Penalized Gaussian Copula mixtures for mixed mode data

Panagou Foteini, Karlis Dimitrios

Department of Statistics
Athens University of Business & Econimics

fwtpanagou@aueb.gr, karlis@aueb.gr

Athens, 26 Oct 2021

Mixtures of Gaussian copulas for mixed mode data

For mixed data sets the joint probability is not always easy to be found. Gaussian copula offers flexibility to describe the dependencies between different types of variables.
Multivariate Gaussian copula is defined as

$$C^N(u_1, u_2, ..., u_p; R) = \boldsymbol{\Phi}_p(\boldsymbol{\Psi}(u_1), ..., \boldsymbol{\Psi}(u_p); R)$$

where: $\boldsymbol{\Phi}_p$ is the p.m.f. of a standard p-variate normal distribution,$\boldsymbol{\Psi}(\cdot) = \boldsymbol{\Phi}^{-1}(\cdot)$ is the inverse p.d.f. of a standard univariate normal distribution.
The correlation matrix R of k continuous & ℓ discrete variables can be split into blocks based on the type of variables:

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right]$$

where e.g. R_{11} is the correlation matrix of the k ciontinuous variables.
Through Choleski factor and angles reparametrization, $R = LL'$, R and Θ are correlated:

$$L_{11} = 1,\ L_{ii} = \prod_{u=1}^{i-1} \sin\theta_{ui},\ L_{ij} = \cos\theta_{ij} \prod_{u=1}^{i-1} \sin\theta_{uj},\ \theta_{ij} = cos^{-1}\left\{ \frac{L_{ij}}{\prod_{u=1}^{i-1} \sin\theta_{uj}} \right\},\ i < j$$

$\theta_{ij} = 0,\ i \geq j$ and angles are measured in radians. We require $\theta_{ij} \in (0, \pi]$ so that the Choleski factor is unique.

Penalized Gaussian Copula

$$\ell = \sum_{g=1}^{G} \log\left\{f(\boldsymbol{x}; \gamma_{\boldsymbol{g}}, \boldsymbol{\Psi})\right\} = \sum_{g=1}^{G}\sum_{t=1}^{k} \log(f_t(\boldsymbol{x}; \gamma_{tg})) - \sum_{g=1}^{G}\sum_{t=1}^{k} \log\left\{\phi(\Phi^{-1}(u_{tg}))\right\}$$

$$+ \sum_{g=1}^{G} \log\left\{\phi_k\left(\Psi(u_1), \cdots, \Psi(u_k)\,; \mathbf{0}, \boldsymbol{R_{11g}}\right)\right\}$$

$$+ \sum_{g=1}^{G} \log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\, \Phi_\ell\left(\Psi(u_{k+1}), \cdots, \Psi(u_{k+l})\,;\, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}\right)\right\} + \boxed{\lambda \sum_{g=1}^{G}\sum_{j=1}^{p(p-1)/2} \sin^2\theta_{gj}}$$

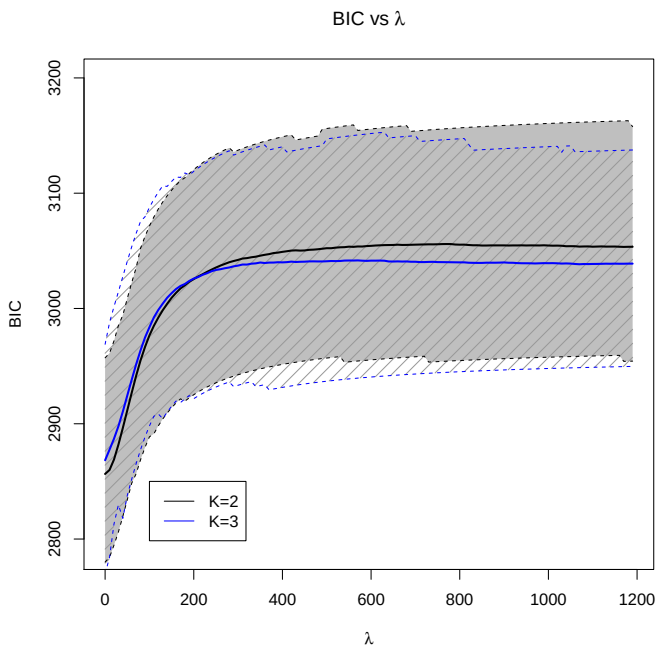
$\boldsymbol{\mu_g} = R_{21g}R_{11g}^{-1}\left[\Psi(u_1), \cdots, \Psi(u_k)\right]'$, $\boldsymbol{\Sigma_g} = R_{22g} - R_{21g}R_{11g}^{-1}R_{12g}$, $u_i = F_i(x_i; \gamma_{\boldsymbol{gi}})$, $p = k + \ell$ and λ takes values into a grid ($\lambda > 0$).

For $\lambda = 0$ the approach is equivalent to the fully parametrized model without any constrain to the correlation matrix.
For $\lambda \to \infty$ the log-likelihood shrinks to the independent model where all θ's are equal to $\pi/2$.

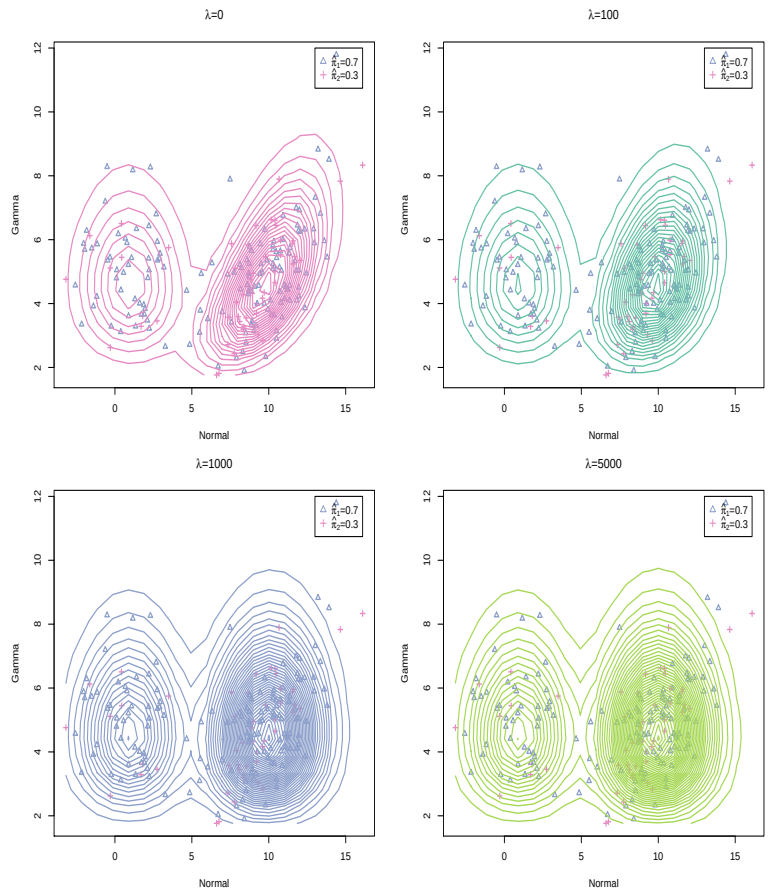
Simulation Study

Assume a 4-variate mixed mode dataset (Normal,Gamma, Poisson & Bernoulli) of length $n = 200$, $G = 2$ components ($\pi_1 = 0.7$).

Model	K=2	K=3
$\lambda = 0$	**2856.46**	2868.34
$\lambda = 10$	**2859.94**	2877.58
$\lambda = 50$	**2911.58**	2924.45
$\lambda = 100$	**2976.66**	2983.74
$\lambda = 500$	3052.28	**3040.95**
$\lambda = 1000$	3054.61	**3039.28**
$\lambda = 5000$	3043.43	**3028.10**



BIC for various values of λ



Contours for various values of λ