumed to be the same.

Modeling of Marginal Distribution

The random variable Y is often related to some covariates $(X_1, X_2, \ldots, X_p$, or in matrix notation $\mathbf{X})$, where the mean $E(Y)$ is linked to the covariates via $E(Y) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$. Therefore, the effect of the covariates can be incorporated into copula models as

$$U_i = F_i(Y_i; g^{-1}(\mathbf{X}_i\boldsymbol{\beta}))$$

Examples

- Probit function: $u_i = \Phi\left(\dfrac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\hat{\sigma}}\right)$
- Logistic function: $u_i = \left(1 + e^{-\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\hat{\sigma}}}\right)^{-1}$

Maximum Likelihood Estimation

As soon as we formulate the marginal distributions and dependence structure, the log-likelihood function is simply

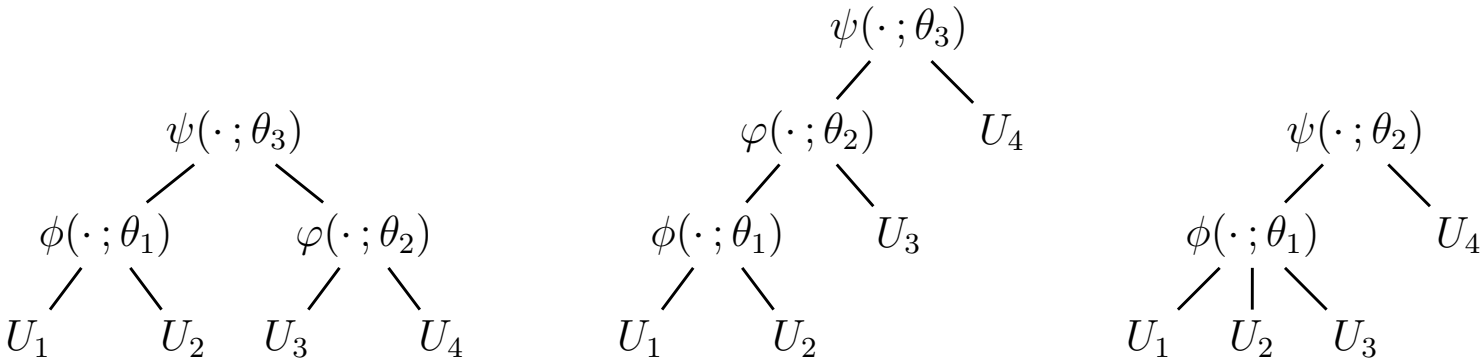
$$\ell = \sum \ln(c(u_1, \ldots, u_k; \boldsymbol{\beta}, \boldsymbol{\theta}))$$

where $c(u_1, \ldots, u_k)$ is the corresponding copula density function.

Optimization needs to be done numerically. R function `optim` and Python function `minimize` will be helpful.

Hierarchical Archimedean Copula

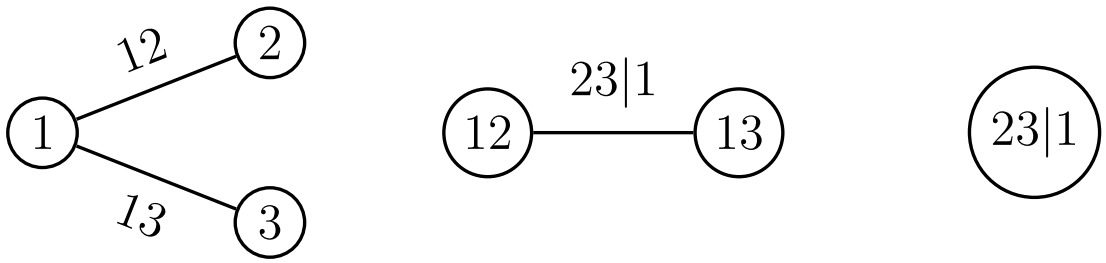
Recall that the dependence in Archimedean copulas is assumed to be the same everywhere. Hierarchical Archimedean copula (HAC) was proposed to account for more complicated dependence structures.



Examples of HAC with four random variables

Vine Copula

A more flexible copula model is vine copula, which builds the dependence hierarchy using "pair copulas".



Tree 1 Tree 2 Tree 3

Example of vine construction with three random variables

Family Data

Blood samples from members of 22 families were collected, erythrocyte adenosine triphosphate (ATP) levels were determined before and after storage at 4°C in acid citrate dextrose solution for 21 days.

famID	Member	Gender	Age	pre-ATP	post-ATP	y
2	Mother	0	62	4.43	2.49	1
2	Father	1	62	3.72	1.79	1
2	Son	1	24	4.18	1.49	1
2	Son	1	41	4.81	2.84	1
2	Daughter	0	31	4.42	2.04	1
2	Daughter	0	38	3.65	1.17	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Source: Dern R. and Wiorkowski J. (1969).

Modeling Discrete Binary Responses

By introducing continuous uniform variables U_i, we categorize Y_i as follows:

$$Y_i = \begin{cases} 1 & \text{if} \quad 0 \leq U_i \leq \eta_i \\ 0 & \text{if} \quad \eta_i < U_i \leq 1 \end{cases}$$

where $\eta_i = g^{-1}(\mathbf{X}\boldsymbol{\beta})$.

We may now model the dependence among continuous variables U_i rather than discrete variables Y_i. And the log-likelihood function to be maximized is

$$\ell = \prod P(Y_i = \{0/1\})$$

The dependence among family members is assumed to be

$$\mathbf{R} = \begin{array}{c} \\ M \\ F \\ Ch_1 \\ Ch_2 \\ Ch_3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} M & F & Ch_1 & Ch_2 & Ch_3 & \dots \end{array} \\ \begin{pmatrix} 1 & \gamma & \rho_1 & \rho_1 & \rho_1 & \dots \\ \gamma & 1 & \rho_2 & \rho_2 & \rho_2 & \dots \\ \rho_1 & \rho_2 & 1 & \alpha & \alpha & \dots \\ \rho_1 & \rho_2 & \alpha & 1 & \alpha & \dots \\ \rho_1 & \rho_2 & \alpha & \alpha & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Evaluation of log-likelihood function is computational intensive since it involves multivariate integration over hyper-rectangle.

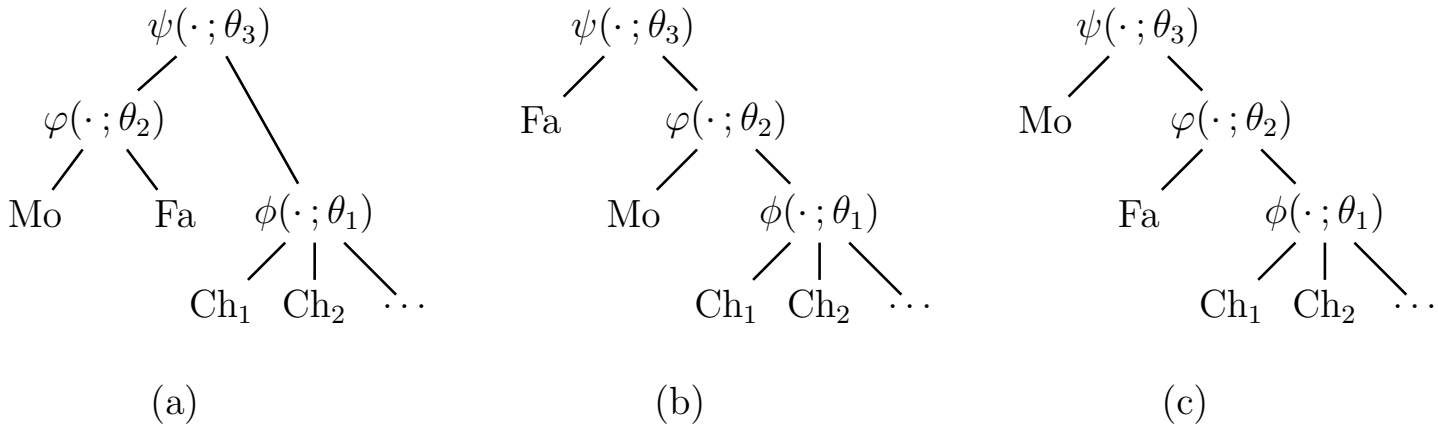
Analysis Result

Parameter	Estimate	S.E.	p-value
Intercept	12.466	1.490	< 0.001
Gender	-0.638	0.556	0.251
Pre-ATP	-2.517	0.292	< 0.001
γ	0.281	0.398	0.480
ρ_1	0.518	0.274	0.059
ρ_2	0.208	0.376	0.580
α	0.568	0.289	0.050

log-likelihood $= -39.195$ with logit link function

HAC Modeling

- Selecting hierarchical dependence structures:



- Selecting Archimedean copula families at each level.

  For simplicity, I used same family for all levels to avoid incompatible issue.

Analysis Result

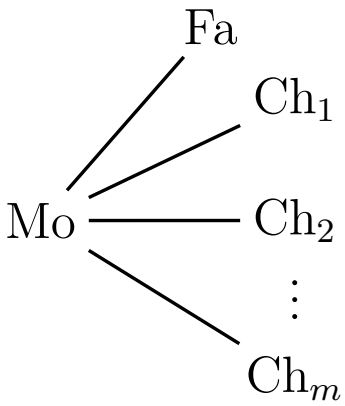
Hierarchy (b) turns out to be the best model, and Frank family is selected.

Parameter	Estimate	S.E.	p-value
Intercept	12.666	3.257	< 0.001
Gender	-0.804	0.548	0.143
Pre-ATP	-2.561	0.671	< 0.001
θ_3	1.316	1.681	0.434
θ_2	2.190	2.610	0.402
θ_1	4.464	3.577	0.212

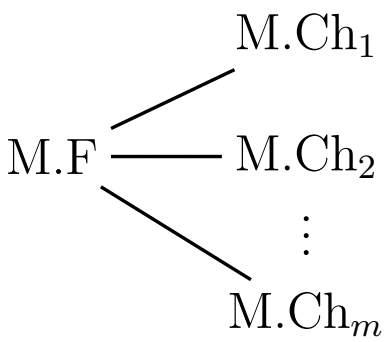
log-likelihood $= -39.588$ with logit link function

Vine Copula Modeling

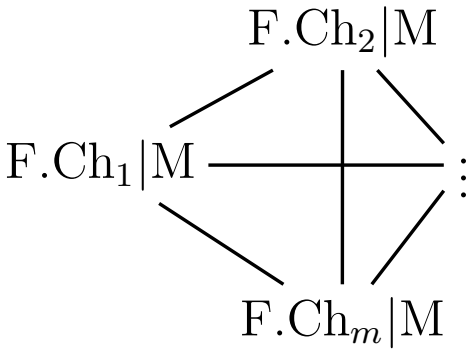
- Pairing processes:



Tree 1



Tree 2



Tree 3

- Selecting pair copulas: find the maximized log-likelihood from all possible combinations.

Analysis Result

Joe family and independent copula are selected for pair copulas.

Parameter	Estimate	S.E.	p-value
Intercept	14.348	3.663	< 0.001
Gender	-0.738	0.566	0.193
Pre-ATP	-2.902	0.738	< 0.001
θ_{12}	1.584	0.689	0.021
θ_{13}	1.837	0.885	0.038
$\theta_{23\|1}$	—	—	—
$\theta_{3\|12}$	2.705	2.163	0.211

log-likelihood $= -38.138$ with logit link function

Thank you!

Selected References

1. Joe H. Multivariate models and dependence concepts. London: Chapman & Hall. 1997.
2. Nelsen R. An introduction to copulas (2nd edition). New York: Springer. 2006.
3. Joe H. Dependence modeling with copulas. Boca Raton: CRC Press. 2015.
4. Kurowicka D, Joe H. Dependence modeling: vine copula handbook. Singapore: World scientific. 2011.
5. Dißmann J, Brechmann E, Czado C, Kurowicka D. Selecting and estimating regular vine copulae and application to financial returns. Computational statistics and data analysis 2013; 59: 52–69.
6. Panagiotelis A, Czado C, Joe H. Pair copula constructions for multivariate discrete data. Journal of the American statistical association 2012; 107: 1063–1072.
7. Panagiotelis A, Czado C, Joe H, Stöber J. Model selection for discrete regular vine copulas. Computational statistics and data analysis 2017; 106: 138–152.

9. Hofert M, Kojadinovic I, Maechler M, Yan J. copula: Multivariate Dependence with Copulas. 2018. R package version 0.999-19.1, https://CRAN.R-project.org/package=copula.
10. Schepsmeier U, Stöber J, Brechmann E, Graeler B, Nagler T, Erhardt T. VineCopula: Statistical Inference of Vine Copulas. 2017. https://CRAN.R-project.org/package=VineCopula.
11. Deng Y. Modeling binary familial data using Gaussian copula. Communications in statistics – theory and methods 2016; 46: 10097–10102.
12. Deng Y, Chaganty N.R. Hierarchical Archimedean copula models for the analysis of binary familial data. Statistics in medicine 2018; 37: 590–597.
13. Dern R, Wiorkowski J. Studies on the preservation of human blood. IV. The hereditary component of pre- and poststorage erythrocyte adenosine triphosphate levels. Journal of laboratory & clinical medicine 1969; 73: 1019–1029.