2016 IEEE 16th International Conference on Data Mining Workshops

Markov Switching Copula Models For Longitudinal Data

Alfredo Cuesta-Infante
School of Computer Science
Universidad Rey Juan Carlos (Spain)
Email: alfredo.cuesta@urjc.es

Kalyan Veeramachaneni
CSAIL, MIT
Cambridge, MA (USA)
Email: kalyan@csail.mit.edu

***Abstract*—In this paper we present a novel Markov Switching generative model for continuous multivariate time series and longitudinal data based on Gaussian copula functions. We assume that the values of the multivariate time series at every time slice are sampled out of a joint probability distribution that is selected by the latent state. The use of Gaussian copula functions give the flexibility of individual marginals for each time series and a common dependence structure given by a correlation matrix. The transition matrix together with all the observation models are learned by means of Gibbs sampling. We also test the method both with synthetic and real data sets, and compare them with other usual techniques. Results show that models assuming normality in real data sets are not a good approach when marginals are not Gaussian, and they are outranked by our proposal.**

I. Introduction

In a machine learning and data analytics setting, researchers usually deal with multiple correlated signals, or *covariates*, such as features and variables. In addition, nowadays is frequent to have multiple variables repeatedly measured over long periods of time, usually referred to as *longitudinal data*. For instance, consider an online education setting, where students interaction with the online platform is monitored over time. Raw click stream data is processed to extract *covariates* for each week of the student engagement for further analysis such as drop out prediction or clustering. An important modeling paradigm in many time series and longitudinal data is the notion of *switching* generative models; where covariates are sampled from one distribution or another, according to the value of a latent variable or hidden state. A first kind of switching generative models assume that covariates are independent given the hidden state; e.g. [22] for continuous covariates, and [25] for discrete or categorical. Otherwise, continuous covariates are usually modeled jointly as multivariate normal (MVN) distributions or as a Dynamic Bayesian Network (DBN) after being discretized. See [10] for a comprehensive work on switching models.

The goal of this paper is to present a framework for constructing and estimating generative models such that it: (i) takes into account the temporal behavior of continuous covariates, (ii) does not assume conditional independence, (iii) does not need to transform the covariates into discrete variables, and (iv) captures the switching hidden state. In order to meet condition (i), we use the structure of a Hidden Markov Model (HMM) with at least two observations. Without loss of generality, let x_t be the hidden state at time t, and let y_t and z_t be two observed covariates at the same time that, according to condition (ii), they shall not be assumed conditionally independent given x_t. Therefore $p(y_t, z_t|x_t) \neq p(y_t|x_t) \cdot p(z_t|x_t)$. Let $c(y_t, z_t|x_t)$ be what the previous expression needs to satify the equality, i.e. the term that models how both covariates are coupled. Moreover, since $p(y_t|x_t)$ and $p(z_t|x_t)$, referred to as the marginals of the joint distribution $p(y_t, z_t|x_t)$, give the individual behaviour of each covariate, the term c can be considered as the dependence structure of the multivariate model. Fortunately, as long as covariates are continuous, $c(y_t, z_t|x_t)$ always exists, is unique, and is called the *copula density function*. In addition, the extension to another continuous covariate is straightforward: just multiply by its marginal and introduce the new variable into the copula density, e.g. with 3 covariates we would have $p(y_t, z_t, w_t|x_t) = p(y_t|x_t) \cdot p(z_t|x_t) \cdot p(w_t|x_t) \cdot c(y_t, z_t, w_t|x_t)$. Thus, by copula modeling we meet condition (iii). Finally, condition (iv) is satisfied by learning a different observation model depending on the state x_t.

All in all, the generative model that we present consists of a switching HMM, usually referred to as Markov Switching Model (MSM), such that the value of state x_t selects the dependence structure of two observed variables y_t and z_t, given by the copula, as well as the individual behavior, given by the marginals. The resulting *Markov Switching Copula Model* (MSCM) is learned by means of Gibbs sampling, following the Bayesian Hidden Markov Model methodology given in [22]. Our method assumes Gaussian copulas and Weibull marginals because of their flexibility to model many stochastic processes but it could be any other. However, when it comes to Gibbs Sampling, Weibull's shape and copula's parameter are known not to have a conjugate likelihood distribution. We overcome this problem by proposing an alternative way for sampling the Weibull's shape and using the results in [11] for sampling the copula's parameter. In addition we test our framework with a synthetic data set and a real data set related to the example given above. Specifically, we attempt to cluster `EdX` students sequentially, i.e. as they progress along the course.

II. Related work

There are a number of real scenarios in which processes involved have two or more dynamics. Usually neither the trigger nor the time when one dynamic gives way to another

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.57

are directly observable. Markov switching models (MSM) are a generative approximation to this problem, which considers that observables are generated according to a hidden state and attempts to construct an observation and a transition model for each possible value of the state. This approach has been used for detection of disease outbreaks [17], prediction of color trends [4] and classification of wind regimes [24], to cite a few in recent times. In econometrics a similar approach, referred to as Markov regime-switching, has been used for assesing the impact of oil shocks on exchange rates [2], or to detect abnormal states in stock market returns [21], to mention two recent works. In addition, [9] reviews bayesian methods for learning the parameters of Markov Switching Processes.

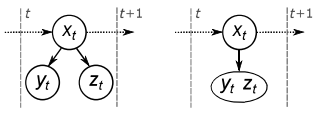
On the other hand, Copula functions are a valuable tool for constructing continuous multivariate models with two or more arbitrary marginals [18]. For this reason they have been broadly applied in Time Series modeling, and more recently are gaining momentum in Machine Learning. In the domain of time series, stationary models based on bivariate copulas have been constructed since the early work by Joe [13] to the recent multivariate approach in [23]. Assuming that the copula family and its parameter are kept constant along time, both a parametric estimation of the model with the classic methods described in [18], and a nonparametric kernel-based estimation [7] are possible. However, many realistic scenarios need not only arbitrary marginals, but also flexibility on the dependence structure provided by the copula. In finance, dynamics of an expansive market and a downturn market are not symmetric [20], and frequently there are collections of time series with different lengths [19]. Fermanian and Wegkamp generalise these two works introducing pseudo-copulas for dynamic dependence structures in [8].

In the domain of stochastic processes and machine learning models, Ibragimov uses them to represent Markov chains of arbitrary order [12]. Gordon and Ghahramani introduce the *copula process* together with Bayesian inference, prediction and modeling of volatility [26]. Efforts have also been done to integrete copulas in bayesian networks in [5] and [6]. This research has also been extended to copula constructions of many variables such as Vines in [15], [16]; as well as proposals with latent variables in [14].

Copulas have also been used jointly with Markov regime-switching processes in [3]; but neither the approach nor the goal are those proposed in this paper. The former include a hidden Markov chain in the equation describing dependence dynamics, allowing the unobserved time-varying dependence parameter to vary according to both a restricted ARMA process and an unobserved two-state markov chain. The latter, on the other hand, present a Markov regime-switching model constructed with a Rotated-Gumbel copula function and a marginal model; and their goal is to calculate the daily lower tail dependency between the chinese stock market and exchange system.

III. Markov Switching Copula Models

A Markov switching model (MSM) is similar to a HMM in the sense of there are two probability distributions: the



(a) (b)

x_t	y_t	z_t
1	$p_y(\cdot\|\ell_y^{(1)}, s_y^{(1)})$	$p_z(\cdot\|\ell_z^{(1)}, s_z^{(1)})$
2	$p_y(\cdot\|\ell_y^{(2)}, s_y^{(2)})$	$p_z(\cdot\|\ell_z^{(2)}, s_z^{(2)})$
$\vdots$	$\vdots$	$\vdots$
N_x	$p_y(\cdot\|\ell_y^{(N_x)}, s_y^{(N_x)})$	$p_z(\cdot\|\ell_z^{(N_x)}, s_z^{(N_x)})$

(a)

x_t	y_t	z_t	Copula
1	$p_y(\cdot\|\ell_y^{(1)}, s_y^{(1)})$	$p_z(\cdot\|\ell_z^{(1)}, s_z^{(1)})$	$c(\cdot,\cdot\|\Sigma^{(1)})$
2	$p_y(\cdot\|\ell_y^{(2)}, s_y^{(2)})$	$p_z(\cdot\|\ell_z^{(2)}, s_z^{(2)})$	$c(\cdot,\cdot\|\Sigma^{(2)})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N_x	$p_y(\cdot\|\ell_y^{(N_x)}, s_y^{(N_x)})$	$p_z(\cdot\|\ell_z^{(N_x)}, s_z^{(N_x)})$	$c(\cdot,\cdot\|\Sigma^{(N_x)})$

(b)

Fig. 1. (a) In a Switching Markov model, the state at time t selects the observation model for each variable Y and Z, parameterized by ℓ and s, which are conditionally independent. (b) In a Switching Markov Copula model Y and Z are jointly observed, so the state at time t also selects the dependence structure given by the parameter of the copula.

transition from state x_t to state x_{t+1}, and the observation at time t given the state x_t. In HMM both transition and observation distributions remain the same along time. On the contrary, in a MSM there is a number of possible observation models, and x_t selects the one to be applied at time t. If there are two or more variables or *covariates*, the observation models will be joint probability distributions. Under the assumption that the covariates are conditionally independent given the state; their joint probability distribution is just the product of every individual observation distribution or *marginal*. Otherwise, the most flexible way to jointly model continuous covariates is using copula functions.

A. Observation model

Copulas of d variables are joint probability distributions with support in $[0,1]^d$, zero-valued wherever at least one of their variables is zero and such that their own marginals are uniform. A comprehensive guide to this subject is [18]. Their importance is due to Sklar's theorem, which states that any given continuous d-variate joint density $p(x_1,\ldots,x_d)$, with marginal probability distributions $P_i(x_i)$ and marginal densities $p_i(x_i)$, for $i = 1,\ldots,d$, can be factorised into the independence density and the copula density; i.e.:

$$p(x_1,\ldots,x_d) = c\,(u_1,\ldots,u_d)\cdot p_I(x_1,\ldots,x_d); \qquad (1)$$

where $u_i = P_i(x_i)$, $c\,(u_1,\ldots,u_d)$ is the copula density, and $p_I(x_1,\ldots,x_d) = \prod_{i=1}^{d} p_i(x_i)$ is the *independence* density. Moreover, such factorization always exists and is unique for continuous variates. Notice that $u_i \sim \mathcal{U}[0,1]$; in other words, the copula, seen as a joint probability function of $U_1,\ldots,U_d$ has always uniform marginals. For this reason, from now on, whenever we refer to marginals, it is always to the individual

probability distribution of the covariates; in the scope of this paragraph that is $P_i(x_i)$.

Thus, modeling the joint distribution using copula functions has a number of advantages. First, if covariates are indeed independent, the copula density $c\left(P_1(x_1),\ldots,P_d(x_d)\right)=1$; in other words, there is a copula for modeling independence. If covariates are continuous we can model each one of them independently with the probability distribution that best fits, and then incorporate a dependence structure given by the copula. For instance, if such a dependence structure was given by a covariance matrix Σ, then we would have a model which is jointly Gaussian but marginally arbitrary. In addition, copulas provide a dual space in which all covariates are marginally uniformly distributed and jointly distributed according to the copula density.

For the sake of clarity, in this paper we only consider two covariates, Y and Z, but it can be extended straightforward to any number. Without loss of generality, we can assume that the hidden state X can only take a value at time t within a finite set, so $x_t\in\{1,2,\ldots,N_x\}$ and observations of Y and Z at that time are denoted as y_t and z_t respectively. In order to construct the joint distribution we have used the Gaussian copula with covariance matrix Σ together with Weibull distributions for the individual behaviour y_t and z_t, with parameter $\ell=\lambda^s$ being λ the scale, and shape s and expression

$$p_W(y;\ell,s)=(s/\ell)\left(y^{s-1}\right)\exp\left(-y^s/\ell\right).$$

We also assume that the copula is always Gaussian and the marginals are always Weibull; so that the switching models come given by variations in their parameters. Hence, the observation model at time t, with hidden state $x_t=i$, is

$$\begin{aligned}p_O(y_t,z_t|x_t=i)=c\left(P_W(y_t),P_W(z_t)|\Sigma^{(i)}\right)\cdot\\ p_W\left(y_t|\ell_y^{(i)},s_y^{(i)}\right)\cdot p_W\left(z_t|\ell_z^{(i)},s_z^{(i)}\right).\end{aligned}\tag{2}$$

In the expression above, the superscript (i) denotes the model switched to the value of x_t, the hidden state at time t. The assumptions made are quite mild and give a great power. With Gaussian copulas, to construct a $d-$dimensional joint distribution functions is just a matter of setting its covariance matrix. On the other hand we choose Weibull marginals because of their flexibility to approximate many distributions. Finally, Figure 1 depicts the building-block graph for MSM and MSCM together with the table of possible models.

B. Transition model

HMM, MSM and MSCM share the same transition model. For a finite number of possible states $x_t\in\{1,2,\ldots,N_x\}$ it consists of a probability mass function (pmf) across the initial state and a transition matrix. Let $\pi=[\pi_1,\ldots,\pi_{N_x}]$ be the pmf across x_1, the initial state; and let $A\in[0,1]^{Nx\times Nx}$ be the transition matrix such that the element $a_{i,j}=p(x_{t+1}=j|x_t=i)$ is the probability of getting state j in time $t+1$ from state i in t; for $i,j\in t\in\{1,2,\ldots,N_x\}$. For the sake of clarity, we introduce $\pi_{(i)}=[a_{i,1},\ldots,a_{i,N_x}]$, the pmf across the transition from $x_t=i$ to every x_{t+1}; in other words the i-th row of A.

IV. Learning Markov Switching Copula Models

According to the decisions made in the previous section, the set of parameters that define the whole model is $\Theta=\{\pi,\pi_{(i)},\ell_y^{(i)},s_y^{(i)},\ell_z^{(i)},s_z^{(i)},\Sigma^{(i)}\}$, for $i=1\ldots N_x$. However, by definition any copula $C(u,v)$ has uniform marginals in $[0,1]$; so it is straighforward to check that the standard deviation $\sigma_u=\sigma_v=1/\sqrt{12}$. Thus, in the particular case of a Gaussian copula, its covariance matrix Σ will be

$$\Sigma=\begin{pmatrix}\sigma_u^2 & \rho\sigma_u\sigma_v\\ \rho\sigma_u\sigma_v & \sigma_v^2\end{pmatrix}=\frac{1}{12}\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}.\tag{3}$$

Therefore, we substitute the covariance matrix $\Sigma^{(i)}$ with $\rho^{(i)}$, the correlation between u_t and v_t when $x_t=i$, as parameter of the copula. Finally, the set of parameters consists of $6N_x+1$ elements,

$$\Theta=\left\{\pi,\pi_{(i)},\ell_y^{(i)},s_y^{(i)},\ell_z^{(i)},s_z^{(i)},\rho^{(i)}\right\};\ i=1\ldots N_x.$$

Let $X_T=x_{1:t}$ be the sequence of hidden states from the very first time until the current time t. Likewise, let Y_T and Z_T be the sequence of observed covariates. In order to estimate the chain of states and the parameters of the model, it would be useful to get a number of samples from the joint distribution $J(X_T,\Theta|Y_T,Z_T)$ enough to approximate their distributions. Recalling Bayes rule $J(X_T,\Theta|Y_T,Z_T)\propto J(Y_T,Z_T|X_T,\Theta)J(X_T,\Theta)$. Hence, although $J(X_T,\Theta|Y_T,Z_T)$ is usually unknown, by making assumptions over the priors $J(X_T,\Theta)$ and the joint likelihood $J(Y_T,Z_T|X_T,\Theta)$, Gibbs sampling provides actual samples from the posterior $J(X_T,\Theta|Y_T,Z_T)$ after a *burn-in* number of iterations that can be discarded. Yet another reason for using Gibbs sampling is the big amount of parameters that have to be estimated.

The prior is usually chosen conjugate to the likelihood so that the posterior is in the same family than the prior but its parameters are updated with observations. However, the decision of using Weibull and Copulas introduces two difficulties. Firstly, Weibull distribution is not conjugate to any true prior for the shape parameter. Secondly, the conjugate prior for a covariance matrix is the Inverse Wishart distribution, but within this framework samples will be valid only if they satisfy the structure given in (3), which is hard to accomplish sampling randomly. Despite of the Inverse Wishart distribution is not a conjugate prior to the Gaussian copula [1], it is still possible to estimate its correlation matrix in a Gibbs sampling by means of the extended rank likelihood [11], which is the method we follow.

For the sake of compactness, we summarize the priors, likelihoods and posterior densities used in Table I. For a clearer notation, throughout the table we denote $\ell_i\equiv\ell_y^{(i)}$, $s_i\equiv s_y^{(i)}$, $\alpha\equiv\alpha_y$, $\beta\equiv\beta_y$, $\zeta\equiv\zeta_y$ and $\xi\equiv\xi_y$, Notice that there are similar priors, likelihoods and posteriors for z_t, $\ell_z^{(i)}$ and $s_z^{(i)}$. Notice also that prior densities selected depend on hyperparameters α,β,ζ and ξ, which appear here for the first time. The derivation of expressions in Table I is tedious but straightforward. Yet, in order to complete the Gibbs sampling

TABLE I
EXPRESIONS OF PRIOR, LIKELIHOOD AND POSTERIOR DENSITIES FOR PARAMETERS π, $\pi_{(i)}$, ℓ, s,

Par.	Prior	Likelihood	Posterior $\propto$	
π	$p_D(\pi;\underline{1})$	$p_M(x_1;\pi)$	$p_D(\pi;\underline{1}+I(x_1))$	(a)
$\pi_{(i)}$	$p_D(\pi_{(i)};\underline{1})$	$\prod_{t=1}^{T} p_M(x_{t+1};\pi_{(i)})$	$p_D(\pi_{(i)};\underline{1}+K(x_t=i))$	(b)
ℓ_i	$p_{IG}(\ell_i;\alpha,\beta)$	$\prod_{t=1}^{T} p_W(y_t;\ell_i,s_i)^{\mathbf{I}(x_t=i)}$	$p_{IG}(\ell_i;\alpha',\beta')$, $\alpha'=\alpha+n_i$, $\beta'=\beta+\sum_{t=1}^{T}\mathbf{I}(x_t=i)y_t^{s_i}$	(c)
s_i	$p_G(s_i;\zeta,\xi)$	$\prod_{t=1}^{T} p_W(y_t;\ell_i,s)^{\mathbf{I}(x_t=i)}$	$f_E\cdot p_G(s,\zeta',\xi')$, $\zeta'=\zeta+n_i$, $1/\xi'=\frac{1}{\xi}+\sum_{t=1}^{T}\ln y_t^{\mathbf{I}(x_t=i)}$, $f_E=\exp\left(-\frac{1}{\ell_i}\sum_{t=1}^{T}\mathbf{I}(x_t=i)y_t^{s_i}\right)$	(d)

The densities above are Dirichlet (p_D), Multinomial (p_M), Gamma (p_G), Inverse Gamma (p_{IG}) and Weibull (p_W). $\{\alpha,\beta,\zeta,\xi\}$ are the hyperparameters of the model. $\underline{1}$ is a vector of N_x ones. n_i=$\sum_{t=1}^{T}\mathbf{I}(x_t=i)$. Function $I(x_1)$ returns a vector of N_x-1 zeros and 1 one in the position $j=x_1$. Function $K(x_t=i)$ returns a row vector which position j is the count of visits from $x_t=i$ to $x_{t+1}=j$, for $t=1$ to T.

it is necessary to devise how to sample the Weibull's shape, the parameter of the copula and the chain of hidden states; together with setting the hyperparameters. In the following we address these issues and provide the full Gibbs sampling procedure.

1) Shape of Weibull across y_t.: Due to the factor f_E in Table I(d), prior and likelihood are not conjugate. Therefore it is necessary to devise a sampling method. Given that it takes values in $(0,1)$ and is monotone, the posterior will have the same aspect than the Gamma. Hence the solution that we propose has the following steps:

1) Locate the peak $\hat{s}=\arg\max p_G(s_i;\zeta',\xi')$, where p_G refers to the term in Table I(d).
2) Locate the values s_a and s_b such that $p(s_a|y_t)=p(s_b|y_t)=10^{-5}$.
3) Evaluate the posterior in Table I(d) for $s\in(s_a,s_b)$.
4) Compute the cumulative sum $P_G(s)$ for $s\in(s_a,s_b)$ and normalized it to 1.
   Let $P_G^{-1}(r)$ be the quantile function of $P_G(s)$.
5) Draw $r\sim\mathcal{U}(0,1)$; then sample from the posterior is $s=P_G^{-1}(r)$.

END

2) Correlation matrix of the copula.: Notice also, that for every different state we have previously sampled the scale and shape of every marginal, so that our given pairs $\{y_t^{(i)},z_t^{(i)}\}_{t=1,\dots,T}$ can be mapped to $\left\{u_t^{(i)}=P_W(y_t^{(i)};\ell_y^{(i)},s_y^{(i)})\ ,v_t^{(i)}=P_W(z_t^{(i)};\ell_z^{(i)},s_z^{(i)})\right\}$. Let Σ have an Inverse Wishart density $p_{IW}(\Sigma;n^0\Sigma^0,n^0)$, with n^0 degrees of freedom and covariance matrix Σ^0, and such that $E[\Sigma^{-1}]=(\Sigma^0)^{-1}$. Using this as a prior distribution it is possible to approximate samples of the correlation $\rho^{(i)}$ out of the posterior distribution [11]. Alike before, we only have to consider those times at which the state is i. Thus, for the sake of clarity we will omit the superscript (i), e.g. $\rho\equiv\rho^{(i)}$ and so on. Then, such a procedure is integrated in our Gibbs sampling as follows.

1) Let n_i be the cardinal of the set of pairs $\{(u_t,v_t):x_t=i\}$ for $t=1,\dots,T$ and $i=1,\dots,N_x$. And let $\Omega=[\nu_t,\omega_t]$ be the $n_i\times 2$ matrix obtained by means of $\nu_t=P_N^{-1}(u_t)$ and $\omega_t=P_N^{-1}(v_t)$, where P_N^{-1} denotes the quantile function of the standard Normal distribution, for every $t=1,\dots,T$.
2) Sample the covariance matrix Σ from the posterior
$$\Sigma\sim p_{IW}(\Sigma;n'\Sigma',n');\quad \text{with}\quad n'=n^0+n_i\ \text{ and }\ \Sigma'=\Sigma^0+\Omega^\top\Omega \tag{4}$$
3) The correlation is
$$\rho=\Sigma_{[1,2]}/\sqrt{\Sigma_{[1,1]}\Sigma_{[2,2]}}. \tag{5}$$

END

3) Updating the hidden chain distribution.: Given both data and parameters, the process proposed in this paper is a non-homogeneous Markov chain with initial distribution
$$p(x_1=i|\cdots)\propto\pi_i\cdot p_O(y_1,z_1|x_1=i)\cdot p(y_{2:N_t},z_{2:N_t}|x_1=i) \tag{6}$$
and transition probabilities
$$p(x_t=j|x_{t-1}=i,\cdots)\propto a_{ij}\cdot p_O(y_t,z_t|x_t=j)\cdot p(y_{t+1:N_t},z_{t+1:N_t}|x_t=j) \tag{7}$$
Notice that both (6) and (7) include a backward recursion step. In addition, probabilities obtained need to be normalised in order to be correct.

4) Setting the hyperparameters.: There are two options for estimating hyperparameters α_y, α_z, β_y, β_z, ζ_y, ζ_z, ξ_y, and ξ_z. The first one is to include them in the set Θ and extend the Gibbs sampling to draw from their posteriors; which leads to assume new priors with new hyperparameters. For instance, with α the posterior would be $p(\alpha|\ell_y^{(i)})\propto p(\alpha;h_\alpha,k_\alpha)p(\ell_y^{(i)};\alpha,\beta)$; where h_α and k_α are hyperparameters with respect to α. The problem then is forwarded to the estimation of h_α and k_α. Hence, adopting this solution leads to a *multi-level* Gibbs sampling that requires to decide beforehand how many levels

it will have. At the bottom level it is necessary to give a fixed value to its hyperparameters. The second option is to bound it in the first level by setting $\alpha_y = \alpha_y^*$, $\beta_y = \beta_y^*$, $\zeta = \zeta_y^*$ and $\xi = \xi_y^*$; and likewise for α_z, β_z, ζ_z, and ξ_z.

We propose to choose the second option with the following procedure for selecting the hyperparameters in terms of observations Y_T and Z_T.

1) Estimate the parameters ℓ_y^* and s_y^* that best fit Y_T to a Weibull distribution.
2) Generate $\vec{\ell}_y{}^*$, a vector of m samples uniformly distributed in $[\ell_y^* - \Delta, \ell_y^* + \Delta]$; where m and Δ are chosen arbitrarily.
3) Estimate parameters α_y^* and β_y^* that best fit $\vec{\ell}_y{}^*$ to an Inverse Gamma distribution.
4) Generate $\vec{s}_y{}^*$, a vector of m samples uniformly distributed in $[s_y^* - \Delta, s_y^* + \Delta]$; where m and Δ are chosen arbitrarily.
5) Estimate parameters ζ_y^* and ξ_y^* that best fit $\vec{s}_y{}^*$ to Gamma distribution.
6) Repeat steps 1 to 5 with α_z^*, β_z^*, ζ_z^*, ξ_z^* and Z_T

END

Finally, since we have two variables we set Σ^0 to the 2×2 identity matrix and $n^0 = 2 + 2$ so that the prior is the less informative one.

5) Sampling procedure.: Gibbs sampling MCMC goes over each marginal of the joint pdf $J(X_T, \Theta | Y_T, Z_T)$ sampling its variable. Thus, for this case, a full iteration of the method consists of:

1) Sample $\pi \sim$ Table I(a)
2) Sample A by sampling its rows $\pi_{(i)} \sim$ Table I(b), for $i = 1, \ldots, N_x$.
3) For $i = 1, \ldots, N_x$:
   a) Sample $\ell_y^{(i)} \sim$ Table I(c).
   b) Sample $s_y^{(i)} \sim$ Table I(d).
   c) Sample $\ell_z^{(i)} \sim$ Table I(c) adapted to z.
   d) Sample $s_z^{(i)} \sim$ Table I(d) adapted to z.
   e) Sample $\rho^{(i)} \sim$ (4) and (5).
4) Update the hidden chain in two steps:
   a) Sample $x_1 \sim$ (6).
   b) Sample $x_{t+1} \sim$ (7) for $t = 2, \ldots, T$.
5) Store $[x_{1:T}, \pi, A, \ell_y^{(i)}, s_y^{(i)}, \ell_z^{(i)}, s_z^{(i)}]_{i=1,\ldots,N_x}$ as a sample from $J(X_T, \Theta | Y_T, Z_T)$.

END

The posterior that is sampled at each step uses the parameters sampled in previous steps of same iteration k, if they are available. Otherwise it takes the samples from the previous iteration $k - 1$.

V. Experiments and discussion

In this section we begin testing our method with synthetic data generated from four different processes that range in difficulty. Then we present two distinctive machine learning tasks with different logitudinal data sets that are tackled with our proposal. Firstly we cluster sequentially students that took 15 week courses in the EdX platform. Then we predict acute hypotensive episodes in ICU patients.

TABLE II
Parameters of processes $S1$a, $S1$b, $S2$a and $S2$b

	$S1$				$S2$			
	a	b	{a,b}		a	b	{a,b}	
i	$s^{(i)}$	$s^{(i)}$	$\lambda^{(i)}$	$\rho^{(i)}$	$s^{(i)}$	$s^{(i)}$	$\lambda^{(i)}$	$\rho^{(i)}$
1	2	6	10	0.1	2	6	10	0.2
2	2	6	20	0.5	2	6	20	−0.5
3	2	6	30	0.8	2	6	30	0.7

$\ell = \lambda^s$. The rest of parameters are $\pi = [1/3, 1/3, 1/3]$, $\pi_{(1)} = [0.6, 0.3, 0.1]$, $\pi_{(2)} = [0.1, 0.8, 0.1]$, $\pi_{(3)} = [0.1, 0.3, 0.6]$.

TABLE III
Comparison of $-\log p(Y_T, Z_T | X_T, \Theta)$, the Neg-Loglikelihood of the three models tested for every data set.

	$S1$a	$S1$b	$S2$a	$S2$b
Ground truth	3.43	2.50	3.46	2.57
Copula model	**3.53**	2.87	**3.63**	**2.81**
MVN model	3.56	**2.65**	4.65	10.26
	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$

the best model is framed

A. Synthetic data set

We first consider four simulated processes, $S1$a, $S1$b and $S2$a and $S2$b; all with timespan $T = 500$, $N_x = 3$ hidden states and two observed random variables Y and Z such that the probability of jointly observe a pair $(y_t^{(i)}, z_t^{(i)})$ is (2). The configuration of each process is given in Table II. In all of them both $y_t^{(i)}$ and $z_t^{(i)}$ are marginally distributed according to a Weibull pdf with shape $s^{(i)}$ and scale $\ell^{(i)} = \lambda^{s^{(i)}}$, for $i = \{1, 2, 3\}$; where λ is the way that scale is represented in Table II. The structure of dependence is a Gaussian copula with correlation $\rho^{(i)}$. We have chosen this four process because all together present positive, negative and almost null correlations and different degrees of density overlapping. In addition the hidden chain is defined by a transition matrix

$$A = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}$$

and an uniformly distributed initial state. Samples from S1a and S2a are quite close, being difficult to tell how many latent states there are; whereas those from S1b and S2b are separeted enough to expect at least three latent states.

The Gibbs sampling has run 2000 iterations and the first half was discarded as *burn-in*. Finally, we validate the results with the following models:

- *Ground truth*: X_T is the actual Markov chain and Θ the actual set of parameters that were used to generate each data set $S1$a, $S1$b, $S2$a and $S2$b.
- *Copula model*: X_T is the last Markov chain obtained by Gibbs sampling, and Θ is the set of parameters estimated.

- *MVN model*: X_T is the last Markov chain and Θ consists of the mean μ and covariance matrix Σ assuming a multivariate normal (MVN) as the observation distribution.

In $S1$a and $S2$a the joint densities given the latent state broadly overlap; being harder to estimate. Notice that the estimated paramters are quite fuzzy. In contrast, with $S1$b and $S2$b both transition matrix A and correlation ρ are well estimated. However $S2$b is easier to estimate than $S1$b because of the negative correlation when $x_t = 2$. In order to validate the estimated models, we compare the negative log-likelihood $\mathcal{L} = -\log p(Y_T, Z_T|X_T, \Theta)$ of the three models with every process in Table II. This comparison is presented in Table III. Results show that data sets $S1$a and $S1$b are quite challenging because both copula and MVN models have a negative log-likelihood close to the ground truth. On the other hand, with data sets $S2$a, and especially $S2$b, copula models clearly explain better the sequential data.

VI. Conclusions

In this paper we have presented a framework for constructing generative models that is able to capture multiple observation dynamics, but with the key difference of incorporating copula functions into it. It consists of a graphical model similiar to a HMM with a discrete set of states, but such that the hidden state at time t switches one observation model out of a set of possibilities on, and switches the rest of them off. In addition we do not consider that the observations are conditionally independent given the state. Instead, our observation model is a joint probability distribution of the covariates constructed with a Gaussian copula for the dependence structure and a Weibull density for each variable observed. Thus, our model gains a great flexibility. We learn all the parameters of the model with a Gibbs sampling.

We want to remark that a key aspect about the contribution of using copulas is that any other marginal or copula could be used as well. In fact Gaussian copula is not a good option for modeling heavy-tail dependencies. For those cases, it is better to use an archimedean copula if there are only two covariates. A solution when there are more is to reduce the dimensionality, with PCA for example. However, in both cases, the cost to pay for not using the Gaussian is that the parameter of the copula can not be straightforward estimated in the Gibbs sampling. Rather an empirical bayes approach should be used.

VII. Acknowledgements

This research has been partially supported by the Spanish Government research funding ref. TIN 2015-69542-C2-1-R (MINECO/FEDER).

References

[1] P. Arbenz. Bayesian copulae distributions, with application to operational risk management. some comments. *Methodology and Computing in Applied Probability*, 15(1):105–108, 2013.

[2] Syed Abul Basher, Alfred A. Haug, and Perry Sadorsky. The impact of oil shocks on exchange rates: A markov-switching approach. *Energy Economics*, 54:11 – 23, 2016.

[3] Luo Changqing, Xie Chi, Yu Cong, and Xu Yan. Measuring financial market risk contagion using dynamic mrs-copula models: The case of chinese and other international stock markets. *Economic Modelling*, 51:657 – 671, 2015.

[4] T. M. Choi, C. L. Hui, S. F. Ng, and Y. Yu. Color trend forecasting of fashionable products with very few historical data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1003–1010, Nov 2012.

[5] E. Eban, G. Rothschild, A. Mizrahi, I. Nelken, and G. Elidan. Dynamic copula networks for modeling real-valued time series. In *JMLR Proc. of Int. Conf. on Artificial Intelligence and Statistics*, volume 31, pages 247–255, 2013.

[6] G. Elidan. Copula bayesian networks. In *Advances in Neural Information Processing Systems 23*, pages 559–567. 2010.

[7] J.D. Fermanian and O. Scaillet. Nonparametric estimation of copulas for time series. Technical report, Int. Center for Financial Asset Management and Engineering. Research paper No.57, 2002.

[8] J.D. Fermanian and M.H. Wegkamp. Time-dependent copulas. *Journal of Multivariate Analysis*, 110(0):19 – 29, 2012.

[9] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Bayesian nonparametric methods for learning markov switching processes. *IEEE Signal Processing Magazine*, 27(6):43–54, Nov 2010.

[10] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics, 2006.

[11] P.D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1(1):265–283, 2007.

[12] R. Ibragimov. Copula-based dependence characterizations and modelling for time series. Technical report, Harvard Institute of Economic Resarch. Discussion paper No.2094., 2005.

[13] H. Joe. *Multivariate models and dependence concepts*, volume Monographs on Statistics and Applied Probability, Vol. 73. 1997.

[14] S. Kirshner. Latent tree copulas. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models (PGM 2012)*, September 2012.

[15] D. Lopez-Paz, J.M. Hernandez-Lobato, and Z. Ghahramani. Gaussian process vine copulas for multivariate dependence. In *JMLR W&CP 28(2): Proceedings of The 30th Int. Conf. on Machine Learning*, pages 10–18. JMLR, 2013.

[16] D. Lopez-Paz, J.M. Hernandez-Lobato, and B. Schölkopf. Semi-supervised domain adaptation with copulas. In *Advances in Neural Information Processing Systems 25*, pages 674–682. Curran Associates Inc., 2013.

[17] H. M. Lu, D. Zeng, and H. Chen. Prospective infectious disease outbreak detection using markov switching models. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):565–577, April 2010.

[18] R. B. Nelsen. *An introduction to copulas*. Springer Series in Statistics, 2nd. edition, 2006.

[19] A.J. Patton. Estimation of multivariate models for time series of possibly different length. *Journal of applied econometrics*, 21:147–173, 2006.

[20] A.J. Patton. Modelling asymetric exchange rate dependence. *Int. Economic Review*, 47(2):527–556, 2006.

[21] Clment Rey, Serge Rey, and Jean-Renaud Viala. Detection of high and low states in stock market returns with {MCMC} method in a markov switching model. *Economic Modelling*, 41:145 – 155, 2014.

[22] T. Rydén. EM versus Markov chain Monte Carlo for estimation of Hidden Markov Models: A Computational Perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

[23] Michael Stanley Smith. Copula modelling of dependence in multivariate time series. *International Journal of Forecasting*, 31(3):815 – 833, 2015.

[24] P. J. Trombe, P. Pinson, and H. Madsen. Automatic classification of offshore wind regimes with weather radar observations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1):116–125, Jan 2014.

[25] L. R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):10–13, 2003.

[26] A. Wilson and Z. Ghahramani. Copula processes. In *Advances in Neural Information Processing Systems 23*, pages 2460–2468. 2010.