Copula modelling of serially correlated multivariate data with hidden structures

Radu Craiu

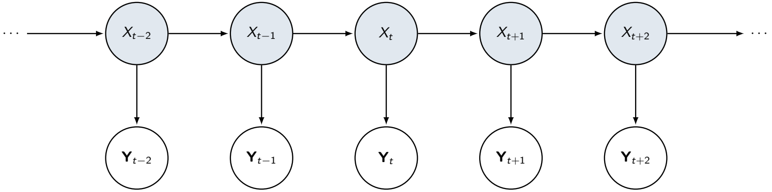
Department of Statistical Sciences
University of Toronto

Joint with Robert Zimmerman and Vianey Leos-Barajas (Toronto)

Ca'Foscari, March 24, 2023

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 $\mathbf{Y}_t = (Y_{t,1}, \dots, Y_{t,d})$ 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 between the components of $\mathbf{Y}_t$ could be informative about the underlying state X_t?

Copulas

- Copula functions are used to model dependence between continuous random variables.

- If $Y_1, Y_2, \ldots Y_d$ are continuous r.v.'s with distribution functions (df) $F_1, \ldots, F_d$, there exists an unique copula function $C : [0,1]^d \to [0,1]$ such that

$$H(t_1, \ldots, t_d) = \mathbb{P}(Y_1 \leq t_1, \ldots, Y_d \leq t_d) = C(F_1(t), \ldots, F_d(t_d)).$$

- The copula bridges the marginal distributions of $Y_1, \ldots, Y_d$ with the joint distribution. It corresponds to a distribution on $[0,1]^d$ with uniform margins.

- This can be extended to conditional distributions and copulas:

$$\mathbb{P}(Y_1 \leq t_1, \ldots, Y_d \leq t_d | X) = C(F_1(t|X), \ldots, F_d(t_d|X)|X).$$

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

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:

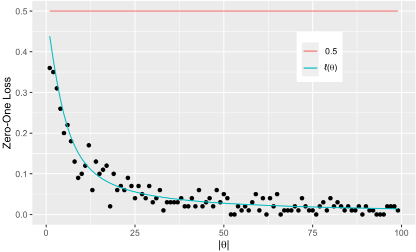


Figure: Zero-one losses for independent margins (red dots) and true model (blue dots)

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

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

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 $\hat{u}_{k,t} = \mathbb{P}(X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$ and $\hat{v}_{j,k,t} = \mathbb{P}(X_{t-1} = j, X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$, 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 and unstable!)
  - 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)

IFM Step

- For each $j \in \mathcal{X}$, estimate the initial distribution and transition probabilities:

$$\delta^{(s+1)} = (\hat{u}_{1,1}^{(s)}, \dots, \hat{u}_{K,1}^{(s)})$$

and

$$\gamma_{j,\cdot}^{(s+1)} = \left(\frac{\sum_{t=2}^{T} \hat{v}_{j,1,t}^{(s)}}{\sum_{k=1}^{K} \sum_{t=2}^{T} \hat{v}_{j,k,t}^{(s)}}, \dots, \frac{\sum_{t=2}^{T} \hat{v}_{j,K,t}^{(s)}}{\sum_{k=1}^{K} \sum_{t=2}^{T} \hat{v}_{j,k,t}^{(s)}} \right).$$

- For each $k \in \{1, \dots, K\}$ and $h \in \{1, \dots, d\}$, estimate the marginal parameters

$$\lambda_{k,h}^{(s+1)} = \arg\sup_{\lambda} \sum_{t=1}^{T} \hat{u}_{k,t}^{(s+1)} \cdot \log\left(f_{k,h}(y_{t,h}; \lambda)\right).$$

- For each $k \in \{1, \dots, K\}$, estimate the copula parameters

$$\tilde{\theta}_{k}^{(s+1)} = \arg\sup_{\theta} \sum_{t=1}^{T} \hat{u}_{k,t}^{(s+1)} \cdot \log\left(c_k \left(F_{k,1}(y_{t,1}; \lambda_{k,1}^{(s+1)}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}^{(s+1)}) \,\middle|\, \theta \right) \right).$$

New problems

- The EIFM algorithm is not an GEM algorithm

$$\sum_{t=1}^{T} \hat{u}_t \cdot \log\left(f_h(y_{t,h}; \lambda_h^{(s)})\right) \leq \sum_{t=1}^{T} \hat{u}_t \cdot \log\left(f_h(y_{t,h}; \lambda_h^{(s+1)})\right), \qquad h \in \{1, \dots, d\} \qquad (2)$$

does not imply

$$\begin{aligned} & \sum_{t=1}^{T} \hat{u}_t \cdot \log\left(c\left(F_1(y_{t,1}; \lambda_1^{(s)}), \dots, F_d(y_{t,d}; \lambda_d^{(s)}) \,\middle|\, \theta^{(s)}\right)\right) \\ \leq \; & \sum_{t=1}^{T} \hat{u}_t \cdot \log\left(c\left(F_1(y_{t,1}; \lambda_1^{(s+1)}), \dots, F_d(y_{t,d}; \lambda_d^{(s+1)}) \,\middle|\, \theta^{(s)}\right)\right). \end{aligned}$$

- The EIFM algorithm will converge (to a local or global maximum).
- The estimator is consistent and asymptotically normal (under mild regularity conditions).
- EIFM as a version of the ES algorithm of Elashoff and Ryan (2004).
- Use asymptotic theory of M-estimators for HMMs Jensen (2011).

Implementation of EIFM

- K is assumed known
- An initial clustering algorithm may be used in which the observed multivariate data follow vine copulas (Sahin and Czado, 2022)
- We consider the k-means algorithm for clustering.
- Initial parameter values for the copula(s) are obtained using a Gaussian copula
- If marginals are Gaussian this means fitting a multivariate normal for each cluster.

Does This Work?

- 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, $\epsilon = \|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}\|_2$
  - 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

Numerical Experiment I

- Generative model:

$$\mathbf{Y}_i \mid (X_i = k) \sim C_k \left(SN(\cdot; \xi_{k,1}, \omega_{k,1}, \alpha_{k,1}), SN(\cdot; \xi_{k,2}, \omega_{k,2}, \alpha_{k,2}) \mid \tau_k \right),$$

for $k \in \{1, \ldots, 4\}$.

State	Copula family	τ_k	$\xi_{k,1}$	$\omega_{k,1}$	$\alpha_{k,1}$	$\xi_{k,2}$	$\omega_{k,2}$	$\alpha_{k,2}$
1	Clayton	0.2	-4	1	5	-1	1	-3
2	B4	0.4	-2	1	3	2	1	-3
3	Gaussian	0.6	0	1	5	3	1	-5
4	$t_{(\nu=5)}$	0.8	2	1	3	4	1	-5

Table: True parameters for the state-dependent distributions.

Numerical Experiment I

T:		500	1000	2500	5000
Stopping Rule Tolerance:	0.01	14	24	23	15
	0.001	17	26	25	17
	0.0001	36	59	62	39
	0.00001	230	115	460	269
Classifier:	k-means	0.9020	0.9090	0.9200	0.9196
	Local state decoding	0.9640	0.9640	0.9696	0.9732

Table: For each $T \in \{500, 1000, 2500, 5000\}$: (Top rows) Number of iterations taken by the EIFM algorithm applied to $\mathbf{Y}_{1:T}$ before stopping using L_1-norm tolerances in $\{0.01, 0.001, 0.0001, 0.00001\}$. (Bottom rows) Classification accuracy of initial k-means clustering and local decoding with parameter estimates provided by the EIFM algorithm.

Numerical Experiment II

- The state-dependent distributions are Markov trees in which all conditional relationships are independent.

Figure: Markov trees for state 1 (left) and state 2 (right)

- The state-dependent distributions have densities supported on $\mathbb{R}^5$ given by

$$h_1(\mathbf{y}) = c_{1,12}(\Phi(y_1 - \mu_{1,1}), \Phi(y_2 - \mu_{1,2}) \mid \tau_{1,12}) \cdot c_{1,23}(\Phi(y_2 - \mu_{1,2}), \Phi(y_3 - \mu_{1,3}) \mid \tau_{1,23})$$

$$\cdot\, c_{1,34}(\Phi(y_3 - \mu_{1,3}), \Phi(y_4 - \mu_{1,4}) \mid \tau_{1,34}) \cdot c_{1,45}(\Phi(y_4 - \mu_{1,4}), \Phi(y_5 - \mu_{1,5}) \mid \tau_{1,45}) \cdot \prod_{h=1}^{5} \varphi(y_h - \mu_{1,h})$$

and

$$h_2(\mathbf{y}) = c_{2,14}(\Phi(y_1 - \mu_{2,1}), \Phi(y_4 - \mu_{2,4}) \mid \tau_{2,14}) \cdot c_{2,43}(\Phi(y_4 - \mu_{2,4}), \Phi(y_3 - \mu_{2,3}) \mid \tau_{2,43})$$

$$\cdot\, c_{2,35}(\Phi(y_3 - \mu_{2,3}), \Phi(y_5 - \mu_{2,5}) \mid \tau_{2,35}) \cdot c_{2,52}(\Phi(y_5 - \mu_{2,5}), \Phi(y_2 - \mu_{2,2}) \mid \tau_{2,52}) \cdot \prod_{h=1}^{5} \varphi(y_h - \mu_{2,h})$$

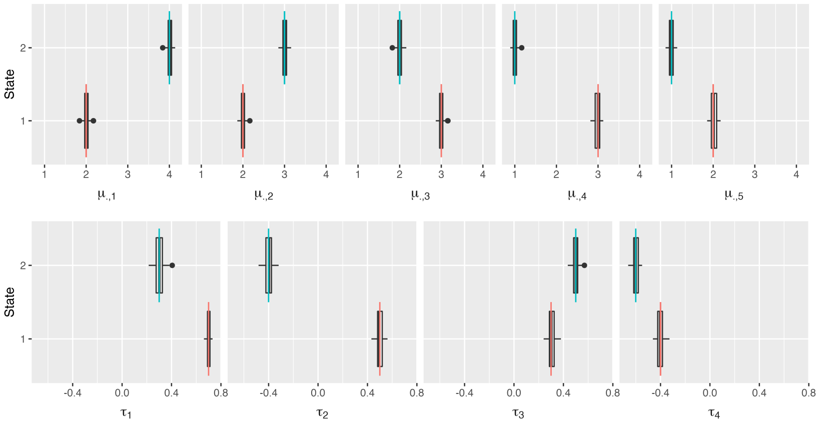


Figure: Parameter estimates based on 100 independent simulations and EIFM algorithm runs for the 5-dimensional 2-state HMMs

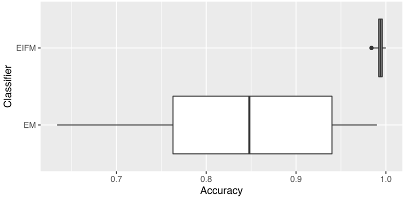


Figure: Accuracy across repetitions using initial independence model (bottom) and copula model (top)

Summary

- When using HMMs to model multivariate time series, ignoring the dependence between observed components can lead to...
  - Inaccurate state classifications
  - Failure to understand the true data-generating process
- The "copula-within-HMM" model integrates state-dependent copulas in order to capture joint information from the observed data, thereby addressing both problems
- The complexity of this model prohibits an application of the standard EM algorithm
- Our IFM-based refinement is faster and much more stable, but still produces estimators with desirable properties that perform as well or better in our experiments

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 families of parametric copulas were tried

Family	AIC (State 1)	AIC (State 2)
Gauss	-370.786	-59.281
t	-437.542	-71.291
Clayton	-474.836	-66.087
Gumbel	-484.252	-72.437
Frank	-273.613	-66.497
Joe	-490.103	-64.995
Galambos	-295.549	-71.031
Hüsler-Reiss	-282.452	-61.547
BB1	-497.013	-70.509
BB6	-495.401	-70.435
BB7	**-518.923**	-68.207
BB8	-490.223	-66.738
Tawn (type 1)	-411.847	**-76.976**
Tawn (type 2)	-422.268	-55.382

Table: AICs for unoccupied (state 1) and occupied (state 2) classifications of the occupancy data.

Occupancy Data

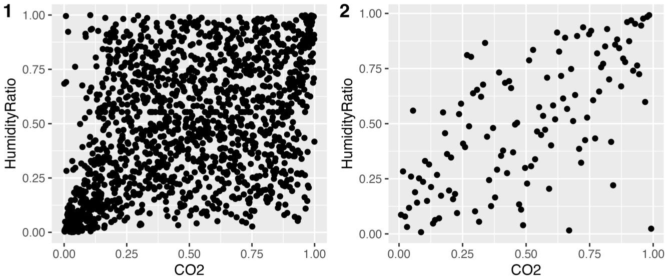


Figure: Pseudo-observations computed from unoccupied (Panel 1) and occupied (Panel 2) subsets.

Occupancy Data

Classifier	Train	Test 1
k-means clustering	0.865	0.818
Independence copulas within HMM	0.895	0.846
BB7/Tawn copulas within HMM	0.900	0.852

Table: Overall state classification accuracy for the training dataset and the test dataset, using *k*-means clustering and local decoding via the HMM with independent margins and the copula-within-HMM model.

Extensions and Future Work

- Can our algorithm be applied to models with continuous-time processes, and/or more general state spaces?
- How do we select the state-dependent copulas and/or the number of states in a fully unsupervised context?

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. *J. Comput. Graph. Stat.* **13** 48–65.

Jensen, J. L. (2011). Asymptotic normality of m-estimators in nonhomogeneous hidden Markov models. *J. App. Probab.* **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. *Can. J. Stat.* **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.

Sahin, Ö. and Czado, C. (2022). Vine copula mixture models and clustering for non-gaussian data. *Econometrics and Statistics* **22** 136–158.

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

Paper available on my homepage: `http://www.utstat.toronto.edu/craiu/`