

Copula Modeling of Serially Correlated Multivariate Data with Hidden Structures

Robert Zimmerman, Radu V. Craiu & Vianey Leos-Barajas

To cite this article: Robert Zimmerman, Radu V. Craiu & Vianey Leos-Barajas (03 Oct 2023): Copula Modeling of Serially Correlated Multivariate Data with Hidden Structures, Journal of the American Statistical Association, DOI: [10.1080/01621459.2023.2263202](https://doi.org/10.1080/01621459.2023.2263202)

To link to this article: <https://doi.org/10.1080/01621459.2023.2263202>




[View supplementary material](#) 



Published online: 03 Oct 2023.




[Submit your article to this journal](#) 



Article views: 186



[View related articles](#) 



[View Crossmark data](#) 



Copula Modeling of Serially Correlated Multivariate Data with Hidden Structures

Robert Zimmerman^a , Radu V. Craiu^a, and Vianey Leos-Barajas^b

^aDepartment of Statistical Sciences, University of Toronto, Toronto, Canada; ^bDepartment of Statistical Sciences/School of the Environment, University of Toronto, Toronto, Canada

ABSTRACT

We propose a copula-based extension of the hidden Markov model (HMM) which applies when the observations recorded at each time in the sample are multivariate. The joint model produced by the copula extension allows decoding of the hidden states based on information from multiple observations. However, unlike the case of independent marginals, the copula dependence structure embedded into the likelihood poses additional computational challenges. We tackle the latter using a theoretically-justified variation of the EM algorithm developed within the framework of inference functions for margins. We illustrate the method using numerical experiments and an analysis of room occupancy. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received August 2022
Accepted September 2023

KEYWORDS

Copulas; EM algorithm;
Hidden Markov models;
Inference functions for
margins

1. Introduction

When developing statistical models that aim to capture the generating mechanism of observed data, one must often consider various forms of *dark data* (Hand 2020), such as missing data, data that have been modified during the sampling stage, or unobservable data that can be injected into the model as latent variables to add meaning and to enhance interpretability. One can make use of the latter approach by setting up a hidden Markov model (HMM) in which the observed data consist of serially correlated observations on each item. The latent aspect of the model is a Markov chain on a discrete state space that evolves on the same time scale as the observed processes. The HMM postulates that the distribution of the data observed at time t depends on the state of the Markov chain at time t . Estimating the hidden structure can illuminate the underlying workings of the system (Rabiner and Juang 1986) and, in some cases, the Markov latent structure can be associated with a real hidden mechanism. Examples of HMMs abound in numerous domains including meteorology, medicine, ecology, and finance (Zucchini, MacDonald, and Langrock 2017).

Several distinct problems can be tackled using the HMM formulation. If one aims to identify the number of hidden states and then to *decode* the state sequence, an unsupervised approach will rely on clustering the observed features and assigning each cluster to its most likely hidden state. Alternatively, a supervised version of the analysis relies on items which were monitored and their hidden states labeled as a complete set of training data, whence one aims to predict the hidden states for observations outside the training sample. In either setting, one might also aim to impute missing values from the observed features in datasets that exhibit missing patterns.

This work is motivated by situations in which the observations recorded at each time for each item are multivariate, and one must *jointly* integrate the information they contain about the hidden states. When the observations in an HMM setup are vector-valued, the state-dependent multivariate distributions can be constructed by assuming either contemporaneous conditional independence or longitudinal conditional independence. In the former case, these distributions are constructed as products of marginal distributions, while in the latter case they are typically assumed to be multivariate normal, which is tractable and easily implemented within the standard array of HMM algorithms (Zucchini, MacDonald, and Langrock 2017). However, data which can be modeled by the multivariate normal distribution are limited; in the bivariate case, for example, the normal distribution cannot capture dependence in the extremes of the upper-right or lower-left tails (a property known as *tail dependence*). One of our main aims here is to propose a general approach to integrate the continuous-valued information provided by multivariate observations in HMMs.

Copulas have become a ubiquitous tool in modeling complex dependence structures. Sklar (1959) provided a theoretical foundation and demonstrated that any multivariate distribution can be represented by its marginal distributions and a copula that fully describes their interdependence, a decomposition which is unique whenever the marginal distributions are continuous. Copulas have been widely used in areas such as actuarial science (Frees and Valdez 1998), hydrology (Genest et al. 2007), and finance (Nasri, Remillard, and Thioub 2020), to name only a few. The introduction of copulas into general HMMs also has broad applications; for example, Härdle, Okhrin, and Wang (2015) model the dynamics of exchange rates using hierarchical Archimedean copulas; Nasri, Remillard, and Thioub (2020)

model option prices using Gaussian copulas; and Ötting, Langrock, and Maruotti (2023) model momentum shifts within football matches using Clayton copulas. In this work, we model the state of occupancy within a room using an HMM with bivariate environmental observations coupled by copulas whose underlying families are state-dependent. Potential future applications include modeling multivariate animal movement data (McClintock et al. 2020), correlating factors in disease dynamics (Sherlock et al. 2013), and detecting stellar flares (Stanislavsky et al. 2020).

This article makes two main contributions. First, we extend the modeling toolbox for HMMs with multivariate observations by considering copula models for the distributions of the observed data. Our experiments show that when the copula is allowed to vary with the hidden state variable, the identification of the hidden states improves in accuracy. Second, because using copulas to model the observed data presents additional computational challenges compared to the case of independent marginals, we develop a new optimization procedure in which we integrate the *inference functions for margins (IFM)* method of Joe and Xu (1996) within the *ES algorithm* of Elashoff and Ryan (2004).

In the next section we introduce the model along with numerical evidence of the gains in decoding precision when using the copula-based joint modeling of the observed data. In Sections 3 and 4, we introduce and theoretically justify the algorithm used for estimation. Section 5 contains our numerical experiments based on simulations as well as an analysis of room occupancy data. We end the article with a summary and a discussion of future work.

2. A Copula-Based HMM

2.1. General Model and Assumptions

Let X_1, X_2, \dots be an unobserved discrete-time first-order Markov process taking values in a finite state space $\mathcal{X} = \{1, 2, \dots, K\}$ with initial distribution $\delta \in \Delta^{K-1} = \{(\delta_1, \dots, \delta_K) : \delta_1, \dots, \delta_K > 0, \sum_{k=1}^K \delta_k = 1\}$ and transition probability matrix $\Gamma = [\gamma_{ij}]_{i,j \in \mathcal{X}}$. Let $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_T$ represent the \mathbb{R}^d -valued observed data with $\mathbf{Y}_t = (Y_{t,1}, \dots, Y_{t,d})$, assumed to satisfy the conditional independence structure $(\mathbf{Y}_s \mid X_s) \perp\!\!\!\perp (\mathbf{Y}_t \mid X_t)$ for $s \neq t$. We assume that for each state $k \in \mathcal{X}$, the multivariate distribution $H_k(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$ is defined by d univariate k 'th state-dependent marginal distributions, $F_{k,h}(\cdot; \lambda_{k,h}) : \mathbb{R} \rightarrow [0, 1]$ for $1 \leq h \leq d$, and a d -dimensional parametric copula $C_k(\cdot, \dots, \cdot \mid \theta_k) : [0, 1]^d \rightarrow [0, 1]$, where the θ_k are state-specific copula parameters and the $\lambda_{k,h}$ are state- and component-specific marginal parameters, such that

$$\mathbf{Y}_i \mid (X_i = k) \sim H_k(\cdot) = C_k(F_{k,1}(\cdot; \lambda_{k,1}), \dots, F_{k,d}(\cdot; \lambda_{k,d}) \mid \theta_k). \quad (1)$$

This is the formulation of a generic finite-state time-homogeneous HMM. We further assume that each C_k and $F_{k,h}$ admit densities c_k and $f_{k,h}$ with respect to Lebesgue measure on \mathbb{R}^d and \mathbb{R} , respectively, whose log densities (and first and second derivatives thereof) can be computed, implying the existence of the k 'th state-dependent joint density

$$h_k(\mathbf{y}) = c_k(F_{k,1}(y_1; \lambda_{k,1}), \dots, F_{k,d}(y_d; \lambda_{k,d}) \mid \theta_k) \cdot \prod_{h=1}^d f_{k,h}(y_h; \lambda_{k,h}), \quad (2)$$

where $\mathbf{y} = (y_1, \dots, y_d)$. Thus, each state $k \in \mathcal{X}$ generates a d -dimensional observation whose joint dependence structure is governed by a particular state-specific copula C_k with associated

parameter θ_k . Note that while we write the $\lambda_{k,h}$'s and θ_k 's as scalars for ease of presentation, our theory also applies to vector parameters associated with any and all marginal distributions and copulas. When C_k belongs to a one-parameter family (i.e., when θ_k is a scalar), it can often be reparameterized in terms of Kendall's tau or Spearman's rho; thus, when it is more convenient, θ_k can also be understood to represent either of these two concordance measures for state k .

In this article, we assume that the number of states K is known a priori, as is often the case for HMMs used to model regime change in financial transactions (Zheng, Li, and Xu 2021), ecology (McClintock et al. 2020) and health status (Kwon et al. 2020), as well as in binary classification problems like the one studied in Section 5.2. In general, determining the number of states in an HMM is a challenging problem to which several solutions have been proposed in the literature; for example, Nasri, Remillard, and Thioub (2020) use a goodness-of-fit test based on Cramér-von Mises test statistics to select the number of regimes within a class of regime-switching models that includes (1) as a special case. Alternatively, one may adopt the general methodology described in Section 4 of Pohle et al. (2017).

2.2. Coupling Benefits: An Illustration

HMMs are generally used to model a data-generating mechanism in which the Markov process serves as proxy for a process of interest, or to classify the hidden states (or both). When the multivariate data regularly switch between different complex dependence structures and hierarchies, one can model the generating process as a finite-state HMM in which the state-dependent marginal distributions are linked together via copulas. The copula-within-HMM framework is highly flexible, because one has the freedom to vary both the copula itself and the marginal distributions between states. Such models have been applied to financial data as *regime-switching copulas* (Nasri, Remillard, and Thioub 2020). In these applications, the focus is on the data-generating mechanism, and the HMM is assessed via measures of goodness of fit and the model's capacity to replicate key features of the data.

When the copula-within-HMM framework is used to classify hidden states, the model is assessed by its predictive capacity. In this paper, we focus on *local state decoding*, in which the prediction \hat{X}_t of the unknown state X_t is the maximizer of the state membership probability of X_t conditioned on observing the data $\mathbf{y}_{1:T}$, where we write $\mathbf{y}_{s:t} = (y_s, y_{s+1}, \dots, y_t) \in \mathbb{R}^{d \times (t-s+1)}$, and we write $x_{s:t} \in \mathcal{X}^{t-s+1}$ and their random counterparts similarly. We denote by η the collection of all model parameters (see Section 3.1), so that $\hat{X}_t = \operatorname{argmax}_{k \in \mathcal{X}} \mathbb{P}_\eta(X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$; Lemma S.1 within Section S.4 of the supplementary material derives an explicit expression for this posterior probability. In contrast, *global state decoding* jointly predicts the entire state sequence as $\hat{X}_{1:T} = \operatorname{argmax}_{x_{1:T} \in \mathcal{X}^T} \mathbb{P}_\eta(X_{1:T} = x_{1:T} \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$. Both methods tend to yield similar predictions in practice (Zucchini, MacDonald, and Langrock 2017).

HMMs can perform poorly when one assumes contemporaneous conditional independence of the observation process, even when the marginal state-dependent distributions are correctly specified. As a simple example, consider the special case of bivariate data arising from a 2-state finite mixture model

of length $T = 100$ with equal persistence between states, where the state-dependent distributions are Frank copulas with standard normal margins, and the state-dependent copula parameters are $\theta_1 = -\theta$ and $\theta_2 = \theta$ for some fixed value of $\theta \in (0, 100)$, representing extreme negative dependence and extreme positive dependence, respectively, when θ is large. In Figure 1(a), we have plotted the zero-one loss for this model (i.e., $\ell_{01} = \frac{1}{T} \sum_{t=1}^T \mathbb{1}_{\hat{X}_t \neq X_t}$, where \hat{X}_t is the state at time t predicted by local state decoding) as a function of θ , as well as the zero-one loss based on an incorrectly specified model with independent marginals. As a function of θ , the expected zero-one loss for the independence model is constant; however, the loss function for the true model is given by $\ell_{01}(\theta) = \frac{1}{2} - \frac{2}{\theta} \log(\cosh \frac{\theta}{4})$. Such measures can be derived for other radially symmetric bivariate copula families within the same finite mixture model (see Corollary S.2 within Section S.4 of the supplementary material); for example, if for some $\rho \in (0, 1)$, Gauss($-\rho$) and Gauss(ρ) copulas replace the Frank copulas under the same setup, it can be shown that $\ell_{01}(\rho) = \cos^{-1}(\rho)/\pi$.

Even when both states feature positive dependence between the components of the observations, the zero-one loss can decrease when the dependence structures induced by the copulas are substantially different. For example, Figure 1(b) shows empirical zero-one losses for a two-state HMM with standard normal margins, a Gauss(ρ) copula in state 1, and a Clayton(θ) copula in state 2, where ρ varies over $(0, 1)$ and $\theta = 100\rho$. Here, the zero-one loss increases slightly as both state-dependent copulas approach the comonotonicity copula, whence the states again become indistinguishable. While a closed-form expression for the expected zero-one loss for the latter model is unavailable, the true expected loss functions are special cases of a far more general expression.

Theorem 2.1. Let $v_{t,k} = \mathbb{P}(X_t = k) = [\delta \Gamma^t]_k$. The expected zero-one loss of the classifications made by local decoding is given by

$$\ell_{01}(\eta) = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^K v_{t,k} \cdot \mathbb{P}_\eta \left(\frac{h_k(\mathbf{Y}_t) \cdot \sum_{x=t} \kappa(x-t, \mathbf{Y}-t) \cdot \gamma_{x_{t-1},k} \cdot \gamma_{k,x_{t+1}}}{\max_{j \neq k} \left\{ h_j(\mathbf{Y}_t) \cdot \sum_{x=t} \kappa(x-t, \mathbf{Y}-t) \cdot \gamma_{x_{t-1},j} \cdot \gamma_{j,x_{t+1}} \right\}} < 1 \mid X_t = k \right) \quad (3)$$

where $\kappa(x-t, \mathbf{y}-t) = \delta_{x_1} \prod_{s \neq t} h_{x_s}(\mathbf{y}_s) \cdot \prod_{s \neq t, t+1} \gamma_{x_{s-1}, x_s}$. Moreover, if the HMM constitutes a finite mixture model, then (3) reduces to

$$\ell_{01}(\eta) = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^K v_{t,k} \int \mathbb{1} \left\{ \frac{v_{t,k} \cdot h_k(\mathbf{y}_t)}{\max_{j \neq k} v_{t,j} \cdot h_j(\mathbf{y}_t)} < 1 \right\} dH_k(\mathbf{y}_t). \quad (4)$$

Here again, H_k refers to the k 'th state-dependent joint distribution, and h_k to the corresponding density function. The complicated form of the HMM posterior state probabilities makes it challenging to derive useful bounds for the zero-one loss of the most general models, but under certain conditions one can deduce upper bounds for individual components of the loss function. For example, if at any time, the Markov chain travels through state k with a higher probability than that of any other state, then the contribution of the k 'th state to the loss function has an upper bound with a more succinct mathematical form.

Proposition 2.2. Suppose $\mathbb{P}_\eta(X_t = k, X_{t+1} = i_1 \mid X_{t-1} = i_2) \geq \mathbb{P}_\eta(X_t = j, X_{t+1} = i_1 \mid X_{t-1} = i_2)$ for all $i_1, i_2, j \in \mathcal{X}$. Then

$$\mathbb{P}_\eta \left(\frac{h_k(\mathbf{Y}_t) \cdot \sum_{x=t} \kappa(x-t, \mathbf{Y}-t) \cdot \gamma_{x_{t-1},k} \cdot \gamma_{k,x_{t+1}}}{\max_{j \neq k} \left\{ h_j(\mathbf{Y}_t) \cdot \sum_{x=t} \kappa(x-t, \mathbf{Y}-t) \cdot \gamma_{x_{t-1},j} \cdot \gamma_{j,x_{t+1}} \right\}} < 1 \mid X_t = k \right) \leq \int_{\mathbb{R}^d} \mathbb{1} \left\{ \frac{h_k(\mathbf{y}_t)}{\max_{j \neq k} h_j(\mathbf{y}_t)} < 1 \right\} dH_k(\mathbf{y}_t). \quad (5)$$

While the additional structure on the Markov chain in Proposition 2.2 is somewhat restrictive, the result provides insight into the general case, which is analytically intractable due to the structure of the HMM. For the case in which the HMM constitutes a finite mixture model, the assumption in Proposition 2.2 simplifies to $\mathbb{P}_\eta(X_t = k) \geq \mathbb{P}_\eta(X_t = j)$ for all $j \in \mathcal{X}$. The effect of the copulas in the HMM can be further analyzed through the lens of the integral (5). In the bivariate case, the following result explicitly relates the integral to the strength of the dependence within any particular state. Below τ_k and ρ_k denote, respectively, Kendall's tau and Spearman's rho of the copula in the k 'th state.

Proposition 2.3. Let $d = 2$ and fix $k \in \mathcal{X}$. As either $|\tau_k| \rightarrow 1$ or $|\rho_k| \rightarrow 1$ while the parameters of the other state-dependent

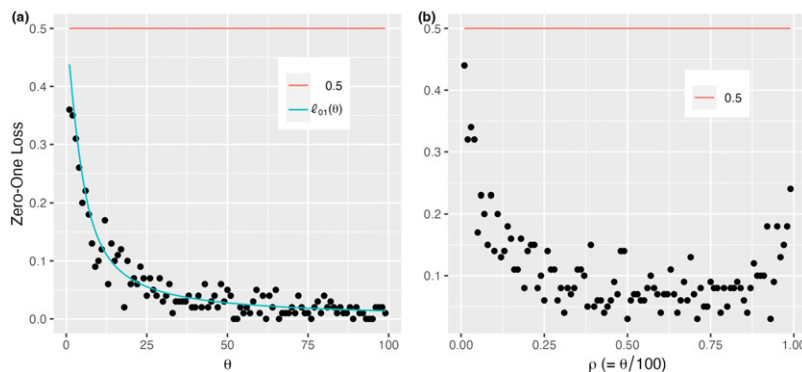


Figure 1. (a) Empirical zero-one losses (black dots) for local decoding of a 2-state HMM with standard normal margins, a Frank(θ) copula in state 1, and a Frank($-\theta$) copula in state 2, for $\theta \in \{1, 2, \dots, 100\}$. (b) The same zero-one losses, but with a Gauss(ρ) copula in state 1 and a Clayton(θ) copula in state 2, for $\rho \in \{0.01, 0.02, \dots, 0.99\}$ and $\theta = 100\rho$.

distributions stay fixed, we have

$$\int_{\mathbb{R}^d} \mathbb{1} \left\{ \frac{\omega_k \cdot h_k(\mathbf{y})}{\max_{j \neq k} \omega_j \cdot h_j(\mathbf{y})} < 1 \right\} dH_k(\mathbf{y}) \rightarrow 0$$

for any $(\omega_1, \dots, \omega_K) \in \Delta^{K-1}$.

In effect, [Theorem 2.1](#), [Propositions 2.2](#), and [2.3](#) clarify that as the copula in a particularly ‘‘common’’ state approaches either of the Fréchet-Hoeffding bounds, the observations produced by that state will become sufficiently distinguished for the local decoding algorithm to detect the state with complete accuracy, thereby eliminating its contribution to the zero-one loss. Proofs of these results are provided in [Section S.4 of the supplementary material](#).

3. Parameter Estimation

3.1. Pitfalls of the Vanilla EM Algorithm

We consider a parametric approach for copula specification, in which the parameter vector characterizes the state-dependent copulas and the marginal distributions. Parameter estimation for HMMs usually relies on the Baum-Welch algorithm (Baum et al. 1970), which is a special case of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) in which the sequence of hidden states plays the role of missing data. For the model proposed here, the E-step calculations are straightforward when all state-dependent densities can be evaluated at the current parameter estimates. Alas, the maximization required to complete the M-step is more difficult because the task requires us to maximize a nonlinear function of all state-dependent parameters; since an exact closed-form solution is usually unavailable, we must resort to numerical optimization. In practice, quasi-Newton methods often bypass the need for analytical score functions when dealing with parametric copulas alone; even so, joint maximization of the entire copula-within-HMM likelihood remains challenging. We present here a succinct version of the relevant equations; complete derivations may be found in Zucchini, MacDonald, and Langrock (2017). When $s < t$, for general continuous $\mathbb{R}^{d \times (t-s+1)}$ -valued $\mathbf{Y}_{s:t}$ and \mathcal{X} -valued X , we employ a standard abuse of notation and write $\mathbb{P}(\mathbf{Y}_{s:t} = \mathbf{y}_{s:t} \mid X = x)$ for the conditional density of $\mathbf{Y}_{s:t} \mid X = x$ and $\mathbb{P}(\mathbf{Y}_{s:t} = \mathbf{y}_{s:t}, X = x)$ for the joint density of $(\mathbf{Y}_{s:t}, X)$, the latter with respect to the product of Lebesgue measure on $\mathbb{R}^{d \times (t-s+1)}$ and counting measure on \mathcal{X} (see Cappé, Moulines, and Ryden (2005) for general definitions).

The set of parameters consists of the initial distribution δ and the vectors of transition probabilities $\boldsymbol{\gamma}_k$, of the latent Markov chain, the parameters of the state-dependent marginal densities $\lambda_{k,h}$, and the copula parameters θ_k . We write $\boldsymbol{\eta}$ for the vector whose entries contain all these parameters, and \mathbf{y}_t for the observed data at time t . Our formulation of the Baum-Welch algorithm follows Zucchini, MacDonald, and Langrock (2017). We consider here a single trajectory $(\mathbf{y}_{1:T}, X_{1:T})$, but the algorithm can easily be extended to handle multiple independent and identically distributed (iid) trajectories $(\mathbf{y}_{1:T_1}, X_{1:T_1}), \dots, (\mathbf{y}_{1:T_n}, X_{1:T_n})$.

The complete-data log-likelihood for one trajectory $(\mathbf{y}_{1:T}, X_{1:T})$ of the copula-within-HMM is given by

$$\begin{aligned} \ell_{\text{com}}(\boldsymbol{\eta} \mid \mathbf{y}_{1:T}, X_{1:T}) &= \log \delta_{X_1} + \sum_{t=2}^T \log \gamma_{X_{t-1}, X_t} \\ &+ \sum_{t=1}^T \left(\sum_{h=1}^d \log f_{X_t, h}(\mathbf{y}_{t, h}; \lambda_{X_t, h}) \right. \\ &\left. + \log c_{X_t}(F_{X_t, 1}(\mathbf{y}_{t, 1}; \lambda_{X_t, 1}), \dots, F_{X_t, d}(\mathbf{y}_{t, d}; \lambda_{X_t, d}) \mid \theta_{X_t}) \right). \end{aligned} \quad (6)$$

For $j, k \in \mathcal{X}$, we define the latent indicators $U_{k,t} = \mathbb{1}_{X_t=k}$ for $t \in \{1, \dots, T\}$, and $V_{j,k,t} = \mathbb{1}_{X_{t-1}=j, X_t=k}$ for $t \in \{2, \dots, T\}$. The complete-data log-likelihood in (6) is then

$$\begin{aligned} \ell_{\text{com}}(\boldsymbol{\eta} \mid \mathbf{y}_{1:T}, X_{1:T}) &= \sum_{k=1}^K U_{k,1} \cdot \log \delta_k + \sum_{j=1}^K \sum_{k=1}^K \left(\sum_{t=2}^T V_{j,k,t} \right) \log \gamma_{j,k} \\ &+ \sum_{k=1}^K \sum_{t=1}^T U_{k,t} \cdot \left(\sum_{h=1}^d \log f_{k,h}(\mathbf{y}_{t,h}; \lambda_{k,h}) \right. \\ &\left. + \log c_k(F_{k,1}(\mathbf{y}_{t,1}; \lambda_{k,1}), \dots, F_{k,d}(\mathbf{y}_{t,d}; \lambda_{k,d}) \mid \theta_k) \right). \end{aligned} \quad (7)$$

Now consider the E- and M-steps at the $(s+1)$ 'th iteration of the algorithm, assuming that the algorithm is initialized at $\boldsymbol{\eta}^{(0)}$. In the E-step, we must compute the conditional expectation $Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(s)}) = \mathbb{E}_{\boldsymbol{\eta}^{(s)}}[\ell_{\text{com}}(\boldsymbol{\eta} \mid \mathbf{y}_{1:T}, X_{1:T}) \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}]$, which relies on the expectations of all $U_{k,t}$'s and $V_{j,k,t}$'s conditioned on the observed data $\mathbf{y}_{1:T}$, given by

$$\begin{aligned} \hat{u}_{j,t}^{(s)} &= \mathbb{P}_{\boldsymbol{\eta}^{(s)}}(X_t = j \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}) \\ &= \frac{\alpha_{j,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}^{(s)}) \cdot \beta_{j,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}{\sum_{l=1}^K \alpha_{l,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}^{(s)}) \cdot \beta_{l,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}, \quad t \in \{1, \dots, T\} \end{aligned} \quad (8)$$

and

$$\begin{aligned} \hat{v}_{j,k,t}^{(s)} &= \mathbb{P}_{\boldsymbol{\eta}^{(s)}}(X_{t-1} = j, X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}) \\ &= \frac{\alpha_{j,t-1}(\mathbf{y}_{1:(t-1)}; \boldsymbol{\eta}^{(s)}) \cdot \gamma_{j,k}^{(s)} \cdot h_k(\mathbf{y}_t; \boldsymbol{\eta}^{(s)}) \cdot \beta_{k,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}{\sum_{l=1}^K \alpha_{l,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}^{(s)}) \cdot \beta_{l,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}, \\ &\quad t \in \{2, \dots, T\} \end{aligned} \quad (9)$$

for all $j, k \in \mathcal{X}$, where $\alpha_{j,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}) = \mathbb{P}_{\boldsymbol{\eta}}(\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}, X_t = j)$ are known as the *forward probabilities* for $t \in \{1, \dots, T\}$, and $\beta_{j,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}) = \mathbb{P}_{\boldsymbol{\eta}}(\mathbf{Y}_{(t+1):T} = \mathbf{y}_{(t+1):T} \mid X_t = j)$ are known as the *backward probabilities* for $t \in \{1, \dots, T-1\}$, with $\beta_{j,T} = 1$. These ingredients are essential in HMM classification algorithms, and are well-known to admit recursive structures that allow for their efficient computation via dynamic programming; this fact is not altered by the addition of the copula density in (7).

The M-step requires finding the maximizer $\boldsymbol{\eta}^{(s+1)}$ of

$$\begin{aligned} Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(s)}) &= \sum_{k=1}^K \hat{u}_{k,1}^{(s)} \cdot \log \delta_k + \sum_{j=1}^K \sum_{k=1}^K \left(\sum_{t=2}^T \hat{v}_{j,k,t}^{(s)} \right) \cdot \log \gamma_{j,k} \\ &+ \sum_{k=1}^K \sum_{t=1}^T \hat{u}_{k,t}^{(s)} \cdot \left(\sum_{h=1}^d \log f_{k,h}(\mathbf{y}_{t,h} \mid \lambda_{k,h}) \right. \\ &\left. + \log c_k(F_{k,1}(\mathbf{y}_{t,1} \mid \lambda_{k,1}), \dots, F_{k,d}(\mathbf{y}_{t,d} \mid \lambda_{k,d}) \mid \theta_k) \right). \end{aligned}$$

Since the parameters in the three groups of sums above are functionally independent, they can be maximized independently. Using Lagrange multipliers, one can easily show that the maximizer for the initial distribution is

$$\delta^{(s+1)} = \operatorname{argmax}_{\delta \in \Delta^{K-1}} \left(\sum_{k=1}^K \hat{u}_{k,1}^{(s)} \cdot \log \delta_k \right) = (\hat{u}_{1,1}^{(s)}, \dots, \hat{u}_{K,1}^{(s)})$$

while that for the vector of transition probabilities from state $j \in \mathcal{X}$ is

$$\begin{aligned} \gamma_{j,\cdot}^{(s+1)} &= \operatorname{argmax}_{\gamma \in \Delta^{K-1}} \left(\sum_{k=1}^K \left(\sum_{t=2}^T \hat{v}_{j,k,t}^{(s)} \right) \cdot \log \gamma_k \right) \\ &= \left(\sum_{k=1}^K \sum_{t=2}^T \hat{v}_{j,k,t}^{(s)} \right)^{-1} \left(\sum_{t=2}^T \hat{v}_{j,1,t}^{(s)}, \dots, \sum_{t=2}^T \hat{v}_{j,K,t}^{(s)} \right). \end{aligned}$$

The maximizer for the vector of parameters involved in the state-dependent distributions consists of the marginal parameters $\lambda_{k,h}$ and copula parameters θ_k that jointly maximize

$$\begin{aligned} &\sum_{k=1}^K \sum_{t=1}^T \hat{u}_{k,t}^{(s-1)} \cdot \left(\sum_{h=1}^d \log f_{k,h}(y_{t,h}; \lambda_{k,h}) \right. \\ &\left. + \log c_k \left(F_{k,1}(y_{t,1}; \lambda_{k,1}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}) \mid \theta_k \right) \right). \quad (10) \end{aligned}$$

The M-step falters here because the numerical optimization of (10) is difficult when d is even moderately high. Even initialization can be challenging; since any copula is grounded, even a *correct* initialization $\lambda_{k,h}^{(0)} = \lambda_{k,h}$ can lead us astray, for we can easily have $F_{k,h}(y_{t,h}; \lambda_{k,h}) \approx 0$ when $X_t \neq k$, and hence $c_k(\dots, F_{k,h}(y_{t,h}; \lambda_{k,h}), \dots) \approx 0$. Unless $\hat{u}_{k,t}^{(0)} \approx 0$ as well—which would be unusual when we have no a priori information about the true underlying states—evaluating (10) will immediately cause a numerical overflow. We emphasize that this can occur even if all parameters are initialized to their true values.

3.2. The EIFM Algorithm

Instead of maximizing all parameters in (10) *jointly*, it is considerably less challenging to estimate the marginal parameters $\lambda_{k,h}$ and the copula parameters θ_k *separately*. For instance, likelihood inference for the five “classical” one-parameter Archimedean families under known marginals was studied by Hofert, Mächler, and McNeil (2012), who derived concise functional representations for the copula densities and their score functions. Altogether, these allow for maximum likelihood estimation using gradient-based optimization methods. Meanwhile, as for standard HMMs with univariate state-dependent distributions, the difficulty of performing likelihood inference for the marginal distributions themselves depends highly on their specific forms. While closed-form solutions (or efficient estimation procedures) for MLEs in the presence of iid data have been long-established for most commonly encountered distributions, the situation is different for our models, in which the objective function is a weighted sum of log-densities, and closed-form expressions for the maximizing values are rare. If the HMM structure were removed from the model so that the data consisted of iid

observations from a multivariate distribution characterized by $C(F_1(\cdot; \lambda_1), \dots, F_d(\cdot; \lambda_d) \mid \theta)$, then (10) would reduce to

$$\sum_{t=1}^T \left(\sum_{h=1}^d \log f_h(y_{t,h}; \lambda_h) + \log c \left(F_1(y_{t,1}; \lambda_1), \dots, F_d(y_{t,d}; \lambda_d) \mid \theta \right) \right). \quad (11)$$

Moreover, if the copula C were known, an estimate for the marginal parameter λ_h could be inferred from the marginal data $y_{1:T,h}$ alone as the standard MLE

$$\tilde{\lambda}_h = \operatorname{arg sup}_{\lambda} \sum_{t=1}^T \log f_h(y_{t,h}; \lambda). \quad (12)$$

This is exactly the first step of the *inference functions for margins* (IFM) method of Joe and Xu (1996). The IFM method first finds the $\tilde{\lambda}_h$ that separately maximize the marginal log-likelihoods (12) for $h \in \{1, \dots, d\}$, and then finds the θ that maximizes the joint log-likelihood (11) in which the marginal parameters are set to their previous estimates:

$$\tilde{\theta} = \operatorname{arg sup}_{\theta} \sum_{t=1}^T \log c \left(F_1(y_{t,1}; \tilde{\lambda}_1), \dots, F_d(y_{t,d}; \tilde{\lambda}_d) \mid \theta \right). \quad (13)$$

Section 10.1 of Joe (1997) shows that under standard regularity conditions, $(\tilde{\lambda}_1, \dots, \tilde{\lambda}_d, \tilde{\theta})$ is a consistent and asymptotically normal estimator of $(\lambda_1, \dots, \lambda_d, \theta)$, although it is generally not as efficient as the usual maximum likelihood estimator.

We propose here a new algorithm, which we call the *expectation-IFM* (EIFM) algorithm, that alleviates the computational difficulties associated with the estimation of η . The main idea is to leave the E-step of the basic EM algorithm unchanged, but embed the IFM method within the M-step by replacing the single-state log-likelihoods in (12) and (13) with weighted averages that are computed over all K states. The log-likelihood for the k 'th state is weighted by the corresponding state membership probability estimates $\hat{u}_{k,t}$. That is, starting with an initial guess $\eta^{(0)}$, we perform the following steps, iterating over $s \geq 1$ until we reach convergence:

Step 1: (E-step) Calculate the conditional expectations:

Step 1a: For each $k \in \mathcal{X}$ and $t \in \{1, \dots, T\}$, calculate

$$\hat{u}_{k,t}^{(s)} = \frac{\alpha_{k,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}^{(s)}) \cdot \beta_{k,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}{\sum_{l=1}^K \alpha_{l,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}^{(s)}) \cdot \beta_{l,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}.$$

Step 1b: For $j, k \in \mathcal{X}$ and $t \in \{2, \dots, T\}$, calculate

$$\hat{v}_{j,k,t}^{(s)} = \frac{\alpha_{j,t-1}(\mathbf{y}_{1:(t-1)}; \boldsymbol{\eta}^{(s)}) \cdot \gamma_{j,k}^{(s)} \cdot h_k(\mathbf{y}_t; \boldsymbol{\eta}^{(s)}) \cdot \beta_{k,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}{\sum_{l=1}^K \alpha_{l,t}(\mathbf{y}_{1:t}; \boldsymbol{\eta}^{(s)}) \cdot \beta_{l,t}(\mathbf{y}_{(t+1):T}; \boldsymbol{\eta}^{(s)})}.$$

Step 2: (IFM-step) Estimate model parameters using the IFM approach:

Step 2a: Set $\delta^{(s+1)} = (\hat{u}_{1,1}^{(s)}, \dots, \hat{u}_{K,1}^{(s)})$ and for each $j \in \mathcal{X}$, estimate the vector of transition probabilities

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

Step 2b: For each $k \in \mathcal{X}$ and $h \in \{1, \dots, d\}$, estimate the marginal parameters

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

Step 2c: For each $k \in \mathcal{X}$, estimate the copula parameters

$$\begin{aligned} \theta_k^{(s+1)} &= \arg \sup_{\theta} \sum_{t=1}^T \hat{u}_{k,t}^{(s+1)} \\ &\cdot \log c_k \left(F_{k,1}(y_{t,1}; \lambda_{k,1}^{(s+1)}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}^{(s+1)}) \mid \theta \right). \end{aligned} \quad (14)$$

We discuss the algorithm's implementation in [Section S.1 of the supplementary material](#), including suggestions for performing the optimizations in Steps 2b and 2c, initializing parameters using simple clustering methods, and assessing convergence of the algorithm.

4. Theoretical Results

4.1. Analysis of the Algorithm

In this section, we demonstrate that the EIFM algorithm does not belong to the class of generalized EM algorithms (Dempster, Laird, and Rubin 1977), and therefore a theoretical analysis is needed to justify its use. We begin with a purely algorithmic perspective, describing the conditions under which the sequence of estimates produced by the EIFM algorithm will converge. To do so, we temporarily set aside its statistical content and view it as a deterministic nonlinear optimization procedure. The algorithm is not an EM algorithm in the traditional sense, because the expectation of the complete-data log-likelihood is not maximized at each iteration. This is mainly a consequence of the fact that

$$\begin{aligned} \sum_{t=1}^T \hat{u}_t \cdot \log f_h(y_{t,h}; \lambda_h^{(s)}) &\leq \sum_{t=1}^T \hat{u}_t \cdot \log f_h(y_{t,h}; \lambda_h^{(s+1)}), \\ h &\in \{1, \dots, d\} \end{aligned} \quad (15)$$

does not imply

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

For a simple counterexample, consider a single bivariate observation (i.e., $T = 1$) with $Y_1 \sim \text{Exp}(\lambda_1)$ and $Y_2 \sim \text{Exp}(\lambda_2)$ such that (Y_1, Y_2) has a Farlie–Gumbel–Morgenstern copula with density $c(u_1, u_2 \mid \theta) = 1 + \theta(2u_1 - 1)(2u_2 - 1)$. Suppose the algorithm were initialized at $\lambda_1^{(0)} = \lambda_2^{(0)} = 1$ and any $\theta^{(0)}$. Step 2b of the algorithm produces the standard univariate MLEs $\lambda_1^{(1)} = y_1^{-1}$ and $\lambda_2^{(1)} = y_2^{-1}$, which satisfy (15) by construction; however, the inequality $\log c \left(F_1(y_1; \lambda_1^{(0)}), F_2(y_2; \lambda_2^{(0)}) \mid \theta^{(0)} \right) \leq \log c \left(F_1(y_1; \lambda_1^{(1)}), F_2(y_2; \lambda_2^{(1)}) \mid \theta^{(0)} \right)$ is equivalent to $(2e^{-y_1} - 1)(2e^{-y_2} - 1) \leq (2e^{-1} - 1)^2$, which immediately fails when $y_1, y_2 > 1$.

This counterexample shows that despite sharing a similar strategy with the ECM algorithm (Meng and Rubin 1993), the EIFM algorithm does not belong to the class of generalized EM algorithms, as the latter increase the objective function at each

iteration. Clearly, one must identify conditions for the EIFM algorithm's convergence, since its sequential updating rule does not offer convergence guarantees without stronger assumptions. Our analysis here follows roughly the one given for the *ES algorithm* by Elashoff and Ryan (2004), whose aim is also to iteratively solve unbiased estimating equations in the presence of missing data. However, the serial dependence inherent in the HMM data precludes a direct application of that algorithm.

We regard the $\hat{u}_{j,t}$ and $\hat{v}_{j,k,t}$ as parameters themselves (in a nonstatistical context), and collect them and the parameter of interest η into a larger parameter vector $\xi = (\hat{u}, \hat{v}, \eta)$ of length $N = TK + (T-1)K^2 + K + K^2 + dK + K$. We define N \mathbb{R} -valued functions of ξ :

$$\begin{aligned} g_{k,t}^{[1]}(\xi) &= \hat{u}_{k,t} - \frac{\alpha_{k,t}(\mathbf{y}_{1:t}; \eta) \cdot \beta_{k,t}(\mathbf{y}_{(t+1):T}; \eta)}{\sum_{l=1}^K \alpha_{l,t}(\mathbf{y}_{1:t}; \eta) \cdot \beta_{l,t}(\mathbf{y}_{(t+1):T}; \eta)}, \quad k \in \mathcal{X}, \\ &t \in \{1, \dots, T\} \\ g_{j,k,t}^{[2]}(\xi) &= \hat{v}_{j,k,t} - \frac{\alpha_{j,t-1}(\mathbf{y}_{1:(t-1)}; \eta) \cdot \gamma_{j,k} \cdot h_k(\mathbf{y}_t; \eta)}{\sum_{l=1}^K \alpha_{l,t}(\mathbf{y}_{1:t}; \eta) \cdot \beta_{l,t}(\mathbf{y}_{(t+1):T}; \eta)}, \quad j, k \in \mathcal{X}, \\ &t \in \{2, \dots, T\} \\ g_k^{[3]}(\xi) &= \delta_k - \hat{u}_{k,1}, \quad k \in \mathcal{X} \\ g_{j,k}^{[4]}(\xi) &= \gamma_{j,k} \cdot \sum_{l=1}^K \sum_{t=2}^T \hat{v}_{j,l,t} - \sum_{t=2}^T \hat{v}_{j,k,t}, \quad j, k \in \mathcal{X} \\ g_{k,h}^{[5]}(\xi) &= \sum_{t=1}^T \hat{u}_{k,t} \cdot \frac{\partial}{\partial \lambda} \log f_{k,h}(y_{t,h}; \lambda), \quad k \in \mathcal{X}, \\ &h \in \{1, \dots, d\} \\ g_k^{[6]}(\xi) &= \sum_{t=1}^T \hat{u}_{k,t} \cdot \frac{\partial}{\partial \theta} \log c_k \left(\mathbf{F}_k(\mathbf{y}_t; \lambda_k) \mid \theta \right), \quad k \in \mathcal{X} \end{aligned}$$

where $\mathbf{F}_k(\mathbf{y}_t; \lambda_k) = (F_{k,1}(y_{t,1}; \lambda_{k,1}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}))$. Next, let $\mathbf{g} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ be defined by $\mathbf{g}(\xi) = (g_1(\xi), \dots, g_N(\xi))^\top$, where g_i is the i 'th function in the above list. From the description of the EIFM algorithm in [Section 3.2](#), it can be seen that the entire $(s+1)$ 'th iteration of the algorithm exactly corresponds to updating $\xi^{(s)} \rightarrow \xi^{(s+1)}$ by setting $\xi_i^{(s+1)}$ as the solution to the univariate problem $g_i(\xi_1^{(s+1)}, \dots, \xi_{i-1}^{(s+1)}, \xi, \xi_{i+1}^{(s)}, \dots, \xi_N^{(s)}) = 0$ for each $i = 1, \dots, N$ (in practice, many of these sub-updates are performed in parallel—for example, the sub-updates of the functions $g_1^{[3]}$ to $g_{K,K}^{[4]}$ correspond to Step 2a of the algorithm). Supposing the sequence $\{\xi^{(s)}\}_{s \geq 1}$ converges to some $\xi^* \in \mathbb{R}^N$, the limiting vector will satisfy $\mathbf{g}(\xi^*) = \mathbf{0}$, and the sub-vector η^* will be taken as our estimator of η ; the statistical properties of η^* are studied in [Section 4.2](#). The EIFM algorithm is thus an example of a *nonlinear Gauss-Seidel method*, or more generally, a *nonlinear successive over-relaxation (SOR) method* (Ortega and Rheinboldt 2000). Its convergence depends on the behavior of the Jacobian $\mathbf{J}_g(\xi)$ of \mathbf{g} in a neighborhood of some solution to $\mathbf{g}(\xi) = \mathbf{0}$, such that a certain transformation $\tilde{\mathbf{g}} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ of \mathbf{g} is (locally) required to be a contraction mapping with ξ^* as a fixed point. The critical theorem is the following:

Theorem 4.1 (Ortega and Rheinboldt (2000), Theorem 10.3.5). Let $\mathbf{g} : D \subset \mathbb{R}^N \rightarrow \mathbb{R}^N$ be continuously differentiable in an open

neighborhood $S_0 \subset D$ of ξ^* such that $\mathbf{g}(\xi^*) = \mathbf{0}$. Decompose the Jacobian $\mathbf{J}_g(\xi) = \mathbf{D}(\xi) - \mathbf{L}(\xi) - \mathbf{U}(\xi)$ where \mathbf{D} , \mathbf{L} and \mathbf{U} are respectively diagonal, strictly lower triangular, and strictly upper triangular matrices, and suppose that $\mathbf{D}(\xi^*)$ is nonsingular. If $\rho([\mathbf{D}(\xi^*) - \mathbf{L}(\xi^*)]^{-1}\mathbf{U}(\xi^*)) < 1$, then there exists an open ball $B(\xi^*, \delta) \subset S_0$ such that for any $\xi^{(0)} \in B(\xi^*, \delta)$, there exists a unique sequence $\{\xi^{(s)}\} \subset B(\xi^*, \delta)$ satisfying the nonlinear SOR prescription, such that $\lim_{s \rightarrow \infty} \xi^{(s)} = \xi^*$.

See Ortega and Rheinboldt (2000) for a proof. Note here that for a general matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ with eigenvalues v_1, \dots, v_N , the function $\rho(\mathbf{A})$ is defined as $\max\{|v_1|, \dots, |v_N|\}$, the *spectral radius* of \mathbf{A} . This theorem implies that the EIFM algorithm, once sufficiently close to a local solution ξ^* of $\mathbf{g}(\xi) = \mathbf{0}$, will produce a sequence converging to ξ^* , provided that the spectral radius of $[\mathbf{D}(\xi^{(s)}) - \mathbf{L}(\xi^{(s)})]^{-1}\mathbf{U}(\xi^{(s)})$ is less than 1. Verifying the nonsingularity of $\mathbf{D}(\xi^{(s)})$ is straightforward; it is easy to see that its first $TK + (T - 1)K^2 + K + K^2$ diagonal entries are equal to 1, while the remaining $dK + K$ entries are of the form

$$\sum_{t=1}^T \hat{u}_{k,t}^{(s)} \cdot \frac{\partial^2}{\partial \lambda^2} \log f_{k,h}(y_{t,h}; \lambda) \Big|_{\lambda=\lambda_{k,h}^{(s)}}$$

or

$$\sum_{t=1}^T \hat{u}_{k,t}^{(s)} \cdot \frac{\partial^2}{\partial \theta^2} \log c_k(F_{j,1}(y_{t,1}; \lambda_{k,1}^{(s)}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}^{(s)}) | \theta) \Big|_{\theta=\theta_k^{(s)}}, \quad (16)$$

which can be calculated directly. Some of the calculations may be unnecessary with prior knowledge of the involved densities; for example, the left-hand term in (16) is certain to be negative if the mapping $\lambda \mapsto f_{k,h}(y; \lambda)$ is known to be strictly concave. The remaining elements of \mathbf{J}_g at $\xi^{(s)}$ and the corresponding spectral radius $\rho([\mathbf{D}(\xi^{(s)}) - \mathbf{L}(\xi^{(s)})]^{-1}\mathbf{U}(\xi^{(s)}))$ can, in principle, be computed exactly following the s 'th iteration of the EIFM algorithm, although in practice numerical approximation of these quantities would be considerably easier than explicitly calculating all of the required derivatives.

4.2. Analysis of the Estimator

We now discuss conditions under which the subvector η^* of the solution ξ^* ultimately produced by the EIFM algorithm—now regarded as a statistical estimator $\eta^*(\mathbf{Y}_{1:T})$ —is consistent for the true parameter η_0 for the model described in Section 2.1. While our analysis of the algorithm's convergence in Section 4.1 uses the same underlying technique as Elashoff and Ryan (2004) for their ES algorithm, we cannot emulate their proof of consistency. Their data, assumed to be iid, allow them to use standard techniques to prove the consistency and asymptotic normality of their estimator; however, the latent variables in our setup—namely, the conditional state membership indicators $U_{k,t}$'s and the $V_{j,k,t}$'s—are not independent due to their strong connection to the underlying Markov structure of the HMM. We instead follow the approach of Jensen (2011), who studied the asymptotic properties of M-estimators used for HMMs.

Specifically, we assume that the EIFM algorithm has produced a vector $\xi^* = (\mathbf{u}^*, \mathbf{v}^*, \eta^*)$ which satisfies $\mathbf{g}(\xi^*) = \mathbf{0}$, where \mathbf{g} is as defined in Section 4.1. To show that this estimator

has desirable asymptotic properties, it suffices to show that the lower $K + K^2 + dK + K$ components of \mathbf{g} (corresponding to the IFM-step of the algorithm) constitute a system of unbiased estimating equations for η conditional on $\mathbf{Y}_{1:T}$ after the latent variables $X_{1:T}$ have been marginalized out. To do this, we construct a new vector-valued function $\psi_T(\eta; \mathbf{Y}_{1:T})$ as follows. First, we define the random \mathbb{R} -valued functions

$$\phi_{k,t}^{[3]}(\eta; X_t) = (\delta_k - U_{k,1}) \cdot \mathbb{1}_{t=1}, \quad k \in \mathcal{X} \quad (17)$$

$$\phi_{j,k,t}^{[4]}(\eta; X_{(t-1):t}) = \left(\gamma_{j,k} \cdot \sum_{l=1}^K V_{j,l,t} - V_{j,k,t} \right) \cdot \mathbb{1}_{t>1}, \quad j, k \in \mathcal{X} \quad (18)$$

$$\phi_{k,h}^{[5]}(\eta; X_t, \mathbf{Y}_t) = U_{k,t} \cdot \frac{\partial}{\partial \lambda} \log f_{k,h}(Y_{t,h}; \lambda), \quad k \in \mathcal{X}, h \in \{1, \dots, d\} \quad (19)$$

$$\phi_k^{[6]}(\eta; X_t, \mathbf{Y}_t) = U_{k,t} \cdot \frac{\partial}{\partial \theta} \log c_k(\mathbf{F}_k(\mathbf{Y}_t; \boldsymbol{\lambda}_k) | \theta), \quad k \in \mathcal{X}, \quad (20)$$

where each $U_{k,t}$ and $V_{j,l,t}$ are implicitly functions of X_t and $X_{(t-1):t}$ respectively, as defined in Section 3.1. Note that only the functions in (17) and (18) depend on t ; the former dependence is required to capture the initial distribution parameters δ , and the latter is merely convenient because $V_{j,k,t}$ is not defined for $t = 1$. The remaining functions are free of t because of the time-homogeneity of the underlying Markov chain. Now, let

$$\phi_t(\eta; X_{(t-1):t}, \mathbf{Y}_t) = \left(\phi_{1,t}^{[3]}(\eta; X_t), \dots, \phi_K^{[6]}(\eta; X_t, \mathbf{Y}_t) \right)^\top \text{ and} \\ \psi_T(\eta; \mathbf{Y}_{1:T}) = \sum_{t=1}^T \mathbb{E}_\eta [\phi_t(\eta; X_{(t-1):t}, \mathbf{Y}_t) | \mathbf{Y}_{1:T}]. \quad (21)$$

From (8) to (9) and the algorithm's convergence to the fixed point ξ^* , we get $u_{k,t}^* = \mathbb{E}_{\eta^*} [U_{k,t} | \mathbf{Y}_{1:T}]$ and $v_{j,k,t}^* = \mathbb{E}_{\eta^*} [V_{j,k,t} | \mathbf{Y}_{1:T}]$, and from these equalities it is easily verified that η^* solves $\psi_T(\eta; \mathbf{Y}_{1:T}) = \mathbf{0}$. Its use as an estimator of η_0 is justified by the following fundamental fact, whose proof is given in Section S.4 of the supplementary material.

Theorem 4.2. Under standard regularity conditions, (21) defines an unbiased estimating equation.

The regularity conditions required in Theorem 4.2 are those that permit the interchange of differentiation and integration in certain cases; details are provided in the proof of the theorem. In the presence of iid data, the vector of state-dependent parameters $(\tilde{\lambda}_1, \dots, \tilde{\lambda}_d, \tilde{\theta})$ which solves $\psi_T(\eta; \mathbf{Y}_{1:T}) = \mathbf{0}$ is known to be a consistent and asymptotically normal estimator of the true data-generating vector of parameters as $T \rightarrow \infty$, whose asymptotic variance is given by the inverse Godambe information (Joe and Xu 1996). However, in order to account for the underlying dependence structure of the HMM, stronger regularity conditions are necessary for the same conclusion to hold. These conditions are encapsulated in the findings of Jensen (2011), which establish a central limit theorem for the sequence of estimators produced by the EIFM algorithm. The relevant

assumptions are restated in Section S.2 of the supplementary material using our notation. The main theorem is presented here:

Theorem 4.3. (Jensen 2011, Theorem 1) Define $\mathbf{G}_T = T^{-1} \text{Cov}_{\eta_0}(\boldsymbol{\psi}_T(\eta_0; \mathbf{Y}_{1:T}))$ and $\mathbf{H}_T = -T^{-1} \mathbb{E}_{\eta_0} [\mathbf{J}_{\boldsymbol{\psi}_T(\cdot; \mathbf{Y}_{1:T})}(\eta_0)]$. Under Assumptions 1–3 of Section S.2 of the supplementary material, there exists a consistent sequence $\{\hat{\eta}_T\}$ solving $\boldsymbol{\psi}_T(\eta; \mathbf{Y}_{1:T}) = \mathbf{0}$ such that $\sqrt{T} \cdot \mathbf{G}_T^{-1/2} \mathbf{H}_T (\hat{\eta}_T - \eta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$ as $T \rightarrow \infty$ under η_0 .

See Jensen (2011) for a proof. In all but the most trivial situations, the variances and expectations needed to calculate \mathbf{G}_T and \mathbf{H}_T (respectively) are analytically intractable. Therefore, we suggest two alternative approaches for estimating standard errors and confidence intervals for η^* . In both cases, we start by sampling n iid copies $(\mathbf{Y}_{1:T}^{(1)}, X_{1:T}^{(1)}), \dots, (\mathbf{Y}_{1:T}^{(n)}, X_{1:T}^{(n)})$ for some n from the model under η^* . The first approach involves estimating \mathbf{G}_T and \mathbf{H}_T via Monte Carlo, using η^* as a plug-in estimator for η_0 . Using the samples, we compute the observed information $\hat{\mathbf{H}}_T = -\frac{1}{nT} \sum_{i=1}^n \left[\mathbf{J}_{\boldsymbol{\psi}_T(\cdot; \mathbf{Y}_{1:T}^{(i)})}(\eta^*) \right]$ and the sample covariance matrix

$$\hat{\mathbf{G}}_T = \frac{1}{(n-1)T} \sum_{i=1}^n \left(\boldsymbol{\psi}_T(\eta^*; \mathbf{Y}_{1:T}^{(i)}) - \bar{\boldsymbol{\psi}}_T(\eta^*; \mathbf{Y}_{1:T}^{(1:n)}) \right) \times \left(\boldsymbol{\psi}_T(\eta^*; \mathbf{Y}_{1:T}^{(i)}) - \bar{\boldsymbol{\psi}}_T(\eta^*; \mathbf{Y}_{1:T}^{(1:n)}) \right)^\top,$$

where $\bar{\boldsymbol{\psi}}_T(\eta^*; \mathbf{Y}_{1:T}^{(1:n)}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\psi}_T(\eta^*; \mathbf{Y}_{1:T}^{(i)})$. The entries of each $\boldsymbol{\psi}_T(\eta^*; \mathbf{Y}_{1:T}^{(i)})$ can be computed by replacing the $U_{k,t}$'s and $V_{j,k,t}$'s in (17)–(20) with their conditional expectations given $\mathbf{Y}_{1:T}^{(i)}$ (i.e., $\hat{u}_{k,t}$ and $\hat{v}_{j,k,t}$ in (8) and (9), respectively) calculated with respect to η^* . While the Jacobian that is required to compute $\hat{\mathbf{H}}_T$ can be challenging to derive using symbolic computation software, it is typically not difficult to approximate numerically by exploiting its block diagonal form. Similarly, one can exploit the block diagonal structure of $\hat{\mathbf{H}}_T$ itself to reduce computational overhead in approximating the inverse Godambe information by $\hat{\mathbf{H}}_T^{-1} \hat{\mathbf{G}}_T \hat{\mathbf{H}}_T^{-1}$; we provide details in Section S.3 of the supplementary material.

The second approach relies on the parametric bootstrap estimator of Efron and Tibshirani (1994), as described by Zucchini, MacDonald, and Langrock (2017) for HMMs. Specifically, for each $i \in \{1, \dots, n\}$, we re-run the EIFM algorithm on $(\mathbf{Y}_{1:T}^{(i)}, X_{1:T}^{(i)})$ to produce a new estimator $\hat{\eta}^{(i)}$, and then estimate the covariance matrix of η^* via

$$\hat{\mathbf{V}}_{\eta^*} = (n-1)^{-1} \sum_{i=1}^n \left(\hat{\eta}^{(i)} - \bar{\eta} \right) \left(\hat{\eta}^{(i)} - \bar{\eta} \right)^\top$$

where $\bar{\eta} = \frac{1}{n} \sum_{i=1}^n \hat{\eta}^{(i)}$. The computation time for both methods is reasonable because the required tasks are parallelizable. Once the first and second order derivatives are computed, however, the Monte Carlo method becomes faster to implement since it does not require re-fitting the model. The latter may also be more suitable to construct confidence intervals when T is relatively large due to the use of η^* as a plug-in estimator, but may also be prone to numerical errors when the true η_0 lies near the boundary of the parameter space.

5. Simulations and Applications

5.1. Simulation Studies

The central aim of our simulation studies was to demonstrate that introducing copulas into the HMM formulation is advantageous when the dependence structure between observed variables is informative about the latent process, and that it is not harmful otherwise. Moreover, our experiments provide a proof of concept for the EIFM algorithm and, in our second simulation, they demonstrate its performance under copula misspecification.

In the first simulation study, we examined the convergence and robustness of the EIFM algorithm by applying it to a data-generating model with state-dependent distributions featuring several Archimedean and non-Archimedean copulas and three-parameter skew-normal marginals, the latter family written generally as $SN(\xi, \omega, \alpha)$. The number of states is $K = 4$, and the model has the stochastic representation

$$\mathbf{Y}_i \mid (X_i = k) \sim C_k \left(SN(\xi_{k,1}, \omega_{k,1}, \alpha_{k,1}), SN(\xi_{k,2}, \omega_{k,2}, \alpha_{k,2}) \mid \tau_k \right) \quad (22)$$

for $k \in \{1, \dots, 4\}$, with copula families C_k and individual state-dependent parameters as specified in Table 3 in Section S.5.1 of the supplementary material. The copula families and the strengths of dependence (as parameterized here by Kendall's tau) are both unique across states, as are the means of the marginal distributions; the state-dependent marginals also take on various combinations of positive and negative skewness in the first and second components, respectively. We chose skewness parameters of relatively high magnitude in order to avoid theoretical issues when these parameters are close to 0; for example, the Fisher information for the general $SN(\xi, \omega, \alpha)$ model is known to be singular at $\alpha = 0$ (Arellano-Valle and Azzalini 2008). The scale parameters $\omega_{k,h}$ were all fixed at 1, as larger values make estimation challenging even in the iid case (in practice, one could rescale the data prior to estimation and then adjust the resulting parameter estimates). The underlying Markov chain had initial distribution $\boldsymbol{\delta} = (1, 0, 0, 0)$ and each k 'th row of its transition matrix was constructed by applying k circular shifts to the vector $(0.1, 0.2, 0.3, 0.4)$.

We generated $T = 1500$ observations from this model using the `VineCopula` (Nagler et al. 2023) and `sn` (Azzalini 2022) packages in R. We adopted the k -means algorithm to obtain initial estimates for the parameters of the state-dependent distributions, initializing $\boldsymbol{\delta}$ and each $\boldsymbol{\gamma}_{k,\cdot}$ at small perturbations of $(1/4, 1/4, 1/4, 1/4)$. We ran the EIFM algorithm using a stopping rule with tolerance $\epsilon = 10^{-3}$, and then generated confidence intervals using both methods described in Section 4.2 with 100 resamples of the $\hat{\eta}$ data for each. The estimates produced by the k -means algorithm and the EIFM algorithm, and the resulting 95% confidence intervals produced by both methods, are shown in Section S.5.2 of the supplementary material; all figures were rounded to four decimal places. The tables also show empirical coverage probabilities of the bootstrap intervals based on 100 independently generated intervals, which generally deviate only narrowly from the expected value of 0.95 (certain off-diagonal elements of $\boldsymbol{\Gamma}$ have somewhat lower coverage, which may be due to some generated chains not transitioning enough to yield robust estimates). We repeated the entire experiment

using tolerance $\epsilon = 10^{-4}$; the corresponding tables may also be found in Section S.5.2 of the supplementary material.

In order to empirically assess the convergence rates of the EIFM algorithm when presented with differing amounts of data and different stopping rule tolerances, we re-generated data $Y_{1:T}$ from the same model with lengths $T \in \{500, 1000, 2500, 3000\}$, and for each of these, we ran k -means clustering followed by the EIFM algorithm with tolerances $\epsilon \in \{0.01, 0.00001\}$. We tracked the number of iterations required before convergence, as well as the classification accuracy from using the initial k -means clustering and local decoding (using the parameter estimates provided by the EIFM algorithm); these figures are given in Table 1. Unsurprisingly, the number of iterations increases as the stopping rule becomes more stringent, but increasing the number of observations T has little effect on either the number of iterations or the accuracy of local state decoding.

Our second simulation study tested the EIFM algorithm on higher dimensional data with two hidden states. The state-dependent distributions were 5-dimensional Markov trees (see Figure 2): special cases of regular vine copulas in which all conditional relationships are independent (Brechmann, Czado, and Aas 2012), whose specifications include 4-dimensional parameters. The extension to more general vine copulas requires the specification and sequential updating of all conditional pair copulas within the vine structure, which can be accomplished using the clustering ideas of Sahin and Czado (2022). The marginal distributions, all of the form $\mathcal{N}(\mu_{k,h}, 1)$, were intentionally chosen “close” to each other to force the copula dependence structure to carry a higher discrimination load. See Section S.5.3 of the supplementary material for the complete model specification. Further experiments suggested that when several marginal means are identical across states, the algorithm struggles to decode accurately and must be compensated for by stronger dependence between all pairs or a higher persistence in each state; this behavior was consistent with the results of the experiments in Section 2.2.

We generated a time series of length $T = 500$ from the model, and assessed the EIFM algorithm’s capacity for parameter estimation and state classification. Optimal values for the $\mu_{k,h}$ ’s are available in closed form at each iteration of the algorithm, thereby limiting potential obstacles in the estimation of the copula parameters in Step 2c of the algorithm. To provide initial values for the algorithm, we followed the

two-step approach described in Section 3.2, choosing as our initial clustering method an EM algorithm that assumed the state-dependent distributions to be multivariate normal with independent marginals. The entire process was replicated 100 times; we show boxplots of the resulting state-dependent parameter estimates in Figure 3. The parameter estimates are all very close to their corresponding true values. We also checked the accuracy of the state classifications (via local state decoding) made by the fitted misspecified independence model and the fitted data-generating model. In every replication, the latter was strictly more accurate than the former, with an average accuracy of 99.41% versus 84.82%. These results are consistent with the fact that the state-dependent marginal distributions are relatively uninformative of the underlying states, similar to the numerical example of Section 2.2. Figure 4 shows associated boxplots; these demonstrate that while the accuracy of the initial classifications varied substantially between replicates, that of the final classifications was highly consistent.

5.2. Occupancy Detection

The ability to detect whether a room is occupied using sensor data such as temperature and CO₂ levels can reduce unnecessary energy consumption by automatically controlling HVAC and lighting systems, without the need for motion detectors or other methods that could constitute an invasion of privacy (Candanedo and Feldheim 2016). We applied the model to three publicly-available labeled datasets presented by Candanedo and Feldheim (2016) consisting of three multivariate time series of four environmental measurements (light, temperature, humidity, CO₂) and one derived metric (the humidity ratio W , described in Candanedo and Feldheim 2016) captured in an office room in Belgium, along with indicators for whether the room was occupied at the time of each measurement. Candanedo, Feldheim, and Deramaix (2017) proposed an analysis that specified various HMMs with bivariate normal distributions and independent margins; model performance was assessed by comparing predictions made by local decoding to the ground-truth labels via the zero-one loss.

We illustrate our model using the same data, processed in the same manner as Candanedo, Feldheim, and Deramaix (2017) by averaging the data across five-minute periods, and then calculating the lagged differences to obtain time series of lengths 1629, 533, and 1951, respectively, with bivariate observations $Y_t = ((CO_2)_t, W_t \cdot 10^6)$. The first dataset was used for training. We used the k -means algorithm to partition the data into initial “unoccupied” and “occupied” subsets, and assigned the former label to the subset with more elements. We then chose state-dependent copula and marginal families. The state-dependent copula families were determined by following the approach outlined by Hofert et al. (2018). First, pseudo-observations were computed for each state (see Figure 5) so that analysis could be performed independently of the margins. Second, statistical

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

T :		500	1000	2500	5000
Tolerance:	0.01	14	24	23	15
	0.00001	230	115	460	269
Classifier:	k -means	0.9020	0.9090	0.9200	0.9196
	Local decoding	0.9640	0.9640	0.9696	0.9732



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

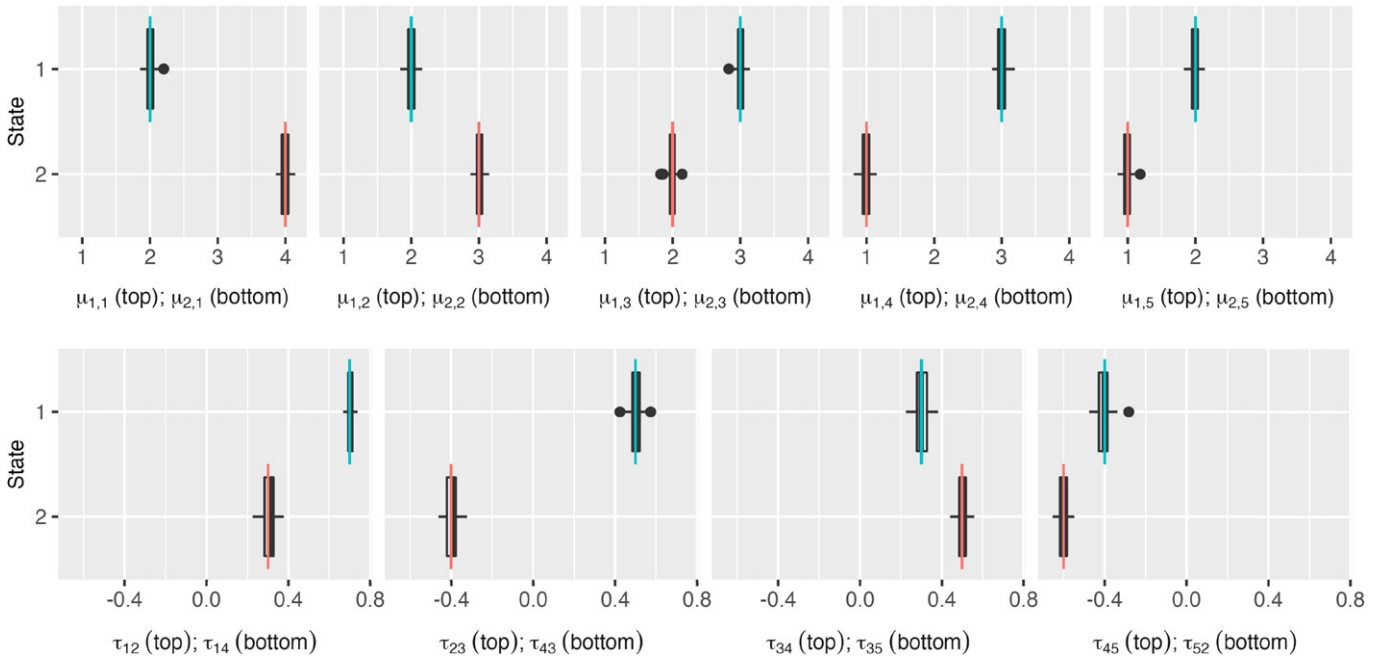


Figure 3. Estimates of marginal parameters (top) and copula parameters (bottom) for the Markov trees study, based on 100 independent simulations and EIFM algorithm runs.

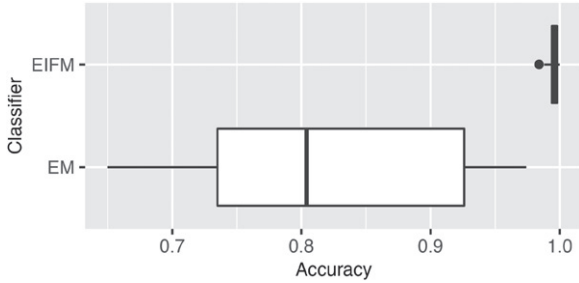


Figure 4. Accuracy across repetitions using the true data-generating model (EIFM) and the initial independence model (EM) for the Markov trees study.

tests of uncorrelatedness, exchangeability, radial symmetry, and extreme-value dependence were performed for each state as implemented in the `copula` package (Hofert et al. 2020); see Chapter 5 of Hofert et al. (2018) for more details. From these tests performed at the 95% confidence level, uncorrelatedness was ruled out in both states, no evidence presented against exchangeability in either state, and strong evidence presented against radial symmetry and extreme-value dependence in the unoccupied state (but no evidence presented against them in the occupied state).

Finally, we selected several common families of bivariate parametric copulas (including four of the five “classical” Archimedean families, the Gaussian and t families, and several extreme-value families and two-parameter families, as well as their rotated versions) not immediately ruled out by the above tests. For each, we computed the Akaike information criterion (AIC) for each state, as shown in Table 13 of Section S.5.4 of the supplementary material (for brevity, we have omitted AICs for the rotated copula families, which were identical to their non-rotated counterparts). The same table shows that the AIC selected the BB7 family for the unoccupied state, and the Tawn (type 1) family for the occupied state; both are two-parameter

families (Joe 2014). We also chose marginal distributions using basic model selection; since marginal selection was not the focus of our work, we limited our choices of marginal distributions to several continuous exponential families and evaluated them using the AIC and basic visual diagnostics. Ultimately, normal distributions yielded the best results for both marginal components within both states.

We labeled the unoccupied state as “1” and the occupied state as “2”, so that the model had the stochastic representation $\mathbf{Y}_t \mid (X_t = k) \sim C_k \left(\mathcal{N}(\mu_{k,1}, \sigma_{k,1}^2), \mathcal{N}(\mu_{k,2}, \sigma_{k,2}^2) \mid \theta_{k,1}, \theta_{k,2} \right)$ for $k \in \{1, 2\}$, where $C_1 = C_{\text{BB7}}$ and $C_2 = C_{\text{Tawn (type 1)}}$. Parameter initializations were based on clustering results and empirical Kendall’s taus produced by the k -means algorithm, as described in Section 3.2; the EIFM algorithm then converged after 14 iterations. Tables of parameter estimates produced by the algorithm, including those for the initial distribution of the Markov chain δ and the chain’s transition probability matrix Γ , are given in Tables 14–15 of Section S.5.4 of the supplementary material, along with an estimate of Kendall’s tau for each state; we also include 95% confidence intervals produced by both methods described in Section 4.2, with standard errors for Kendall’s tau produced using the delta method. Interestingly, 0 is contained in both intervals for $\hat{\theta}_{1,2}$, suggesting that the B5 subfamily (obtained as $\theta_{1,2} \rightarrow 0$ (Joe 2014)) may be a suitable model for the unoccupied state; similarly, 1 is contained in the intervals for $\hat{\theta}_{2,2}$, suggesting the Gumbel subfamily for the occupied state.

While the primary purpose of our study was to understand the data-generating mechanism by estimating the copula-within-HMM, the estimated model also allowed us to classify all three datasets and compare the model’s predictive performance with that of a more basic model featuring independent marginals, fitted using the standard EM algorithm for HMMs. We used local decoding to classify all three datasets, and assessed

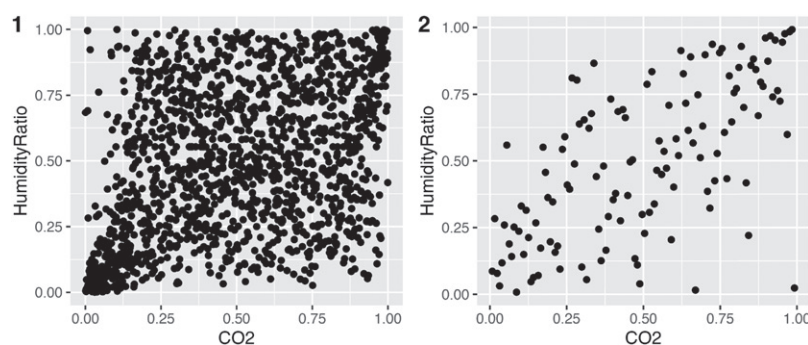


Figure 5. Pseudo-observations computed from unoccupied (Panel 1) and occupied (Panel 2) subsets of the data, as determined by an initial k -means clustering algorithm.

Table 2. Overall state classification accuracy for the occupancy data, using k -means clustering and local decoding via independence- and copula-based HMMs.

Classifier	Train	Test 1	Test 2
k -means clustering	0.865	0.818	0.788
HMM with independent marginals	0.895	0.846	0.679
HMM with BB7 and Tawn (type 1) copulas	0.900	0.852	0.690

the accuracy of these classifications via the zero-one loss. The performance of both models is shown in Table 2, with results similar to those in Candanedo, Feldheim, and Deramaix (2017). In all cases, as expected, the performance of the copula model exceeded that of the independence model. For comparison, the top row of Table 2 also displays the performance of k -means clustering, which outperformed both independence- and copula-based HMMs on the third dataset; this is to be expected, however, since Candanedo and Feldheim (2016) note that these data were collected under different conditions from the first two datasets.

6. Discussion

This article proposes a copula-based approach for integrating information from multivariate observations in an HMM setting. Dependence between the components of the observed vectors is captured using copulas, and we show that ignoring this dependence can deteriorate the predictive capacity of the statistical model. We meet the computational challenges of fitting the copula-within-HMM with a new iterative algorithm, the EIFM algorithm, which is broadly similar to the EM algorithm but differs in essential ways; most importantly, it allows for computation when other traditional methods fail.

The EIFM algorithm allows for development of other common HMM extensions when considering copula-based state-dependent distributions. One of the most common extensions is to assume a time-inhomogeneous state process such that the dynamics of the transition probability matrix Γ depend on time-varying covariates of interest. A related extension is to consider state-dependent parameters that vary over time. However, once the copula is introduced, variations of the HMM will usually require application-specific customizations; for example, one might want to allow for situations in which components of the vector of observations are observed at different times.

Overall, we aim to challenge the assumption of contemporaneous conditional independence often made in HMM set-

tings when working with multivariate data. We demonstrate the utility of several Archimedean and non-Archimedean copulas in an HMM setting and also demonstrate that with the aid of labeled data, we can choose different copula families across states. Our work here is limited to continuous-valued observation processes; discrete margins pose a fundamental problem in copula modeling, for it is known that any copula applied to discrete margins cannot be uniquely identified (Genest and Nešlehová 2007). Moreover, Trivedi and Zimmer (2017) have shown that in the iid setting, the nonuniqueness of the copula leads to biased estimation of the copula parameter θ ; we have observed that such behavior carries over to the EM and EIFM settings as well (although the EIFM algorithm still appears to produce unbiased marginal parameter estimates, as one might expect). Further, the support of the observation process may not always be the same across dimensions. For example, having a multivariate process where one dimension is discrete-valued, another continuous-valued and a third circular-valued is not uncommon in ecological and environmental applications of HMMs (Hodel and Fieberg 2022). Such situations will require customizing copula techniques that have been developed for mixed or discrete variables.

Supplementary Materials

The supplementary materials include R code and data used in the paper, details regarding the implementation of the EIFM algorithm (Section S.1), a list of regularity conditions sufficient for the Theorem 4.3 (Section S.2), details regarding the Monte Carlo approximation of the inverse Godambe information matrix described in Section 4.2 (Section S.3), proofs of results (Section S.4), and additional material related to Section 5 (Section S.5).

Acknowledgments

The authors thank the Associate Editor and two anonymous reviewers for their helpful comments and suggestions. RZ thanks Michaël Lalancette and Yanbo Tang for helpful discussions, and Amy Goldwater for proofreading the manuscript.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

RZ was supported by an Ontario Graduate Scholarship. RVC's and VLB's work has been supported by the Natural Sciences and Engineering Research Council of Canada.

ORCID

Robert Zimmerman  <http://orcid.org/0000-0002-0062-0926>

References

- Arellano-Valle, R. B., and Azzalini, A. (2008), “The Centred Parametrization for the Multivariate Skew-Normal Distribution,” *Journal of Multivariate Analysis*, 99, 1362–1382. [8]
- Azzalini, A. (2022), *The R package sn: The Skew-Normal and Related Distributions such as the Skew-t and the SUN*. Università degli Studi di Padova, Italia. R package version 2.1.0. [8]
- Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains,” *The Annals of Mathematical Statistics*, 41, 164–171. [4]
- Brechmann, E. C., Czado, C., and Aas, K. (2012), “Truncated Regular Vines in High Dimensions with Application to Financial Data,” *Canadian Journal of Statistics*, 40, 68–85. [9]
- 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 Building*, 112, 28–39. [9,11]
- Candanedo, L. M., Feldheim, V., and Deramaix, D. (2017), “A Methodology based on Hidden Markov Models for Occupancy Detection and a Case Study in a Low Energy Residential Building,” *Energy Building*, 148, 327–341. [9,11]
- Cappé, O., Moulines, E., and Ryden, T. (2005), *Inference in Hidden Markov Models*. Springer Series in Statistics. Berlin, Heidelberg: Springer-Verlag. [4]
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B*, 39, 1–22. [4,6]
- Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [8]
- Elashoff, M., and Ryan, L. (2004), “An EM Algorithm for Estimating Equations,” *Journal of Computational and Graphical Statistics*, 13, 48–65. [2,6,7]
- Frees, E. W., and Valdez, E. A. (1998), “Understanding Relationships Using Copulas,” *North American Actuarial Journal*, 2, 1–25. [1]
- Genest, C., Favre, A.-C., Béliveau, J., and Jacques, C. (2007), “Metaelliptical Copulas and Their Use in Frequency Analysis of Multivariate Hydrological Data,” *Water Resources Research*, 43, 1–12. [1]
- Genest, C., and Nešlehová, J. (2007), “A Primer on Copulas for Count Data,” *ASTIN Bulletin*, 37, 475–515. [11]
- Hand, D. J. (2020), *Dark Data: Why What You Don't Know Matters*, Princeton: Princeton University Press. [1]
- Härdle, W. K., Okhrin, O., and Wang, W. (2015), “Hidden Markov Structures for Dynamic Copulae,” *Econom. Theory*, 31, 981–1015. [1]
- Hodel, F. H., and Fieberg, J. R. (2022), “Circular–Linear Copulae for Animal Movement Data,” *Methods in Ecology and Evolution*, 13, 1001–1013. [11]
- Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2018), *Elements of Copula Modeling with R*, Cham: Springer. [9,10]
- Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2020), *copula: Multivariate Dependence with Copulas*, R package version 1.0-1. [10]
- Hofert, M., Mächler, M., McNeil, A. J. (2012), “Likelihood Inference for Archimedean Copulas in High Dimensions under known Margins,” *Journal of Multivariate Analysis*, 110, 133–150. [5]
- Jensen, J. L. (2011), “Asymptotic Normality of m-estimators in Nonhomogeneous Hidden Markov Models,” *Journal of Applied Probability*, 48, 295–306. [7,8]
- Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*, Boca Raton, FL: CRC Press. [5]
- (2014), *Dependence Modeling with Copulas*, Boca Raton, FL: CRC Press. [10]
- Joe, H., and Xu, J. J. (1996), “The Estimation Method of Inference Functions for Margins for Multivariate Models,” Technical Report 166, Department of Statistics, University of British Columbia. [2,5,7]
- Kwon, B. C., Anand, V., Severson, K. A., Ghosh, S., Sun, Z., Frohnert, B. I., Lundgren, M., and Ng, K. (2020), “Dpvis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways,” *IEEE Transactions on Visualization and Computer Graphics*, 27, 3685–3700. [2]
- McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R., and Patterson, T. A. (2020), “Uncovering Ecological State Dynamics with Hidden Markov Models,” *Ecology Letters*, 23, 1878–1903. [2]
- Meng, X.-L., and Rubin, D. B. (1993), “Maximum Likelihood Estimation via the ECM Algorithm: A General Framework,” *Biometrika*, 80, 267–278. [6]
- Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., and Erhardt, T. (2023), *VineCopula: Statistical Inference of Vine Copulas*, R package version 2.4.5. [8]
- 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. [1,2]
- Ortega, J. M., and Rheinboldt, W. C. (2000), *Iterative Solution of Nonlinear Equations in Several Variables*, Philadelphia: SIAM. [6,7]
- Ötting, M., Langrock, R., and Maruotti, A. (2023), “A Copula-based Multivariate Hidden Markov Model for Modelling Momentum in Football,” *AStA Advances in Statistical Analysis*, 107, 9–27. [2]
- Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017), “Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement,” *Journal of Agricultural, Biological and Environmental Statistics*, 22, 270–293. [2]
- Rabiner, L., and Juang, B. (1986), “An Introduction to Hidden Markov Models,” *IEEE ASSP Magazine*, 3, 4–16. [1]
- Sahin, Ö., and Czado, C. (2022), “Vine Copula Mixture Models and Clustering for Non-Gaussian Data,” *Economic Statistics*, 22, 136–158. [9]
- Sherlock, C., Xifara, T., Telfer, S., and Begon, M. (2013), “A Coupled Hidden Markov Model for Disease Interactions,” *Journal of the Royal Statistical Society, Series C*, 62, 609–627. [2]
- Sklar, A. (1959), “Fonctions de Répartition à n Dimensions et leurs Marges,” *Publications de l'Institut de statistique de l'Université de Paris*, 8, 229–231. [1]
- Stanislavsky, A., Nitka, W., Mał ek, M., Burnecki, K., and Janczura, J. (2020), “Prediction Performance of Hidden Markov Modelling for Solar Flares,” *Journal of Atmospheric and Solar-Terrestrial Physics*, 208, 105407. [2]
- Trivedi, P., and Zimmer, D. (2017), “A Note on Identification of Bivariate Copulas for Discrete Count Data,” *Econometrics*, 5, 10. [11]
- Zheng, K., Li, Y., and Xu, W. (2021), “Regime Switching Model Estimation: Spectral Clustering Hidden Markov Model,” *Annals of Operations Research*, 303, 297–319. [2]
- Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R*, Boca Raton, FL: CRC Press. [1,2,4,8]