Bayesian Copula-based Latent Variable Models

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Luis Nieto-Barajas (ITAM, Mexico) and Ruyi Pan (University of Toronto)

ICSDS, Sevilla, December 2025

Copulas: The Joys

- Copulas are mathematical devices used to model dependence between random variables regardless of their marginals.
- If $Y_1, Y_2, \ldots, Y_K$ are continuous r.v.'s with cdfs $F_1, F_2, \ldots, F_k$, there is an unique copula $C : [0,1]^K \to [0,1]$ that links the joint cdf with the marginal ones (Sklar's Theorem),

$$F(t_1, \ldots, t_k) = C(F_1(t_1), \ldots, F_k(t_k)).$$

- Copulas are useful for data fusion/integration because they lead to coherent joint models, even when the marginals are in different families (e.g., Gaussian, Poisson, Student, etc) or of different types (e.g, discrete, continuous).
- Copulas unlock information contained in the dependence part of the distribution (second-order) that complements the information in the marginals.
- Simply put, copulas allow us to extend statistical methods beyond the use of a multivariate Gaussian or Student.

Motivation

- Parametric copulas can capture a wide range of dependence patterns for lower values of k, $k \leq 2, 3$.
- In higher dimensions, say $k > 5$ a parametric family (usually indexed by a parameter of dimension < 3) will not be able to capture complex dependencies
- Vine copulas offer additional flexibility but can be difficult to fit.
- Today: Bayesian nonparametric mixtures of copulas to extend their flexibility in higher dimensions

Big Picture

- Bayesian estimation for copula models is desirable as the posterior incorporates uncertainty due to marginals and dependence structure (Levi and Craiu, 2018)
- We consider Bayesian nonparametric mixtures of Archimedean copulas in which the mixing distribution is the Poisson-Dirichlet process, introduced by Pitman and Yor (1997).
- It extends the range of dependence patterns that can be modeled.
- In the case of heterogeneous populations, it clusters the sample based on information contained in the marginals AND the dependence structure.

Archimedean copulas

- An Archimedean family is characterized by a continuous, decreasing and convex generator function ϕ such that $\phi : [0,1] \to \mathbb{R}^+$, $\phi(0) = \infty$, $\phi(1) = 0$.
- The copula with generator ϕ is defined as

$$C(u_1, \ldots, u_p) = \phi^{-1}\{\phi(u_1) + \cdots + \phi(u_p)\}. \quad (1)$$

- Generators ϕ usually belong to families that are parameterized in terms of a single parameter $\theta \in \Theta$
- Archimedean copulas assume that the variables are exchangeable and the mixture will preserve this.

BNP mixtures of copulas

- G has a Poisson-Dirichlet prior with scalar parameters $a \in [0, 1)$, $b > -a$ and mean parameter G_0, denoted as $G \sim \mathrm{PD}(a, b, G_0)$, when

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}(\cdot), \tag{2}$$

  where $\omega_1 = \nu_1$ and $\omega_k = \nu_k \prod_{j<k}(1 - \nu_j)$ for $k = 2, 3 \ldots$, with $\nu_k \overset{\text{ind}}{\sim} \mathrm{Be}(1 - a, b + ka)$ independent of the weights, locations $\theta_k \overset{\text{iid}}{\sim} G_0$ for $k = 1, 2, \ldots$, and δ_θ is the Dirac measure at θ.
- The functional parameter G_0 is known as the centering measure

BNP mixtures of copulas

- Bayesian nonparametric mixture model for Archimedean copulas $C(\mathbf{u} \mid \theta)$ is obtained when using the Poisson-Dirichlet process as the mixing distribution for the parameter θ

$$C(\mathbf{u}) = \int C(\mathbf{u} \mid \theta) G(\mathrm{d}\theta) = \sum_{k=1}^{\infty} \omega_k C(\mathbf{u} \mid \theta_k), \tag{3}$$

- The Bayesian nonparametric mixture copula model can also be defined hierarchically
- For $i = 1, \ldots, n$:

$$\begin{aligned} (U_{1i}, \ldots, U_{pi}) \mid \theta_i &\sim f_C(\mathbf{u}_i \mid \theta_i) \\ \theta_i \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim \mathrm{PD}(a, b, G_0), \end{aligned} \tag{4}$$

where f_C is the copula density.

BNP mixtures of copulas

- The centering measure G_0 has density, g_0, with support in the parameter space Θ.
- Pitman (1995) showed

$$f(\theta_i \mid \boldsymbol{\theta}_{-i}) = \frac{b + am_i}{b + n - 1} g_0(\theta_i) + \sum_{j=1}^{m_i} \frac{n^*_{i,j} - a}{b + n - 1} \delta_{\theta^*_{i,j}}(\theta_i), \qquad (5)$$

with $\boldsymbol{\theta}_{-i}$ being the set of all θ_i's excluding the ith element and $(\theta^*_{i,1}, \ldots, \theta^*_{i,m_i})$ denoting the distinct values in $\boldsymbol{\theta}_{-i}$, each with frequencies $n^*_{i,j}$, for $i = 1, \ldots, n$, $j = 1, \ldots, m_i$.

Posterior sampling - brief discussion

- The posterior conditional distributions for each θ_i are given by

$$f(\theta_i \mid \mathbf{u}, \boldsymbol{\theta}_{-i}) \propto (b+am_i) f(\mathbf{u}_i \mid \theta_i) g_0(\theta_i) + \sum_{j=1}^{m_i} (n^*_{i,j} - a) f_C(\mathbf{u}_i \mid \theta_i) \delta_{\theta^*_{i,j}}(\theta_i)$$

- We initialize the sampler using random draws θ_i from the prior g_0, for $i = 1, \dots, n$.
- Given the chain's state at time t, $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \dots, \theta_n^{(t)})$, compute the unique values $(\theta^*_1, \dots, \theta^*_m)$ in $\boldsymbol{\theta}^{(t)}$ and re-sample each θ^*_j, $j = 1, \dots, m$ from

$$f(\theta_j \mid c.c.) \propto g_0(\theta_j) \prod_{\{i:\theta_i = \theta^*_j\}} f_C(\mathbf{u}_i \mid \theta_j),$$

where $c.c.$ stands for clustering configuration.

Posterior sampling - brief discussion

- Draw $\theta_i^{(t+1)}$, $i = 1, \ldots, n$, from

$$f(\theta_i \mid \mathbf{u}, \mathbf{v}, \boldsymbol{\theta}_{-i}, \boldsymbol{\theta}^\star) = \frac{1}{k_i}\left[\sum_{j=1}^{m_i}(n_{i,j}^* - a) f_C(\mathbf{u}_i \mid \theta_{i,j}^*)\delta_{\theta_{i,j}^*}(\theta_i) + \sum_{j=m_i+1}^{m_i+r}\{(b + a m_i)/r\} f_C(\mathbf{u}_i \mid \theta_j^\star)\delta_{\theta_j^\star}(\theta_i)\right],$$

where $\boldsymbol{\theta}^\star = \{\theta_{m_i+1}^\star, \ldots, \theta_{m_i+r}^\star\} \overset{iid}{\sim} g_0$,

$$k_i = \sum_{j=1}^{m_i}(n_{i,j}^* - a) f_C(\mathbf{u}_i \mid \theta_{i,j}^*) + \sum_{j=m_i+1}^{m_i+r}\{(b + a m_i)/r\} f_C(\mathbf{u}_i \mid \theta_j^\star).$$

Posterior sampling - brief discussion

- The hyperparameters (a, b) are important in determining the number of components in the mixture
- We assign hyperpriors instead of keeping them fixed
- Since $a \in [0, 1)$ and $b \in (-a, \infty)$, the joint hyperprior is factorized as $f(a, b) = f(b \mid a)f(a)$
- $f(a, b) = \text{Gamma}(b + a \mid c_b, d_b)\text{Beta}(a \mid c_a, d_a)$

Additional considerations

- Goodness-of-fit is evaluated using the logarithm of the pseudo-marginal likelihood $LPML = \sum_{i=1}^{n} \log(CPO_i)$ where $CPO_i = f(\mathbf{u}_i \mid \mathbf{u}_{-i})$ and can be estimated using

$$\widehat{CPO}_i = \left[\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f_C(\mathbf{u}_i \mid \theta_i^{(l)})}\right]^{-1}.$$

- The uncertainty about the number of mixture components is quantified by the posterior distribution but... there are multiple cluster configurations with the same number of components $\tilde{m}$.
- In order to select a single clustering configuration and produce further inferences, we used the search algorithm presented in Dahl et al (2022) and implemented in the R package `salso`.

Simulation experiment - Sanity check

Data / Model	AMH	CLA	FRA	GUM	JOE	BSA
AMH	0.34	-0.08	0.02	-1.07	**2.48**	-4.32
CLA	14.75	**147.42**	74.91	78.37	53.23	7.4
FRA	0.13	7.11	**10.00**	1.62	3.09	-3.01
GUM	77.85	199.23	238.08	**276.27**	270.27	110.14
JOE	37.76	75.85	98.33	118.10	**120.91**	55.33

Table: Simulation study: Bivariate data from a mixture of Archimedean copulas. LPML statistics when taking a sample of size $n = 200$ and fitting the five models. Bayesian semiparametric Archimedean copula competing model in the last column.

Simulation experiment - Sanity check

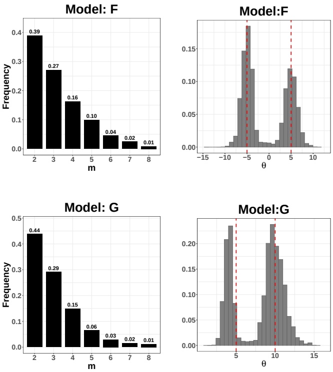
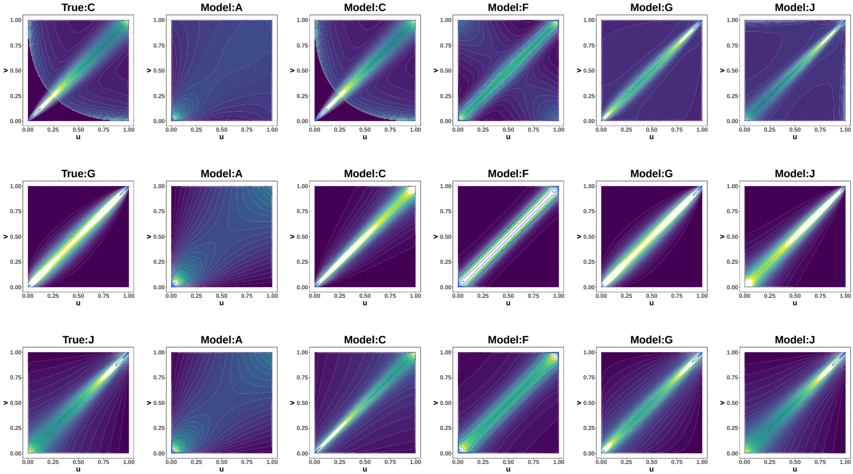


Figure: Simulation: Bivariate data from a mixture of Archimedean copulas with $n = 500$. Posterior distributions when using the same kernel used to sample the data. Vertical dotted lines correspond to the true values.

Simulation experiment - Sanity check



Wine analysis

- We consider the red wine data of **?** which consists of several physicochemical tests of the red variants of $1,599$ Portuguese wines (*Vinho verde*)
- We concentrate on three variables: fixed acidity (X_1), citric acid (X_2) and density (X_3)
- The MCMC was run for 15,000 iterations with a burn-in of 5,000
- The LPML scores are:

Copula	AMH	Clayton	Frank	Gumbel	Joe
LPML	279	604	685	711	413

Wine analysis

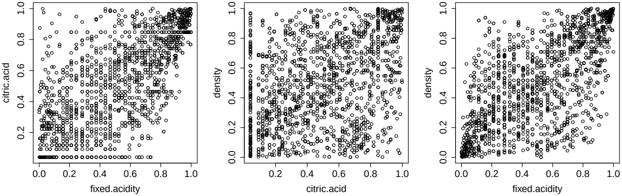
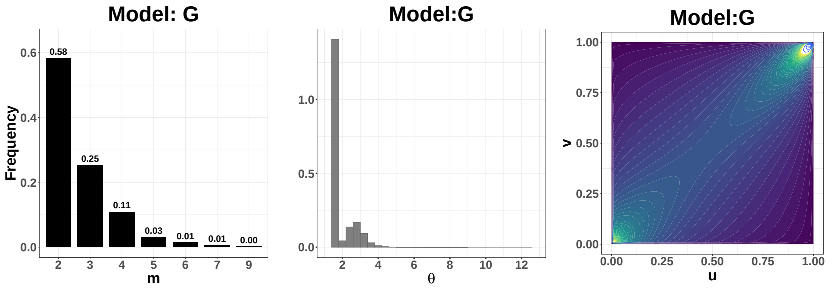


Figure: Red wine data: fixed acidity, citric acid, and density. Pairwise scatterplots of the three variables.

Wine analysis



Wine analysis

- The posterior estimation of Kendall tau for the best-fitting model, Gumbel, is 0.361 with a 95% CI of $(0.24, 0.70)$.
- The CI contains the three empirical Kendall tau values that correspond to pairwise association parameters among the three characteristics of the wine.
- Implementing our clustering selection procedure, we obtain two groups with copula parameters estimated at: 1.52 with a 95% CI $(1.48, 1.56)$ and weight 0.98; and at 6.08 with 95% CI $(4.94, 7.41)$ and weight 0.02.

What I would like to see

- A mixture model where the symmetry is not required $Dep(x_i, x_j) \neq Dep(x_{i'}, x_{j'})$.
- Integration of asymmetric Archimedean copulas in the current framework.
- Can we identify copula families whose mixtures are dense in the space of copulas of dimension d?
- Papers available here: `https://www.utstat.toronto.edu/craiu/Papers/index.html`