Statistical modelling for and with copulas

Radu Craiu

Department of Statistical Sciences
University of Toronto

Charles University, November 30, 2022

Outline

Motivation - part 0

- The smoking cessation study of Liu, Daniels and Marcus (JASA '09):

  Q = smoking cessation (0=No, 1=Yes)
  W = weight change
  X = time spent exercising

- Does exercise weaken the association between smoking status and weight gain?
- We are interested in understanding the covariate effect on the dependence pattern between responses.

Motivation - part 1

- Joint models for multivariate data.
- If the joint distribution of Y_1, Y_2, Y_3 ($Y_i \sim f_i$, $1 \leq i \leq 3$) then

$$\begin{aligned} f(y_1, y_2, y_3) &= c_{12}(F_1(y_1), F_2(y_2))c_{23}(F_2(y_2), F_3(y_3)) \\ &\times c_{13|2}(F_{1|2}(y_1|y_2), F_{3|2}(y_3|y_2); y_2)f_1(y_1)f_2(y_2)f_3(y_3) \end{aligned}$$

- As dimension increases, the bivariate conditional copulas depend on increasing number of variables.
- Useful in prediction of one (expensive) response given the other (cheaper) ones.

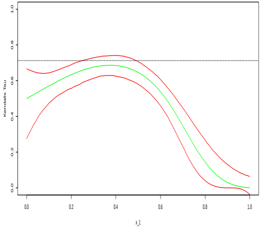
$$f(y_1|y_2, y_3) = c_{12}(F_1(y_1), F_2(y_2))c_{13|2}(F_{1|2}(y_1|y_2), F_{3|2}(y_3|y_2); y_2)f_1(y_1)$$

Motivation - part 2

- $Y_i|x \sim N(f_i(x), \sigma_i)$ $x \in \mathbb{R}^2$
- 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

Motivation - part 2

- $Y_i | x \sim N(f_i(x), \sigma_i)$ $x \in \mathbb{R}^2$
- 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



Motivation - part 2

- Simplified setting:

$$Y_i|X_1, X_2 \sim N(f_i(X_1, X_2), 1) \; i = 1, 2$$

$$\text{Cor}(Y_1, Y_2|X_1, X_2) = \rho$$

- Set $W_i = Y_i - E[Y_i|X_1]$ for $i = 1, 2$

$$\text{Cov}(W_1, W_2|X_1) = \text{Cov}(Y_1, Y_2|X_1), \tag{1}$$

and

$$\begin{aligned}\text{Cov}(Y_1, Y_2|X_1) &= E_{X_2}[\text{Cov}(Y_1, Y_2|X_1, X_2)] + \\ &+ \text{Cov}_{X_2}(E[Y_1|X_1, X_2], E[Y_2|X_1, X_2]) \\ &= \rho + \underbrace{\text{Cov}_{X_2}(f_1(X_1, X_2), f_2(X_1, X_2))}_{\text{constant in } X_1 \text{ if } f_i(X_1,X_2)=f_{i1}(X_1)+f_{i2}(X_2),\ i=1,2}\end{aligned}$$

- Omission of covariates $\Rightarrow$ Non-constant calibration.

Conditional Copulas

- If $\mathbf{X} \in \mathbb{R}^d$ is a covariate

$$H(t_1, \ldots, t_d|\mathbf{X}) = C_{\mathbf{X}}(F_1(t_1|\mathbf{X}), \ldots, F_d(t_d|\mathbf{X})).$$

- The parametric CC model assumes there is a family $\{C_\theta : \theta \in \Theta\}$ s.t.

$$C_{\mathbf{X}}(F_1(t_1|\mathbf{X}), \ldots, F_d(t_d|\mathbf{X})) = C_{\eta(\mathbf{X})}(F_1(t_1|\mathbf{X}), \ldots, F_d(t_d|\mathbf{X})).$$

- η is the unknown calibration function we are interested in.

- The simplifying assumption (SA):

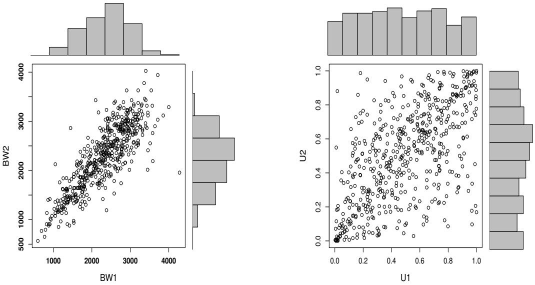
$$C_{\mathbf{X}}(F_1(t_1|\mathbf{X}), \ldots, F_d(t_d|\mathbf{X})) = C(F_1(t_1|\mathbf{X}), \ldots, F_d(t_d|\mathbf{X}))$$

or

$$\eta(\mathbf{X}) = const.$$

Semiparametric and Nonparametric

- Semiparametric and nonparametric methods (Acar, C. and Yao 2011; Veraverbeke, Omelka & Gijbels 2011)
- Estimate $\eta(X)$ after fitting the models for the marginals.
  - When marginals are unknown $\rightarrow$ propagation of errors
  - Testing for $H_0 : \eta(x) = const$ is cumbersome and lacks power.
  - Scales poorly with dimension of X.



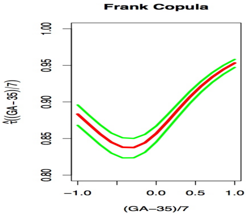
Bayesian Cubic Splines

- Joint Bayesian modelling of marginals and calibration regression models using cubic splines (C. & Sabeti 2012)
  - Suitable for $X \in \mathbb{R}$
  - Model comparisons between $M_0 : \eta(x) = const$ and $M_1 : \eta(x) \neq const$ favours M_1.
- Use additive models when $X \in \mathbb{R}^d$ (Sabeti, Wang & C. 2014; Chavez-Demoulin & Vatter 2015)

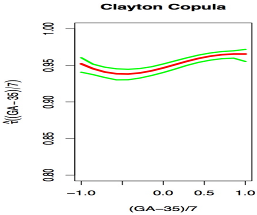
$$\eta(x_1, \ldots, x_d) = \eta_0 + \sum_{i=1}^{d} \eta_i(x_i).$$

  - Results are sensitive to violations of additivity.
  - Additivity is not preserved when changing dependence measure.

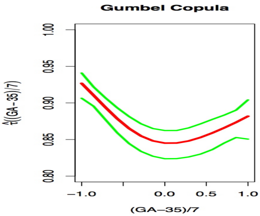
Twin Data Example



DIC=10449



DIC=14810



DIC=6.97×10^7

GP Prior with Single Index Models

- When $X \in \mathbb{R}^d$ we consider the SIM model (Levi & C 2018)

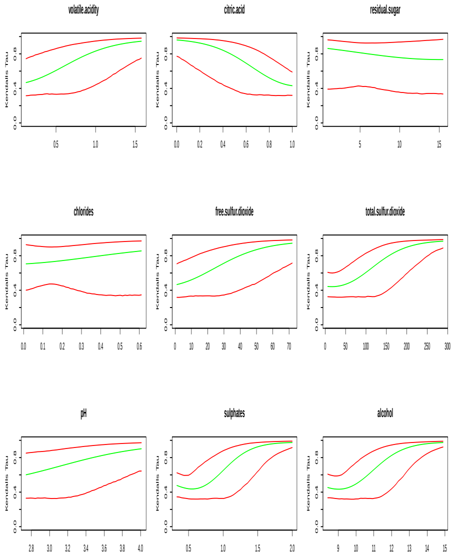
$$\eta(X) = f(\beta^T X). \text{ with } f : \mathbb{R} \to \mathbb{R} \text{ smooth.}$$

- f is estimated using a sparse Gaussian process prior
- This is invariant to nonlinear one-to-one transformations $\tau(\theta)$.
- Marginals are fitted also using GP-SIM models, but other models are possible.
- The parameter β is unidentifiable up to a constant so we assume $||\beta|| = 1$.
- Allows for variable selection.

Red Wine

- 11 Physiochemical properties of 1599 varieties of red "Vinho Verde" Portuguese wine.
- We consider the dependence between fixed acidity and density. The former is strongly associated with quality of wine, while the latter is used as a measure of grape quality.
- Covariates: volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol.

Red Wine


volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
pH
sulphates
alcohol
Kendalls Tau

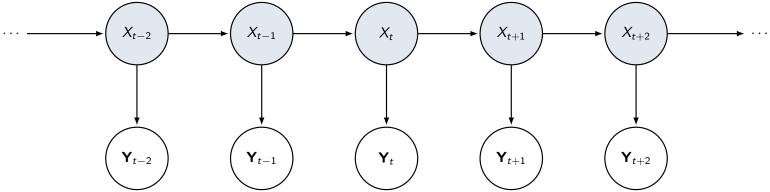
Part II: Copulas in Statistical Models

Based on "Copula Modelling of Serially Correlated Multivariate Data with Hidden Structures"

Co-authors: Robert Zimmerman and Vianey Leos-Barajas

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 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 $s \neq t$:



- $\mathbf{Y}_{t,h}|\{X_t = k\} \sim f_{k,h}(\cdot|\lambda_{k,h}) \; \forall h = 1, \ldots, d$
- $\{X_t\}$ is a Markov process (finite state space $\mathcal{X}$) with initial probability mass distribution $\{\pi_i\}_{i\in\mathcal{X}}$ and transition probabilities $\{\gamma_{i,j}\}_{i,j\in\mathcal{X}}$

Inferential aims for HMMs

- Typically, the chain $\{X_t\}_{t \geq 1}$ is partially or completely unobserved.
- The hidden states can correspond to a precise variable (occupancy data) or might be postulated (psychology, ecology, etc)
- Aim 1: Model the data generating mechanism Nasri et al. (2020)
- Aim 2: Decode (i.e., classify) or predict the X_t's from the observed data.

Examples

- A tri-axial accelerometer captures a shark's acceleration with respect to three positional axes depending on the shark's activity (resting, hunting, attacking). For short periods some of the sharks are filmed.
- Stock exchanges keep track of real-time prices for hundreds of stocks within an industry, depending on market conditions/states (stagnant, growing, shrinking).
- In-game team statistics like shots on goal and ball touches in a ~~soccer~~ football match are changing with the "momentum" of the team (defensive, offensive, passive) Ötting et al. (2021)

Fusion of Multiple Data Sources

- In the real-world applications above, various 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)
- The corresponding assumption for HMMs – that of contemporaneous conditional independence Zucchini et al. (2017) – is often violated
- Instead, 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 $\{(\mathbf{Y}_t, X_t)\}_{t\geq1} \subseteq \mathbb{R}^d \times \mathcal{X}$ in which the state-dependent distributions are copulas:

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

- $C_k(\cdot, \ldots, \cdot \mid \theta_k)$ is a d-dimensional parametric copula
- $\{X_t\}_{t\geq1}$ 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

Information in the dependence

- For a range of $\theta \in [0, 100)$, we simulated a bivariate time series of length $T = 100$ from the 2-state HMM

$$\mathbf{Y}_t \mid (X_t = k) \sim C_{\text{Frank}} \left(\mathcal{N}(0,1), \mathcal{N}(0,1) \mid (-1)^k \cdot |\theta| \right), \quad k = 1, 2$$

and then separately assessed the accuracy of a standard decoding algorithm, first assuming independent margins and then the true model:

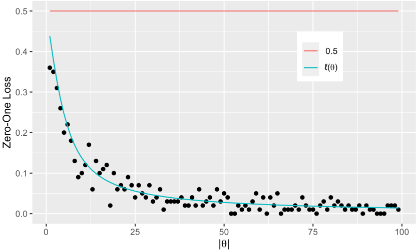


Figure: Zero-one losses for independent margins (red dots) and true model (blue dots)

Stronger Dependence Leads to Better Accuracy

- In fact, $\ell_{01}(\theta) = \frac{1}{2} - \frac{2}{\theta} \log\left(\cosh \frac{\theta}{4}\right) \to 0$ as $\theta \to \infty$
- Similar formulas hold for other radially symmetric copulas
- Much more generally, we have the following:

Theorem

Let $\nu_{t,k} = \mathbb{P}(X_t = k)$. The expected zero-one loss of the classifications made by local decoding is given by

$$\ell_{01}(\boldsymbol{\eta}) = 1 - \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} \nu_{t,k} \int_{\mathbb{R}^d} \mathbb{1}\left\{ \frac{\nu_{t,k} \cdot h_k(\mathbf{y})}{\max_{j \neq k} \nu_{t,j} \cdot h_j(\mathbf{y})} > 1 \right\} \mathrm{d}H_k(\mathbf{y}).$$

where $h_k(\mathbf{Y})$ is the joint density of $\mathbf{Y}|X = k$.

- Corollary: as the copula in any particular state approaches either of the Fréchet-Hoeffding bounds, the observations produced by that state will be detected with complete accuracy

Estimation with missing data

- Data consist in observed $\mathbf{Y}_{1:T}$ and missing $X_{1:T}$
- Parameters are $\eta = \{\lambda_{h,k}\}_{\substack{h=1:d \\ k=1:T}} \cup \{\theta_k\}_{k=1:T} \cup \{\gamma_{i,j}\}_{\substack{i=1:K \\ j=1:K}}$ $\cup\{\pi_j\}_{j=1:K}$.
- The complete-data log-likelihood for one trajectory of the copula HMM is given by

$$\ell_{\text{com}}\left(\boldsymbol{\eta} \mid \mathbf{y}_{1:T}, X_{1:T}\right) = \pi_{X_1} + \sum_{t=2}^{T} \log \gamma_{X_{t-1},X_t} + \sum_{h=1}^{d} \log f_{X_t,h}(y_{t,h}; \lambda_{X_t,h})$$
$$+ \sum_{t=1}^{T} \log c_{X_t}\left(F_{X_t,1}(y_{t,1}; \lambda_{X_t,1}), \ldots, F_{X_t,1}(y_{t,d}; \lambda_{X_t,d}) \mid \theta_{X_t}\right). \tag{2}$$

Inference for HMMs Via the EM Algorithm

- Without copula, the estimation is done via the EM algorithm (aka Baum-Welch)

E-step Compute $Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(s)}) = E[l_{com}(\boldsymbol{\eta}|\mathbf{Y}_{1:T}, X_{1:T})|\boldsymbol{\eta}^{(s)}, \mathbf{Y}_{1:T}]$

M-step Set $\boldsymbol{\eta}^{(s+1)} = \arg\max_{\eta} Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(s)})$

- The complete-data log-likelihood is written in terms of the state membership indicators $U_{k,t} = \mathbb{1}_{X_t=k}$ and $V_{j,k,t} = \mathbb{1}_{X_{t-1}=j, X_t=k}$
- In the **E-Step**, these indicators are estimated by the conditional probabilities $\hat{u}_{k,t} = \mathbb{P}\left(X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}\right)$ and $\hat{v}_{j,k,t} = \mathbb{P}\left(X_{t-1} = j, X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}\right)$, which are computed based on current parameter estimates
- This only requires evaluating the state-dependent densities at each of the observations $\mathbf{y}_1, \ldots, \mathbf{y}_T$ (this is "OK")

The M-Step Is Hard

- In the **M-Step**, the resulting complete-data log-likelihood is maximized with respect to all parameters in the model simultaneously
  - Only for the simplest univariate models do the state-dependent MLEs exist in closed form; otherwise, one must resort to numerical methods (this is hard!)
  - Evaluating a copula density $c_k(\cdot, \dots, \cdot \mid \theta_k)$ in high dimensions is slow
  - When the state-dependent distributions in an HMM are copulas, performing the M-Step directly requires the evaluation of

$$\underset{\{\theta_k\},\{\lambda_{k,h}\}}{\operatorname{argmax}} \left\{ \sum_{k=1}^{K} \sum_{t=1}^{T} \hat{u}_{k,t} \left[\log c_k\left(F_{k,1}(y_{t,1}; \lambda_{k,1}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}) \,\middle|\, \theta_k \right) + \sum_{h=1}^{d} \log f_{k,h}(y_{t,h}; \lambda_{k,h}) \right] \right\}$$

  - This is very unstable (and slow)

Inference Functions for Margins

- Likelihood-based inference for copulas is easier when the goal is to estimate θ alone in the presence of known margins

- Why not perform inference on the marginal distributions first, and then on the copula itself?

- In the context of iid data, this is exactly the inference functions for margins (IFM) approach of Joe and Xu (1996):
  - First estimate each λ_h by its "marginal MLE" $\hat{\lambda}_h$ given $\{Y_{t,h}\}_{t\geq 1}$, for $h \in \{1,\ldots,d\}$
  - Then estimate θ assuming fixed marginals $F_1(\cdot;\hat{\lambda}_1),\ldots,F_d(\cdot;\hat{\lambda}_d)$

- One can show that the IFM estimator is consistent and asymptotically normal (although relatively less efficient than the MLE)

A Better Approach

- Replace the M-Step in the EM algorithm with an IFM iteration to create an "EFM algorithm"

- For $T \in \{100, 1000, 5000\}$ and $d \in \{2, 5, 10\}$, we simulated a d-dimensional time series of length T from the 2-state HMM

$$\mathbf{Y}_t \mid (X_t = 1) \sim C_{\text{Frank}}\left((\mathcal{N}(\mu_{1,h} = -h, 1))_{h=1}^{d} \mid \theta_1 = 3\right)$$
$$\mathbf{Y}_t \mid (X_t = 2) \sim C_{\text{Clayton}}\left((\mathcal{N}(\mu_{2,h} = h, 1))_{h=1}^{d} \mid \theta_2 = 3\right)$$

  and estimated $\boldsymbol{\eta} = (\mu_{1,1}, \ldots, \mu_{2,d}, \theta_1, \theta_2)$ using both approaches

- Applied to the basic EM algorithm, R's `optim` with `L-BFGS-B` (i.e., quasi-Newton with box constraints) typically fails as soon as $d \geq 3$
  - The procedure is extremely sensitive to initial values and requires $\hat{\boldsymbol{\eta}}^{(0)} \approx \boldsymbol{\eta}$ just to avoid overflow
  - This kind of tuning is very tedious or impossible in high dimensions

Does This Work?

- We keep track of the time (in seconds) until the algorithm converges, and the L_2 error of the resulting estimate, $\epsilon = \|\eta - \hat{\eta}\|_2$
  - We used the `lbfgsb3c` package, which is more stable than `optim`

	$d = 2$	$d = 5$	$d = 10$
$T = 100$	111.9 s, $\epsilon = 0.14$	123.4 s, $\epsilon = 299.98$	111.8 s, $\epsilon > 10^9$
$T = 1000$	166.6 s, $\epsilon = 0.63$	169.5 s, $\epsilon > 10^{11}$	418.23 s, $\epsilon = 725.06$
$T = 5000$	?	?	?

Table: EM Algorithm

	$d = 2$	$d = 5$	$d = 10$
$T = 100$	5.1 s, $\epsilon = 2.9$	3.0 s, $\epsilon = 0.94$	4.2 s, $\epsilon = 0.58$
$T = 1000$	34.4 s, $\epsilon = 0.57$	22.9 s, $\epsilon = 0.60$	34.4 s, $\epsilon = 0.80$
$T = 5000$	172.6 s, $\epsilon = 0.13$	106.2 s, $\epsilon = 0.12$	168.7 s, $\epsilon = 0.19$

Table: EFM Algorithm

This Works

- R has no problem with the EFM algorithm
- The algorithm is considerably less sensitive to starting values than the vanilla EM algorithm, and terminates much faster
- It is also theoretically justified
  - We show that the sequence of estimates produced by our algorithm will converge, and the resulting estimator is consistent and asymptotically normal (under mild regularity conditions)
  - Accomplished by viewing our method as an adaptation of the ES algorithm of Elashoff and Ryan (2004) and using established asymptotic theory of M-estimators for HMMs Jensen (2011)

Occupancy Data

- The ability to detect whether a room is occupied using sensor data (such as temperature and CO_2 levels) can potentially reduce unnecessary energy consumption by automatically controlling HVAC and lighting systems, without the need for motion detectors
- Consider three publicly-available labelled datasets presented by Candanedo and Feldheim (2016) which contain multivariate time series of four environmental measurements (light, temperature, humidity, CO_2) and one derived metric (the humidity ratio)
- Data contain binary indicators for whether the room was occupied or not at the time of measurement

Occupancy Data

- Several common families of parametric copulas (the Frank, Clayton, Gumbel, Joe, and Gauss families), and for each we carried out a goodness-of-fit test based on the pseudo-observations using the multiplier bootstrap method (Kojadinovic et al., 2011)
- The parametric family based on the lowest corresponding Cramér-von Mises test statistic is selected.
- This process yielded a Clayton copula for State 1 and a Frank copula for State 2

State	Frank	Clayton	Gumbel	Joe	Gauss
1	0.356	**0.255**	0.423	0.770	0.345
2	**0.018**	0.433	0.038	0.206	0.045

Table: Cramér-von Mises test statistics based on pseudo-observations computed from unoccupied (Row 1) and occupied (Row 2) subsets.

Occupancy Data

- Denote the unoccupied state as '1' and the occupied state as '2'

$$\mathbf{Y}_t \mid (X_t = 1) \sim C_{\text{Clayton}}\left(\mathcal{N}(\mu_{1,1}, \sigma^2_{1,1}), \mathcal{N}(\mu_{1,2}, \sigma^2_{1,2}) \mid \theta_1\right)$$
$$\mathbf{Y}_t \mid (X_t = 2) \sim C_{\text{Frank}}\left(\mathcal{N}(\mu_{2,1}, \sigma^2_{2,1}), \mathcal{N}(\mu_{2,2}, \sigma^2_{2,2}) \mid \theta_2\right).$$

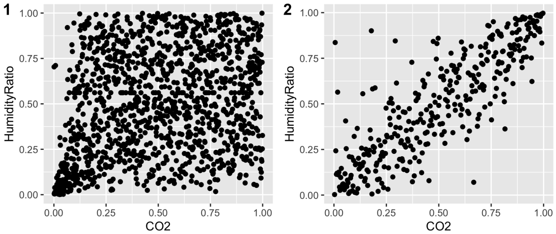


Figure: Pseudo-observations computed from unoccupied (Panel 1) and occupied (Panel 2) subsets.

Occupancy Data

Copula Model	Train	Test 1	Test 2
Independence	0.895	0.846	0.680
Clayton/Frank	0.899	0.852	0.696

Table: Overall state classification accuracy for the training dataset and the two test datasets using either the independence or Clayton/Frank copula.

Discussion

- Copulas are interesting objects that can play useful roles in statistical models
- Allow data fusion in many contexts
- Customized computation is often required
- Future: computation and model selection for/with vine copulas

References

Candanedo, L. M. and Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models. *Energy and Buildings* **112** 28–39.

Elashoff, M. and Ryan, L. (2004). An em algorithm for estimating equations. *Journal of Computational and Graphical Statistics* **13** 48–65.

Jensen, J. L. (2011). Asymptotic normality of m-estimators in nonhomogeneous hidden markov models. *Journal of Applied Probability* **48** 295–306.

Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. Tech. Rep. 166, Department of Statistics, University of British Columbia.

Kojadinovic, I., Yan, J. and Holmes, M. (2011). Fast large-sample goodness-of-fit tests for copulas. *Statistica Sinica* 841–871.

Nasri, B. R., Remillard, B. N. and Thioub, M. Y. (2020). Goodness-of-fit for regime-switching copula models with application to option pricing. *Canadian Journal of Statistics* **48** 79–96.

Ötting, M., Langrock, R. and Maruotti, A. (2021). A copula-based multivariate hidden markov model for modelling momentum in football. *AStA Advances in Statistical Analysis* 1–19.

Zucchini, W., MacDonald, I. L. and Langrock, R. (2017). *Hidden Markov models for time series: an introduction using R*. CRC press.

My own papers are available on my homepage: `http://www.utstat.toronto.edu/craiu/`