Statistical models for and with copulas

Radu Craiu

Department of Statistical Sciences
University of Toronto

BIRS, November 1, 2023

Copulas

- A copula is a function that joins multivariate distribution functions to their 1-dimensional marginals.

- **(Sklar's Theorem - bivariate case)** Let H be a joint distribution function with continuous margins F and G. Then there exists a unique **copula** C s.t. for all $Y_1, Y_2 \in \mathbb{R}$

$$H(Y_1, Y_2) = C(F_1(Y_1), F_2(Y_2))$$

- The copula C binds the marginals into the joint dist'n.
- It characterizes the dependence structure in the model.
- The copula $C : [0,1] \times [0,1] \rightarrow [0,1]$ itself is a bivariate distribution

Copula models

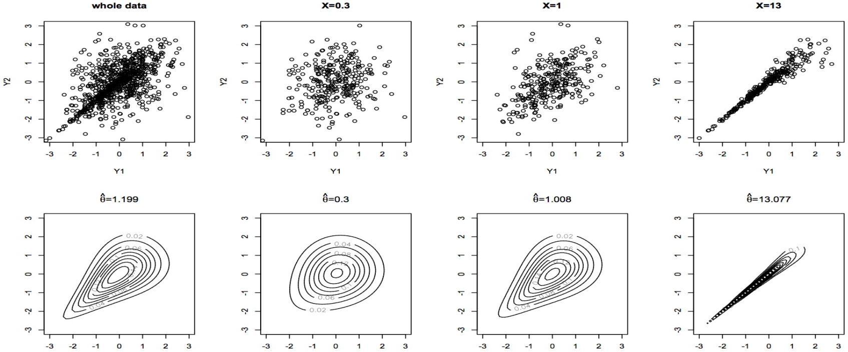
- A copula model involves:
  - Specifying marginal distributions for Y_1 and Y_2, say $F_1(Y_1|\eta)$ and $F_2(Y_2|\zeta)$
  - Specifying a (parametric) copula distribution C_θ
- Estimation can be done in two stages (propagation of errors)
  - First estimate the marginals $F_1(Y|\hat{\eta})$ and $F_2(Y|\hat{\zeta})$
  - Second fit copula $C_{\hat{\theta}}$ to $U_1 = F_1(Y_1|\hat{\eta})$ and $U_2 = F_2(Y_2|\hat{\zeta})$
- Bayesian estimation in one stage!

Copulas: What for?

- Flexible modelling that goes beyond multivariate Gaussianity
- Scientific interest in understanding dependence structure
- Prediction of Y_1 from $Y_2, Y_3, \ldots$
- Imputation of missing data
- Study of extremes (tail dependence, extreme value theory, etc)
- A technique for data fusion

Conditional Copulas

Example It is known that there is a dependence between blood pressure (BP) and body mass index (BMI). What if dependence varies with subject's age? Can we still use copulas to model this dependence?



The Model

- Consider a random sample $\{x_i, y_{1i}, y_{2i}\}_{1\le i\le n}$ and suppose $F_{1|\eta(X)}$ and $F_{2|\zeta(X)}$ are the conditional marginal distributions.
- The conditional copula (CC) model links the conditional joint and the conditional marginal distributions

$$H(Y_1, Y_2|X) \sim C(F_{1|\eta(X)}(Y_1|X), F_{2|\zeta(X)}(Y_2|X)|\theta(X)),$$

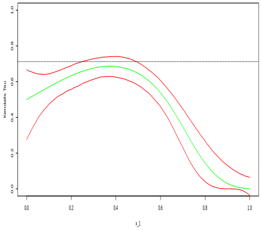
  and $\eta(X)$, $\zeta(X)$, $\theta(X) \in \mathbb{R}^p$ are of interest.
- The CC model can be estimated non-parametrically, semi-parametrically (marginals are parametric, $\theta(X)$ is NP), Bayesian spline model, additive models, or GP with a SIM twist ($\theta(X) = f(\beta^T X)$).

Hmm.. that's "funny"!

- $Y_i|x \sim N(f_i(x), \sigma_i)\ x \in \mathbb{R}^2$
- True marginal means:
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- Copula: $\tau(x) = 0.71$
- Suppose x_2 is not observed so inference is based only on x_1

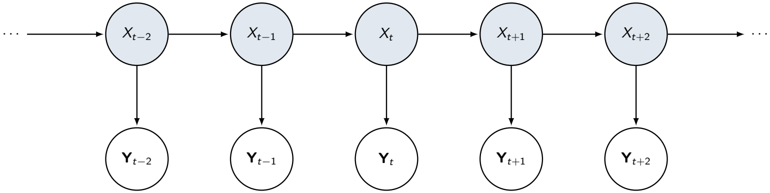
Hmm.. that's "funny"!

- $Y_i | x \sim N(f_i(x), \sigma_i)$ $x \in \mathbb{R}^2$
- True marginal means:
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- Copula: $\tau(x) = 0.71$
- Suppose x_2 is not observed so inference is based only on x_1



Hidden Markov Models: A Primer

- A hidden Markov model (HMM) pairs an observed time series $\{\mathbf{Y}_t\}_{t\geq 1} \subseteq \mathbb{R}^d$ with a latent Markov chain $\{X_t\}_{t\geq 1}$ on some state space $\mathcal{X}$, such that the distribution of $\mathbf{Y}_s \mid X_s$ is independent of $\mathbf{Y}_t \mid X_t$ for $s \neq t$:



Fusion of Multiple Data Sources

- In many applications, sensors capture multiple streams of data, which are "fused" into a multivariate time series $\{\mathbf{Y}_t\}_{t\geq 1}$
- In such situations, the components of any $\mathbf{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})$ cannot be assumed independent (even conditional on X_t)
- It is common to assume that $\mathbf{Y}_t$ follows a multivariate Gaussian distribution, but this places limits on marginals and dependence structures
- What if the strength of dependence – or even the "kind" of dependence – between the components of $\mathbf{Y}_t$ could be informative about the underlying state X_t?

Copulas Within HMMs

- Our model consists of an HMM $\{(\mathbf{Y}_t, X_t)\}_{t\geq 1} \subseteq \mathbb{R}^d \times \mathcal{X}$ in which the state-dependent distributions are copulas:

$$\mathbf{Y}_t \mid (X_t = k) \sim H_k(\cdot) = \underbrace{C_k\Big(F_{k,1}(\cdot\,;\lambda_{k,1}), \ldots, F_{k,d}(\cdot\,;\lambda_{k,d}) \Big| \theta_k\Big)}_{\text{depends on the hidden state value } k}.$$

- $C_k(\cdot, \ldots, \cdot \mid \theta_k)$ is a d-dimensional parametric copula
- $\{X_t\}_{t\geq 1}$ is a Markov process on finite state space $\mathcal{X} = \{1, 2, \ldots, K\}$ and K is known
- In this model, virtually all aspects of the state-dependent distributions are allowed to vary between states

References

Papers are available here: http://www.utstat.toronto.edu/craiu/