MCMC methods for continuous time multi-state models and high dimensional copula models

Ph.D. candidate: Rosario Barone
supervisor: Prof. Andrea Tancredi

Thesis in Statistics

XXXII Ph.D. cycle

Department of Methods and Models for Economics, Territory and Finance

Sapienza University of Rome

Contents

Introduction

In this Thesis we propose Markov chain Monte Carlo (MCMC) methods for several classes of models. We consider both parametric and nonparametric Bayesian approaches, proposing either alternatives in computation to already existent methods or new computational tools.

In particular, we consider continuous time multi-state models (CTMSM), that is a class of stochastic processes useful for modelling several phenomena evolving continuously in time, with a finite number of states. Inference for these models is straightforward if the processes are fully observed, while it presents some computational difficulties if the processes are discretely observed and there is no additional information about the state transitions. In particular, in the semi-Markov models case the likelihood function is not available in closed form and approximation techniques are required. In the first Chapter we provide a uniformization based algorithm for simulating continuous time semi-Markov trajectories between discretely observed points and propose a Metropolis within Gibbs algorithm in order to sample from the posterior distributions of the parameters of that class of processes. As it will be shown, our method generalizes the Markov case.

In the second Chapter we present a novel Bayesian nonparametric approach for inference on CTMSM. We propose a Dirichlet Process Mixture with continuous time Markov multi-state kernels, providing a Gibbs sampler which exploit the conjugacy between the Markov CTMSM density and the chosen base measure. The method, that is applicable with fully observed and discretely observed data, represents a flexible solution which avoid parametric assumptions on the process and allows to get density estimation and clustering.

In the last Chapter we focus on copulas, a class of models for dependence between random variables. The copula approach allows for the construction of joint distributions as product of marginals and copula function. In particular, we focus on the modelling of the dependence between more than two random variables. In that case, assuming a multidimensional copula model for the multivariate data implies that paired data dependencies are assumed

to belong to the same parametric family. This constraint makes this class of models not very flexible. A proposed solution to this problem is the vine copula constructions, which allows us to rewrite the multivariate copula as product of pair-copulas which may belong to different copula families. Another solution may be the nonparametric approach. We present two Bayesian nonparametric methods for inference on copulas in high dimensions. The first proposal is an alternative to an already existent method for high dimensional copulas. The second method is a novel Dirichlet Process Mixture of conditional multivariate copulas, which accounts for covariates on the dependence between the considered variables.

Applications with both simulated and real data are provided in the last section of the first and the second Chapters, while in the last Chapter there are only application with simulated data.

Chapter 1

MCMC methods for discretely observed continuous-time multi state models

Continuous time multi-state models (CTMSM) represent a widely applicable class of models useful for modelling phenomena - assuming a finite number of states - evolving continuously in time. Inference for these models present some computational difficulties when the process is only observed at discrete time points with no additional information about the state transitions. In particular, when transitions between states may depend on the time since entry into the current state, and semi-Markov models should be fitted to the data, the likelihood function is neither available in closed form. We are interested in providing an efficient Bayesian method for making inference when the likelihood function of a CTMSM is intractable. In fact we relax the assumption of Markovianity of the process considering the more general class of semi-Markov CTMSM.

In this Chapter we first introduce the class of processes we are working on, starting from the Markov CTMSM and extending the definition for the semi-Markov case. We explain the properties of the processes, therefore we discuss the computational difficulties when observations are discrete. In the second section we show how to bypass the likelihood calculation in the Markov case by simulating trajectories via uniformization based algorithm [Hobolth and Stone, 2009]. Then we introduce a Metropolis Hastings algorithm in order to generalize the paths simulation for the wider class of semi-Markov CTMSM. In the last section we set up a Markov Chain Monte Carlo algorithm for simulating the posterior distribution of the model parameters when sojourn times are assumed to be Weibull distribuited.

1.1 Markov and semi-Markov Continuous-time multi state processes

A continuos time stochastic process $\{X(t), t \in \mathcal{T}\}$ is a Markov process on the state space $\mathcal{S} = \{1, 2, \dots S\}$ and time space $\mathcal{T} \subset \mathbb{R}^+$ if

$$P(X(t+u = s|X(t) = r, \mathcal{F}_t) = P(X(t+u) = s|X(t) = r),$$

when $\mathcal{F}_t$ is the past history up to time t, for all $r, s \in \mathcal{S}$ and $u \in \mathbb{R}^+$.

The process $\{X(t), t \in \mathcal{T}\}$ is a *homogeneous* Markov process if

$$P(X(t+u) = s|X(u) = r) = P(X(u) = s|X(0) = r).$$

Let the *transition probabilities* for a time-homogeneous process be defined as

$$p_{rs}(t) = P(X(t) = s|X(0) = r).$$

Moreover let $P(t)$ be the transition probability matrix whose generic element is $p_{rs}(t)$.

The Markov process $X(t)$ can be defined via the transition intensity functions

$$q_{rs}(t) = \lim_{\delta t \to 0} \frac{P\left(X(t+\delta t) = s|X(t) = r\right)}{\delta t} \tag{1.1}$$

representing the instantaneous transition rate from state r to state s at time t.

For the time homogeneous Markov process we have

$$P\left(X(t+\delta t) = s|X(t) = r\right) = \begin{cases} \gamma_{rs}\delta t + o(\delta t) & s \neq r \\ 1 + \gamma_{rr}\delta t + o(\delta t) & s = r \end{cases} \tag{1.2}$$

where $\gamma_{rs} \geq 0$ and $\gamma_{rr} = -\sum_{s \neq r} \gamma_{rs} = -\gamma_r$, γ_{rs} and γ_r denote respectively the r-s and the diagonal element of the the *transition rates matrix*, or *generator matrix*

$$A = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1S} \\ \vdots & \ddots & \vdots \\ \gamma_{S1} & \cdots & \gamma_{SS} \end{pmatrix},$$

which satisfies the three following conditions:

1. $0 \leq -\gamma_{ii} < \infty \quad \forall i$;
2. $\gamma_{ij} > 0 \quad \forall i \neq j$;

3. $\sum_{j \in \mathcal{S}} \gamma_{ij} = 0 \quad \forall i.$

The *Chapman–Kolmogorov* equations for a continuous-time Markov chain are

$$P(t + \delta t) = P(t)P(\delta t)$$

for any $t \in \mathcal{T}$. Transition probabilities for a continuous-time Markov chain are functions of time t, derivable from the transition intensity functions of the chain.

The link between the generator matrix and the transition probability matrix for the time t is given by the *Kolmogorov differential equations*.
The forward equations and backward equations are

$$\frac{d}{dt}P(t) = P(t)A \quad \text{and} \quad \frac{d}{dt}P(t) = AP(t)$$

respectively, with $P(0) = I$.
Given A, the solution for $P(t)$ is the exponential matrix

$$P(t) = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} = \exp(tA).$$

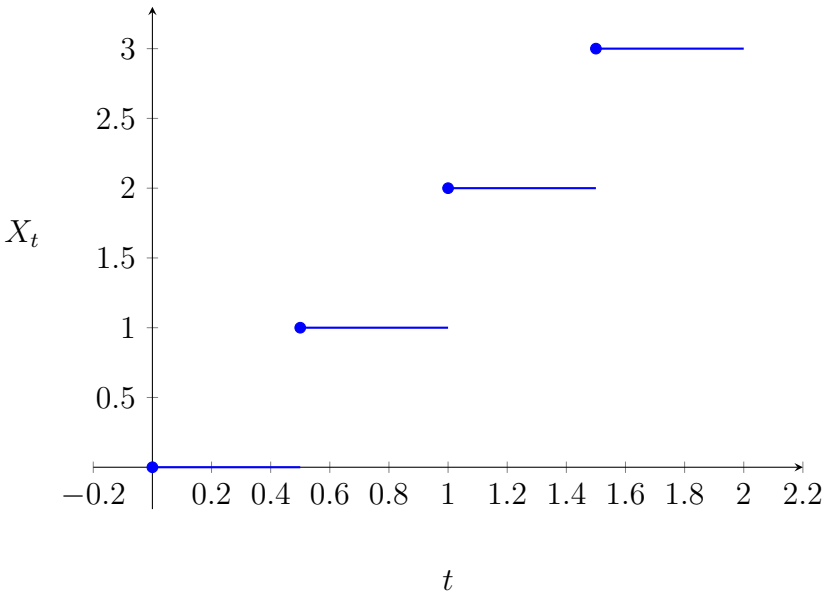
For the proof see Norris [1998] or Davison [2003].

If the process we are considering has one absorbing state, the row of the generator matrix A corresponding to the absorbing state has all entries equal to zero. Once the process is in an absorbing state r, the rate γ_{rs} of moving to another state s is zero. As a consequence, the diagonal entry γ_{rr} is zero as well. The row in the transition probability matrix $P(t)$ that corresponds to an absorbing state has the diagonal entry equal to one and all off-diagonal entries equal to zero, meaning that the probability of being in the same state in the future is one.

1.1.1 Poisson process as Markov chain

The Poisson process will have an important role in our work. In fact it will be used for simulating CTMSM. Moreover it represents a simple example of continuous time Markov process although with a countable state space. The

Figure 1.1: The path of a Poisson Process.



rate matrix for the Poisson process is then the infinite matrix

$$A = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots & \dots \\ \vdots & 0 & -\lambda & \lambda & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

We now give the definition of Poisson process in terms of jump chain and sojourn times (or holding times).

A right continuous process $(X_t, t \in \mathcal{T})$ which takes values in $\{1, 2, \dots\}$ is a *Poisson process* of rate $\lambda > 0$ if

- its state chain i.e. the state sequence generated by the process is given by $Y_n = n$.
- for each visited state the sojourn times $w_1, w_2, ...$ are Exponential random variables with rate λ;

The picture shows an example of the path of a Poisson Process. The Markovianity of X_t is guaranteed from the memoryless of the exponential

sojourn times, which leads to the memoryless property of the whole Poisson process.

1.1.2 State chain and sojourn times

We have introduced the continuous-time Markov processes by the transition intensities functions. Moreveover, we analyzed the relationship between the generator matrix A and the transition probability matrix P(t). Here we focus on another - equivalent - representation, which allows us to define the continuous-time Markov process as a tuple of states and sojourn times.

Let $\mathcal{S}$ be a countable state space. The rate matrix A provides all the information we need for the continuous-time Markov process construction. Given a rate matrix A, we may define the jump matrix $P = (p_{ij} : i, j \in \mathcal{S})$, where

$$p_{ij} = \begin{cases} \gamma_{ij}/\gamma_i & \text{if } j \neq i \text{ and } \gamma_i \neq 0 \\ 0 & \text{if } j \neq i \text{ and } \gamma_i = 0 \end{cases},$$

$$p_{ii} = \begin{cases} 0 & \text{if } \gamma_i \neq 0 \\ 1 & \text{if } \gamma_i = 0 \end{cases}.$$

In each row we take the off diagonal entries and scale them so they add up to 1, while the diagonal element is 0. If there are no entries, we put 1 on the diagonal element of A.

Then, a continuous-time Markov process can be defined in terms of its jump chain and holding times. Precisely, a discrete state space continuous time stochastic process $(X_t, t \in \mathcal{T})$ is a continuous-time Markov process with transition rate matrix A

- if its jump chain $(Y_n, n > 0)$ is discrete-time Markov chain with transition probability matrix P;
- if for each $n>1$, conditional on $Y_0, Y_1, \ldots, Y_{n-1}$, its holding times $w_1, \ldots, w_n$ are independent exponential random variables of parameters $\gamma_{(Y_0)}, \ldots, \gamma_{(Y_{n-1})}$ respectively [Norris, 1998].

1.1.3 Semi-Markov CTMSM

The definitions of Markov process given in (1.1) and (1.2) can be extended to the more general case of continuous-time semi-Markov processes.

We define a semi-Markov CTMSM via its transition intensity function,

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \to 0} \frac{P\{X(t+\delta t) = s | X(t) = r, \mathcal{F}_t\}}{\delta t},$$

where $\mathcal{F}_t$ represents the natural filtration associated to the process at time t. More precisely, we rewrite

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \to 0} \frac{P\{X(t+\delta t) = s | X(t) = r, T^* = t-u\}}{\delta t},$$

where T^* denotes the entry time in the last state assumed before time t. Unlike the Markovian case, here the transition intensity functions also depend on the time spent in the current state. In particular, setting

$$P\{X(t+\delta t) = s | X(t) = r, T^* = t-u\} = \begin{cases} q_{rs}(u)\delta t + o(\delta t) & s \neq r \\ 1 - \sum_{l \neq r} q_{rl}(u)\delta t + o(\delta t) & s = r \end{cases}$$

we describe a time-homogeneous semi-Markov process $X(t)$.

Note that the semi-Markov processes can be expressed by the properties of the state sequence and the sojourn times. Let $F_r(u)$ be the distribution with hazard function $F_r'(u)/(1 - F_r(u)) = \sum_{l \neq r} q_{rl}(u)$. Consider $p_{rs} = \int_0^\infty q_{rs}(u)(1 - F_r(u))du$ and $F_{rs}(u) = \frac{1}{p_{rs}} \int_0^u q_{rs}(v)(1 - F_r(v))dv$, for $s \neq r$. Then, $X(t)$ is the result of the state sequence generated by the Markov chain with transition probabilities p_{rs} and sojourn times depending on the departure and arrival states generated independently with distributions F_{rs}.

1.2 Inference and simulation for CTMSM

After having introduced the models from the mathematical point of view, we now focus on the computational and inferential aspects. CTMSM represent a useful class of stochastic processes for analyzing event history data [Lawless, 2013]. Practical applications of these models can be found in many fields. For example in biostatistics they are often used for modelling both disease progression and patient recovery after medical tratment, see for example Gentleman et al. [1994], O'Keeffe et al. [2011] and Ieva et al. [2017]. Other applications can be found in econometrics where, for example, CTMSM have ben adopted for modeling individual labor market status [Joutard et al., 2012] or credit rating transitions [Bladt and Sørensen, 2009].

Inference details for modeling data generated by specific classes of CTMSM differ if the sample paths are continuously observed or if the data only consist of the states observed at discrete time points, with no information about the state sequence and times of events between observation times. In fact, the former case does not present particular issues while the latter, usually referred to as panel data framework, apart from specific models may present considerable computational issues.

In fact, the problem with the panel data framework is that the likelihood function is available only for the Markov case or its simpler extensions and also in these cases it must be evaluated via numerical approximations. Kalbfleisch and Lawless [1985] were the first to introduce appropriate numerical techniques for the standard Markov CTMSM. Relaxing the Markov assumptions, for example by assuming a semi-Markov process where transitions between states may depend on the time since entry into the current state, leads generally to an intractable likelihood problem. To bypass the problem, Kang and Lagakos [2007] assume time-homogeneous transition intensities from at least one of the states while Armero et al. [2012] discuss a Weibull progressive disability model. Titman and Sharples [2010] focus on the tractable class of phase-type sojourn distributions while Titman [2014] suggests using phase-type distributions to approximate the likelihood of CTMSM with Gamma or Weibull sojourn time distributions.

All the proposals discussed above do not consider the possibility to reconstruct the whole sample paths in order to make inference via ordinary missing data techniques. However, note that a missing data formulation has been frequently adopted in the Markov case starting from Bladt and Sørensen [2005] where both an EM and a Gibbs sampler were proposed to estimate the parameters of a discretely observed Markov CTMSM. Anyway the Gibbs sampler was performed by a naive rejection sampling, i.e. by drawing unconditioned trajectories and discarding them that do not hit the right states. The limitations of the rejection sampling are discussed in Hobolth and Stone [2009], where it is also introduced a different sampling strategy based on the uniformization technique that permits to simulate directly Markov trajectories with fixed starting and ending states. The uniformization algorithm was used also to draw the distribution of a Markov sample path conditionally on a sequence of observed points by Fearnhead and Sherlock [2006] and was also used by Pfeuffer et al. [2018] for implementing a stochastic version (SEM) of the EM for the Markov case. A SEM algorithm was recently proposed also for the semi-Markov case by Aralis and Brookmeyer [2019] but the reconstruction of the sample paths was performed by a naive rejection sampling.

Finally Tancredi [2019] proposes approximate Bayesian computation (ABC) techniques for Markov and Semi-Markov cases by approximately matching the observed and simulated state transition matrices between different observation times.

1.2.1 Inference for fully observed continuous-time semi-Markov processes

When the whole process trajectory is observed, inference for semi-Markov continuous-time multi state models is straightforward. We first have to assume a probability distribution for the sojourn times, then we may write down the likelihood function of the process. Assuming as sojourn time density a function z depending on a vector of parameters θ, the only condition that must hold is that z is the density of a positive random variable.

We suggest to choose a generalization of the exponential distribution in order to include also the Markovian case, allowing to relax the assumption of independence of the time spent in the current state.

Let $(X_t, t \in \mathcal{T})$ be a continuous-time semi-Markov model of parameters $\theta = (\zeta, P)$ with discrete state space $\mathcal{S}$. Let $z(\cdot|\zeta_r)$ be the density function for the sojourn time in the state r. Note that we do not assume that sojourn times depend on arrival state. Moreover, let p_{rs} be the probability of a transition from state r to state s. Let $\zeta = (\zeta_1, \dots, \zeta_S)$ be the vector of sojourn times parameters and let P be the transition probability matrix. Suppose to observe the full trajectory between two points, 0 and T. Let $s = (s_0, s_1, \dots, s_\ell)$ be the state sequence and $w = (w_1, \dots, w_\ell)$ be the sojourn times sequence. Note that s_ℓ represents the last visited state, while w_ℓ represents the last sojourn time, truncated at the end point T, so that $\sum_{i=1}^{\ell} w_i = T$. Notice that $X_t \leftrightarrow (s, w)$. Finally, we assume that s_0 is observed.

Let $n_{rs} = \sum_{j=0}^{\ell-1} \mathbf{1}(s_j = r, s_{j+1} = s)$ be the transition counts from the state r to the state s, let $n_r = \sum_{s \neq r} n_{rs}$ be the total number of complete sojourns into the state r, i.e. excluding the truncated sojourn times. Moreover, let $w_{r\cdot} = (w_{r1}, \dots, w_{rn_r})$ be the sequence of observed sojourn times in the state r. We can write down the density of the single trajectory

$$p(s, w|\theta) = \prod_{rs} (p_{rs})^{n_{rs}} \prod_{r} \prod_{i=1}^{n_r} z(w_{ri}|\zeta_r) \cdot (1 - Z(w_\ell|\zeta_{s_\ell})) \,. \tag{1.3}$$

where $1 - Z(w_\ell|\zeta_{s_\ell})$ is the survival function for the last observation, due to

the truncation of the trajectory. Moreover, note that this factor does not appear if T is exactly the entry time in an absorbing state.

Let us indicate $(s, w)_j$ for $j = 1, \ldots, N$ the state and sojourn times sequences for the j-th trajectory. The likelihood function is then

$$\mathcal{L}(\theta) = \prod_{j=1}^{N} p((s, w)_j|\theta), \tag{1.4}$$

and it is available in closed form. Inference, both frequentist and Bayesian, does not present particular issues.

Problems arise if the trajectory is observed at discrete points. In that case, we do not know the exact jump times, hence we do not have information about the sojourn times in each state and the likelihood function is numerically available only for the Markov case, i.e. when the sojourn times are exponentially distributed.

Our goal is to provide an MCMC sampler which remedy to the problem of discrete observations of a continuous-time process and avoid the Markovian assumption of the analyzed processes.

1.2.2 Uniformization based algorithm

The uniformization is a sampling strategy for continuous-time Markov processes based on the methodology proposed by Jensen [1953]. This methods draws the continuous-time Markov process $(X_t, t \in \mathcal{T})$ trajectories via construction of an auxiliary stochastic process $(Y_t, t \in \mathcal{T})$.

Let A be the rate transition matrix of X_t and consider the transition probability matrix

$$R = I + \frac{1}{\lambda} A, \tag{1.5}$$

where $\lambda = \max A_{ii}$. Note that via the transition matrix R the same state may be visited several times in a row. Consider also a Poisson process with rate λ. Then, the auxiliary process Y_t is obtained by taking the points drawn from the Poisson process as jump points and the state sequence generated by the transition probability matrix R and is called *Markov process subordinated to a Poisson Process*. Y_t is equivalent to the original continuous-time Markov

process X_t, indeed the transition matrix of X_t can be written as

$$P(t) = e^{At} = e^{\lambda(R-I)t} = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t R)^n}{n!} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} R^n. \tag{1.6}$$

It follows that the transition probability matrix $P_{ab}(t)$ of the Markov process X_t can be written as

$$P_{ab}(t) = P(X_t = b | X_0 = a) = e^{-\lambda t} \mathbb{I}_{a=b} + \sum_{n=1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} R^n_{ab}. \tag{1.7}$$

Now, without loss of generality, suppose to know the start point $X_0 = a$ and the end point $X_T = b$. The number of state changes for the auxiliary process conditional on his starting point a and endpoint b is given by

$$P(N = n | X_0 = a, X_T = b) = e^{-\lambda T} \frac{(\lambda T)^n}{n!} R^n_{ab} / P_{ab}(T). \tag{1.8}$$

Conditionally on the number of state changes $N = n$, the jump times are Uniform$(0, T)$ both also conditionally on the endpoints and unconditionally. Finally the state-sequence conditioned on the endpoints can be obtained by drawing the first n steps of a Markov chain with transition matrix R, initial state a and end state b.

Thus, we can simulate trajectories of the continuous-time Markov process X_t, given two observed points $X_0 = a$ and $X_T = b$, with the uniformization algorithm described Algorithm 1.

Moreover we can draw Markovian trajectories trajectories on a sequence of observations at times $0 = t_0, t_1, \ldots, t_m = T$ by iterating the uniformization algorithm between t_{i-1} and t_i for $i = 1, \ldots, m$. Our idea of generalization of Uniformization for sampling semi-Markov CTMSM finds application in Bayesian inference and it is basically quite intuitive. In a Markov Chain Monte Carlo framework, our intention is to implement a Metropolis Hastings to simulate trajectories. The idea is to propose via Uniformization Markovian paths, choosing as target a semi-Markov density.

1.2.3 The base measure for CTMSM

In this section we give the theoretical details for the base measure of CTMSM. This applies for both Markov and semi Markov CTMSM.

Algorithm 1 Uniformization, Hobolth and Stone

Input: A, $X_0 = a$, $X_T = b$.
Output: (s, w).

- Compute $\lambda = \max A_{ii}$
- Compute $R = I + \frac{1}{\lambda}A$
- Simulate the number of state changes n from the distribution (2.8):

◦ If the number of state changes is 0, **stop**: $X_t = a$, $0 \leq t \leq T$;
◦ If the number of state changes is 1 and a=b, **stop**: $X_t = a$, $0 \leq t \leq T$;
◦ If the number of state changes is 1 and $a \neq b$, simulate t_1 from a Uniform$(0, T)$ and **stop**: $X_t = a$, $t \leq t_1$ and $X_t = b$ $t \geq t_1$;
◦ If the number of state changes n is at least 2:

1. simulate independently n independent uniform random variables in [0,T] and sort the number in increasing order to obtain the times of state changes $0 < t_1 < ... < t_n < T$ in order to get the *jump times*;

2. simulate X_{t_1},...,$X_{t_{n-1}}$ from a discrete time Markov chain with transition matrix R, conditional on $X_0 = a$ and $X_T = b$;

3. discard virtual state changes, compute $w_i = t_i - t_{i-1}$ and return the two-tuple containing the sequences of state changes and sojourn times (s, w).

For $m > 2$ observations of the same individual, the whole trajectory is generated simulating independently the paths between each pair of observed points in row.

Let X_t be a CTMSM with parameters $\theta = (\zeta, P)$. Let $\mathcal{S}$ and $\mathcal{T}$ be respectively the discrete state space and the time space. The process X_t takes values in the product union space $\bigcup_n \mathcal{S}^n \times \mathcal{T}^n$.

Consider the finite measure spaces $(\mathcal{S}^n, \Sigma_{\mathcal{S}^n}, \nu^n)$ and $(\mathcal{T}^n, \Sigma_{\mathcal{T}^n}, \mu^n)$, respectively the measure space of the states and the measure space of the sojourn times for a sequence of n jumps.

We define the product measure space $(\mathcal{M}^n, \Sigma_{\mathcal{M}^n}, \eta^n)$, where η^n is the product measure $\eta^n = \nu^n \times \mu^n$ and $\Sigma_{\mathcal{M}^n}$ is the product σ-algebra. Then, we define

$$\mathcal{M}^{\cup} \equiv \bigcup_{n=0}^{\infty} \mathcal{M}^n \equiv \bigcup_{n=0}^{\infty} \mathcal{S}^n \times \mathcal{T}^n$$

a *union space*, where each $\mathcal{M}^n$ is a product space.
Moreover, let $\Sigma_{\mathcal{M}}^{\cup}$ be the union product σ-algebra, where each measurable set $A \in \Sigma_{\mathcal{M}^{\cup}}$ can be expressed as

$$A = \bigcup_{n=0}^{\infty} A^n \quad \text{with} \quad A^n = A \cap \mathcal{M}^n \in \Sigma_{\mathcal{M}}^n.$$

Therefore, we assign this space the measure $\eta^{\cup}$, defined as

$$\eta^{\cup}(A) = \sum_{n=0}^{\infty} \eta^n(A^n).$$

Hence, $X_t \in \mathcal{M}^{\cup}$ has density w.r.t. $\eta^{\cup}$.

Thus, in general CTMSM have densities w.r.t. the base measure $\eta^{\cup}$. Let p be the density function of a CTMSM X_t, defined on the measure space $(\mathcal{M}^{\cup}, \Sigma_{\mathcal{M}}^{\cup}, \eta^{\cup})$. Then p is such that

$$\int_{\mathcal{M}^{\cup}} p(s, w) d\eta^{\cup}(s, w) = 1.$$

1.2.4 A Metropolis-Hastings for semi-Markov CTMSM

We now show how to simulate semi-Markov CTMSM trajectories conditional on the observed points. We have already seen how to simulate Markov CTMSM paths via the uniformization algorithm. Then, our idea is to use a Metropolis-Hastings algorithm having as a proposals distribution that of a Markov CTMSM and as a target distribution that of a semi-Markov CTMSM.

The choice for the semi-Markov CTMSM sojourn time distribution is quite arbitrary. Hence, since in Markov CTMSM sojourn times are exponentially distributed, we propose to use as the sojourn time density for the semi-Markov model a generalization of the exponential distribution.

Suppose we need to simulate the path of a semi-Markov CTMSM conditionally on the observed points $x = (x_0, \dots, x_m)$ at the times $t_0 = 0, \dots, t_m = T$. Let p_{SM} generally denotes the density under the semi-Markov model. The conditional density of the path is

$$\begin{aligned} \Pi_{SM}(s, w|x) &= \frac{p_{SM}(s, w, x)}{p_{SM}(x)} = \frac{p_{SM}(s, w)}{p_{SM}(x)} \\ &\propto \left(\prod_{rs} p_{rs} \right) \left(\prod_{r} \prod_{i=1}^{n_r} z(w_{ri}|\zeta_r) \right) \cdot (1 - Z(w_{\ell}|\zeta_{s_\ell})) \,. \end{aligned}$$

Since we are not able to simulate directly from $p_{SM}(s, w|x)$, we propose a sample path which matches the points exactly but relies on the Markov CTMSM. This sample path will be generated via the uniformization algorithm. In particular we iterate the simulation of end points conditional path between $[t_{i-1}, t_i]$ with states $[x_{i-1}, x_i]$ for $i = 1, \dots, m$. The proposal density is then

$$\begin{aligned} Q(s, w|x) &= \frac{p_M(s, w, x)}{p_M(x)} = \frac{p_M(s, w)}{p_M(x)} \\ &\propto \prod_{rs} \tilde{p}_{rs}^{n_{rs}} \prod_{r} \tilde{\gamma}_r^{n_r} e^{-\tilde{\gamma}_r \sum_{j=1}^{n_r} w_{rj}} \cdot e^{-\tilde{\gamma}_{s_\ell} w_{s_\ell}} \end{aligned}$$

where $\tilde{\gamma}_{rs}$ are the rate parameters of a Markov proposal process, $\tilde{p}_{rs} = \tilde{\gamma}_{rs}/\tilde{\gamma}_r$, while n_{rs} indicates the number of jumps from the state r to the state s; z is an exponential model with rate parameter $\tilde{\gamma}$.

The target density Π is proportional to the density of a continuous-time semi-Markov process. Instead, the sojourn time density is not exponential anymore. The only condition which must hold is that z is a positive random variable. Since our goal is to generalize the uniformization algorithm, we suggest to choose a probability model which contains as special case the exponential: Weibull and gamma distributions both respect our conditions.

Thus, given a last accepted path (s, w) and the uniformization output (s^*, w^*) acceptance ratio is

$$\alpha_{acc} = \min \left(1; \frac{\Pi(s^*, w^*|x)}{Q(s^*, w^*|x)} \cdot \frac{Q(s, w|x)}{\Pi(s, w|x)} \right) . \tag{1.9}$$

Note that by construction in (1.9)

$$\frac{\frac{p_{SM}(s^*,w^*)}{p_{SM}(x)}}{\frac{p_M(s^*,w^*)}{p_M(x)}} \cdot \frac{\frac{p_M(s,w)}{p_M(x)}}{\frac{p_{SM}(s,w)}{p_{SM}(x)}} = \frac{p_{SM}(s^*,w^*)}{p_M(s^*,w^*)} \cdot \frac{p_M(s,w)}{p_{SM}(s,w)}.$$

Moreover, if z is chosen to be Weibull or Gamma, the closer the analyzed process will be to the Markov CTMSM, the closer α will be to 1. The algorithm is summarized in Algorithm 2.

Algorithm 2 Uniformization based semi-Markov CTMSM sampler

Input: A, $X_0 = a$, $X_T = b$, $(s,w)_{(t)}$.
Output: $(s,w)_{(t+1)}$

- In the iteration $(t+1)$, let $(s,w) = (s,w)_{(t)}$ be the last accepted path;
- let (s^*,w^*) be the proposed path from uniformization;
- set $\alpha = \min\left(1; \frac{\Pi(s^*,w^*)}{Q(s^*,w^*)} \cdot \frac{Q(s,w)}{\Pi(s,w)}\right)$;

- draw $\omega \sim$ Unif(0,1):

  - **if** $\omega < \alpha$, $(s,w)_{(t+1)} = (s^*,w^*)$
  - **else** $(s,w)_{(t+1)} = (s,w)$

1.3 Bayesian inference for discretely observed semi-Markov CTMSM

We now present an MCMC method based on the modified uniformization sampler presented in the previous section, which allows us to simulate the posterior distribution of the parameters of a semi-Markov CTMSM when it is discretely observed and there is a panel data setting, We first introduce theoretically the method, then we show the behavior of the algorithm assuming a Weibull distribution for the sojourn times.

Let X_t be a semi-Markov CTMSM with probability transition matrix P and sojourn times distribution $z(\cdot|\zeta_r)$ for $r = 1,\dots,S$; let $\zeta = \zeta_1,\dots,\zeta_S$. We set $\theta = (\zeta, P)$. When X_t is only observed at discrete points, we propose to simulate the posterior distribution of θ by the following Metropolis within

Gibbs sampler algorithm.

Consider a panel structure, with at least two observations for each individual. Let n be number of individuals, i.e. the number of partially observed trajectories; let m_i be the number of observed points for each trajectory; let $0 = t_{i_0}, t_{i_1}, \ldots, t_{i_{m_i}} = T_i$ be the observation times and $x_i = (x_{i_1}, \ldots, x_{i_{m_i}})$ be the observed states at these times.

For each iteration of the Metropolis within Gibbs algorithm, we first simulate the whole trajectories $(s, w)_i$ for $i = 1, \ldots, n$ conditioned on the observed points x_i. For this task we use the Metropolis-Hastings step described in the previous section, constructing the rate matrix A of the Markov proposal distribution on the base of the parameters θ of the semi-Markov model. More details of this step will be given later. The updating of the vector of parameters ζ depends on the distribution we assume for the sojourn times.

Anyway note that conditionally on all the other parameters and the whole trajectories $(s, w)_i$, the parameters $\zeta_1, \ldots, \zeta_S$ are independent. Let $n_{r^{(i)}}$ be the number of complete sojourn times for the r-th state and the i-th trajectory, and let $w_{ir\cdot} = (w_{ir1} \ldots, w_{irn_{r(i)}})$. Moreover, let ℓ_i be the end state for the i-th trajectory. The conditional distribution of ζ_r given all the other unknown quantities is

$$\pi(\zeta_r | \ldots) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_{r(i)}} z(w_{irj}|\zeta_r) \cdot (1 - Z(w_{\ell_i}|\zeta_r))^{\delta(\ell_i, r)},$$

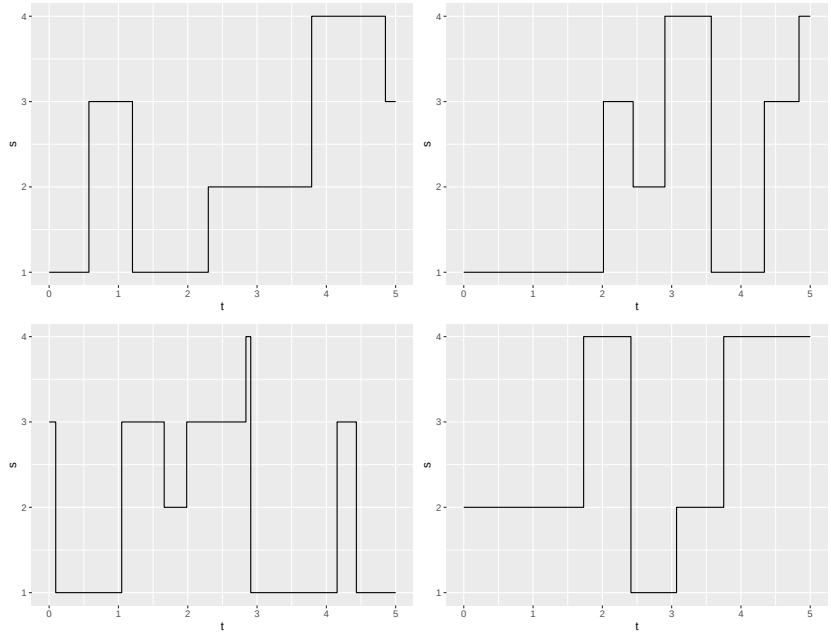
where

$$\delta(\ell_i, r) = \begin{cases} 1 & \text{if} \quad \ell_i = r \\ 0 & \text{if} \quad \ell_i \neq r \end{cases}$$

and generally can be simulated via a Metropolis-Hastings step.

The simulation of the conditional distribution of the transition probability matrix P is straightforward. In fact let $s = (s_1, \ldots, s_\ell)$ be the sequence of visited states. Let n_{rs} be the transition counts from the state r to the state state s. For each state r, the sequence of jumps into each other state represents a Multinomial$_{(k-1)}$ likelihood. If the state space $\mathcal{S} = 1, \ldots, S$ is such that $S > 2$, choosing as prior for the transition probabilities $p_r = (p_1, \ldots, p_{r-1}, p_{r+1}, \ldots, p_S)$ a Dirichlet$_{(k-1)}(m)$ we get

Figure 1.2: Simulated paths via semi-Markov CTMSM sampler.



$$\pi(p_r|s) \propto \left(\prod_{s \in \mathcal{S} \setminus r} p_{rs}^{n_{rs}} \right) \cdot \left(p_{r1}^{m_{r1}} \cdots p_{r(k-1)}^{m_{r(k-1)}} \right) = \prod_{s \in \mathcal{S} \setminus r} p_{rs}^{m_{rs}+n_{rs}},$$

where by conjugacy $p_r|s \sim \text{Dirichlet}_{(k-1)}(m + n_{rs})$. After simulating transition probabilities p_{rs}, the rate matrix A is updated setting $\tilde{\gamma}_{rs} = p_{rs} \cdot \gamma_r$. Note also that P and ζ are conditionally independent given the reconstructed trajectories.

1.3.1 Bayesian inference for Weibull sojourn times

We now assume that the sojourn times are Weibull distributed, that is the density for a complete sojourn time w in state r is

$$f(w|\alpha_r, \gamma_r) = \alpha_r(\gamma_r w)^{\alpha_r - 1} \cdot e^{-\gamma_r w^{\alpha_r}} \quad r = 1, \ldots, S.$$

We first focus on the path sampler. The target density w.r.t. $\eta^{\cup}$ is

$$\Pi((s,w)_i|x_i) \propto \prod_{rs} p_{rs}^{n_{rs(i)}} \prod_r \alpha_r^{n_{r(i)}} \gamma_r^{\alpha_r n_{r(i)}} \left(\prod_{j=1}^{n_{r(i)}} w_{rij} \right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n_{r(i)}} w_{rij}^{\alpha_r}} e^{-\gamma_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}} w_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}}}. \tag{1.10}$$

Algorithm 3 MCMC sampler for semi-Markov CTMSM

Iteration $(t+1)$; **n** individuals; **m** observations of the process for each individual; **d** observed covariates.

Input: $A^{(t)}$, $(s,w)^{(t)}$, $\theta^{(t)}$, $x = x_1, \ldots, x_m$, $0 = t_0, \ldots, t_m = T$.
Output: $A^{(t+1)}$, $(s,w)^{(t+1)}$, $\theta^{(t+1)}$

• for each individual **i** in **n**
◦ Get a semi-Markov trajectory $(s,w)_i^{(t+1)}$ via Uniformization based CTMSM sampler;
• update the parameters $\theta = (P, \zeta)$ drawing from $\pi(\theta|(s,w)^{(t+1)})$;
• for each individual **i** in **n**
◦ update the individual rate matrix $A_i^{(t+1)}$.

Thus, assuming z to be Weibull(α,γ) we define $\zeta = (\alpha, \gamma)$; our parameters of interest are $\theta = (P, \alpha, \gamma)$, with $\alpha = (\alpha_1, \ldots, \alpha_S)$ and $\gamma = (\gamma_1, \ldots, \gamma_S)$.

As a proposal distribution we take a Markov CTMSM distribution whose parameters are $\tilde{\theta} = (\tilde{P}, \tilde{\gamma})$ where $\tilde{P}$ is the transition probability matrix and $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_S)$ is the rate parameters vector. The proposal density w.r.t. $\eta^{\cup}$ for the i-th trajectory is

$$Q((s,w)_i|x_i) \propto \prod_{rs} \tilde{p}_{rs}^{n_{rs(i)}} \prod_r \tilde{\gamma}_r^{n_{r(i)}} e^{-\tilde{\gamma}_r \sum_{j=1}^{n_{r(i)}} w_{rij}} \cdot e^{-\tilde{\gamma}_{s_{\ell_i}} w_{s_{\ell_i}}}.$$

As a proposal density parameters we take $\tilde{\gamma}_{rs} = p_{rs}\gamma_r$ where p_{rs} and γ_r are the current parameters of the semi-Markov model. Note that the transition probabilities do not enter in the acceptance ratio of the Metropolis-Hastings step for the trajectories simulation. Let $(w^*, s^*)_i$ be the proposed trajectory for the i-th observation. Then the acceptance probability is such that $\alpha_{acc} = \min(1, r_{acc})$, where r_{acc} is

$$r_{acc} \propto \frac{\prod_{rs} p_{rs}^{n^*_{rs(i)}} \prod_r \alpha_r^{n^*_{r(i)}} \gamma_r^{\alpha_r n^*_{r(i)}} \left(\prod_{j=1}^{n^*_{r(i)}} w^*_{rij}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n^*_{r(i)}} w^{*\alpha_r}_{rij}} e^{-\gamma_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}} w^{*\alpha_{s_{\ell_i}}}_{s_{\ell_i}}}}{\prod_{rs} p_{rs}^{n^*_{rs(i)}} \prod_r \gamma_r^{n^*_{r(i)}} e^{-\gamma_r \sum_{j=1}^{n^*_{r(i)}} w^*_{rij}} \cdot e^{-\gamma_{s_{\ell_i}} w^*_{s_{\ell_i}}}} \times$$
$$\frac{\prod_{rs} p_{rs}^{n_{rs(i)}} \prod_r \gamma_r^{n_{r(i)}} e^{-\gamma_r \sum_{j=1}^{n_{r(i)}} w_{rij}} \cdot e^{-\gamma_{s_{\ell_i}} w_{s_{\ell_i}}}}{\prod_{rs} p_{rs}^{n_{rs(i)}} \prod_r \alpha_r^{n_{r(i)}} \gamma_r^{\alpha_r n_{r(i)}} \left(\prod_{j=1}^{n_{r(i)}} w_{rij}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n_{r(i)}} w^{\alpha_r}_{rij}} e^{-\gamma_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}} w^{\alpha_{s_{\ell_i}}}_{s_{\ell_i}}}} =$$
$$= \frac{\prod_r \alpha_r^{n^*_{r(i)}} \gamma_r^{\alpha_r n^*_{r(i)}} \left(\prod_{j=1}^{n^*_{r(i)}} w^*_{rij}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n^*_{r(i)}} w^{*\alpha_r}_{rij}} e^{-\gamma_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}} w^{*\alpha_{s_{\ell_i}}}_{s_{\ell_i}}}}{\prod_r \gamma_r^{n^*_{r(i)}} e^{-\gamma_r \sum_{j=1}^{n^*_{r(i)}} w^*_{rij}} \cdot e^{-\gamma_{s_{\ell_i}} w^*_{s_{\ell_i}}}} \times$$
$$\frac{\prod_r \gamma_r^{n_{r(i)}} e^{-\gamma_r \sum_{j=1}^{n_{r(i)}} w_{rij}} \cdot e^{-\gamma_{s_{\ell_i}} w_{s_{\ell_i}}}}{\prod_r \alpha_r^{n_{r(i)}} \gamma_r^{\alpha_r n_{r(i)}} \left(\prod_{j=1}^{n_{r(i)}} w_{rij}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n_{r(i)}} w^{\alpha_r}_{rij}} e^{-\gamma_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}} w^{\alpha_{s_{\ell_i}}}_{s_{\ell_i}}}}.$$

Secondly, we need to update the parameters θ. Note that we assume (α_r, γ_r) for $j = 1, \ldots, S$ to be independent a priori, with prior distribution

$$\pi(\alpha_r) = \frac{1}{\alpha_r \sqrt{2\pi}} \exp\left(-\frac{1}{2}\log(\alpha_r)^2\right),$$
$$\pi(\gamma_r) = \frac{1}{\gamma_r}.$$

The joint density of the sequence of sojourn times for the single individual i is

$$\mathbf{z}(w_{i\cdot}|\alpha, \gamma) \propto \prod_r \alpha_r^{n_{ir}} \gamma_r^{\alpha_r n_{ir}} \left(\prod_{i=1}^{n_{ir}} w_{rij}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{i=1}^{n_{ir}} w^{\alpha_r}_{rij}} e^{-\gamma_{s_{\ell_i}}^{\alpha_{s_{\ell_i}}} w^{\alpha_{s_{\ell_i}}}_{s_{\ell_i}}}.$$

With n observed individuals, we can write down the likelihood function and then the joint posterior density of the sojourn times parameters

$$\pi(\alpha, \gamma|w) \propto \prod_{i=1}^{n} \mathbf{z}(w_{i\cdot}|\alpha, \gamma) \cdot \pi(\alpha)\pi(\gamma) = \mathcal{L}(\alpha, \gamma) \cdot \pi(\alpha)\pi(\gamma).$$

Therefore the posterior distribution of γ_r is

$$\pi(\gamma_r| \ldots) \propto \gamma_r^{\alpha_r n_r} e^{-\gamma_r^{\alpha_r} T(W_{ir}; \alpha_r)},$$

where $T(W_{ir}; \alpha_r) = \sum_j w_{ir_j}^{\alpha_r}$ is a function of the total time spent in the state r and n_r is the number of visits in the state r. The posterior distribution of

α_r is

$$\pi(\alpha_r|\dots) \propto \alpha_r^{n_r-1}\gamma_r^{\alpha_r}\prod_{i=1}^{n_r} w_{r(i)}^{\alpha_r-1} e^{-\left(\frac{1}{2}\log(\alpha_r)^2+\gamma_r^{\alpha_r}T(W_{ir};\alpha_r)\right)}.$$

Next, transition probabilities $p_r = (p_1, \dots, p_{r-1}, p_{r+1}, \dots, p_S)$ are updated by drawing from $p_{r\cdot}|s \sim \text{Dirichlet}_{(k-1)}\left(m + \sum_{n_r s} \mathbf{1}\left(s_{r,s}\right)\right)$ and the rate matrix A is updated by setting $\tilde{\gamma}_{rs} = p_{rs} \cdot \gamma_r$.

Note that we may extend this structure in order to evaluate the impact of covariates on sojourn times. Suppose to have n observations and for each observation we observe d covariates. We assume that covariates affect the rates of the sojourn time distribution γ_r. Therefore, we reparametrize γ_r as

$$\log \gamma_r = \beta_{r0} + \beta_{r1} \cdot C_1 + \dots + \beta_{rd} \cdot C_d.$$

This further extension increases the computational complexity. We need to generate for each individual the whole trajectory between the discretely observed points. Then we update the unknown parameters $\theta = (P, \alpha, \gamma)$ as in the previous case. Unlike the previous case, here the rate matrix for the Markov proposal does not depend only on θ: now the rates also have dependence on the observed covariates, meaning that we will need n generator matrices A_j, setting for each individual j

$$\tilde{\gamma}_{rs} = \left(\exp\left(\beta_{r0} + \beta_{r1} \cdot C_{1j} + \dots + \beta_{rd} \cdot C_{dj}\right)\right) \cdot p_{rs},$$

with p_{rs} element of $p_r|s$ representing the probability of a transition from the state r to the state s.

1.4 Applications

Multi-state models allows for an extremely flexible approach that can model almost any kind of longitudinal failure time data. Particularly, this class of models have been widely applied in economics and biology. In this section we present results from applications in both the fields. We show the behavior of the algorithm with both simulated data and real data. In particular, we focus on two applications: the first one comprises credit rating data from Standard and Poor's, the second one is a medical application analyzing the ambulatory status for a set of woman affected by the breast cancer. In each of the experiments, for the parameters of interest we use the prior setting defined in Section 1.3.1.

Table 1.1: Simulated data: mean and standard deviation (in parentheses) of the posterior means across 100 samples of size $n = (50, 100, 500, 1000)$ under a three states Weibull semi-Markov model with one absorbing state, $\theta = (\gamma_{12}, \alpha_1, \gamma_{13}, \gamma_{21}, \alpha_2, \gamma_{23}) = (0.25, 1.4, 0.05, 0.04, 0.7, 0.1)$ and follow-up times equal to 0,3,6,12,24,60. Upper table: death time unknown. Lower table: death time exactly known.

n	γ_{12}	α_1	γ_{13}	γ_{21}	α_2	γ_{23}
50	0.25 (0.06)	1.43 (0.30)	0.07 (0.02)	0.05 (0.03)	0.91 (0.20)	0.09 (0.03)
100	0.25 (0.04)	1.42 (0.18)	0.06 (0.02)	0.05 (0.02)	0.81 (0.14)	0.10 (0.03)
500	0.25 (0.02)	1.39 (0.10)	0.05 (0.01)	0.04 (0.01)	0.73 (0.07)	0.10 (0.01)
1000	0.25 (0.02)	1.40 (0.08)	0.05 (0.01)	0.04 (0.01)	0.72 (0.05)	0.10 (0.01)

50	0.25 (0.05)	1.43 (0.23)	0.07 (0.02)	0.05 (0.03)	0.86 (0.17)	0.09 (0.03)
100	0.25 (0.04)	1.40 (0.17)	0.06 (0.02)	0.05 (0.02)	0.76 (0.11)	0.10 (0.03)
500	0.25 (0.02)	1.40 (0.07)	0.05 (0.01)	0.04 (0.01)	0.72 (0.06)	0.10 (0.01)
1000	0.25 (0.01)	1.40 (0.07)	0.05 (0.01)	0.04 (0.01)	0.71 (0.04)	0.10 (0.01)

1.4.1 Simulation study

Finally to assess the proposed methodology, we applied the MCMC algorithm to simulated data sets, partially replicating the experiment conducted by Titman [2014] and Tancredi [2019]. In particular, data were generated from a model with three states: healthy, ill, dead. All patients start in the healthy state and can recover from the ill state according to a Weibull model with transition intensity functions $q_{rs}(u) = \gamma_{rs}\alpha_r(u\gamma_r)^{\alpha_r - 1}$ where $\gamma_r = \sum_{s \neq r} \gamma_{rs}$. The exact model parameters are fixed to $\theta = (\gamma_{12}, \alpha_1, \gamma_{13}, \gamma_{21}, \alpha_2, \gamma_{23}) = (0.25, 1.4, 0.05, 0.04, 0.7, 0.1)$, corresponding to a process where the hazard of the transition out from the state is increasing with time for the healthy state and decreasing for the ill state. Moreover, the transition probability towards the dead state is greater under the ill state ($p_{23} = \gamma_{23}/\gamma_2 = 0.71$) than with the healthy state ($p_{13} = \gamma_{13}/\gamma_1 = 0.167$). Finally note that the follow-up times are set equal to (0,3, 6,12,24,60) months and that we consider both the cases with the death times unknown and known. We set the sample size at $n = 50, 100, 500$ and 1000 and for each sample size we generated 100 data sets running the MCMC algorithm for 10000 iterations. In Table 1.1 we present the empirical averages and standard deviations of the posterior means obtained for each simulated data set. Note that as the sample size increases, the Bayesian estimators to concentrate on the true values of the parameters both when the detah time is unknown (upper table) and when

it is know. We notice also that, as expected, the information introduced by assuming that the death time is exactly known provides always a smaller mean square error with respect to the unknown death time scenario.

1.4.2 Modelling rating classes with Standard and Poor's data

Multi state models are widely used in economics and finance. More precisely, in credit rating modelling they play an important role. A credit rating is an evaluation of the credit risk of a prospective debtor (an individual, a business, company or a government), predicting their ability to pay back the debt, and an implicit forecast of the likelihood of the debtor defaulting [Kronwald, 2010]. The credit rating represents an evaluation of a credit rating agency of the qualitative and quantitative information for the prospective debtor, including information provided by the prospective debtor and other non-public information obtained by the credit rating agency's analysts.

In the first application, we consider a data set of 205 institutions from all over the world, each one observed at least two times. These institutions are almost all independent countries. We summarize the data in four rating classes (A,B,C,D), with the first class (A) representing solvent institutions, while the fourth class (D) represents the default. For these data we fitted a continuous time Weibull semi-Markov model, assuming the default state as absorbing state. We assumed as prior distribution for the rate parameter $\pi(\gamma_j) = 1/\gamma_j$, while for the shape parameter we set $\pi(\alpha_j)$ as a log-normal distribution centered on 1. Via the MCMC sampler for discretely observed data described in the previous sections we generated the posterior distributions of the semi-Markov process parameters.

As already remarked, semi-Markov CTMSM differs from Markov processes since sojourn times also depends on the past history of the process. In Markov models sojourn times are exponential distributed and only depends on the rate of the process. Moreover, the expected sojourn time in the state r is $1/\gamma_r$. Here sojourn times are assumed to be Weibull; for each state r, the average sojourn time is

$$\bar{w}_r = \frac{1}{\gamma_r}\Gamma\left(1+\frac{1}{\alpha_r}\right) = \frac{1}{\gamma_r\alpha_r}\Gamma\left(\frac{1}{\alpha_r}\right)$$

where $\Gamma(\cdot)$ represents the gamma function.

Table 1.2: Credit rating modelling: posterior mean and standard deviation for the model parameters.

Semi-Markov model

	p_{12}	p_{13}	p_{14}
$E(\cdot\|y)$	0.8580	0.0760	0.0660
$SD(\cdot\|y)$	0.0907	0.0716	0.0617

	p_{21}	p_{23}	p_{24}
$E(\cdot\|y)$	0.2607	0.5705	0.1688
$SD(\cdot\|y)$	0.1012	0.1012	0.0976

	p_{31}	p_{32}	p_{34}
$E(\cdot\|y)$	0.1060	0.2106	0.6834
$SD(\cdot\|y)$	0.0946	0.1243	0.1398

	γ_1	γ_2	γ_3
$E(\cdot\|y)$	0.0086	0.0440	1.3522
$SD(\cdot\|y)$	0.0073	0.0151	1.2780

	α_1	α_2	α_3
$E(\cdot\|y)$	0.6868	1.2676	0.7679
$SD(\cdot\|y)$	0.1995	0.2836	0.3121

Markov model

	p_{12}	p_{13}	p_{14}
$E(\cdot\|y)$	0.8648	0.0687	0.0665
$SD(\cdot\|y)$	0.0862	0.0644	0.0621

	p_{21}	p_{23}	p_{24}
$E(\cdot\|y)$	0.2796	0.5028	0.2176
$SD(\cdot\|y)$	0.0971	0.1118	0.0919

	p_{31}	p_{32}	p_{34}
$E(\cdot\|y)$	0.0941	0.1942	0.7117
$SD(\cdot\|y)$	0.0853	0.1167	0.1324

	γ_1	γ_2	γ_3
$E(\cdot\|y)$	0.0205	0.0294	0.4992
$SD(\cdot\|y)$	0.0057	0.0068	0.1670

The traces of the posterior distributions are represented in Figure 1.3 and Figure 1.4. Results are shown in Table 1.2 and Table 1.3. The difference in terms of estimated average sojourn times shows the sensitivity of the CTMSM with respect to the model assumptions. In particular the semi-Markov do not assume the memoryless of the sojourn times distributions and this fact seems to produce very different results with respect to the Markov case. Anyway, in both cases we observe that the rate of the process decreases as the class increases, with average sojourn times increasing as the rating class increases, meaning that Countries in higher classes show a greater stability in terms of solvency. The increasing rate is justified from market dynamics, which tends to reduce investments in countries which have been downgraded.

Figure 1.3: MCMC samples of sojourn times parameters for S&P data.

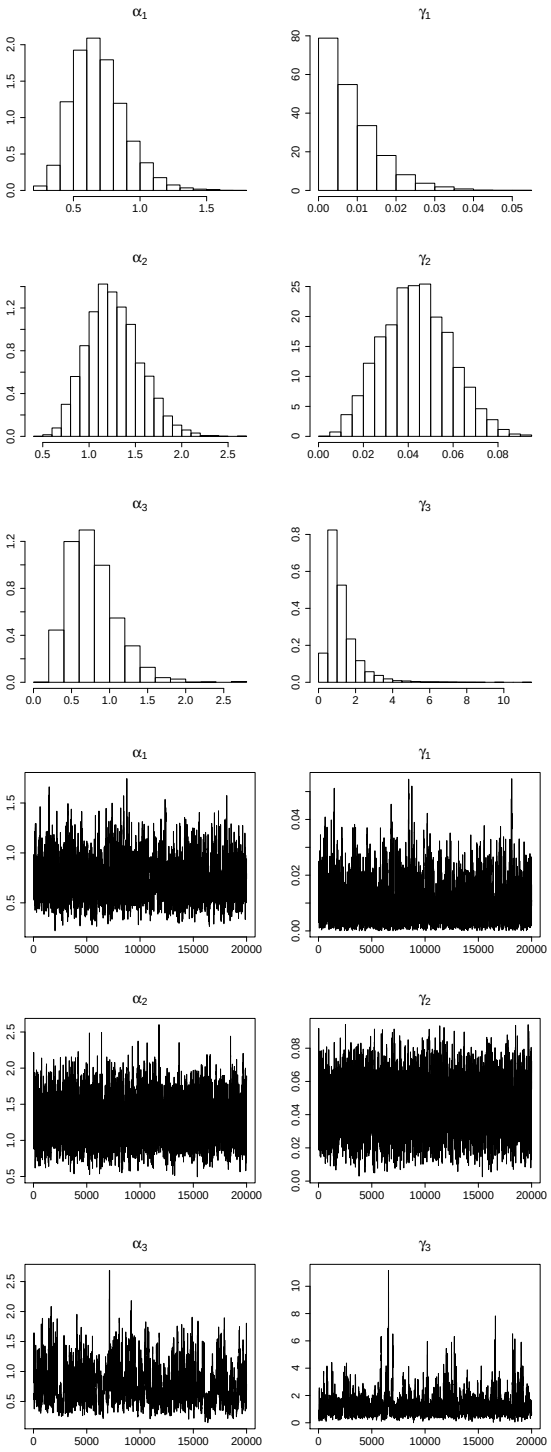


Figure 1.4: MCMC samples of transition probabilities posterior parameters for S&P data.

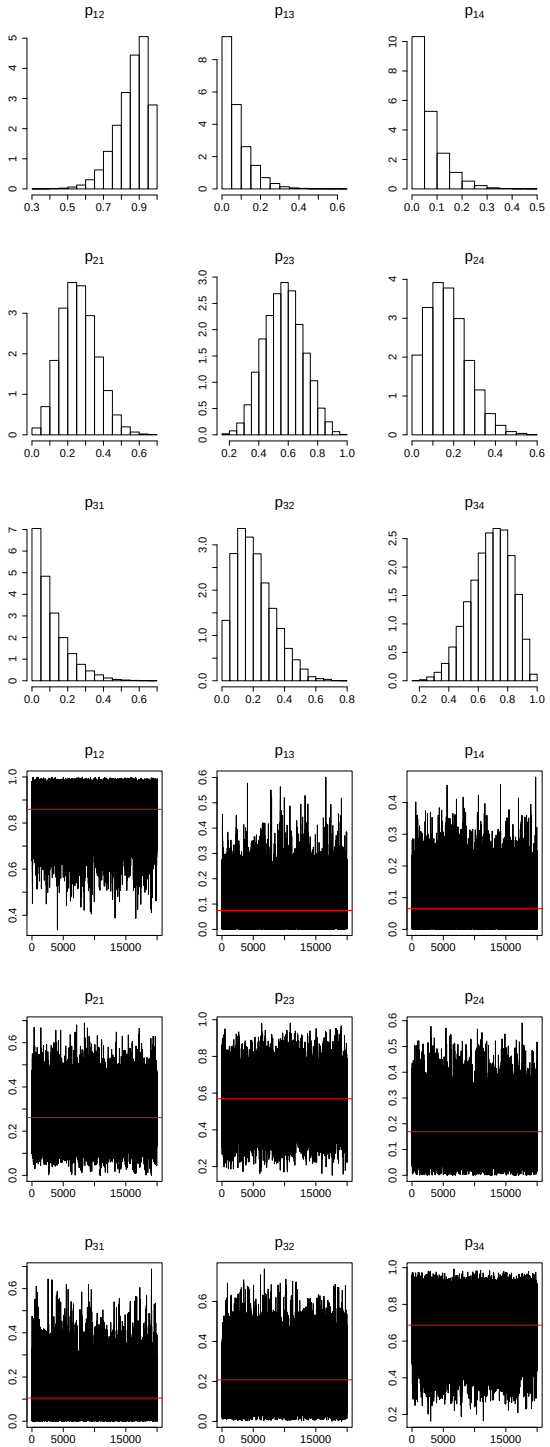


Table 1.3: Credit rating modelling: median of the average sojourn times expressed in years.

	$Me(\bar{w}_1)$	$Me(\bar{w}_2)$	$Me(\bar{w}_3)$
Semi-Markov	201.2190	21.4278	1.2477
Markov	50.1734	34.6002	2.0816

Table 1.4: Breast cancer data: mean and standard deviation of the posterior model parameters and the average sojourn times expressed in years.

Semi-Markov model

	p_{12}	p_{13}	p_{21}	p_{23}	γ_1	γ_2	α_1	α_2
$E(\cdot\|y)$	0.87	0.13	0.20	0.80	0.14	0.36	0.80	0.68
$SD(\cdot\|y)$	0.11	0.11	0.10	0.10	0.05	0.21	0.17	0.18

Markov model

	p_{12}	p_{13}	p_{21}	p_{23}	γ_1	γ_2
$E(\cdot\|y)$	0.67	0.33	0.13	0.87	0.06	0.10
$SD(\cdot\|y)$	0.13	0.13	0.07	0.07	0.02	0.02

	$\mathbb{E}(\bar{w}_1)$	$\mathbb{E}(\bar{w}_2)$	$\mathrm{sd}(\bar{w}_1)$	$\mathrm{sd}(\bar{w}_2)$
Semi-Markov	8.91	4.32	2.58	1.18
Markov	17.86	11.03	3.84	2.60

1.4.3 Breast Cancer Data

In the second application, we consider a data set comprising 37 women with breast cancer treated for spinal metastases; see De Stavola [1988], Davison [2003] and the supplementary material of Tancredi [2019] for previous analysis of these data. The ambulatory status of the women, defined as ability to walk unaided or not, was recorded when the treatment began and then 3, 6, 12, 24, and 60 months after treatment. The three states are: able to walk unaided (1) unable to walk unaided (2) and dead (3). We fitted the semi-Markov Weibull model with death as absorbing state. The model parameters are $\theta = (p_{12}, p_{13}, p_{21}, p_{23}, \gamma_1, \gamma_2, \alpha_1, \alpha_2)$. Figure 1.5 shows the posterior distributions for the shape parameters α_1 and α_2 under a vague prior distribution for θ. From Table 1.4 we may observe a difference in terms of posterior distributions and average sojourn times between the Markov and the semi-Markov model. The expected values of the posterior shape parameters suggest that the Markov model assumption is quite restrictive for this data.

Figure 1.5: MCMC samples of sojourn times parameters for breast cancer data.

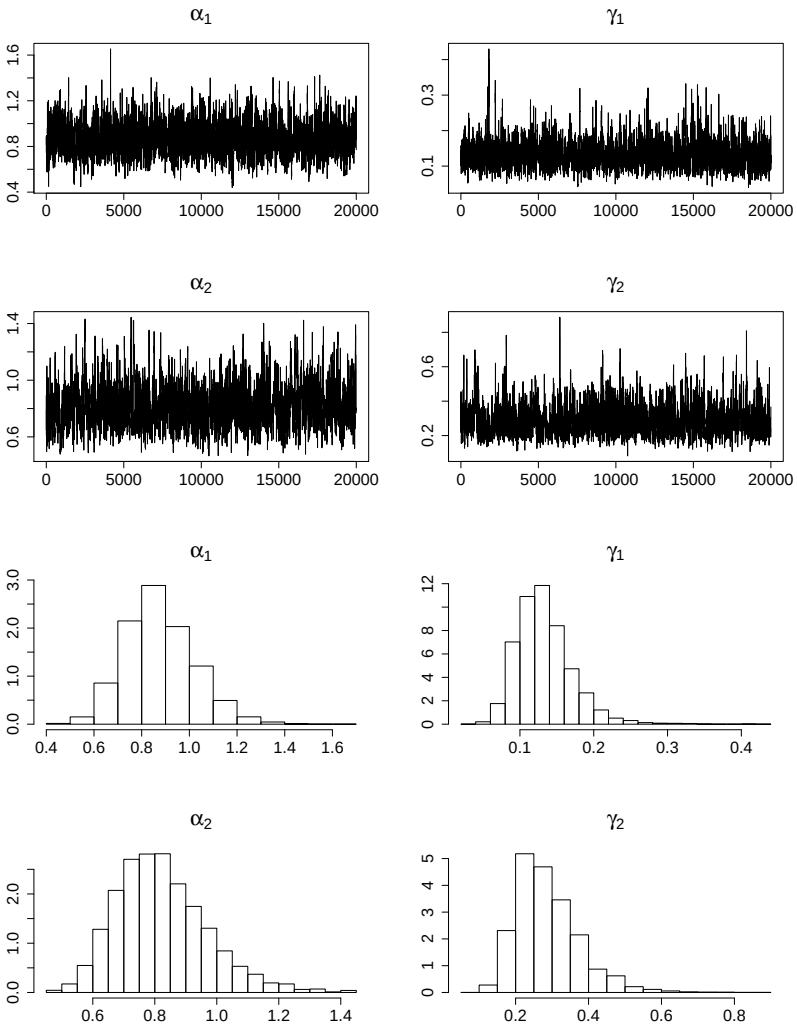
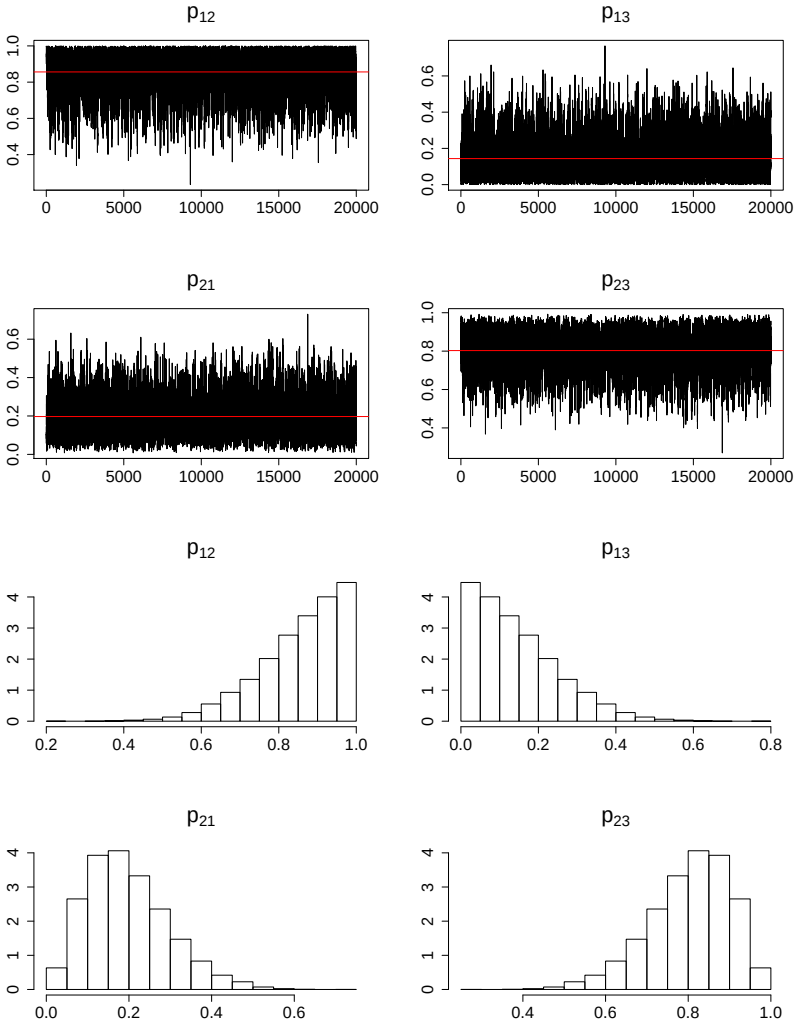


Figure 1.6: MCMC samples of transition probabilities posterior parameters for breast cancer data.



Chapter 2

Bayesian nonparametric inference for continuous time multi-state models

Clustering techniques may be required in order to find groups of similar time series in a sample of time series that are unlabeled a priori. This would allow to determine subsets of similar time series within the sample. However, Liao [2005] showed that it is difficult to define an appropriate distance-measure for time series data. As opposed to that, Frühwirth-Schnatter and Kaufmann [2008] demonstrated that choosing an appropriate clustering kernel density, model-based clustering based on finite mixture models extends to time series data in quite a natural way. The adequacy of the selected mixture kernel allows to capture salient features of the observed time series. Various clustering kernels were suggested for panels with real-valued time series observations. In particular, we focus on continuous time Markov processes clustering, which could be viewed as fitting a dynamic multinomial model with cluster-specific parameters to each time series in the panel. Frydman [2005] consider finite mixtures of time-homogeneous Markov chains both in continuous and discrete time, with an application to bond ratings migration. While such a model allows the transition behavior to be different across clusters, Fougère and Kamionka [2003] considered a mover-stayer model in continuous time which is a constrained mixture of two Markov chains to incorporate a simple form of heterogeneity across individual labor market transition data. Also, Pamminger and Frühwirth-Schnatter [2010] proposed a finite mixture of random-effects models designed specifically to capture unobserved heterogeneity in the transition behavior across time series within the same cluster from a Bayesian perspective. Cardot et al. [2018] generalize the previous methods by estimating finite mixtures of semi-Markov chains, in discrete or

continuous time. However, all these approaches are related to finite mixtures and consider completely observed processes. Instead, in this Chapter we propose a novel Bayesian method which allows to model such data as an infinite mixture of continuous-time Markov processes: with a Dirichlet process prior on the mixing measure, we get a Dirichlet Process Mixture model having as kernel density a continuous-time Markov process. In the first Section we introduce Dirichlet process and Dirichlet process mixtures, following Müller et al. [2015]. Next, we present a Dirichlet process mixture of Markov continuous time multi-state models for both discretely and fully observed data. In the last Section we show applications with simulated and real data.

2.1 Introduction to Bayesian nonparametric inference

Generally a Bayesian nonparametric (BNP) model is a Bayesian model where the functional form of the data generating the process is unknown. Hence, for defining a nonparametric Bayesian model, we should define the prior probability distribution on an infinite-dimensional space.

Before introducing the Dirichlet process we observe that BNP inference may be seen as a generalization of the standard parametric inference. Let $y_1, \ldots, y_n$ be a sample of independent observations coming from the unknown density f. Let $\pi(df)$ be a prior distribution on a suitable space of density functions. The first observation provides information about f, which in turn provides information about the second observation, and so on. Similarly to the parametric models, after n observations the posterior distribution of f can be written as

$$\pi(df|y_1, \ldots, y_n) = \frac{\prod_{i=1}^{n} f(y_i)\pi(df)}{\int \prod_{i=1}^{n} f(y_i)\pi(df)},$$

providing information about the future observation y_{n+1} via the predictive density

$$f(y_{n+1}|y_1, \ldots, y_n) = \int f(y_{n+1})\pi(df|y_1, \ldots, y_n).$$

In Bayesian nonparametrics there are two main categories of priors depending on the infinite dimensional parametric space featuring the inferential problem at hand. If we have to estimate functions we need to take as prior random functions (stochastic processes, random basis expansion and random densities) as prior distributions. While in spaces of probability measures we have random probability measures. In this chapter we focus on the latter class of priors, particularly on the Dirichlet process [Ferguson, 1973] whose sample paths are almost surely discrete distributions. Particularly, in Dirichlet Process Mixture (DPM) models [Lo, 1984] the Dirichlet process is the mixing distribution generating random density functions. With the improvement of Bayesian computation techniques the DPM has become one of the most important tool for Bayesian nonparametrics.

2.1.1 Dirichlet Process

One of the most popular BNP models is the Dirichlet process (DP) prior, introduced by Ferguson [1973] as a prior on the space of probability measures.

Suppose to observe an i.i.d. sample

$$y_i|G \overset{iid}{\sim} G, \quad i = 1, \ldots, n. \tag{2.1}$$

In order to carry out Bayesian modelling, we need to assume a prior probability model Π for the unknown infinite-dimensional parameter G.

Definition 2.1.1.1 (Dirichlet Process, Ferguson [1973]) *Let $M > 0$ and G_0 be a probability measure defined on S. A DP with parameters (M, G_0) is a random probability measure G defined on S which assigns probability $G(B)$ to every (measurable) set B such that for each (measurable) finite partition $\{B_1, \ldots, B_k\}$ of S, the joint distribution of the vector $G(B_1), \ldots, G(B_k)$ is the Dirichlet distribution with parameters*

$$(MG_0(B_1), \ldots, MG_0(B_k)).$$

We denote the process as $\mathrm{DP}(MG_0)$, where M is the precision parameter and G_0 is the centering measure; as M tends to infinite, G concentrates around G_0, and product MG_0 is called base measure of the DP.

An important property of the DP is the discrete nature of the random probability measure G, which allows us to rewrite it as a weighted sum of point masses: $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot)$ where w_h are probability weights and δ_x denotes the Dirac measure at x. Sethuraman [1994] introduced an equivalent representation of a DP random probability measure based on the discrete nature of the process G. Let $w_h = v_h \prod_{\ell<h}(1 - v_\ell)$ with $v_h \overset{iid}{\sim} Beta(M, 1)$ and $m_h \overset{iid}{\sim} G_0$, where $\{v_h\}$ and $\{m_h\}$ are independent for $h = 0, \ldots, \infty$. Then

$$G(\cdot) = \sum_{h=1}^{\infty} \omega_h \delta_{m_h}(\cdot), \tag{2.2}$$

with $\sum_{h=1}^{\infty} \omega_h = 1$, defines a $\mathrm{DP}(MG_0)$ random probability measure. This representation of a DP random measure is known as "stick-breaking", since it may be represented as successively breaking fractions v_h of a stick of initially unit length.

Another property of the DP is its large weak support: let Q be any probability measure with $Q \ll G_0$ and $\epsilon > 0$; for any finite number of measurable sets $B_1, \ldots, B_m$,

$$\pi\left\{|G(B_i) - Q(B_i)| < \epsilon, \quad \text{for i} = 1, \ldots, m\right\} > 0,$$

where $\pi\{\dots\}$ refers to the probability model π of G. Roughly speaking, any distribution with the same support as G_0 can be well approximated weakly by a DP random probability measure.

Another important feature of the DP is the conditioning property: if A is a (measurable) set with $G_0(A) > 0$ (which implies that $G(A) > 0$ a.s.), then the random measure $G|_A$, the restriction of G to A defined by $G|_A(B) = G(B|A) = G(A \cap B)/G(A)$ is also a DP with parameters M and $G_0|_A$, and is independent of $G(A)$. Extending the argument to more than one set, we observe that the DP locally splits into numerous independent DP's.

Posterior Distribution Let $y_1, \dots, y_n$ be i.i.d. from a random variable with distribution G and let G be a realization from a DP(MG_0). The DP is conjugate with respect to i.i.d. sampling. Thus, with a DP prior on G, the posterior distribution for G is again a DP, with base measure adding a point mass to the prior base measure at each observed data point y_i, so that

$$G|y_1, \dots, y_n \sim DP\left(MG_0 + \sum_{i=1}^{n} \delta_{y_i}\right). \tag{2.3}$$

Marginal Distribution Consider a random sample as in (2.1). Blackwell et al. [1973] represented the marginal distribution

$$f(y_1, \dots, y_n) = \int \prod_{i=1}^{n} G(y_i) d\pi(G)$$

with the Polya urn scheme, exploiting the discreteness of the random probability measure G which implies a positive probability of ties among the y_i. They specified the marginal distribution as a product of a sequence of increasing conditionals $f(y_1, \dots, y_n) = f(y_1) \prod_{i=2}^{n} f(y_i|y_1, \dots, y_{i-1})$ with

$$f(y_i|y_1, \dots, y_{i-1}) = \frac{1}{M+i-1} \sum_{h=1}^{i-1} \delta_{y_h}(y_i) + \frac{M}{M+i-1} G_0(y_i) \tag{2.4}$$

for $i = 2, 3, \dots$ and $y_1 \sim G_0$. Since the y_i are i.i.d. given G the marginal joint distribution of $(y_1, \dots, y_n)$ is exchangeable. Thus, the complete conditional $f(y_i|y_h, h \neq i)$ has the same form as (2.4) for y_n, while the (posterior) predictive for a future observation y_{n+1} given data $y_1, \dots, y_n$ takes the form of (2.4) for $i = n + 1$.

2.1.2 Dirichlet Process Mixtures

The discreteness of the Dirichlet process makes it uncomfortable for density estimation. This limitation can be remedied by convolution of its paths with a continuous kernel, i.e. by using a DP random measure as the mixing distribution in a mixture over some simple parametric forms. It is hardly possible to derive the resulting posterior distribution analytically. However, there are many efficient algorithm which - by exploiting the properties of the DP - allow us to compute it numerically. This approach has been introduced by Ferguson [1983], Lo [1984], Escobar [1990], Escobar [1994], and Escobar and West [1995].

Let f_θ be a continuous probability density function, with $\theta \in \Theta$, and let G be a probability distribution on Θ. We define the density function of a mixture f_θ with respect to G as

$$f_G(y) = \int f_\theta(y) dG(\theta) \tag{2.5}$$

With a DP prior on the mixing distribution G, which induces a prior on kernel densities, we get a DP mixtures (DPM) model. The mixture model (2.5), together with a DP prior on the mixing measure G can be also represented in a hierarchical form:

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} f_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(MG_0). \end{aligned} \tag{2.6}$$

Under this model, the posterior distribution $\pi(G|y_1, \ldots, y_n)$ is a mixture of DP models, mixing with respect to new latent variables θ_i specific to each experimental unit:

$$G|\mathbf{y} \sim \int DP\left(MG_0 + \sum_{i=1}^{n} \delta_{\theta_i}\right) d\pi(\boldsymbol{\theta}|\mathbf{y}), \tag{2.7}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. Therefore, marginalizing with respect to $\boldsymbol{\theta}$, the posterior distribution $\pi(G|\mathbf{y})$ becomes a mixture over (2.3) with respect to the posterior distribution on $\boldsymbol{\theta}$.

The choice of the suitable kernel is based on the support of the underlying density function or more generally on the problem at hand. If it is defined

on $\mathbb{R}$, a location-scale kernel is appropriate; on the unit interval, beta distributions may be used; on $\mathbb{R}^+$, mixtures of gamma, Weibull or lognormal distributions may be used. Naturally, if the underlying model is a stochastic process, also the kernel should be a stochastic process. In particular in this chapter we will illustrate the use of DPM by adopting a Markov CTMSM kernel.

Clustering Since the discreteness of the DP implies a positive probability for ties among the latent θ_i, the DPM model induces a probability model on clusters. Formally, let θ_j^*, $j = 1, \ldots, k$, denote $k \leq n$ unique values, $\Psi_j = \left\{i : \theta_i = \theta_j^*\right\}$, and let $n_j = |\Psi_j|$ denote the number of θ_i tied with θ_j^*. Since the θ_i are random, also the Ψ_j are random. Thus, the DPM implies a model on the random partition $\xi_n = \{\Psi_1, \ldots, \Psi_k\}$ of the experimental units $\{1, \ldots, n\}$; the posterior model $\pi(\boldsymbol{\psi}|\mathbf{y})$ reports posterior inference on clustering of the data. We represent the clustering by an equivalent set of cluster membership indicators, $\psi_i = j$ if $i \in \Psi_j$, with clusters labeled by order of appearance.

Let k_i denote the number of unique θ_ℓ among $\{\theta_1, \ldots, \theta_i\}$, with n_{ij} representing the multiplicity of the j-th unique value $\left(\sum_{j=1}^{k_i} n_{ij} = i \text{ by definition}\right)$. Then, by the properties of the DPM we have:

$$\pi(\psi_i = j|\psi_1, \ldots, \psi_{i-1}) = \begin{cases} \frac{n_{i-1,j}}{M+i-1} & \text{for } j = 1, \ldots, k_{i-1} \\ \frac{M}{M+i-1} & \text{for } j = k_{i-1}+1 \end{cases} \tag{2.8}$$

Moreover, by exchangeability of θ, the prior conditional probability $\pi(\psi_i|\boldsymbol{\psi}_{-i})$ - with $\boldsymbol{\psi}_{-i} = (\psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_n)$ - takes the same form as (2.8) for $i = n$.

Therefore, we may write the prior $\pi(\boldsymbol{\psi})$ as

$$\pi(\boldsymbol{\psi}) = \prod_{i=2}^{n} \pi(\psi_i|\psi_1, \ldots, \psi_{i-1}) = \frac{M^{k-1} \prod_{j=1}^{k} (n_j - 1)!}{(M+1) \cdots (M+n-1)} \tag{2.9}$$

with $\psi_1 = 1$ by definition, since we label clusters in order of appearance.

Let $\theta_{i,j}^*$ denote the j-th unique value among $\{\theta_1, \ldots, \theta_i\}$. If $\psi_i = j$, then $\theta_i = \theta_{i-1,j}^*$, while if $\psi_i = k_{i-1} + 1$ then $\theta_i \sim G_0$. Then:

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}) \propto \sum_{j=1}^{k_{i-1}} n_{i-1,j} \delta_{\theta_{i-1,j}^*}(\theta_i) + M G_0(\theta_i). \tag{2.10}$$

Since the DPM is exchangeable, we get

$$\pi(\theta_i|\boldsymbol{\theta_{-i}}) \propto \sum_{j=1}^{k^-} n_j^- \delta_{\theta_j^{*-}}(\theta_i) + MG_0(\theta_i). \tag{2.11}$$

where $\boldsymbol{\theta_{-i}} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k\}$, k^- represents the number of unique values in $\boldsymbol{\theta_{-i}}$, θ_j^{*-} denotes the j-th unique element and n_j^- is the number of observations lying in the j-th cluster excluding y_i if $\psi_i = j$.

Posterior Simulation One of the main advantages of using nonparametric methods is the ability to reduce uncertainty avoiding distributional assumptions. In Bayesian framework this flexibility increases the computational cost. It was not by chance that much of the development of BNP methods has been consequence of the improvement of simulation-based computational methods.

In DPM models there are two levels of conjugacy: between the DP random measure $\pi(G)$ and its posterior $\pi(G|\mathbf{y})$, and between the kernel element $f_\theta(y)$ and the centering measure G_0 which takes the role of prior on θ. When talking about nonconjugate DP mixtures, we always refer to the case where f_θ and G are not conjugate; a sampler for this case is the no gaps sampler and it still relies on conjugacy of the DP posterior.

Conjugate DPM The first Gibbs sampler for DPM model has been proposed by Escobar [1990]. It is based on the updating of θ_i by drawing from the complete posterior distribution $\pi(\theta_i|\theta_{-i}, \mathbf{y})$. Nevertheless, this sampler suffers from slow mixing of the resulting Markov Chain.

Bush and MacEachern [1996] introduced the most currently used posterior MCMC methods for DPM models, using two types of transition probabilities.

1. Sampling from $\pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y})$:

   θ_j^* conditional on the imputed partition is updated using

   $$\pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y}) \propto G_0(\theta_j^*) \prod_{i \in \Psi_j} f_{\theta_j^*}(y_i). \tag{2.12}$$

   The posterior $\pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y})$ is derived as the posterior on a parametric model with prior $G_0(\theta_j^*)$ and sampling model $f_\theta(y_i)$, for y_i with $i \in \Psi_j$.

Let $y_j^* = (y_i, i \in \Psi_j)$ denote y_i arranged by cluster. Therefore, the conditioning on $\boldsymbol{\psi}$ is implicit in the selection of the elements in y_j^*, that is if we know Ψ then we also know y_j^* and

$$\pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y}) = \pi(\theta_j^*|\mathbf{y}_j^*).$$

To obtain the distribution of $\psi_i|\boldsymbol{\psi}_{-i}, y$ we first calculate the conditional distribution $\pi(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{y})$. Then we show the expression for $\pi(\theta_i, \psi_i|\boldsymbol{\theta}_{-i}, \mathbf{y})$. We finally integrate $\pi(\psi_i, \boldsymbol{\theta}|\boldsymbol{\psi}_{-i}, \mathbf{y}) = \pi(\psi_i, \theta_i|\boldsymbol{\theta}_{-i}, \mathbf{y}) \cdot \pi(\boldsymbol{\theta}_{-i}|\boldsymbol{\psi}_{-i}, \mathbf{y})$ with respect to $\boldsymbol{\theta}$ in order to obtain $\pi(\psi_i|\boldsymbol{\psi}_{-i}, \mathbf{y})$.

2. Sampling from $\pi(\psi_i|\boldsymbol{\psi}_{-i}, \mathbf{y})$:

We now get the posterior distribution $\pi(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{y})$ by multiplying the prior $\pi(\theta_i|\boldsymbol{\theta}_{-i})$ (2.11) with the sampling distribution $f_{\theta_i}(y_i)$.

$$\pi(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{y}) \propto \sum_{j=1}^{k^-} n_j^- f_{\theta_j^{*-}}(y_i)\delta_{\theta_j^{*-}}(\theta_i) + M f_{\theta_i}(y_i)G_0(\theta_i). \tag{2.13}$$

In the second term $f_{\theta_i}(y_i)G_0(\theta_i)$ is not normalized. Anyway, if we let $H_0(\theta_i) \propto f_{\theta_i}(y_i)G_0(\theta_i)$ with normalization constant $h_0(y_i) \equiv \int f_\theta(y_i)G_0(d\theta)$, recalling that if $\theta_i = \theta_j^{*-}$ then $\psi_i = j$ and if $\theta_i \neq \theta_j^{*-}$ for $j = 1, \ldots, k^-$ then $\psi_i = k^- + 1$, we can write the (2.13) as

$$\pi(\theta_i, \psi_i|\boldsymbol{\theta}_{-i}, \mathbf{y}) \propto \sum_{j=1}^{k^-} n_j^- f_{\theta_j^{*-}}(y_i)\delta_j(c_i)\delta_{\theta_j^{*-}}(\theta_i) + M h_0(y_i)\delta_{k^-+1}(c_i)H_0(\theta_i).$$

Finally to obtain the distribution $\psi_i|\boldsymbol{\psi}_{-i}, \mathbf{y}$ we marginalize with respect to $\boldsymbol{\theta}$, that is with respect to θ_i and $\boldsymbol{\theta}_{-i}$, calculating the integral

$$\int \pi(\theta_i, \psi_i|\boldsymbol{\theta}_{-i}, \mathbf{y})\pi(\boldsymbol{\theta}_{-i}|\boldsymbol{\psi}_{-i}, \mathbf{y})d\theta_i d\boldsymbol{\theta}_{-i}.$$

For the first k^- terms we get

$$\int f_{\theta_j^*}(y_i)d\pi(\theta_j^{*-}|\mathbf{y}_j^{*-}) = f(y_i|\mathbf{y}_j^{*-}),$$

where $\mathbf{y}_j^{*-} = \mathbf{y}_j^* \setminus \{y_i\}$. For the last term we get

$$\int h_0(y_i)H_0(\theta_i)d\theta = h_0(y_i).$$

To clarify the expression of $\pi(\theta_j^{*-}|y_j^{*-})$ and the differences with respect to (2.12) where in the conditioning subset we have also y_i, note that θ_j^{*-} need not to be the same as θ_j^*. Moreover, when i is a singleton cluster, then removing the i-th unit from the partition might change the indices of other clusters. Therefore, we have

$$\pi(\theta_j^{*-}|\mathbf{y}_j^{*-}) \propto G_0(\theta_j^{*-}) \prod_{\ell \in \Psi_j \setminus \{i\}} f_{\theta_j^{*-}}(y_\ell). \tag{2.14}$$

Finally, we get

$$\pi(\psi_i = j|\boldsymbol{\psi}_{-i}, \mathbf{y}) \propto \begin{cases} n_j^- f(y_i|\mathbf{y}_j^{*-}) & \text{for } j = 1, \dots, k^- \\ M h_0(y_i) & \text{for } j = k^- + 1. \end{cases} \tag{2.15}$$

This posterior Gibbs sampler is only practicable for DPM models with conjugate G_0 and f_θ, otherwise the evaluation of h_0 would be analytically intractable.

Algorithm 4 Gibbs sampler for conjugate DPM

• **Clustering**:
– for $i = 1 \dots, n$ draw $\psi_i \sim \pi(\psi_i = j|\boldsymbol{\psi}_{-i}, \mathbf{y})$;
• **Cluster parameters**:
– for $j = 1, \dots, k$ draw $\theta_j^* \sim \pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y})$.

Non-Conjugate DPM If G_0 and f_θ are not conjugate, the evaluation of h_0 would not be analytically tractable and sampling from $\pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y})$ may result challenging.

In order to evaluate the integral $\int f_\theta(y_i)G(d\theta)$, West et al. (1994) suggested using either numerical quadrature or a Monte Carlo approximation. By approximating the required integral with an average over m θ draws from G_0, it is also possible to approximate a draw from $\pi(\theta_j^*|\boldsymbol{\psi}, \mathbf{y})$ by sampling from among these m points with probabilities proportional to their likelihood.

MacEachern and Müller [1998] propose the "no-gaps" algorithm. Let k be the number of distinct elements in $\boldsymbol{\theta}$. They proposed a model augmentation

$$\underbrace{\{\theta_1^*, \dots, \theta_k^*}_{\theta_F^*} \underbrace{\theta_{k+1}^*, \dots, \theta_n^*\}}_{\theta_E^*}$$

with $\theta_j^* \sim G_0$ for $j = k+1, \ldots, n$; note that $n_j > 0$ for $j \leq k$, while $n_j = 0$ for $j = k+1, \ldots, n$. In practice, the augmentation includes the constraint that there will be no gaps in the values of the ψ_j: the θ_F^* values represents the locations of the full clusters, while θ_E^* represents potential locations of the empty clusters. In this augmented model the evaluation of integrals is replaced by simple likelihood evaluations. As in the conjugate case, assume $\psi_i = j$ in the current imputation, we need to distinguish two cases:

- $n_j > 1$, then

$$p(\psi_i = j | \boldsymbol{\psi}_{-i}, \theta_j^{*-}, \mathbf{y}) \propto \begin{cases} n_j^- f_{\theta_j^*}(y_i) & \text{for } j = 1, \ldots, k^- \\ \frac{M}{k^-+1} f_{\theta^*_{k^-+1}}(y_i) & \text{for } j = k^- + 1 \end{cases} \tag{2.16}$$

- if $n_j = 1$, $\psi_i = j$ is imputed to form a singleton cluster. Then:
  - with probability $(k-1)/k$, leave ψ_i in the j-th cluster;
  - otherwise, remove ψ_i form the j-th cluster, relabel θ_j^* in order to have no-gaps and assign ψ_i to another cluster by (2.16).

For details on the derivation of the probabilities in (2.16) see MacEachern and Müller (1998).

Algorithm 5 No-gaps sampler for non-conjugate DPM

• **Clustering**:
– for $i = 1 \ldots, n$ draw $\psi_i \sim \pi(\psi_i = j | \boldsymbol{\psi}_{-i}, \theta^*, \mathbf{y})$;
• **Cluster parameters**:
– for $j = 1, \ldots, k$ draw $\theta_j^* \sim \pi(\theta_j^* | \boldsymbol{\psi}, \mathbf{y})$.
– for $j = k+1, \ldots, n$ draw $\theta_j^* \sim G_0$

The key feature of this algorithm is that it does not require evaluation of the integral h_0 and it can be implemented for any model as long as we can generate from G_0 and compute $f_\theta(y_j)$. There is no need for G_0 to be the conjugate prior for F_θ.

2.2 Dirichlet Process Mixtures of CTMSM

After a brief introduction on Dirichlet process and Dirichlet process mixtures, we now extend the methodology to a particular class of stochastic processes. In this section we first introduce mixtures of CTMSM and define a general DPM model for multi state processes in continuous time. Then we assume Markov CTMSM as kernel density of the mixture, providing a Gibbs sampler for DPM models for fully observed continuous-time multi state models. This section is introductory for the next one, in which we will extend the framework to the more complex context of discretely observed trajectories.

2.2.1 Infinite mixtures of CTMSM

As in the first chapter, let X_t be a CTMSM, with $\mathcal{S} = \{1, 2, \dots S\}$ finite-discrete state space and $\mathcal{T}$ time space. Again, X_t takes values in the union of product spaces $\mathcal{M}^{\cup} \equiv \mathcal{S}^{\cup} \times \mathcal{T}^{\cup}$.

We have defined the finite measure space of the states $(\mathcal{S}^{\cup}, \Sigma_{\mathcal{S}}^{\cup}, \nu^{\cup})$, the measure space of the sojourn times $(\mathcal{T}^{\cup}, \Sigma_{\mathcal{T}}^{\cup}, \mu^{\cup})$, and the product measure space $(\mathcal{M}^{\cup}, \Sigma_{\mathcal{M}}^{\cup}, \eta^{\cup})$, where $\eta^{\cup}$ is the union product measure $\eta^{\cup} = \mu^{\cup} \times \nu^{\cup}$.

Let f density function of a CTMSM $X_t = (s, w)$ w.r.t. $\eta^{\cup}$, where s and w represent respectively the states and the sojourn times sequences, such that

$$\int_{\mathcal{M}} f(s, w) d\eta^{\cup}(s, w) = 1.$$

Recall that if all the event rates are finite, the process will have almost surely a finite number of state-changes. Therefore, the support of f has finite dimension by construction.

Let f_θ be a probability density function of a multi-state model where $\theta \in \Theta$. For instance, in the Markov case, Θ is the space of the rate matrices

$$A = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1S} \\ \vdots & \ddots & \vdots \\ \gamma_{S1} & \cdots & \gamma_{SS} \end{pmatrix}$$

with $\gamma_{rr} = -\sum_{s \neq r} \gamma_{rs}$. Let G be a probability distribution defined on the parameter space Θ. We define the density function f_G of a mixture with multi-state model kernels f_θ with respect to the mixing measure G as

$$f_G(s, w) = \int f_\theta(s, w) dG(\theta). \tag{2.17}$$

Let $X_i = (s, w)_i$, for $i = 1, \ldots, n$, be n fully observed paths on $[0, T_i]$. Note that we assume that the observation times T_i can be different across the paths. We may rewrite the model in a hierarchical form, assuming a DP prior for the mixing measure G. Then:

$$\begin{aligned} X_i|\theta_i &\overset{ind}{\sim} f_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(MG_0). \end{aligned} \tag{2.18}$$

The posterior distribution $\Pi(G|\mathbf{X})$ is a mixture of DP models, mixing with respect to latent variables θ_i specific to each observed path X_{t_i} for $i = 1, \ldots, n$:

$$G|\mathbf{X} \sim \int DP(MG_0 + \sum_{i=1}^{n} \delta_{\theta_i}) d\Pi(\boldsymbol{\theta}|\mathbf{X}), \tag{2.19}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\mathbf{X} = (X_{t_1}, \ldots, X_{t_n})$.

Regarding the kernel choice there are several issues to consider. We may choose between Markov and semi-Markov multi-state models in continuous time. The latter is a generalization of the former, meaning that it should have more flexibility. However there is a trade-off between the model flexibility and the computation time: in addition to a further parameter with respect to the Markov processes, semi-Markov does not have conjugate priors for the model's parameters. This would involve the lack of a conjugate G_0, increasing the computational cost. Instead, with Markov CTMSM as kerneld density f_θ we may exploit a conjugate prior density for G_0. These are the reasons why in this thesis we implement only DPM with Markov CTMSM density kernel.

2.2.2 Assuming Markov density kernel

Consider an observation $X = (s, w)$ observed on $[0, T]$. As in Chapter 1, we define the generic CTMSM density as product of two densities defined on the two underlying spaces

$$f_\theta(s, w) = h_p(s) z_\zeta(w),$$

with h_p representing the state transitions distribution density and z_ζ representing the sojourn times density, defined respectively on the state space $\mathcal{S}^\cup$ and on the time space $\mathcal{T}^\cup$.

More specifically, let n_r for $r = 1, \ldots, S$ represent the number of completed sojourn times $w_{r_1} \ldots, w_{r_{n_r}}$ in each state r during the trajectory (s, w).

Let n_{rs} be the number of transitions from the state r to the state s. Moreover, let $W_r = \sum_j w_{r_j}$ be the total sojourn time spent in the state r considering also the truncated part of the trajectory. Finally, let assume the sojourn time distribution in the state r for $r = 1, \dots, S$ to be $\text{Exp}(\gamma_r)$ such that $\zeta = \gamma$. Then, we may write down the density as

$$f_\theta(s, w) = \prod_{rs} p_{rs}^{n_{rs}} \prod_r \gamma_r^{n_r} e^{-\gamma_r W_r}, \tag{2.20}$$

where $p_{rs} = \gamma_{rs}/\gamma_r$ for $r, s \in \mathcal{S}$.

Note that the mixing measure G, defined on the parameter space Θ, has to be a $S \times (S-1)$-dimensional probability distribution if we do not have absorbing state, otherwise is a $(S-1) \times (S-1)$-dimensional probability distribution. Moreover note that

$$f_\theta(s, w) = \underbrace{\prod_{rs} \tilde{p}_{rs}^{n_{rs}}}_{1} \underbrace{\prod_r \gamma_r^{n_r} e^{-\gamma_r W_r}}_{2}$$

can be seen as the product of two factors. The first is proportional to the product of S Multinomial likelihood functions of dimension $S-1$, while the second component is the product of S exponential likelihood functions with a truncated observation.

2.2.3 Posterior Computation

In order to exploit the conjugacy, we chose G_0 to be a product measure

$$G_0 \equiv \text{Ga}_{(S)}(\mathbf{a}, \mathbf{b}) \times \text{Dir}_{(S\times(S-1))}(\boldsymbol{\alpha}), \tag{2.21}$$

where $\text{Ga}_{(S)}(\mathbf{a}, \mathbf{b})$ represents the product of S independent gamma distributions with parameters a and b, while $\text{Dir}_{(S\times(S-1))}(\boldsymbol{\alpha})$ is the product of S Dirichlet distributions of dimension $S-1$ with parameter $\boldsymbol{\alpha}$.

We now present the Gibbs Sampler for DPM of Markov CTMSM, following the scheme of Bush and MacEachern [1996] presented in the first section of the current Chapter. We update the MCMC algorithm by drawing from the cluster complete conditional posterior probability distribution of the cluster membership, after marginalizing with respect to θ. Then we update the parameters θ^* of the different values θ conditional on their cluster observations.

Let $X_i = (s, w)_i$, $i = 1, \ldots, N$, be i.i.d. continuous-time fully observed paths. The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f_\theta((s, w)_i). \tag{2.22}$$

Let $\mathbf{X} = (s, w)$ be the set of observed trajectories, let $\Psi_h = (i : \theta_i = \theta_h^*)$ indicate that if $\psi_i = h$, the i-th observation belongs to the h-th cluster. Following (2.20), let $\theta_h^* = (\boldsymbol{\gamma}_h^*, \mathbf{p}_h^*)$, where $\boldsymbol{\gamma}_h^* = (\gamma_{h_1}^*, \ldots, \gamma_{h_S}^*)$ are vectors of rate parameters of dimension S and $\mathbf{p}_h^* = (p_{h_1}^* \ldots, p_{h_S}^*)$ are transition probabilities, where each element of $\mathbf{p}_h^*$ is a vector of dimension $S - 1$. We may write down the Markov CTMSM likelihood function for the cluster h

$$\begin{aligned} \mathcal{L}(\theta_h^*) &= \prod_{i \in \Psi_h} \left(\prod_{rs} {p^*}_{h_{rs}}^{n_{rs,i}} \prod_{r} {\gamma^*}_{h_r}^{n_{r,i}} e^{-{\gamma^*}_{h_r} W_{ir}} \right) = \\ &= \left(\prod_{i \in \Psi_h} \prod_{rs} {p^*}_{h_{rs}}^{n_{rs,i}} \right) \left(\prod_{i \in \Psi_h} \prod_{r} {\gamma^*}_{h_r}^{n_{r,i}} e^{-{\gamma^*}_{h_r} W_{ir}} \right). \end{aligned} \tag{2.23}$$

In order to define the Gibbs sampler, we first derive the posterior density of the cluster parameters $\pi(\theta_h^* | \boldsymbol{\psi}, \mathbf{X})$ and then the posterior conditional density of the clusters $\pi(\psi_i = h | \boldsymbol{\psi}_{-i}, \mathbf{X})$.

1. We compute the posterior density for θ_h^* as

$$\pi(\theta_h^* | \boldsymbol{\psi}, \mathbf{X}) = G_0(\theta_h^*) \cdot \mathcal{L}(\theta_h^*). \tag{2.24}$$

From the mixing measure definition in (2.21) we may rewrite (2.24) as

$$\begin{aligned} \pi(\theta_h^* | \boldsymbol{\psi}, \mathbf{X}) = \prod_{r=1}^{S} \frac{\Gamma(\alpha_{(r)})}{\Gamma(\alpha_{1r})\Gamma(\alpha_{2r}) \ldots \Gamma(\alpha_{(S-1)r})} {p^*}_{h_{1r}}^{\alpha_{1r}-1} {p^*}_{h_{2r}}^{\alpha_{2r}-1} \ldots {p^*}_{h_{(S-1)r}}^{\alpha_{(S-1)r}-1} \\ \left(\prod_{i \in \Psi_h} \prod_{rs} {p^*}_{h_{rs}}^{n_{rs,i}} \right) \times \prod_{r=1}^{S} \frac{b_r^{a_r}}{\Gamma(a_r)} {\gamma^*}_{h_r}^{a_r - 1} e^{-b_r {\gamma^*}_r} \left(\prod_{i \in \Psi_h} \prod_{r} {\gamma^*}_{h_r}^{n_{r,i}} e^{-{\gamma^*}_r W_{ir}} \right). \end{aligned}$$

By conjugacy, the posterior distribution $\pi(\theta_h^* | \boldsymbol{\psi}, \mathbf{X})$ is still a product measure of Gamma and Dirichlet. With an abuse of notation we may write

$$\pi(\theta_h^* | \boldsymbol{\psi}, \mathbf{X}) = \prod_{r=1}^{S} \mathrm{Ga}(a_r + \mathbf{n}_{h_r}^*, b_r + \mathbf{w}_{h_r}^*) \cdot \prod_{r=1}^{S} \mathrm{Dir}_{(S-1)}(\alpha_r + \mathbf{s}_{h_{rs}}^*), \tag{2.25}$$

where:

- a_r and b_r are elements of the S-dimensional vectors of prior hyperparameters a and b of the Gamma component;
- $\mathbf{n}^*_{h_r}$ and $\mathbf{w}^*_{h_r}$ are respectively the sum of the number of visits to the r-th state and the sum of the sojourn times into the r-th state, for each of the N_h trajectories lying inside the cluster h, i.e $\mathbf{n}^*_{h_r} = \sum_{i=1}^{N_h} n_{h_{r,i}}$ and $\mathbf{w}^*_{h_r} = \sum_{i=1}^{N_h} w_{h_{r,i}}$, for $r = 1, \ldots, S$;
- α_r is a $(S-1)$-dimensional vector of prior hyperparameters Dirichlet component;
- $\mathbf{s}^*_{h_{rs}}$ is the $(S-1)$-dimensional vector of transition counts for the trajectories lying inside the h-th cluster.

2. Recall that θ^{*-}_h denote the h-th of the k^- unique values among $\boldsymbol{\theta}_{-i}$, which represents the vector $\boldsymbol{\theta}$ without the i-th element θ_i. Also, let $\mathbf{X}^{*-}_h = \mathbf{X}^*_h \setminus (s,w)_i$ be the set of the observations lying inside the cluster h with the exclusion of the i-th trajectory and let N_h be the number of units (complete paths) inside the cluster h. Now, in order to complete the Gibbs sampling scheme, we have to define the full conditional $\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{X})$ for $h = 1 \ldots k^-$ and $h = k^- + 1$ which denotes the creation of a new cluster. From (2.15) it follows that the probability for the i-th element to belong to the h-th cluster is

$$\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{X}) \propto \begin{cases} N^-_h f((s,w)_i|\mathbf{X}^{*-}_h) & \text{for } h = 1, \ldots, k^- \\ M f((s,w)_i) & \text{for } h = k^- + 1 \end{cases}, \qquad (2.26)$$

where N^-_h is the number of elements in the h-th cluster with exclusion of the i-th observation, M represents the precision parameter of the DP and

$$f((s,w)_i) \equiv \int f_{\theta^*_h}((s,w)_i) G_0(d\theta^*_h),$$

while

$$f((s,w)_i|\mathbf{X}^{*-}_h) = \int f_{\theta^*_h}((s,w)_i) d\pi(\theta^*_h|\mathbf{X}^{*-}_h = (\mathbf{s}^{*-}_h, \mathbf{w}^{*-}_h)),$$

where $\pi(\theta^*_h|\mathbf{X}^{*-}_h)$ with $\mathbf{X}^{*-}_h = (\mathbf{s}^{*-}_h, \mathbf{w}^{*-}_h)$ is the posterior distribution of θ^*_h conditional on the paths of the h-th cluster without the i-th observation, namely:

$$\int f_{\theta_h^*}((s,w)_i)d\pi(\theta_h^*|\mathbf{X}_h^{*-}) = \int\cdots\int\prod_{r=1}^{S}\left(\gamma^{*}{}_{h_r}^{n_{i_r}}e^{-\gamma^{*}{}_{h_r}W_{ir}}\prod_{s\neq r}p^{*}{}_{h_{rs}}^{n_{i_{rs}}}\right)\times$$

$$\prod_{r=1}^{S}\left(\frac{\left(b_r+\mathbf{W}_{hr}^{*-}\right)^{a_r+N_{h_r}^{-}}}{\Gamma(a_r+N_{h_r}^{-})}\gamma^{*}{}_{h_r}^{N_{h_r}^{-}}e^{-\gamma^{*}{}_{h_r}(b_r+\mathbf{W}_{hr}^{*-})}\cdot\Gamma\left(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}^{-})\right)\prod_{s\neq r}\frac{p^{*}{}_{h_{rs}}^{\left(\alpha_{rs}+N_{h_{rs}}^{-}-1\right)}}{\Gamma(\alpha_{rs}+N_{h_{rs}}^{-})}\right)d\mathbf{p}^{*}d\boldsymbol{\gamma}^{*}=$$

$$=\prod_{r=1}^{S}\left(\frac{\left(b_r+\mathbf{W}_{hr}^{*-}\right)^{a_r+N_{h_r}^{-}}}{\Gamma(a_r+N_{h_r}^{-})}\frac{\Gamma(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}^{-}))}{\prod_{s\neq r}\Gamma(\alpha_{rs}+N_{h_{rs}}^{-})}\right)\int\cdots\int\prod_{r=1}^{S}\left(\gamma^{*}{}_{h_r}^{n_{i_r}}e^{-\gamma^{*}{}_{h_r}W_{ir}}\right)\times$$

$$\left(\gamma^{*}{}_{h_r}^{N_{h_r}^{-}}e^{-\gamma^{*}{}_{h_r}\left(b_r+\mathbf{W}_{hr}^{*-}\right)}\prod_{s\neq r}p^{*}{}_{h_{rs}}^{(\alpha_{rs}+n_{i_{rs}}-1)}p^{*}{}_{h_{rs}}^{\left(\alpha_{rs}+N_{h_{rs}}^{-}-1\right)}\right)d\mathbf{p}^{*}d\boldsymbol{\gamma}^{*}=$$

$$=\prod_{r=1}^{S}\left(\frac{\left(b_r+\mathbf{W}_{hr}^{*-}\right)^{a_r+N_{h_r}^{-}}}{\Gamma(a_r+N_{h_r}^{-})}\frac{\Gamma(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}^{-}))}{\prod_{s\neq r}\Gamma(\alpha_{rs}+N_{h_{rs}}^{-})}\right)\times$$

$$\int\cdots\int\prod_{r=1}^{S}\left(\gamma^{*}{}_{h_r}^{\left(n_{i_r}+N_{h_r}^{-}\right)}e^{-\gamma^{*}{}_{h_r}\left(b_r+W_{ir}+\mathbf{W}_{hr}^{*-}\right)}\prod_{s\neq r}p^{*}{}_{h_{rs}}^{\left(\alpha_{rs}+n_{i_{rs}}+N_{h_{rs}}^{-}-1\right)}\right)d\mathbf{p}^{*}d\boldsymbol{\gamma}^{*}=$$

$$=\prod_{r=1}^{S}\left(\frac{\left(b_r+\mathbf{W}_{hr}^{*-}\right)^{a_r+N_{h_r}^{-}}}{\Gamma(a_r+N_{h_r}^{-})}\frac{\Gamma(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}^{-}))}{\prod_{s\neq r}\Gamma(\alpha_{rs}+N_{h_{rs}}^{-})}\right)\times$$

$$\prod_{r=1}^{S}\int\cdots\int\left(\gamma_{h_r}^{\left(n_{i_r}+N_{h_r}^{-}\right)}e^{-\gamma^{*}{}_{h_r}\left(b_r+W_{ir}+\mathbf{W}_{hr}^{*-}\right)}\prod_{s\neq r}p^{*}{}_{h_{rs}}^{\left(\alpha_{rs}+n_{i_{rs}}+N_{h_{rs}}^{-}-1\right)}\right)dp^{*}{}_{rs}d\gamma^{*}{}_{h_r}=$$

$$=\prod_{r=1}^{S}\left(\frac{\left(b_r+\mathbf{W}_{hr}^{*-}\right)^{a_r+N_{h_r}^{-}}}{\Gamma(a_r+N_{h_r}^{-})}\frac{\Gamma(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}^{-}))}{\prod_{s\neq r}\Gamma(\alpha_{rs}+N_{h_{rs}}^{-})}\right)\times$$

$$\prod_{r=1}^{S}\int\gamma^{*}{}_{h_r}^{(N_{h_r})}e^{-\gamma_{h_r}\left(b_r+\mathbf{W}_{hr}^{*}\right)}d\gamma^{*}{}_{h_r}\prod_{s\neq r}\int p^{*}{}_{h_{rs}}^{\left(\alpha_{rs}+N_{h_{rs}}-1\right)}dp^{*}{}_{rs}.$$

By integration we get

$$f((s,w)_i|\mathbf{X}_h^{*-}) = \prod_{r=1}^{S}\left(\frac{\left(b_r+\mathbf{W}_{hr}^{*-}\right)^{a_r+N_{hr}^-}}{\Gamma(a_r+N_{h_r}^-)}\frac{\Gamma(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}^-))}{\prod_{s\neq r}\Gamma(\alpha_{rs}+N_{h_{rs}}^-)}\right)\times$$
$$\prod_{r=1}^{S}\left(\frac{\Gamma(a_r+N_{h_r})}{(b_r+\mathbf{W}_{hr}^{*})^{a_r+N_{h_r}}}\frac{\prod_{s\neq r}\Gamma(\alpha_{rs}+N_{h_{rs}})}{\Gamma(\sum_{s\neq r}(\alpha_{rs}+N_{h_{rs}}))}\right),$$

where $\Gamma(\cdot)$ denotes the gamma function, while $\mathbf{W}_{hr}^{*-}$ represent the sum of the sojourn times in the state r for the elements lying in the cluster h with exclusion of the i-th observation. Moreover n_{i_r} represents the number of visits of the state r and $n_{i_{rs}}$ represents the number of transitions from the state r to the state s, such that $n_{i_r} = \sum_{s\neq r} n_{i_{rs}}$; finally $N_{h_r}^-$ represents the number of visits of the state r for the trajectories in lying in the cluster h with exclusion of the i-th trajectory, while $N_{h_{rs}}^-$ represents the number of transition from the state r to the state s for the trajectories lying in the cluster h excluding the i-th observation, such that $N_{h_r}^- = \sum_{s\neq r} N_{h_{rs}}^-$.

Algorithm 6 Gibbs sampler for DPM of Markov models

• **Clustering**:
– for $i = 1\ldots, N$ draw $\psi_i \sim \pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{X})$ from (2.26);
• **Cluster parameters**:
– for $h = 1, \ldots, k$ draw $\theta_h^* \sim \pi(\theta_h^*|\boldsymbol{\psi}, \mathbf{X})$ from (2.25).

2.2.4 BNP inference for discretely observed CTMSM

As we have already seen in Chapter 1, inference for multi-state models in continuous time become hard when observations when the process is observed at discrete points. The likelihood function is not available, therefore approximations are required. In Chapter 1 we presented a parametric MCMC method for inference in semi-Markov models based on the Uniformization-based algorithm, proposed by Hobolth and Stone [2009]. Here we discuss the Bayesian nonparametric extension of the methodology presented in Chapter 1. For each cluster, we implement the Uniformization algorithm in order to get full observed path using the corresponding cluster rate matrix, eliminating the problem of having discrete observations and extending the DPM for CTMSM also to discretely observed case.

In Algorithm 1 we describe in detail how Uniformization works. The input for the algorithm are start-point, end-point and a rate matrix A. After simulating the trajectories, we follow the same scheme as in the fully observed case (Algorithm 7).

Let $X_{(t_i,i)} = (\hat{s}, \hat{w})_{(t_i,i)} = (\hat{s}, \hat{w})_i$, $i = 1, \dots, N$, be discretely observed paths with not necessarily the same length. Let $(s, w)_i$ be the i-th discretely observed unit lying inside the cluster h, such that $\psi_i = h$. Then, recalling that $\theta_h^* = (\boldsymbol{\gamma}_h^*, \tilde{\mathbf{p}}_h^*)$, we construct the rate matrix for the h-cluster A_h^* as

$$\begin{cases} A_{h,rr}^* = -\gamma_{h,rr}^* \\ A_{h,rs}^* = \gamma_{h,rr}^* \cdot p_{h,rs}^*. \end{cases} \tag{2.27}$$

By Uniformization function we simulate between each two observed points the continuous time trajectory, getting as output the path $(s, w)_i$.

After having got the full trajectory, we can follow the previous scheme: first updating the cluster membership, drawing ψ_i from $\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{X})$ (2.26) and then updating the cluster parameters θ_h^*, drawing from $\pi(\theta_h^*|\boldsymbol{\psi}, \mathbf{X})$ (2.25).

Algorithm 7 Gibbs sampler for DPM of discretely observed continuous time Markov models

• **Path simulation**:
– for $i = 1, \dots, N$ draw $(s, w)_i$ from Uniformization$((\hat{s}, \hat{w})_i, A_h^*)$
• **Clustering**:
– for $i = 1 \dots, N$ draw $\psi_i \sim \pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{X})$ from (2.26);
• **Cluster parameters**:
– for $h = 1, \dots, k$

$$\begin{cases} \text{draw } \theta_h^* \sim \pi(\theta_h^*|\boldsymbol{\psi}, \mathbf{X}) \text{ from (2.25)} \\ \text{set } A_{h,rr}^* = -\gamma_{h,rr}^* \\ \text{set } A_{h,rs}^* = \gamma_{h,rr}^* \cdot p_{h,rs}^*. \end{cases}$$

Table 2.1: True values of the parameters of the two mixture components for fully observed data.

ψ	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	γ_1	γ_2	γ_3
1	0.87	0.13	0.13	0.87	0.13	0.87	4	4	4
2	0.33	0.66	0.33	0.66	0.66	0.33	0.3	0.3	0.3

Table 2.2: Simulated paths with $n = 100$: posterior mean and standard deviation for the model parameters.

	ψ	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	γ_1	γ_2	γ_3
$E(\cdot\|x)$	1	0.83	0.17	0.12	0.88	0.14	0.86	3.75	4.29	3.70
$SD(\cdot\|x)$	1	0.03	0.03	0.02	0.02	0.02	0.02	0.30	0.20	0.18
$E(\cdot\|x)$	2	0.28	0.72	0.43	0.57	0.63	0.37	0.52	0.85	0.11
$SD(\cdot\|x)$	2	0.17	0.17	0.26	0.26	0.26	0.26	0.27	0.89	0.13

2.3 Applications

After having introduced the Gibbs sampler for DPM of Markov models, we have extended the methodology to the more interesting case of discretely observed processes. In this section we discuss the behavior of the proposed method with both fully observed and discretely observed trajectories. We first show results with simulated data, then we present an application to the Standard and Poor's rating data as in Chapter 1.

2.3.1 Simulation study

We divide the simulation study in two parts: in the first we show the behavior of the algorithm with fully observed trajectories, while in the second part we have discretely observed data. We have simulated $n = 100$ trajectories, each one with time length 5, from a finite mixture of three-states Markov Processes. The true values of the generating parmeters are presented in Table 2.1. We assume the concentration parameter to be $M = 1$ and the centering measure to be a product measure of Gamma and Dirichlet as in (2.21); we ran the MCMC sampler for 10000 iterations with a burnin of 2000 iterations. Results are shown in Table 2.2. Although the sample size is small, the model captures the information and the posterior densities of most of the models parameters concentrates around the true values. In Figure 2.1 and Figure 2.2 are shown the traceplots of the model parameters for each cluster.

Table 2.3: True values of the parameters of the two mixture components for discretely observed data.

ψ	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	γ_1	γ_2	γ_3
1	0.4	0.6	0.6	0.4	0.8	0.2	1.5	1.5	1.5
2	0.5	0.5	0.67	0.33	0.83	0.17	3	3	3

Table 2.4: Discretely observed data with $n = 200$: posterior mean and standard deviation for the model parameters.

	ψ	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	γ_1	γ_2	γ_3
$E(\cdot\|x)$	1	0.44	0.56	0.49	0.51	0.84	0.16	0.61	1.94	1.66
$SD(\cdot\|x)$	1	0.14	0.14	0.23	0.23	0.16	0.16	0.45	0.96	0.94
$E(\cdot\|x)$	2	0.49	0.51	0.67	0.33	0.88	0.12	2.65	3.15	2.34
$SD(\cdot\|x)$	2	0.15	0.15	0.21	0.21	0.17	0.17	1.19	1.37	1.28

Table 2.5: Discretely observed data with $n = 500$: posterior mean and standard deviation for the model parameters.

	ψ	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	γ_1	γ_2	γ_3
$E(\cdot\|x)$	1	0.37	0.63	0.64	0.36	0.80	0.20	1.40	1.71	1.16
$SD(\cdot\|x)$	1	0.16	0.17	0.25	0.25	0.21	0.21	1.15	1.28	0.89
$E(\cdot\|x)$	2	0.42	0.58	0.81	0.19	0.94	0.06	2.18	3.45	2.34
$SD(\cdot\|x)$	2	0.08	0.08	0.12	0.12	0.08	0.08	1.17	0.93	0.78

For the discretely observed case, we have simulated two samples of size $n = 200$ and $n = 500$, where each individual has been observed 10 times in a interval of length 2. Observations have been generated from a mixture of three-states Markov processes with parameters presented in Table 2.3. The hyperparameters setting is the same as in the fully observed case. Results for both the simulations are shown in Table 2.4 and Table 2.5, while traces of the posterior model parameters for each cluster are shown in Figure 2.3, Figure 2.5 and Figure 2.6. Posterior estimates are concentrated around the real values of the model parameters, although there is higher standard error comparing to the fully observed data case, since there is less information from data. Naturally, posterior estimates concentrates around the mean as sample size increases, interval length reduces and number of observed interval increases.

Figure 2.1: MCMC traces for first and second clusters of posterior rate parameters for n=100 and fully observed trajectories.

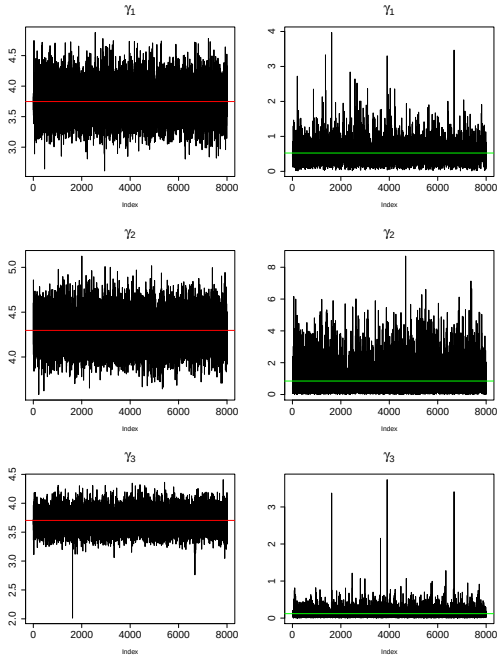


Figure 2.2: MCMC traces for first and second clusters of transition probabilities posterior parameters for n=100 and fully observed trajectories.

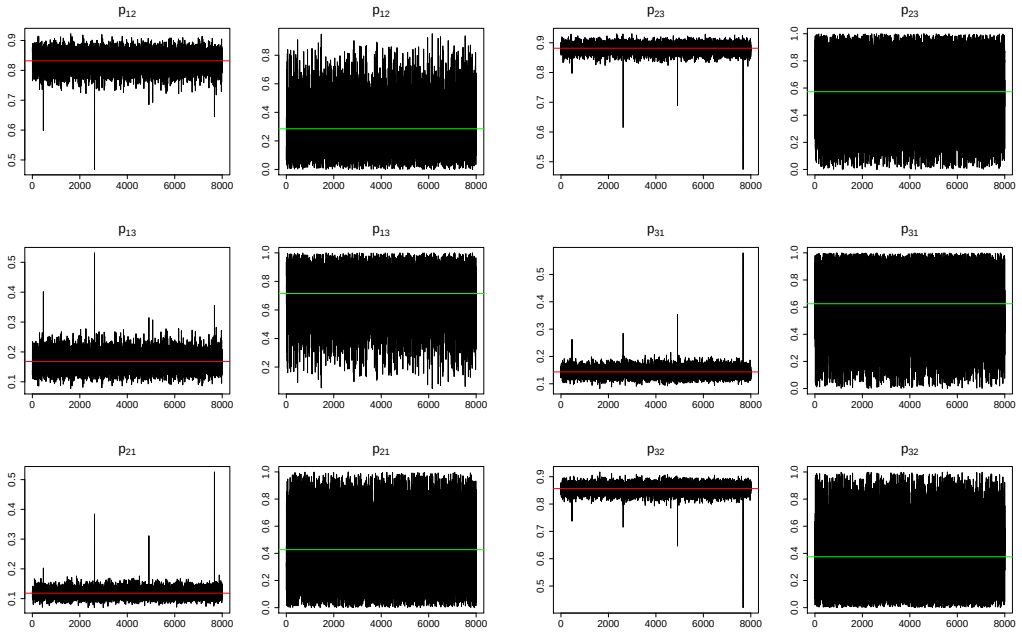


Figure 2.3: MCMC traces for first and second clusters of posterior rate parameters with sample size n=200.

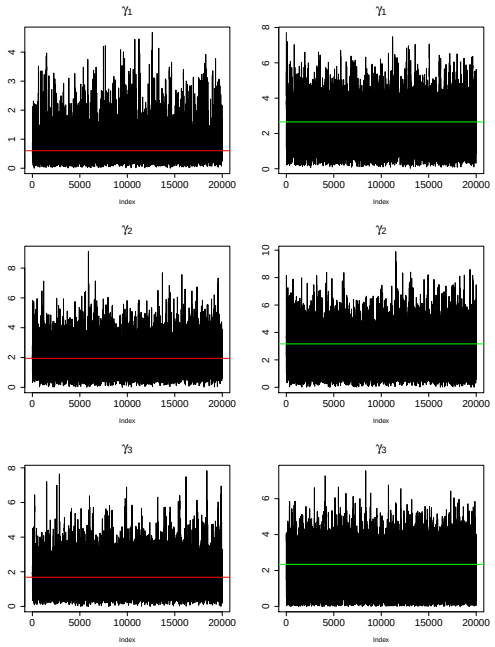


Figure 2.4: MCMC traces for first and second clusters of posterior rate parameters with sample size n=500.

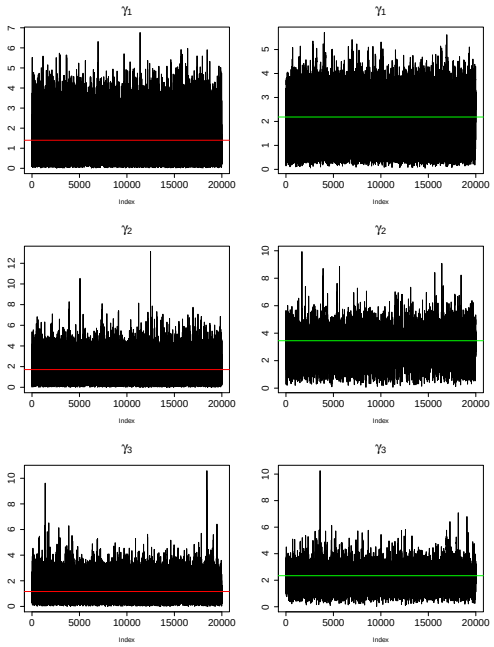


Figure 2.5: MCMC traces for first and second clusters of transition probabilities posterior parameters with sample size n=200.

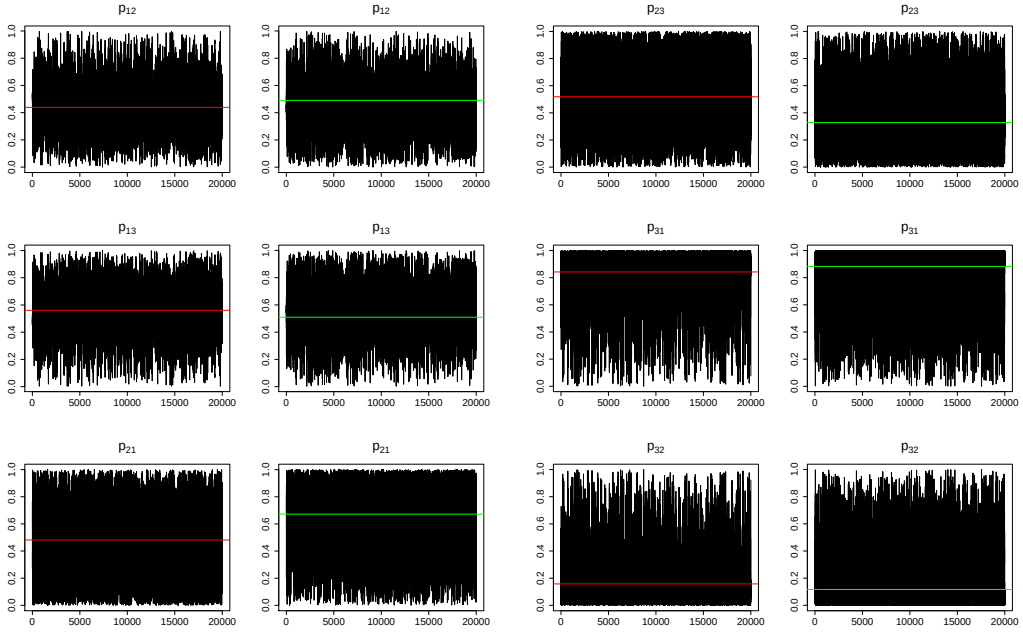
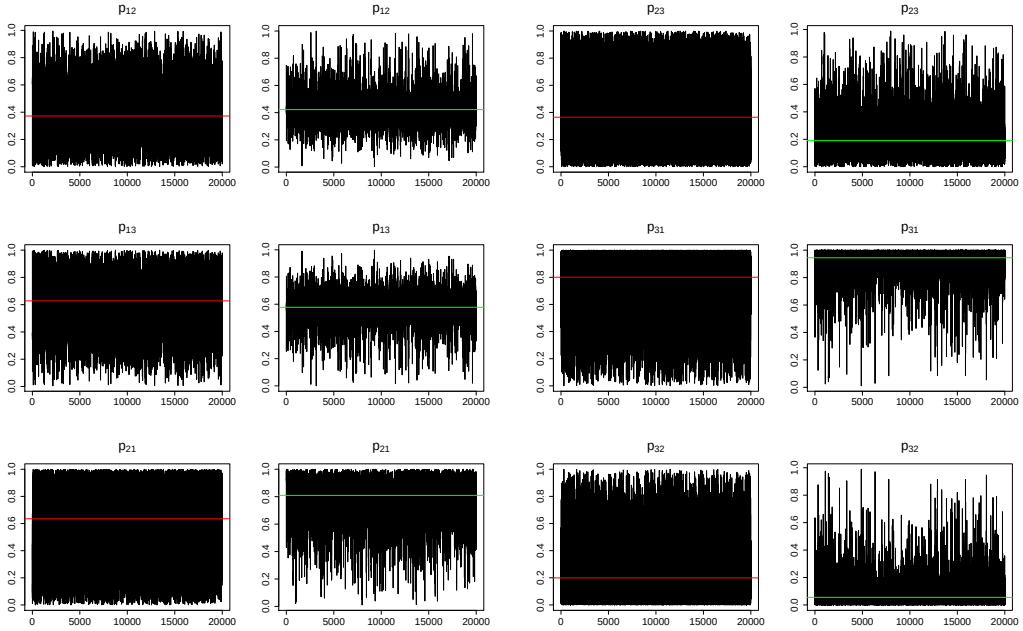


Figure 2.6: MCMC traces for first and second clusters of transition probabilities posterior parameters with sample size n=500.



2.3.2 BNP modelling of rating classes with Standard and Poor's data

As in Chapter 1, here we model Standard and Poor's rating classes data, considering 205 institutions, each one observed at least two times. Again, we summarize the data in four rating classes (A,B,C,D), with the first class (A) representing solvent institutions and the last class (D) which is assumed to be the absorbing state and represents the default. We model data via Dirichlet Process mixture for discretely observed data, assuming the concentration parameter M to be 1, while the centering measure is again assumed to be a product measure of Gamma and Dirichlet as in (2.21). We ran 25000 iterations with a burnin of 5000 iterations. A summary of the output is shown in Table 2.6. From Figure 2.7 we note that observations concentrates in the first two clusters, which have almost the same size. In Figure 2.8 we show the traceplots of the posterior rate parameters in each of mixture components: there is strong evidence of the difference in terms of expected values between the parameters of the two estimated mixture components. This result proves the adequacy of the model for this data. Furthermore, it leads to a deeper intuition regarding the nature of the process compared to the MCMC algorithm for semi-Markov presented in Chapter 1. In that case, we have observed that there is a decreasing rate as the rating class increases. Here we can observe that the insitutions belonging to the first cluster show an higher rate in the first (A) and third (C) states, meaning that they are based in less stable economies and they are rated most of the times as class B institutions. On the contrary, the institutions lying in the second cluster reside in stable economies and they are rated most of the times as class A insitutions. The result is in line with the theory of economic stability: an economy with absence of excessive fluctuations in the macroeconomy, with fairly constant output growth and low and stable inflation would be considered economically stable. An economy with frequent large recessions, a pronounced business cycle, very high or variable inflation, or frequent financial crises would be considered economically unstable.

Table 2.6: Standard and Poor's rating data: posterior mean and standard deviation for the model parameters and average sojourn times expressed in years.

	ψ	p_{12}	p_{13}	p_{14}	p_{21}	p_{23}	p_{24}	p_{31}	p_{32}	p_{34}
$E(\cdot\|x)$	1	0.85	0.08	0.07	0.36	0.42	0.22	0.25	0.26	0.49
$SD(\cdot\|x)$	1	0.10	0.08	0.07	0.21	0.22	0.18	0.19	0.20	0.23
$E(\cdot\|x)$	2	0.41	0.29	0.30	0.26	0.52	0.22	0.13	0.23	0.64
$SD(\cdot\|x)$	2	0.25	0.22	0.22	0.11	0.14	0.11	0.12	0.15	0.17

	ψ	γ_1	γ_2	γ_3
$E(\cdot\|x)$	1	0.38	0.008	1.59
$SD(\cdot\|x)$	1	0.13	0.01	1.08
$E(\cdot\|x)$	2	0.006	0.32	0.89
$SD(\cdot\|x)$	2	0.005	0.11	0.43

	ψ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$
$E(\cdot)$	1	9.71	361.89	1.50
$Me(\cdot)$	1	2.76	157.68	0.72
$E(\cdot)$	2	3579.53	4.58	1.45
$Me(\cdot)$	2	115.54	3.27	1.22

Figure 2.7: Lables distribution across the MCMC iterations.

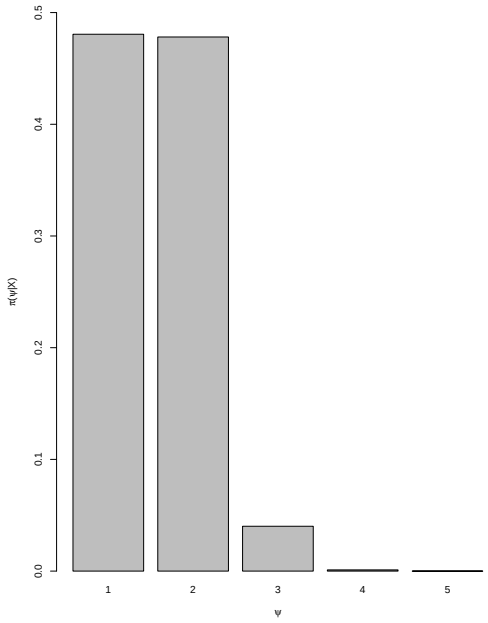
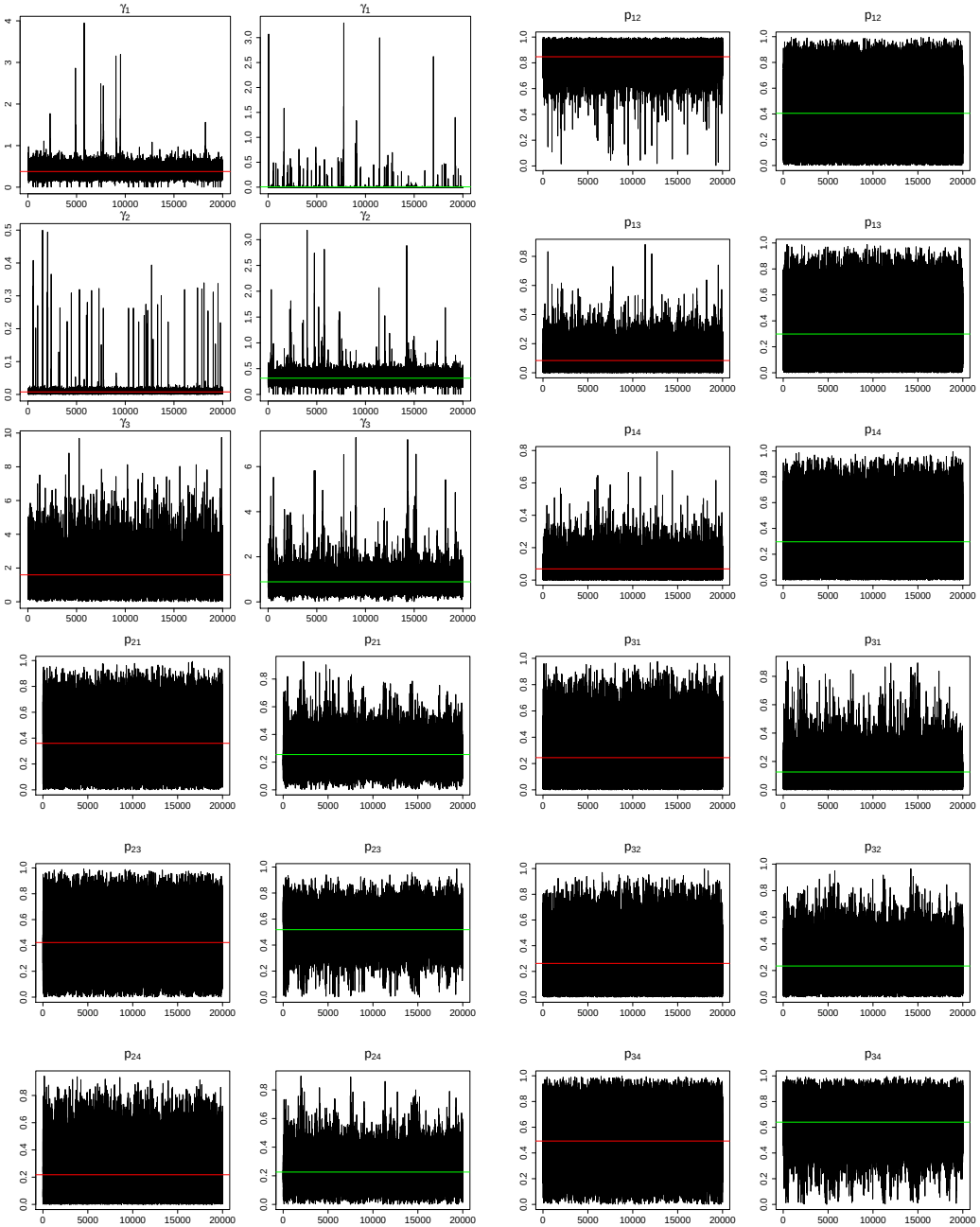


Figure 2.8: MCMC traces for first and second clusters of posterior rates and transition probabilities for Standard and Poor's data.



Chapter 3

MCMC methods for high dimensional copulas

Modelling a joint distributions of two variables, let we say Y_1 and Y_2, described by a function $F(Y_1, Y_2) = P(Y_1 \leq y_1, Y_2, \leq y_2)$ would make it possible to fully describe the dependence between these variables. Nevertheless, relations between data are often difficult to describe by simple bivariate distributions. Ideally, a statistical procedure for the estimation of a joint distribution would provide first the estimates of the marginal distribution using the univariate data and prior information, then the estimates the dependence structure, using the multivariate data. The copula approach allows for the construction of joint distributions as product of marginals and copula function.

Even if Bayesian methods have proven successful in both formulating and estimating multivariate models, the most popular estimation methods are full or two-stage maximum likelihood estimation [Joe, 2005] and method of moments style estimators in low dimensions [Genest and Rivest, 1993]. In the Bayesian literature we find Huard et al. [2006], which suggested a method to select between different bivariate copulas, and dos Santos Silva and Lopes [2008] who use Markov chain Monte Carlo methods to estimate low dimensional parametric copula functions. Hoff [2007] proposed a Gibbs sampler for semiparametric Gaussian copulas, estimating marginals via rank likelihood. Min and Czado [2010] considered methods to estimate so called "vine" copulas with continuous margins using MCMC. Ausin and Lopes [2010] presented a Bayesian estimation of multivariate time series with copula-based time varying cross-sectional dependence, while Smith and Khaled [2012] suggested a Bayesian data augmentation for estimating copulas with discrete marginals. Wu et al. [2014] presented a Bayesian nonparametric method for

multivariate copula models using a Dirichlet Process Mixture of multivariate skew-Normal copulas; they also proposed a Dirichlet Process Mixture of bivariate Gussian copulas [Wu et al., 2015]. Grazian and Liseo [2017] described an approximate Bayesian Monte Carlo method for inference on semiparametric copulas. Finally, Dalla Valle et al. [2018] presented a Bayesian nonparametric method for bivariate conditional copulas, accounting for covariates in copula density estimation.

In this Chapter we first describe the mathematical structure of these models and show the vine copula construction which facilitates the treatment of high dimensional joint distributions; then we introduce the main inferential aspects, from both frequentist and Bayesian perspectives. In the last section we present two Bayesian nonparametric methods for estimating multidimensional copula densities, also including the conditional multivariate copula case, desribing in detail the inferential and computational aspects. We follow Shemyakin and Kniazev [2017] in the introduction to copula models.

3.1 Copula models

A function $C : \mathbf{I}^2 \to \mathbf{I}$ is called a *copula* if it satisfies the following conditions:

1. for any $u, v \in [0,1]$, $C(0,v) = C(u,0) = 0$;
2. for any $u, v \in [0,1]$, $C(1,v) = u$, $C(u,1) = v$;
3. for any $0 < u_1 < u_2 < 1$ and $0 < v_1 < v_2 < 1$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) > 0.$$

A bivariate copula is then the distribution function of a bivariate random variable with Uniform(0,1) marginal distributions. For any copula $C(u,v)$, partial derivatives $\frac{\partial C}{\partial u}$ and $\frac{\partial C}{\partial v}$ exist for almost all $u, v \in [0,1]$. Let $\frac{\partial^2 C}{\partial u \partial v}$ and $\frac{\partial^2 C}{\partial v \partial u}$ exist and be continuous on $\mathbf{I}^2$. Then, we define the *copula density* as

$$c(u,v) = \frac{\partial^2 C}{\partial u \partial v} = \frac{\partial^2 C}{\partial v \partial u}.$$

If $u = F(x)$ and $v = G(y)$ are two distribution functions, any function $C(F(x), G(y))$ is a valid bivariate distribution function. The converse is true as well: every joint probability distribution function is a copula. We may rewrite the bivariate joint density function of (X, Y) as

$$f(x,y) = \frac{\partial^2 C}{\partial u \partial v} \cdot \frac{dF}{dx} \cdot \frac{dG}{dy} = f(x) \cdot f(y) \cdot c(u,v), \tag{3.1}$$

while

$$F(x|y) = \frac{\partial C(F(x), G(y))}{\partial G(y)}. \tag{3.2}$$

The most important statement in copulas theory is without any doubt the *Sklar's theorem* [Sklar, 1959].

Theorem 3.1 *Let J be a joint distribution function with margins F and G. Then there exists a copula C such that for all x, y*

$$J(x,y) = C(F(x), G(y)). \tag{3.3}$$

If F and G are continuous, then C is unique.

The theorem states that every valid bivariate distribution can be represented as copula of its marginals. It allows to separate margins from the copula density, simplifying the computation. Hence, thanks to this theorem,

if we have to build a model for bivariate distribution given marginals, we only have to find the proper copula model.

Moreover let $C(u, v)$ be a copula defining a joint distribution with marginal distributions u and v. Then, the function

$$\bar{C}(u, v) = u + v - 1 + C(1 - u, 1 - v) \tag{3.4}$$

is called *survival* copula.

3.1.1 Elliptical copulas

We now consider one popular class of copulas often used in applications, named *elliptical*, which includes Gaussian and student t-copulas.

We first define the main property of these copulas, the *elliptical symmetry*: a copula $C(u, v)$ is elliptically symmetric if it is both symmetric with respect to the main diagonal of the unit square, $u = v$, and with respect to the diagonal $u = 1 - v$. It turns out the identity

$$C(u, v) = \bar{C}(u, v)$$

meaning that C coincides with its survival version $\bar{C}$.

Let $Q_\rho(s, t)$ represent the class of bivariate elliptical distributions, with density function

$$q_\rho(s, t) = \frac{k^2}{\sqrt{1 - \rho^2}} g\left(\frac{s^2 - 2\rho st + t^2}{1 - \rho^2}\right), \tag{3.5}$$

where $\rho \in (-1, 1)$, $g : \mathbb{R} \to \mathbb{R}^+$ with $\int_{\mathbb{R}} g(t)dt < \infty$ and k is the normalizing constant. Moreover, we define

$$q(s) = \int_{\mathbb{R}} q_0(s, t)dt = \int_{\mathbb{R}} q_0(t, s)dt$$

and

$$q(t) = \int_{\mathbb{R}} q_0(s, t)ds = \int_{\mathbb{R}} q_0(t, s)ds$$

as the marginal density functions of the components of the vector (s, t) with $\rho = 0$.

The most popular method of constructing elliptically symmetric copulas using an elliptic distribution $Q_\rho(s,t)$ is the *method of inverses*. Considering U and V independently uniformly distributed on the line [0,1], the inverse transforms $Q^{-1}(U)$ and $Q^{-1}(V)$ are two independent random variables with the same c.d.f. Q. Hence, we define an elliptical copula for any $u, v \in [0,1]$ as

$$C_\rho(u,v) = Q_\rho(Q^{-1}(u), Q^{-1}(v)).$$

Moreover, denoting $s = Q^{-1}(u)$ and $t = Q^{-1}(v)$, the resulting copula density is

$$c_\rho(u,v) = \frac{q_\rho(s,t)}{q(s)q(t)}.$$

The most attractive property of this construction is the effective separation of the margins $u = F(x)$ and $v = G(y)$ from the structure of dependence. Thus, we can rewrite the bivariate distribution $J(x,y)$ with margins $F(x)$ and $G(y)$ as

$$J(x,y) = Q_\rho\left(Q^{-1}\left(F\left(x\right)\right), Q^{-1}\left(G\left(y\right)\right)\right),$$

where the parameter ρ characterizes the intensity of dependence.

We now present two families of copulas belonging to the elliptical class. The first one is the Gaussian copula: we get this copula in the following way

$$C_\rho(u,v) = \Phi_\rho\left(\Phi^{-1}\left(u\right), \Phi^{-1}\left(v\right)\right),$$

where Φ_ρ is the distribution function of the bivariate normal distribution with zero means, unit variances and correlation ρ, while Φ is the standard normal c.d.f..

Consequently, denoting $s = \Phi^{-1}(u)$ and $t = \Phi^{-1}(v)$, the density of the Gaussian copula is

$$c_\rho(u,v) = \frac{\partial^2 C_\rho(u,v)}{\partial u \partial v} = \frac{\partial^2 \Phi_\rho(s,t)}{\partial s \partial t} \cdot \frac{\partial s}{\partial u} \cdot \frac{\partial t}{\partial v} = \frac{\phi_\rho(s,t)}{\phi(s)\phi(t)},$$

where ϕ_ρ represents the density of the bivariate normal distribution with zero mean vector, unit variances and correlation ρ, while ϕ represents the density of the standard normal distribution. In explicit form we have

$$c(u,v) = \frac{1}{\sqrt{(1-\rho^2)}} \exp\left(-\frac{\rho^2 s^2 + \rho^2 t^2 - 2\rho s t}{2(1-\rho^2)}\right). \tag{3.6}$$

Basically, the idea beyond the Gaussian copula is to transform two random variables X and Y with c.d.f respectively F and G into standard normal

variables $S = \Phi^{-1}(F(x))$ and $T = \Phi^{-1}(G(x))$, such that the whole dependence between X and Y is expressed thorough the linear dependence of their standard normal transforms.

In applications it often happens that data requires heavy-tailed multivariate distributions and the Student t-copula may be used. Following the previous scheme, we can obtain this copula as

$$C_{\eta\rho}(u, v) = T_{\eta\rho}\left(T_{\eta}^{-1}(u), T_{\eta}^{-1}(v)\right),$$

where $T_{\eta\rho}$ is the bivariate Student t-distribution c.d.f. with η degrees of freedom and correlation ρ, while T_{η} is the Student t-distribution c.d.f. with η degrees of freedom.

Again, denoting $s = T_{\eta}^{-1}(u)$ and $t = T_{\eta}^{-1}(v)$, we get the copula density as follows

$$c_{\eta\rho}(u, v) = \frac{\partial^2 C_{\eta\rho}(u, v)}{\partial u \partial v} = \frac{\psi_{\eta\rho}(s, t)}{\psi_{\eta}(s)\psi_{\eta}(t)}$$

where $\psi_{\eta\rho}$ is the bivariate student t-distribution density with η degrees of freedom and correlation ρ, while ψ_{η} is the Student t-distribution density with η degrees of freedom. Thus, the explicit formula for the density is

$$c_{\eta\rho}(u, v) = \frac{\Gamma\left(\frac{\eta+2}{2}\right)\Gamma\left(\frac{\eta}{2}\right)}{\sqrt{1-\rho^2}\cdot\Gamma^2\left(\frac{\eta+1}{2}\right)} \cdot \frac{\left(\left(1+\frac{s^2}{\eta}\right)\left(1+\frac{t^2}{\eta}\right)\right)^{\frac{\eta+1}{2}}}{\left(1+\frac{s^2+t^2-2\rho st}{\eta(1-\rho^2)}\right)^{\frac{\eta+2}{2}}}. \tag{3.7}$$

For modelling joint distributions for both Gaussian and Student t-copula, we combine the copula with any marginal distributions $u = F(x)$ and $v = G(y)$.

3.1.2 Archimedean copulas

Suppose again to have a copula $C(u, v)$. Another way of thinking of dependence between the margins u and v may be as "deviation" from the independence case, corresponding to the copula $C(u, v) = uv$. Since transformations ξ which support additivity are easier to manipulate, the idea is to determine a subclass of copulas such that $\xi(C) = \xi(u) + \xi(v)$.

Let $\xi(t)$ be a continuous, strictly decreasing function from $[0, 1]$ to $[0, \infty)$ such that $\xi(1) = 0$. We determine the *pseudo-inverse* function of ξ as follows:

$$\xi^{[-1]}(t) = \max\left\{\xi^{-1}(t), 0\right\}.$$

Pseudo-inverses extend the inverse transformation to functions of limited range. If $\xi(t) \to \infty$ when $t \to 0$, then the pseudo-inverse function coincides with the inverse function.

Theorem 3.2 *Let $\xi : \mathbf{I} \to [0, \infty)$ be a continuous, strictly decreasing function such that $\xi(1) = 0$. Then the function*

$$C_\xi(u, v) = \xi^{[-1]}(\xi(u) + \xi(v))$$

is a copula if and only if ξ is convex.

If $C_\xi(u, v)$ satisfies these conditions, it is called *Archimedean copula* and the function $\xi(t)$ is its *generator*. We can check whether a copula is Archimedean denoting by $\delta_c(u) = C(u, u)$ the diagonal projection of a copula, if for all $u \in (0, 1)$ $\delta_C(c) < u$, then $C(u, v)$ is Archimedean. Moreover, while it is easy to verify that Archimedean copulas are commutative, hence $C(u, v) = C(v, u)$, it is much harder to prove that they are associative: $C(C(u, v), w) = C(u, C(v, w))$.

If the second derivative ξ'' exists, the Archimedean copula density may be expressed through its generator and its derivatives as

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v} = -\frac{\xi''(C(u, v))\xi'(u)\xi'(v)}{\left(\xi'(C(u, v))\right)^3}.$$

We now briefly introduce some Archimedean copula families, showing how to get copula densities from different generators.

The first copula we introduce is the *Clayton* copula. Let $\xi(t)$ be the generator such that

$$\xi(t) = \frac{1}{\alpha}\left(t^{-\alpha} - 1\right),$$

thus the pseudo-inverse

$$\xi^{[-1]}(s) = \max\left\{(1 + \alpha s)^{-1/\alpha}, 0\right\}$$

defines the Clayton's class of copulas

$$C_\alpha(u, v) = \max\left\{\left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}, 0\right\}, \alpha \in [-1, 0) \cup (0, \infty).$$

Since this copula is typically applied for $\alpha > 0$, we get the density as

$$c_\alpha(u, v) = \frac{(\alpha + 1)(uv)^\alpha}{\left(u^\alpha + v^\alpha - (uv)^\alpha\right)^{\frac{1}{\alpha}+2}}, \alpha > 0. \tag{3.8}$$

The Clayton copula is used in situations where the dependence between low values of u and v is stronger than the dependence between values close to 1.

The second family we consider is the *Frank* copula. Again, let the generator be

$$\xi(t) = -\log \frac{e^{-\alpha t} - 1}{e^{-\alpha} - 1},$$

then the pseudo-inverse function is

$$\xi^{[-1]}(s) = -\frac{1}{\alpha} \log \left(1 + e^{-s}(e^{-\alpha} - 1)\right).$$

Thus, we get the Frank Copula

$$C_\alpha(u, v) = -\frac{1}{\alpha} \log \left(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{e^{-\alpha} - 1}\right), \quad \text{with} \quad \alpha \neq 0,$$

where the density is

$$c_\alpha(u, v) = \frac{\alpha (1 - e^{-\alpha}) e^{-\alpha(u+v)}}{\left(e^{-\alpha} - 1 + (e^{-\alpha u} - 1)(e^{-\alpha v} - 1)\right)^2}. \tag{3.9}$$

This copula is used when the strength dependence is relatively similar for values of u and v.

The last Archimedean copula class we present is the *Gumbel-Hougaard* class. Let the generator be

$$\xi(t) = (-\log t)^\alpha$$

with pseudo-inverse

$$\xi^{[-1]}(s) = e^{-s^{1/\alpha}}.$$

Thus, the Gumbel-Hougaard copula is

$$C_\alpha(u, v) = \exp\left(-\left((-\log u)^\alpha + (-\log v)^\alpha\right)^{1/\alpha}\right), \alpha \geq 1,$$

with density

$$c_\alpha(u, v) = (uv)^{-1} (\log u \cdot \log v)^{\alpha - 1} \left(w^{2/\alpha - 2} + (\alpha - 1) w^{1/\alpha - 2}\right) C_\alpha(u, v), \tag{3.10}$$

where $w = (-\log u)^{-\alpha} + (-\log v)^{-\alpha}$.

The Gumbel-Hougaard copula is used when the strongest dependence is for values of u and v close to 1 (right tail).

The popularity of these three classes of copulas derives from the flexibility in applications, since their structures allow to model various types of non-linear dependence, especially the tail-dependence. Survival copulas can simply be obtained from each of these three copulas.

3.1.3 Multidimensional copulas and vine construction

All discussions so far were restricted to bivariate copulas, even if most interesting applications are in dimension $d \geq 3$ (multivariate copulas). Construction for elliptical and Archimedean copulas can be naturally extended to higher dimensions.

Elliptical distributions $Q_{d,R}$ of a random vector $t = (t_1, \ldots, t_d)$ can be defined by its joint density function

$$Q_{d,R} = |\Sigma|^{-1/2} k((t-1)^T \Sigma^{-1} (t-1)))$$

where μ is a d-dimensional vector of means, Σ is the $d \times d$ positive definite covariance matrix, $k(t)$ is some non negative function of a variable integrable over $\mathbb{R}$, while R with elements $R_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ is the correlation matrix determining all pairwise associations between the components of t.

Let $Q_i(t_i)$ be the marginal distribution on t_i. We may define the elliptical copula as

$$Q_{d,R} = Q_{d,R}(Q_1^{-1}(u_1), \ldots, Q_d^{-1}(u_d)).$$

The most popular multidimensional elliptical copula is the Gaussian. Considering data with marginal distributions $u_i = F(y_i)$, we define

$$C_{\Sigma_\rho}(u_1, \ldots, u_d) = \Phi_{d,\Sigma_\rho}(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_d(y_d))).$$

where Φ is standard normal distribution and Φ_{d,Σ_ρ} is d-variate normal with zero mean, unit variances, and covariance (and correlation) matrix Σ_ρ. Thanks to the multivariate normal properties, the off-diagonal elements of matrix Σ_ρ describe pairwise associations, so the strength of association may differ for different pairs of components of the vector Y.

A formal extension of Archimedean copula construction to dimension $d > 2$ is straightforward:

$$C_\alpha(u_1, \ldots, u_d) = \xi^{[-1]}(\xi_\alpha(u_1), \ldots, \xi_\alpha(u_d)),$$

is a copula with generator ξ_α. The problem here is the exchangeability or symmetry requirement, which suggests that one parameter α describes all pairwise associations, thus all these associations have to be of equal strength. This is a substantial limitation of the modelling process. Therefore, to address nonexchangeable situations, this construction has to be modified.

Moreover, with "classical" multivariate copula modelling there are parametric constraints which provide poor flexibility, in both elliptical and Archimedean cases, since the dependencies between each variable are assumed to belong to the same parametric family.

These problems were noted by Aas et al. [2009], which proposed to use a wider class of multivariate copulas, based on a technique of construction of multivariate copulas introduced by Joe [1996] and also treated by Bedford and Cooke [2001] and Kurowicka and Cooke [2006]. The method consists in rewriting a d-variate copula as product of $d(d-1)/2$ bivariate copulas and it is called *pair-copula* (or *vine copula*) construction. This approach is more flexible, as we can select bivariate copulas from a wide range of (parametric) families.

Let us begin with dimension $d = 3$ considering the problem of modelling the joint distribution $P(X \leq x, Y \leq y, Z \leq z)$, where the marginals are $u = F(x)$, $v = G(y)$ and $w = H(z)$, while Y is designated as the central variable, meaning that its associations with X and Z are most important and modelling the association between X and Z will have a lower priority. The choice of the central variable is inevitable, and is established either from context or by preliminary estimation of the strength of pairwise associations.

Let us denote by $f(x)$, $g(x)$ and $h(x)$ respective marginal densities of X, Y and Z. We can write down double and triple joint densities as

$$fg(x, y) = g_{Y|X}(y|x)f(x)$$

the joint density between x and y and

$$fgh(x, y, z) = h_{Z|X,Y}(z|x, y)fg(x, y).$$

the joint density between x, y and z. Thus we obtain the joint density of all three variables as a "chain" of conditional densities

$$fgh(x, y, z) = h_{Z|X,Y}(z|x, y)g_{Y|X}(y|x)f(x). \tag{3.11}$$

Using two pair-copulas, we first model associations (X, Y) and (Y, Z):

$$C_1(u, v) = C_1(F(x), G(y)), \quad C_2(v, w) = C_2(G(y), H(z)).$$

with densities respectively $c_1(F(x), G(y))$ and $c_2(G(y), H(z))$. We rewrite the bivariate joint densities as:

$$fg(x, y) = c_1(F(x), G(y))f(x)g(y), \quad gh(y, z) = c_2(G(y), H(z))g(y)h(z).$$

Conditional distribution may be expressed as

$$g_{Y|X}(y|x) = c_1(F(x), G(y))g(y), \quad h_{Z|Y}(z|y) = c_2(G(y), H(z))h(z). \quad (3.12)$$

We can rewrite the joint density of all three variables as

$$fgh(x, y, z) = \frac{h_{Z|X,Y}(z|x, y)}{h_{Z|Y}(z|y)} c_1(F(x), G(y))c_2(G(y), H(z))f(x)g(y)h(z).$$

We introduce the copula $C_3(F_{X|Y}(x|y), H_{Z|Y}(z|y))$ having density

$$c_3(F_{X|Y}(x|y), H_{Z|Y}(z|y)) = \frac{h_{Z|X,Y}(z|x, y)}{h_{Z|Y}(z|y)},$$

assuming the same copula structure to link two conditional distributions. We may finally rewrite the joint density of the three variables as

$$fgh(x, y, z) = c_1(F(x), G(y))c_2(G(y), H(z))c_3(F_{X|Y}(x|y), H_{Z|Y}(z|y))f(x)g(y)h(z),$$

meaning that

$$c(F(x), G(y), H(z)) = c_1(F(x), G(y))c_2(G(y), H(z))c_3(F_{X|Y}(x|y), H_{Z|Y}(z|y)). \quad (3.13)$$

This construction is known as vine copula or *pair copula* construction.

$$X, Y \quad Y, Z$$

$$X, Z|Y$$

For dimension $d = 4$, we introduce a further variable W, thus we have to add a new link. We distinguish vine diagrams in two types: C-Vine and D-Vine.

- Three levels of association for a C-vine.

$$X, Y \quad Y, Z \quad Z, W$$

$$X, Z|Y \quad Z, W|Y$$

$$X, W|Z, Y$$

- Three levels of association for a D-vine.

$$X,Y \quad Y,Z \quad Z,W$$
$$X,Z|Y \quad Y,W|Z$$
$$X,W|Z,Y$$

Generalizing, we say that with the pair-copula construction, a d-variate copula may be rewritten as the product of $d(d-1)/2$ bivariate copulas.

3.2 Bayesian inference for copulas

We now briefly present classical inferential techniques for copula models and then we focus on Bayesian methods illustrating both the parametric and the nonparametric approaches.

In particular we will first introduce the standard inferential approaches for d dimensional copulas and then we will concentrate on pair-copula analysis. In fact using vine construction we can deal in higher dimensional space by pair-copula estimation. Once we have chosen a particular family of copulas, the dimension of the parametric space will depend on the number of parameters of the marginals and of the selected copula model. Note also that in parametric copula estimation we have to determine whether estimate all the parameters in *one step* or we prefer to choose a property of copulas which allows us to proceed in *two steps* (also *IFM: inference functions from margins*), first estimating margins parameters and then estimating the association given the estimated marginal distributions. Both methods are consistent, while the efficiency depends on the estimation procedure in the margins [Joe, 2005].

Specifically, suppose to have the sample of size n $(y_{1_i}, \ldots, y_{d_i})$, $i = 1, \ldots, n$. We can choose a *semiparametric* approach, assuming margins to be absolutely continuous and choosing a parametric approach for the copula. In that case, the likelihood function for the copula model is

$$\mathcal{L}(\alpha) = \prod_{i=1}^{n} c_\alpha \left(\hat{F}(y_{1_i}), \ldots, \hat{F}(y_{d_i}) \right) \cdot \prod_{i=1}^{n} \prod_{j=1}^{d} f(y_{j_i}), \tag{3.14}$$

where $\hat{F}$ represents the empirical distribution function. Via likelihood maximization we get

$$\hat{\alpha} = \arg\max_{\alpha=0} \mathcal{L}$$

where $\hat{\alpha}$ is the maximum likelihood estimator for the association parameter of the copula C.

If we choose a full parametric approach, we should first assume to model each Y_{j_i} with $j = 1 \dots d$ as a probability distribution $F_j(y_{j_i}; \theta_j)$; then, we select a copula model C depending on the association parameter α. Thus, the likelihood function is

$$\mathcal{L}(\alpha, \theta) = \prod_{i=1}^{n} c_\alpha \left(F_1(y_{1_i}; \theta_1), \dots, F_d(y_{d_i}; \theta_d) \right) \cdot \prod_{i=1}^{n} \prod_{j=1}^{d} f_j \left(y_{j_i}; \theta_j \right). \tag{3.15}$$

Maximizing the likelihood function we get

$$(\hat{\alpha}, \hat{\theta}) = \arg\max \mathcal{L}(\alpha, \theta),$$

where $\hat{\alpha}$ still represents the maximum likelihood estimator for the association parameter of the copula C, while $\hat{\theta}$ represents the d dimensional vector of maximum likelihood estimators for the marginals parameters.

3.2.1 Parametric estimation: Bayesian inference

In the Bayesian framework we have more stable procedures, exploiting the benefits of integration instead of the optimization and using the entire posterior distribution. For complex marginal structures or copula functions, the likelihood can be hard to maximise directly. One solution is to use a two stage estimator [Joe, 2005]; another solution is to use to an iterative scoring algorithm to maximise the likelihood, as suggested by Song et al. [2005]. However, an attractive Bayesian alternative in this circumstance is to construct inference from the joint posterior distribution of the association and marginal parameters evaluated in a Monte Carlo manner. Furthermore, when estimating a copula model, the objective is often to construct inference on measures of dependence, quantiles or functionals of the random variable vector. Evaluation of the posterior distribution of these quantities is often straightforward using MCMC methods.

Let $(y_{1_i}, \dots, y_{d_i})$, $i = 1, \dots, n$ be our sample. We may choose a semiparametric approach, modelling marginals nonparametrically. In that case we might have either a fully Bayesian approach, specifying the prior process on the marginal parameters, or we can have a frequentist nonparametric estimation of the marginals e.g. via empirical distribution functions. In both

cases, we write down the likelihood function as

$$\mathcal{L}(\alpha) = \prod_{i=1}^{n} c_\alpha \left(\hat{F}(y_{1_i}), \ldots, \hat{F}(y_{d_i}) \right).$$

Assuming $\pi(\theta)$ as the prior distribution of the association parameter, the posterior distribution is

$$\pi(\alpha | \hat{F}(y_{1_i}), \ldots, \hat{F}(y_{d_i})) \propto \mathcal{L}(\alpha) \cdot \pi(\alpha). \tag{3.16}$$

Choosing a parametric approach for marginal modelling, we have to define the prior distribution for the marginals and the copula parameters. Let $\pi(\alpha)$ be the prior distribution for the association parameter and let $\pi(\theta) = \prod_{j=1}^{d} \pi(\theta_j)$ the joint d-dimensional prior distribution for the vector of marginal parameters. Assuming the marginals to have densities $f_1, \ldots, f_d$, the likelihood function is

$$\mathcal{L}(\alpha, \theta) = \prod_{i=1}^{n} c_\alpha \left(F_1(y_{1_i}|\theta_1), \ldots, F_d(y_{d_i}|\theta_d) \right) \cdot \prod_{i=1}^{n} \prod_{j=1}^{d} f_j \left(y_{j_i} | \theta_j \right).$$

The joint posterior distribution for the parameters of interest is

$$\pi(\alpha, \theta | F_1(y_1), \ldots, F_d(y_d)) \propto \mathcal{L}(\alpha, \theta) \cdot \pi(\alpha) \cdot \pi(\theta), \tag{3.17}$$

by marginalization we get:

$$\pi(\theta_j | F_1(y_1), \ldots, F_d(y_d)) = \int \cdots \int \pi(\alpha, \theta) d\alpha d\theta_1, \ldots, d\theta_{j-1} d\theta_{j+1}, \ldots, d\theta_d$$

and

$$\pi(\alpha | F_1(y_1), \ldots, F_d(y_d)) = \int \cdots \int \pi(\alpha, \theta) d\theta_1, \ldots, d\theta_d.$$

Posterior computation In order to draw samples from the correct posterior distribution, MCMC methods are required. The choice between Metropolis-Hastings and Gibbs sampler depends on the conjugacy of the model. Clearly, in case of parametric assumptions on the margins, in order to implement a Gibbs sampler also the marginal parameters prior distributions have to be conjugate. In any case, as already said, we can use both one-step and two-steps procedures.

3.2.2 Dealing with high dimensions: inference for vine copulas

In Section 3.1.3 we presented the vine copula construction, helpful when the dimension of the copula is $d \geq 3$. Again, we consider the $d = 3$ case, modelling the joint distribution $P(X \leq x, Y \leq y, Z \leq z)$ with marginals $u = F(x)$, $v = G(y)$ and $w = H(z)$, where Y is designated as the central variable. From (3.13) we have

$$c(F(x), G(y), H(z)) = c_1(F(x), G(y))c_2(G(y), H(z))c_3(F_{X|Y}(x|y), H_{Z|Y}(z|y)).$$

Inference on the parameters of the first two elements is quite intuitive, since they are independent from each other and there is no problem in writing down the likelihood functions

$$\mathcal{L}(\alpha_1; x, y) = \prod_{i=1}^{n} c_1(F_X(x_i), G_Y(y_i)) \text{ and } \mathcal{L}(\alpha_2; y, z) = \prod_{i=1}^{n} c_2(G_Y(y_i), H_Z(z_i)).$$

Estimates are provided depending on the chosen approach.

Regarding the last element, the likelihood function is

$$\mathcal{L}(\alpha_3; x, y, z) = \prod_{i=1}^{n} c_3(F_{X|Y}(x_i|y_i), H_{Z|Y}(z_i|y_i)). \tag{3.18}$$

From (3.2), we rewrite (3.18) as

$$\mathcal{L}(\alpha_3; x, y, z, \hat{\alpha}_1, \hat{\alpha}_2) = \prod_{i=1}^{n} c_3 \left(\frac{\partial C_1(F(x_i), G(y_i)|\hat{\alpha}_1)}{\partial G(y_i)}, \frac{\partial C_2(G(y_i), H(z_i)|\hat{\alpha}_2)}{\partial G(y_i)} \right),$$

with $\hat{\alpha}_1$ and $\hat{\alpha}_2$ having different meaning depending on the statistical approach.

In frequentist inference they represent respectively the MLE estimates for the association parameters of $c_1(F(x), G(y))$ and $c_2(G(y), H(z))$, with

$$\hat{\alpha}_3 = \arg\max_{\alpha_3 = 0} \mathcal{L}(\alpha_3; x, y, z, \hat{\alpha}_1, \hat{\alpha}_2).$$

In Bayesian inference, the posterior density

$$\pi(\alpha_3|x, y, z, \hat{\alpha}_1, \hat{\alpha}_2) \propto \pi(\alpha_3) \cdot \mathcal{L}(\alpha_3; x, y, z, \hat{\alpha}_1, \hat{\alpha}_2), \tag{3.19}$$

also depends on the values $\hat{\alpha}_1$ and $\hat{\alpha}_2$. Thus, in computing (3.19) we need to further distinguish two cases:

1. in case of one-step estimation, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are set as the simulated parameter values of $c_1(F(x), G(y))$ and $c_2(G(y), H(z))$ for the current iteration;

2. in case of two-steps estimation, they represent the expected value of the posterior density of the association parameters $\hat{\alpha}_1 = \mathbb{E}(\alpha_1)$ and $\hat{\alpha}_2 = \mathbb{E}(\alpha_2)$.

Hence, by construction, in both frequentist and Bayesian frameworks we need a hierarchical approach, with the extension to the $d > 3$ case following naturally. The Bayesian approaches for inference on vine copulas are summarized in Algorithm 8 and Algorithm 9.

Algorithm 8 One-step Bayesian inference for vine copulas

• for each iteration t
− draw $\alpha_{1_t} \sim \pi(\alpha_1|x, y)$ and $\alpha_{2_t} \sim \pi(\alpha_2|y, z)$;
∘ set $\hat{\alpha}_1 = \alpha_{1_t}$ and $\hat{\alpha}_2 = \alpha_{2_t}$
− draw $\alpha_{3_t} \sim \pi(\alpha_3|x, y, z, \hat{\alpha}_1, \hat{\alpha}_2)$;
end

Algorithm 9 Two-steps Bayesian inference for vine copulas

• for each iteration t
− draw $\alpha_{1_t} \sim \pi(\alpha_1|x, y)$ and $\alpha_{2_t} \sim \pi(\alpha_2|y, z)$;
end
∘ set $\hat{\alpha}_1 = \mathbb{E}(\alpha_1)$ and $\hat{\alpha}_2 = \mathbb{E}(\alpha_2)$
• for each iteration t
− draw $\alpha_{3_t} \sim \pi(\alpha_3|x, y, z, \hat{\alpha}_1, \hat{\alpha}_2)$;
end

3.3 Bayesian nonparametric inference for multidimensional copulas

For the two-dimensional case, there exists a wide collection of parametric copula models. However, in higher dimensions, the number of families of parametric copulas is more limited. Furthermore, assuming a parametric model for multivariate data implies the assumption of the same parametric model for each subset of paired variables. A solution to this problem may be the

pair-copula constructions, which allows us to rewrite the multivariate copula as product of pair-copulas. Alternatively, nonparametric methods may overcome the problem giving flexibility.

Wu et al. [2014] presented a method for Bayesian nonparametric inference on multidimensional copulas mixing over a skew-Normal multivariate copula. Moreover, in estimating densities of two-dimensional copulas by mixing over Gaussian copulas, Wu et al. [2015] showed that any bivariate copula density can be arbitrarily accurately approximated by an infinite mixture of Gaussian copula density functions. Dalla Valle et al. [2018] extended the methodology to the conditional bivariate copula estimation, accounting for covariates in the two dimensional case.

In this Section we focus on Bayesian nonparametric techniques for multivariate copula density estimation. We first present an alternative DPM of multivariate copulas with respect to the Wu et al. [2014] model. By choosing a Gaussian copula as kernel density of the DPM and adopting an Inverse Wishart centering measure, we exploit the conjugacy between G_0 and the mixture components. Then, we present a Bayesian nonparameteric method for multivariate copulas accounting for covariates.

3.3.1 Infinite mixtures of multivariate copulas

Let G be a probability distribution defined on the parameter space Θ. Let $(u_{1_i}, \dots, u_{d_i}) = (F_1(y_{1_i}), \dots, F_d(y_{d_i}))$ with $i = 1, ..., n$ be observations of d variables. Given the set of parameters $R \in \Theta$, we define the d-variate copula density c_G of a mixture of d-variate copulas with kernels c_R with respect to the mixing measure G as

$$c_G(u_1, \dots, u_d) = \int c_R(u_1, \dots, u_d) dG(R).$$

With a DP prior on G, we get a DPM model, which can be rewritten in a hierarchical form. We define a DP prior on G such that

$$(u_{1_i}, \dots, u_{d_i}) | R_i \overset{ind}{\sim} c_{R_i}(u_{1_i}, \dots, u_{d_i})$$

$$R_i | G \overset{iid}{\sim} G,$$

$$G \sim DP(M, G_0).$$

with M total mass parameter and G_0 is the centring measure, and R_i is the i-th parameters set defined on the parameter space Θ.

The posterior distribution $\Pi(G|u_1, \ldots, u_d)$ is a mixture of DP models, mixing with respect to latent variables R_i specific to each observation $(u_{1_i}, \ldots, u_{d_i})$ for $i = 1, \ldots, n$:

$$G|(u_1, \ldots, u_d) \sim \int DP(MG_0 + \sum_{i=1}^{n} \delta_{\theta_i}) d\Pi(R|u_1, \ldots, u_d).$$

Wu et al. [2014] presented a DPM model having as kernel density the multivariate skew-Normal copula. They showed that the Gaussian copula, that is one of the most widely used copulas because of its attractive properties and mathematical tractability, has the symmetric property which makes it difficult to deal with skewed data. However, even if estimating marginals nonparametrically, the presence of a further parameter in the copula model, in addition to the lack of conjugacy between G_0 and the kernel density, of the mixture make the computational cost increase. Therefore, if summary statistics show symmetry of the data, their approach results quite expansive in terms of computation.

3.3.2 Dirichlet Process Mixtures of multidimensional Gaussian copulas

We present a DPM of multivariate Gaussian copulas, which will overcome most of the problems related to multidimensional copula modelling. Our method relaxes distributional assumptions and allows to get good approximation of any symmetric copula family.

Let $(u_{1_i}, \ldots, u_{d_i})$, $i = 1, \ldots, n$, be i.i.d. observations defined in $\mathbf{I}^d$. Let $\Sigma \equiv R \in \Theta$ be the $d \times d$ covariance matrix of the d-dimensional Gaussian copula kernel density c_Σ.

In order to exploit the conjugacy with the multidimensional Gaussian copula model, we define the centering measure as a d-dimensional Inverse Wishart distribution

$$G_0 \equiv \mathcal{W}_d^{-1}(S_0, \eta),$$

that is in explicit form

$$G_0(\Sigma) = \frac{|S_0|^{\eta/2}}{2^{\eta d/2}\Gamma_d(\eta/2)} |\Sigma|^{-(\eta+d+1)/2} \exp\left(-\frac{1}{2} S_0 \cdot \Sigma^{-1}\right),$$

where S_0 is a positive-definite scale matrix and $\eta > d-1$ are the degrees of freedom.

Let Σ_h^{*-} denote the h-th of the k^- unique values among Σ_{-i}, which represents the set of covariance matrices Σ without the i-th element Σ_i. Also, let $(\mathbf{u_1}, \ldots, \mathbf{u_d})_h^{*-} = (\mathbf{u_1}, \ldots, \mathbf{u_d})_h^* \setminus (u_{1_i}, \ldots, u_{d_i})$ and $\Psi_h = (i : \Sigma_i = \Sigma_h^*)$, so that if $\psi_i = h$, the i-th observation belongs to the h-th cluster and n_h represents the number of observations lying inside the cluster h. Thus, we write down the likelihood function for the cluster h as

$$\mathcal{L}(\Sigma_h^*) = \prod_{i=1}^{n_h} \left(|\Sigma|^{1/2} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_{1_i}) \\ \vdots \\ \Phi^{-1}(u_{d_i}) \end{pmatrix}^T \cdot \left(\Sigma_h^{-1} - I\right) \cdot \begin{pmatrix} \Phi^{-1}(u_{1_i}) \\ \vdots \\ \Phi^{-1}(u_{d_i}) \end{pmatrix} \right) \right),$$

$$\mathcal{L}(\Sigma_h^*) = |\Sigma|^{n/2} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(\mathbf{u_1}_h^*) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}_h^*) \end{pmatrix}^T \cdot \left(\Sigma_h^{-1} - I\right) \cdot \begin{pmatrix} \Phi^{-1}(\mathbf{u_1}_h^*) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}_h^*) \end{pmatrix} \right).$$

We derive the posterior density of the cluster correlation matrices $\pi(\Sigma_h^*|\boldsymbol{\psi}, \mathbf{u_1}, \ldots, \mathbf{u_d})$ and the posterior conditional density of the clusters $\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{u_1}, \ldots, \mathbf{u_d})$.

1. We compute the posterior density for Σ_h^* as

$$\pi(\Sigma_h^*|\boldsymbol{\psi}, \mathbf{u_1}, \ldots, \mathbf{u_d}) = G_0(\Sigma_h^*) \cdot \mathcal{L}(\Sigma_h^*), \tag{3.20}$$

rewriting in explicit form

$$\pi(\Sigma_h^*|\boldsymbol{\psi}, \mathbf{u_1}, \ldots, \mathbf{u_d}) = \frac{|S_0|^{\eta/2}}{2^{\eta d/2}\Gamma_d(\eta/2)} |\Sigma_h|^{-(\eta+d+1)/2} \exp\left(-\frac{1}{2} S_0 \cdot \Sigma_h^{-1} \right) \times$$

$$|\Sigma_h|^{n/2} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(\mathbf{u_1}_h^*) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}_h^*) \end{pmatrix}^T \cdot \left(\Sigma_h^{-1} - I\right) \cdot \begin{pmatrix} \Phi^{-1}(\mathbf{u_1}_h^*) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}_h^*) \end{pmatrix} \right),$$

where $\Gamma_d(\cdot)$ is the d-dimensional multivariate gamma function. We note that $\pi(\Sigma_h^*|\boldsymbol{\psi}, \mathbf{u_1}, \ldots, \mathbf{u_d})$ is still an Inverse Wishart

$$\pi(\Sigma_h^*|\boldsymbol{\psi}, \mathbf{u_1}, \ldots, \mathbf{u_d}) \equiv \mathcal{W}^{-1} \left(S_0 + \begin{pmatrix} \Phi^{-1}(\mathbf{u_1}_h^*) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}_h^*) \end{pmatrix}^T \begin{pmatrix} \Phi^{-1}(\mathbf{u_1}_h^*) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}_h^*) \end{pmatrix}, \eta + n_h \right) \tag{3.21}$$

where:

- S_0 is the prior scale matrix and η represents the prior on the degrees of freedom;
- $(\mathbf{u_1}^*_h, \ldots, \mathbf{u_d}^*_h)$ is the set of observations lying inside the cluster h;
- $\begin{pmatrix} \Phi^{-1}(\mathbf{u_1}^*_h) \\ \vdots \\ \Phi^{-1}(\mathbf{u_d}^*_h) \end{pmatrix}$ is a $n_h \times d$ matrix.

2. We now define the posterior probabilities $\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{u}, \mathbf{v})$ for $h = 1 \ldots k^-$ and $h = k^- + 1$ which denotes the creation of a new cluster. From (2.15) it follows that the probability of the i-th element of belong to the h-th cluster is

$$\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{u_1}, \ldots, \mathbf{u_d}) \propto \begin{cases} n_h^- f(u_{1_i}, \ldots, u_{d_i}|(\mathbf{u_1}, \ldots, \mathbf{u_d})^{*-}_h) & \text{for } h = 1, \ldots, k^- \\ Mc(u_{1_i}, \ldots, u_{d_i}) & \text{for } h = k^- + 1 \end{cases} \tag{3.22}$$

where n_h^- is the number of elements in the h-th cluster with exclusion of the i-th observation, M represents the precision parameter of the DP and $c(u_{1_i}, \ldots, u_{d_i}) \equiv \int c_{R^*_h}(u_{1_i}, \ldots, u_{d_i}) G_0(d\Sigma^*_h)$, while

$$f(u_{1_i}, \ldots, u_{d_i}|(\mathbf{u_1}, \ldots, \mathbf{u_d})^{*-}_h) = \int c_{\Sigma^{*-}_h}(u_{1_i}, \ldots, u_{d_i}) d\pi(\Sigma^{*-}_h|\mathbf{u_1}, \ldots, \mathbf{u_d}^{*-}_h),$$

where $\pi(\Sigma^{*-}_h|(\mathbf{u_1}, \ldots, \mathbf{u_d})^{*-}_h)$ is the posterior density of Σ in the h-th cluster excluding the i-th observation. In particular, in order to simplify the notation we define

$$T(u|\Sigma_h) = \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^T \cdot \left(\Sigma_h^{-1} - I\right) \cdot \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}$$

and

$$T(u) = \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^T \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}.$$

Then, we calculate

$$f(u_{1_i},\dots,u_{d_i}|(\mathbf{u_1},\dots,\mathbf{u_d})_h^{*-}) = \int |\Sigma_h|^{-1/2} e^{\left(-\frac{1}{2}T(u_i|\Sigma_h)\right)} \times$$

$$\frac{\left|S_0+T(\mathbf{u}_h^{*-})\right|^{(\eta+n_h^-)/2}}{2^{(\eta+n_h^-)d/2}\cdot\Gamma_d\left((\eta+n_h^-)/2\right)} |\Sigma_h|^{-(\eta+n_h^-+d-1)/2} e^{\left(-\frac{1}{2}T(\mathbf{u}_h^{*-}|\Sigma_h)\right)} d\Sigma_h =$$

$$= \frac{\left|S_0+T(\mathbf{u}_h^{*-})\right|^{(\eta+n_h^-)/2}}{2^{(\eta+n_h^-)d/2}\cdot\Gamma_d\left((\eta+n_h^-)/2\right)} \int |\Sigma_h|^{-(\eta+n_h+d-1)/2} e^{\left(-\frac{1}{2}T(\mathbf{u}_h^{*}|\Sigma_h)\right)} d\Sigma_h =$$

$$= \frac{\left|S_0+T(\mathbf{u}_h^{*-})\right|^{(\eta+n_h^-)/2}}{2^{(\eta+n_h^-)d/2}\cdot\Gamma_d\left((\eta+n_h^-)/2\right)} \cdot \frac{2^{(\eta+n_h)d/2}\cdot\Gamma_d\left((\eta+n_h)/2\right)}{|S_0+T(\mathbf{u}_h)|^{(\eta+n_h)/2}},$$

where $\Gamma_d(\cdot)$ represents the d-dimensional multivariate gamma function, $\mathbf{u}_h^{*-}$ represents the observations in the h-th cluster with the exclusion of the i-th observation and $\mathbf{u}_h^{*}$ is got including the i-th element in the set $\mathbf{u}_h^{*-}$. The number of elements in $\mathbf{u}_h^{*-}$ and $\mathbf{u}_h^{*}$ are respectively n_h^- and n_h.

Thus, we may define the Gibbs sampler for DPM of multidimensional Gaussian copulas as in the following scheme.

Algorithm 10 Gibbs sampler for DPM of multidimensional Gaussian copulas

• **Clustering**:
− for $i = 1\dots,N$ draw $\psi_i \sim \pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{u_1},\dots,\mathbf{u_d})$ from (3.22);
• **Cluster parameters**:
− for $h = 1,\dots,k$ draw $\Sigma_h^* \sim \pi(\Sigma_h^*|\boldsymbol{\psi}, \mathbf{u_1},\dots,\mathbf{u_d})$ from (3.21).

3.3.3 Bayesian nonparametric conditional multidimensional copulas

We finally present a method which allows to make inference on multidimensional copulas accounting for covariates. Suppose to have $(u_{1_i},\dots,u_{d_i}) = (F(y_{1_i}),\dots,F(y_{d_i}))$ with $i = 1,...,n$. We also observe a $n \times v$ matrix of covariates X. Dalla Valle et al. [2018] provided a Bayesian nonparametric method for bivariate conditional copulas. We propose a Bayesian nonparametric method for conditional multidimensional copulas, which allows us to detect the impact of each covariate X_j $j = 1,\dots,v$ on the dependence structure of the multivariate copula, basing on the vine construction.

Let G be a probability distribution on defined on the parameter space Θ. Let $(u_{1_i}, \ldots, u_{d_i}) = (F(y_{1_i}), \ldots, F(y_{d_i}))$ with $i = 1, ..., n$ be observations of d variables and let X be the $n \times v$ matrix of observed covariates. We define the density c_G as a mixture of d-variate conditional copulas with kernels $c_{\boldsymbol{\theta}(X)}$ with respect to the mixing measure G, that is

$$c_G(F_1(y_1), \ldots, F_d(y_d)) = \int c_{\boldsymbol{\theta}(X)}(F_1(y_1), \ldots, F_d(y_d)) dG(\boldsymbol{\theta}(X)).$$

Remark that, since we need to specify a vine structure, we consider both c_G and $c_{\boldsymbol{\theta}(X)}$ as products of $p = d(d-1)/2$ pair-copulas. Let

$$\boldsymbol{\theta}(X) = g(X; \boldsymbol{\beta}), \tag{3.23}$$

be a p-dimensional vector of functions. Then, defining a DP prior on G we rewrite

$$(F_1(y_{1_i}), \ldots, F_d(y_{d_i}))|\boldsymbol{\theta}(X)_i \overset{ind}{\sim} c_{\boldsymbol{\theta}(X)_i}(F_1(y_{1_i}), \ldots, F_d(y_{d_i}))$$

$$\boldsymbol{\theta}(X)_i|G \overset{iid}{\sim} G,$$

$$G \sim DP(M, G_0).$$

with M total mass parameter and G_0 is the centring measure. Again, $\boldsymbol{\theta}(X)_i = (\theta(X)_{i_1}, \ldots, \theta(X)_{i_p})$ has dimension $p \times 1$, with $p = \frac{d \cdot (d-1)}{2}$, since by pair-copula construction we rewrite $c(F_1(y_1), \ldots, F_d(y_d))$ as product of $d(d-1)/2$ bivariate copulas and each θ_{i_j} is a v-variate function of the covariates

$$\theta_{i_j} = g(X, \beta_{i_j}) \tag{3.24}$$

with β_{i_j} vector of dimension v.

The posterior distribution $\Pi(G|F_1(y_1), \ldots, F_d(y_d))$ is a mixture of DP models, mixing with respect to latent variables $\boldsymbol{\theta}_i(X)$ specific to each observation $(F_1(y_{1_i}), \ldots, F_d(y_{d_i}))$ for $i = 1, \ldots, n$:

$$G|(F_1(y_1), \ldots, F_d(y_d), X) \sim \int DP(MG_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i(X)}) d\Pi(\boldsymbol{\theta}(X)|F_1(y_1), \ldots, F_d(y_d)).$$

Therefore, by vine copulas construction we rewrite d-variate copula densities as the product of $d(d-1)/2$ bivariate copulas, reducing the kernel choice to the biavriate copulas class of densities whatever d is.

As kernel density of our DPM model, we need a bivariate copula that is able to capture any kind of dependence and may approximate each copula family. The reparametrization of $\boldsymbol{\theta}$ as function fo X makes impossible to exploit the conjugacy between G_0 and the kernel density. Hence, we should choose only basing on the properties of the kernel. Wu et al. [2015] showed that that bivariate density functions on the real plain can be arbitrarily well approximated by a mixture of a countably infinite number of bivariate normal distributions.

3.3.4 Dirichlet Process Mixtures of conditional Gaussian vine copulas

We now present a DPM of conditional vine copulas which allows to estimate multivariate copula densities in presence of covariates. Let $(F(y_{1_i}), \dots, F(y_{d_i}))$ be $i = 1, \dots, n$ observations defined in the hypercube $\mathbf{I}^d$. Let X be a $n \times p$ matrix of covariates. Assuming as kernel density of our mixture the product of Gaussian copulas $c_{\rho_i} = (u, v)$, with $i = 1, \dots, p$ and $p = d(d-1)/2$, each one having association parameter $\rho \in (-1, 1)$ assumed to be function of v covariates X. Thus, considering

$$\boldsymbol{\rho}(X) = g(X; \boldsymbol{\beta}), \tag{3.25}$$

where $\boldsymbol{\rho}$ is a $p \times 1$ vector of functions each one depending on v-dimensional vectors β, we need to define a $p \times v$ dimensional centering measure. In particular, we chose G_0 to be multivariate Normal

$$G_0 \equiv \mathbf{N}_{p \times v}(\mu, \Sigma). \tag{3.26}$$

We present the details of the case with $d = 3$. Thus, let $(F(y_{1_i}), F(y_{2_i}), F(y_{3_i}))$, $i = 1, \dots, n$, be i.i.d. observations defined in $\mathbf{I}^3$, with X $n \times p$ matrix of covariates. Let we choose y_2 as central variable, so that:

$$c(F_1(y_1), F_2(y_2), F_3(y_3)) = c_1(F_1(y_1), F_2(y_2)) \cdot c_2(F_2(y_2), F_2(y_3)) \cdot c_3(F_1(y_1|y_2), F_3(y_3|y_2)).$$

Consider ρ for each pair-copula as function of unknown parameters β and covariates X

$$\rho = g(X; \beta),$$

where g is the link function and β has dimension depending on the number of covariates v. Let β_h^{*-} denote the h-th of the k^- unique values among $\boldsymbol{\beta}_{-i}$, which represents the vector $\boldsymbol{\beta}$ without the i-th element β_i. Let $\Psi_h = (i : \beta_i = \beta_h^*)$, so that if $\psi_i = h$, the i-th observation belongs to the h-th cluster and n_h represents the number of observations lying inside the cluster h.

1. We first define $\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \beta_h^*, X, F_1(y_1), F_2(y_2), F_3(y_3))$ for $h = 1 \dots k^-$ and $h = k^- + 1$. From (2.16) it follows that the probability of the i-th element of belong to the h-th cluster is

- for $h = 1, \dots, k^-$

$$n_h^- c_{\beta^*_{h,1}}(F_1(y_1), F_2(y_2)) \cdot c_{\beta^*_{h,2}}(F_2(y_2), F_2(y_3)) \cdot c_{\beta^*_{h,3}}(F_1(y_1|y_2), F_3(y_3|y_2)),$$

- for $h = k^- + 1$

$$\frac{M}{k^- + 1} c_{\beta^*_{k^-+1,1}}(F_1(y_1), F_2(y_2)) \cdot c_{\beta^*_{k^-+1,2}}(F_2(y_2), F_2(y_3)) \cdot c_{\beta^*_{k^-+1,2}}(F_1(y_1|y_2), F_3(y_3|y_2)),$$

where n_h^- is the number of elements in the h-th cluster with exclusion of the i-th observation, M represents the precision parameter of the DP and c is the Gaussian bivariate copula density of parameters β_h^*.

2. We compute the conditional distribution functions $F_1(y_1|y_2)$ as

$$F_1(y_1|y_2) = \frac{\partial c_{\beta_h^*}(F_1(y^*_{1_h}), F_2(y^*_{2_h}))}{\partial F_2(y^*_{2_h})} \quad \text{for } h = 1, \dots, k, \tag{3.27}$$

where in deriving for each observation we consider the belonging cluster h. We do the same in order to get $F_3(y_3|y_2)$.

3. Finally, denoting by $s_i = \Phi^{-1}(F(y_{1_i}))$ and $t_i = \Phi^{-1}(F(y_{2_i}))$, the likelihood function for the cluster h is

$$\mathcal{L}(\beta_h^*) = \prod_{i=1}^{n_h} \frac{1}{\sqrt{(1 - g(\beta_h; X)^2)}} \exp\left(- \frac{g(X; \beta_h)^2 s_i^2 + g(X; \beta_h)^2 t_i^2 - 2 g(X; \beta_h) s_i t_i}{2(1 - g(X; \beta_h)^2)} \right).$$

We compute the posterior density for β_h^* as

$$\pi(\beta_h^*|\boldsymbol{\psi}, X, F_1(y_1), F_2(y_2)) = G_0(\beta_h^*) \cdot \mathcal{L}(\beta_h^*), \tag{3.28}$$

which is a complex v-dimensional density, approximated with a Metropolis Hastings step with v-variate Normal proposal. We repeat the point with $s_i = \Phi^{-1}(F(y_{2_i}))$ and $t_i = \Phi^{-1}(F(y_{3_i}))$ and with $s_i = \Phi^{-1}(F(y_{1_i}|y_{2_i}))$ and $t_i = \Phi^{-1}(F(y_{2_i}|y_{2_i}))$.

The algorithm is summarized below.

Algorithm 11 No-gaps sampler for DPM of conditional vine copulas

• **Clustering**:
– for $i = 1 \dots, n$ draw $\psi_i \sim \pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \beta_h^*, X, F_1(y_1), F_2(y_2), F_3(y_3))$;
• **Cluster parameters**:
◦ compute the conditioned distribution function from (3.25);
◦ for each pair-copula:
– for $h = 1, \dots, k$ draw $\beta_h^* \sim \pi(\beta_h^*|\boldsymbol{\psi}, X, F_j(y_j), F_\ell(y_\ell))$;
– for $h = k+1, \dots, n$ draw $\beta_h^* \sim G_0$.

3.4 Simulation study

In this section we provide a simulation study for both the DPM of conditional multivariate copulas (*Conditional model*) and the DPM of multivariate Gaussian copulas (*Unconditional model*). We show the efficiency of the methods in clustering and density estimation. We also provide a comparison of the models. In each of the presented cases we simulate data from copula functions, treating the marginals as given.

For the implementation of the DPM of conditional multivariate vine copulas, in the choice of the function $\rho(X)$ for each pair-copula, we follow the approach proposed by Abegaz et al. [2012], which models the dependence of the parameter of interest, with respect to the covariate, through a calibration function $\lambda(X;\beta)$. Note that in many copula families the parameter space is restricted. In contrast, the calibration function $\lambda(X;\beta)$ can assume any value on the real line. Since with a bivariate Gaussian copula kernel the parameter space is restricted to the interval $(-1, 1)$, we need a transformation which can link the calibration function $\lambda(X;\beta)$ to $\rho(X)$. We chose two links:

- Following Dalla Valle et al. [2018], we adopt:

$$\rho(X) = \frac{2}{|\lambda(X;\beta)| + 1} - 1; \tag{3.29}$$

- The second link is the *Inverse Fisher Transform*:

$$\rho(X) = \frac{e^{\lambda(X;\beta)} - 1}{e^{\lambda(X;\beta)} + 1}. \tag{3.30}$$

In both cases λ is set to be

$$\lambda(X;\beta) = \beta_0 + \beta_1 X.$$

Table 3.1: Results for 3-dimensional copula with $n = 200$ observations: posterior mean and standard deviation for the model parameters.

ψ		ρ_{12}	ρ_{23}	ρ_{13}
1	$E(\cdot\|y)$	0.5041	0.5086	0.4681
1	$SD(\cdot\|y)$	0.0834	0.0855	0.0889
2	$E(\cdot\|y)$	-0.5045	-0.4813	-0.4062
2	$SD(\cdot\|y)$	0.0782	0.0818	0.0864

For the Unconditional model, we chose as concentration parameter of the DPM $M = 1$ and as centering measure an Inverse Wishart distribution. Instead, for the Conditional model we set as concentration parameter $M = 1$ and chose as centering measure a multivariate Normal distribution as defined in (3.26).

3.4.1 Clustering

We first show how the proposed models works for clustering. First, for the implementation of the DPM of multivariate Gaussian copulas we have simulated 200 observation from an equally weighted mixture of two Gaussian copulas with covariance matrices

$$\Sigma_1 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{pmatrix}.$$

We ran 15000 MCMC iterations with a burnin of 1000 iterations. A summary on the results is presented in Table 3.1, while traceplots of the estimated clusters parameters and the distribution of the labels across the MCMC iterations are shown in Figure 3.1. The model shows a good behavior, the posterior distributions of the parameters are centered around the true values even with a small sample; there is also a good estimation of the size and the number of clusters.

As second example, we show how both models work on the same data. We have simulated data from a finite mixture of $d = 3$ conditional vine copulas, with one covariate X simulated from a $N(1, 0.2)$. The true value of the parameters are presented in Table 3.2. For the Conditional model we ran 25000 iterations with a burnin of 5000, adopting as link function the Inverse Fisher Transform (3.30). Instead, for the Unconditional model we ran 15000 MCMC iterations, with a burnin of 1000 iterations. From Figure 3.2 we may

Figure 3.1: Labels distribution across the MCMC iterations and traceplots of the off-diagonal elements of the posterior correlation matrix for the estimated clusters.

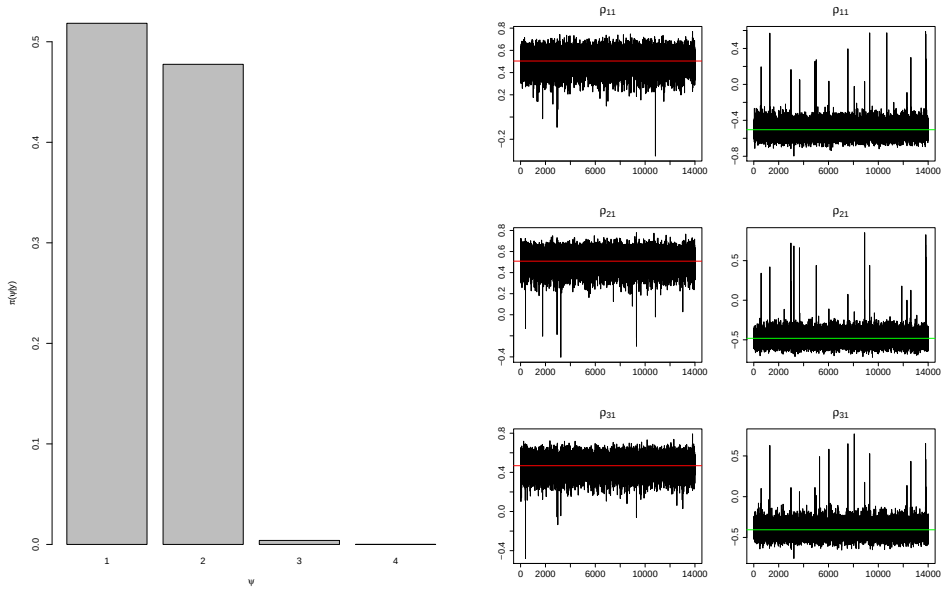


Table 3.2: True values of the pair-copula parameters. In simulating vine copula data, we chose as central variable Y_2.

ψ	β_0	β_1
1	0.5	1.5
2	-0.5	-1.5

note that for the Unconditional model the frequencies of the labels are close to the real values, while the Conditional model seems to be less consistent, showing also the presence of a third cluster. From Figure 3.3 we observe that the posterior densities of the Unconditional model parameters for the first two pair associations have a low standard deviation, proving the very good performance of this model for clustering. In Figure 3.4 we observe that the posterior densities of the Conditional model parameters are concentrated around the true values (red line for the first cluster and green line for the second cluster). However, especially in the second cluster, some parameters show an high standard deviation. Therefore, we may say that for clustering the Unconditional model seems to perform better than the Conditional one.

Figure 3.2: Labels distribution across the MCMC iterations for the Conditional (left) and the Undonditional model (right).

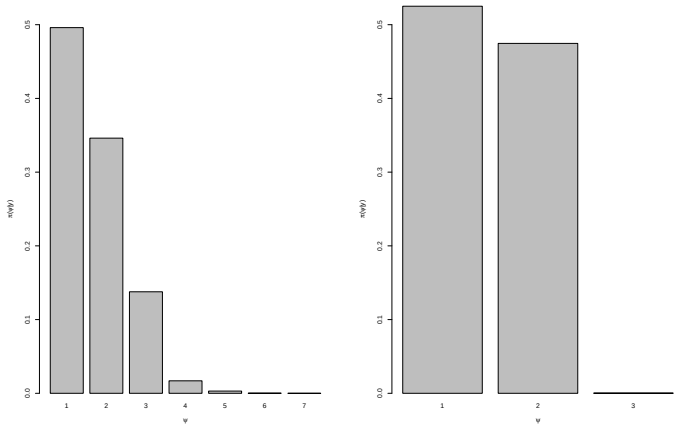


Figure 3.3: Unconditional model: posterior densities of the parameters for the first two clusters.

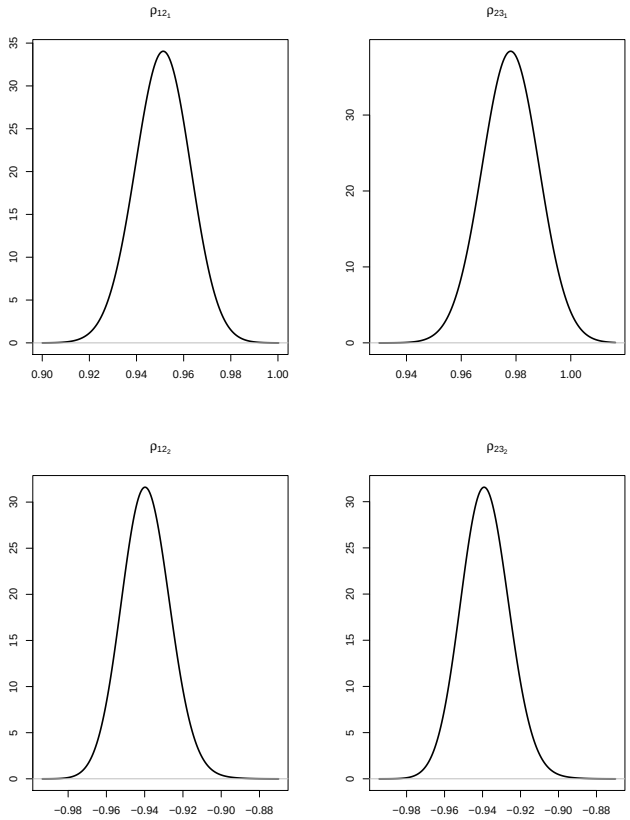


Figure 3.4: Conditional model: posterior densities of the parameters for the first two clusters.

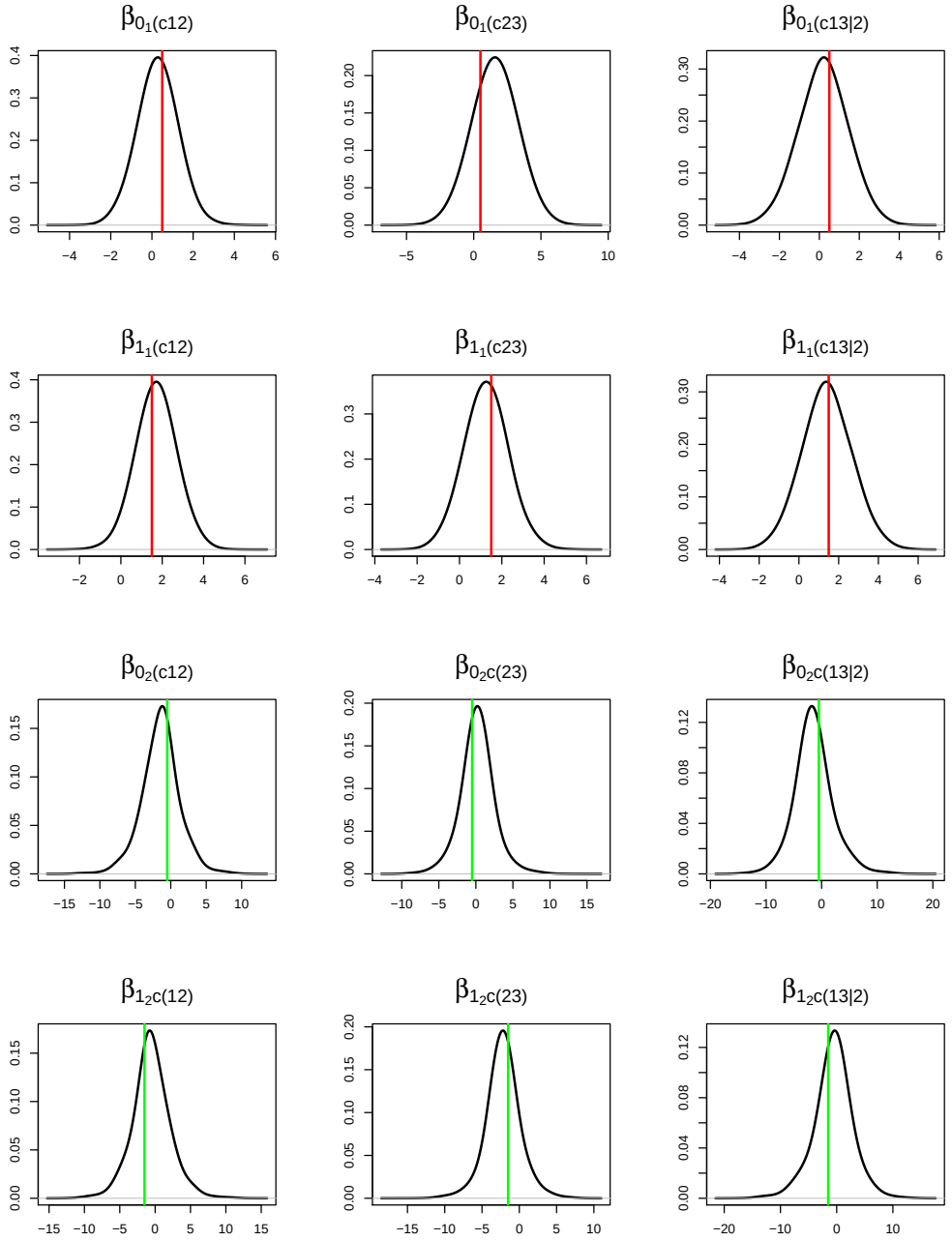


Table 3.3: Results for 3-dimensional vine copulas mixture with $n = 500$ observations: posterior mean and standard deviation for the models (Conditional and Unconditional) parameters.

ψ		ρ_{12}	ρ_{23}	ρ_{13}
1	$E(\cdot\|y)$	0.95	0.95	0.98
1	$SD(\cdot\|y)$	0.006	0.003	0.006
2	$E(\cdot\|y)$	-0.94	-0.94	0.82
2	$SD(\cdot\|y)$	0.008	0.008	0.003

ψ		$\beta_{0(c12)}$	$\beta_{1(c12)}$	$\beta_{0(c23)}$	$\beta_{1(c23)}$	$\beta_{0(c13\|2)}$	$\beta_{1(c13\|2)}$
1	$E(\cdot\|y)$	0.39	1.89	1.08	1.35	0.42	1.41
1	$SD(\cdot\|y)$	0.89	0.87	1.12	0.96	0.91	0.81
2	$E(\cdot\|y)$	-0.72	-1.03	-0.34	-1.71	-0.85	-1.05
2	$SD(\cdot\|y)$	2.49	2.62	1.37	1.34	2.07	2.02

3.4.2 Density estimation

We now show how both the proposed models works for density estimation. Focusing on the Unconditional model, we have simulated data from a mixtures of Frank, Gumbel and Gaussian copulas. We ran 15000 MCMC iterations with a burnin of 1000 iterations. Results are shown in Figure 3.5, Figure 3.6 and Figure 3.7.. In each of these cases the algorithm provides good density estimates.

In order to test how the Conditional model works for density estimation, we have simulated data from different scenarios, choosing as central variable in the vine construction Y_2. In Figure 3.8 we show results for data simulated from a multivariate (d=3) Gaussian copula with covariate X generated from a $N(1, 0.2)$. In the second case (Figure 3.9) we have simulated data from a mixture of multivariate (d=3) Frank copulas with covariate X generated from a $N(0, 0.2)$. In Figure 3.10 we show results for simulated data simulated from a mixture of multivariate (d=3) Gaussian copulas with covariate X generated from a $N(0.5, 0.2)$. Comparing the observed data with the density estimates, we may note that the model gives good estimates of any pair-copula density.

Figure 3.5: Scatterplot of the mixture of Frank copulas data and density estimation via Unconditional model with sample size n=200.

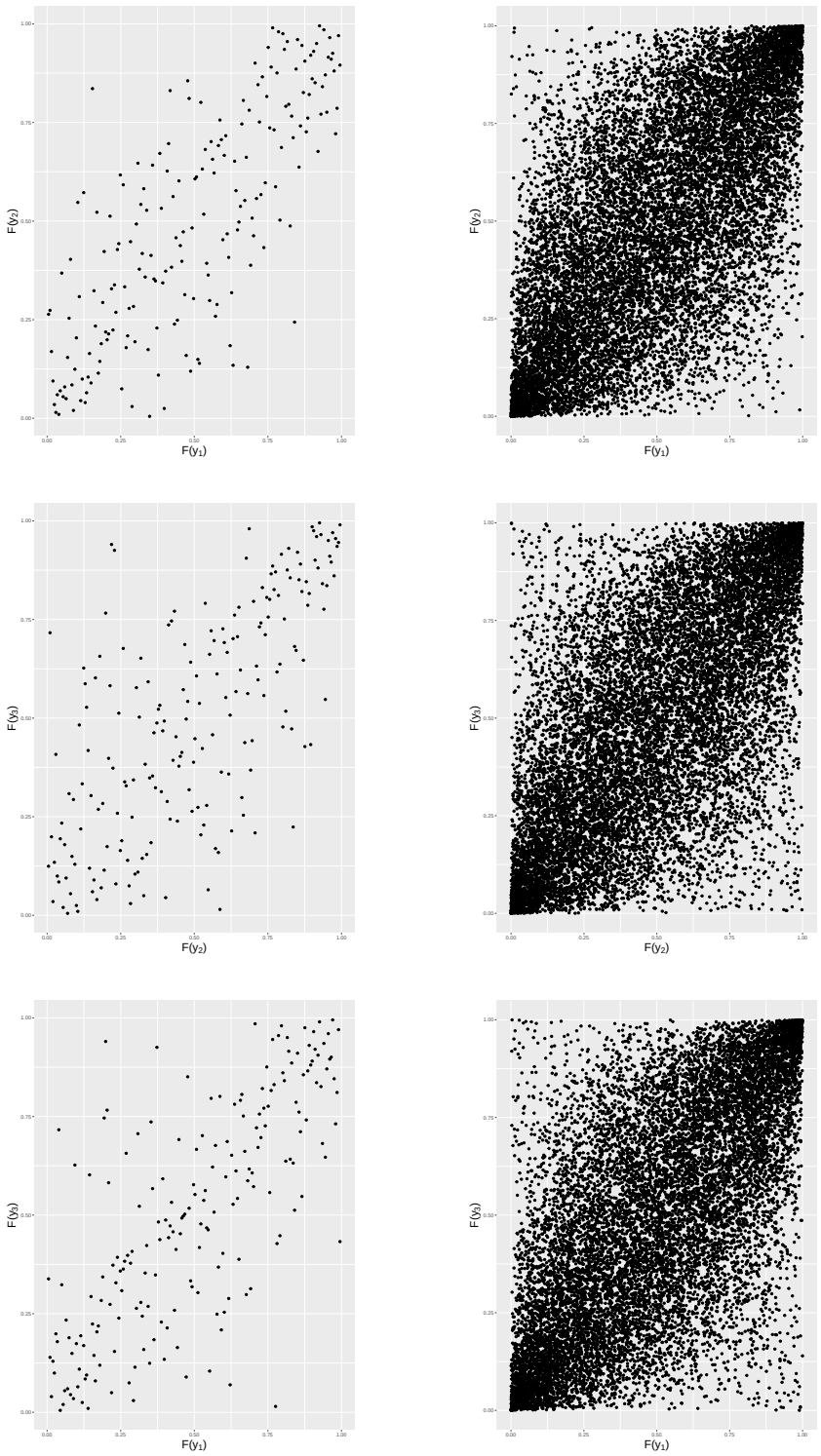


Figure 3.6: Scatterplot of the mixture of Gumbel copulas data and density estimation via Unconditional model with sample size n=200.

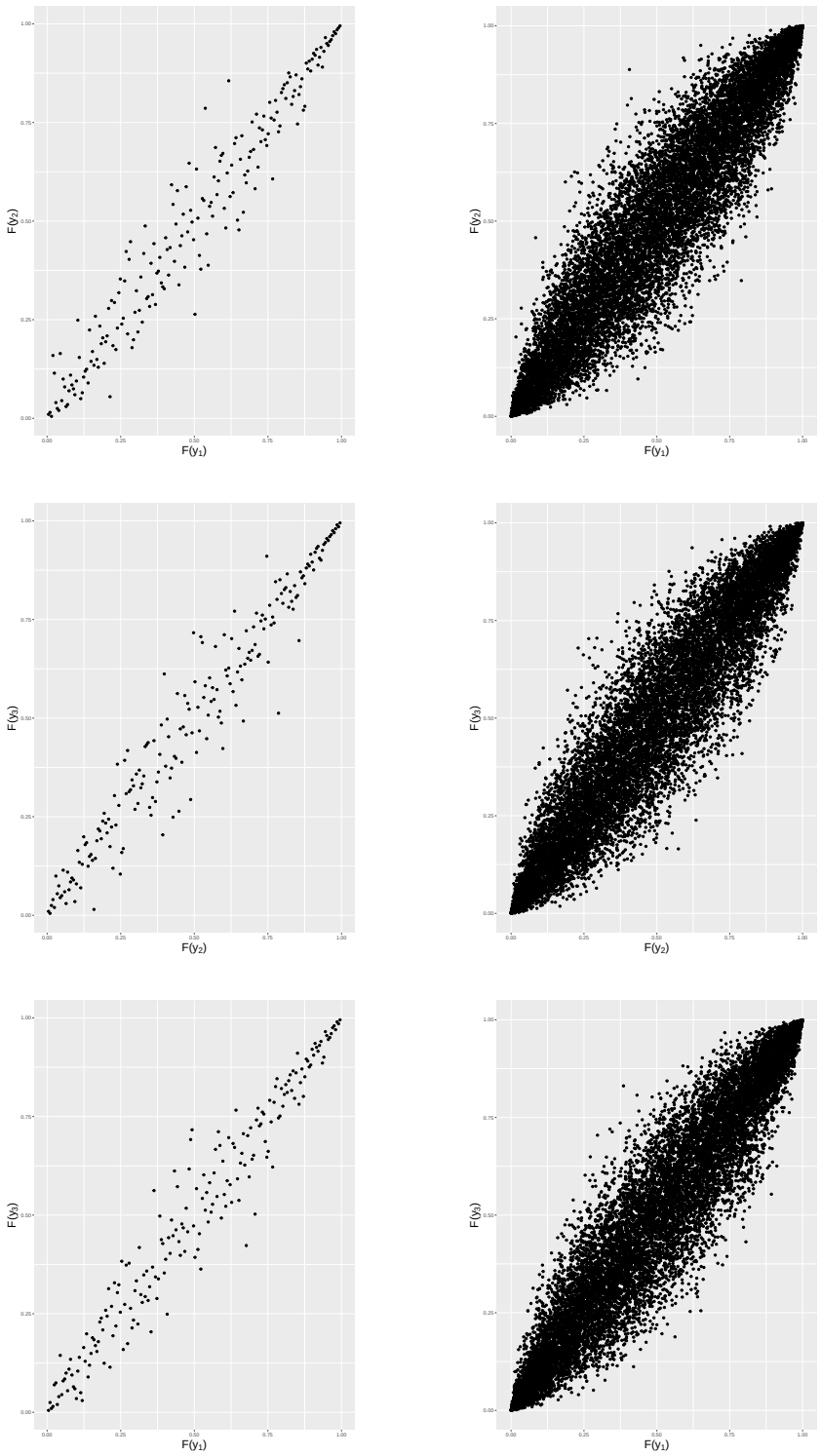


Figure 3.7: Scatterplot of the mixture of Gaussian copulas data and density estimation via Unconditional model with sample size n=200.

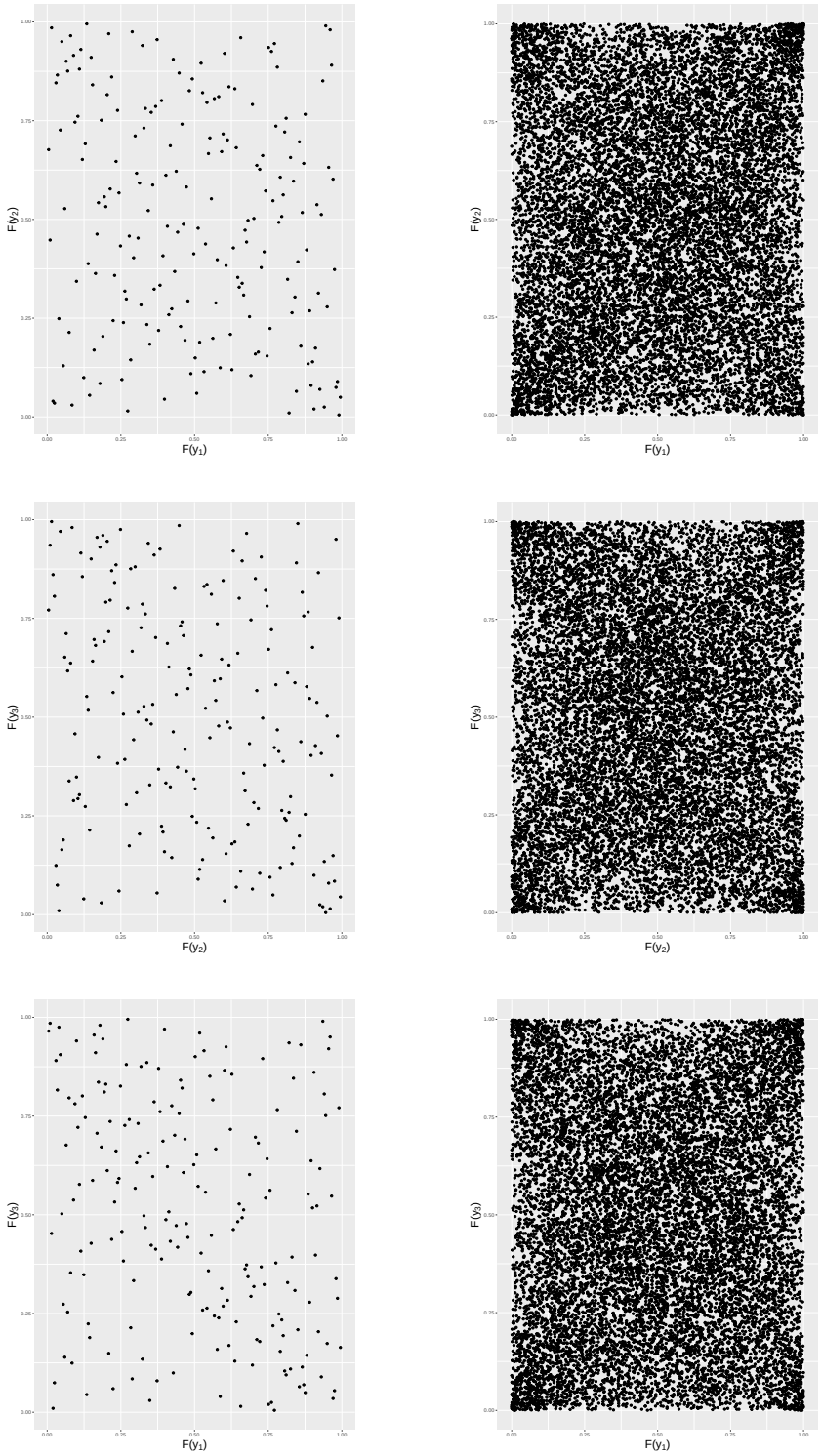
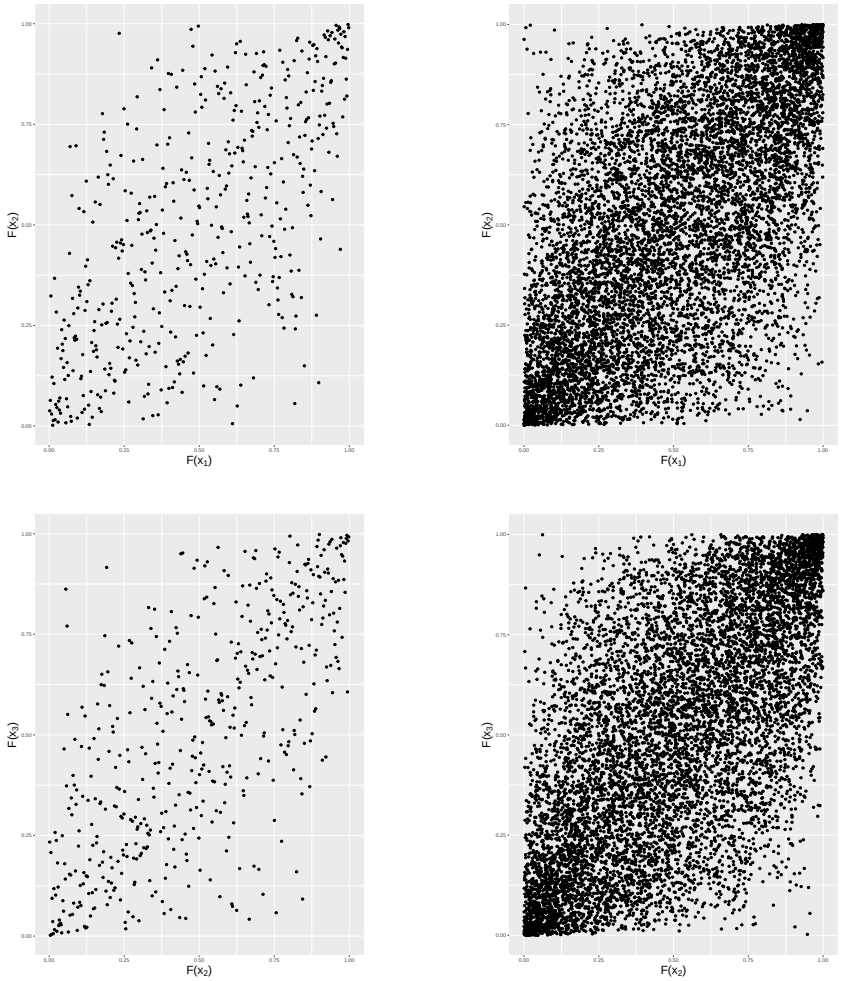


Figure 3.8: Scatterplot of the Gaussian copula data and density estimation via Conditional model with sample size n=500 and link function (3.29).



Lastly, we show how both the models estimate the pair-copula densities for the observed data, extending the second example presented in 3.4.1. From Figure 3.11 we observe that both models have a good performance in terms of density estimation, with the Conditional model performing slightly better.

Figure 3.9: Scatterplots of the mixture of Frank copulas data and density estimation via Conditional model with sample size n=500 and link function (3.29).

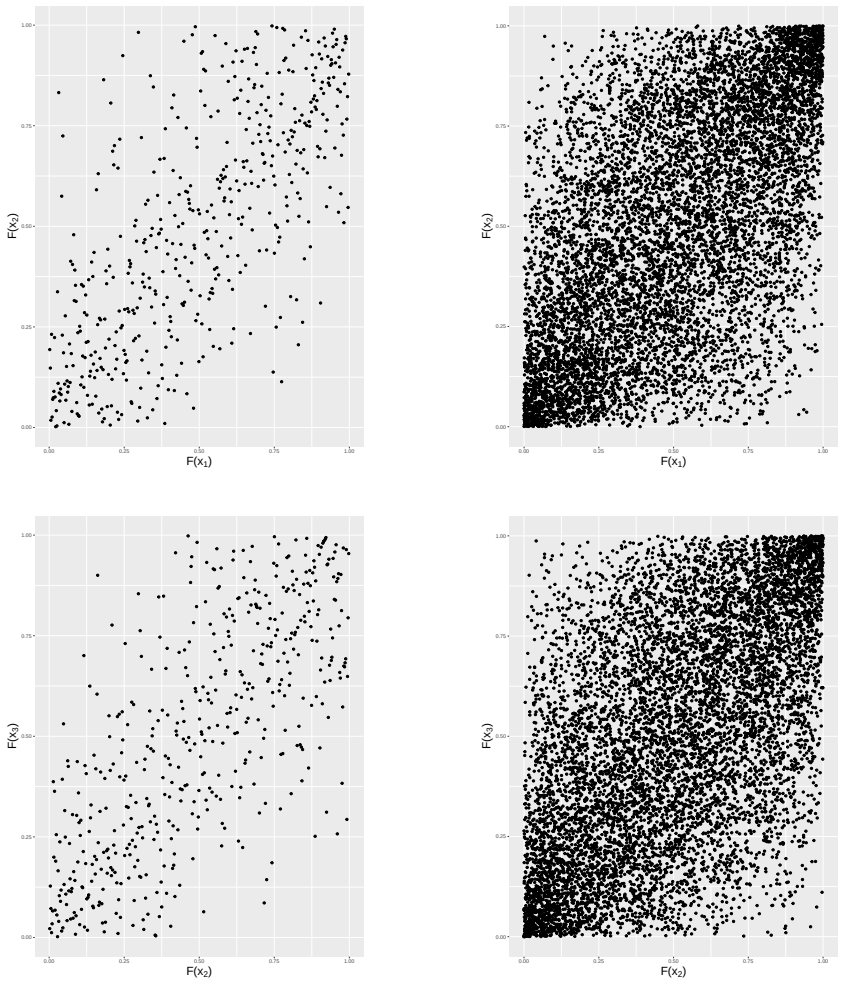


Figure 3.10: Scatterplots of the mixture of Gaussian copulas data and density estimation via Conditional model with sample size n=500 and link function (3.30).

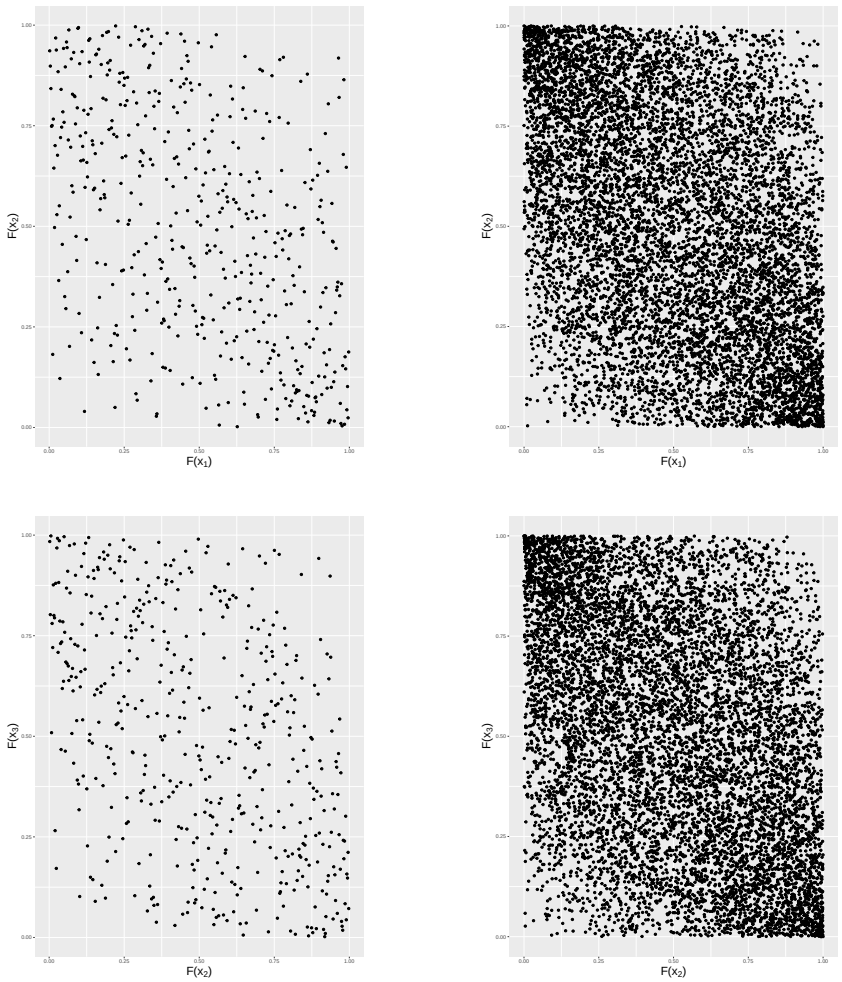
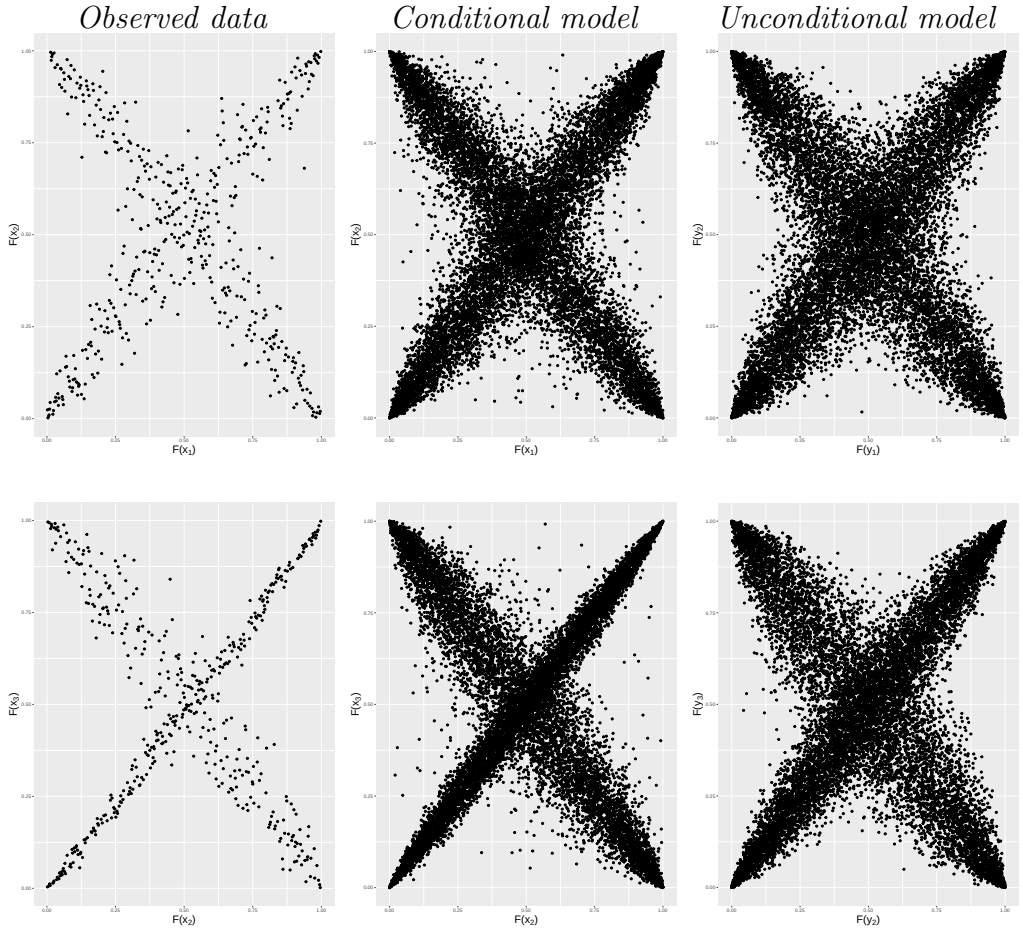


Figure 3.11: Scatterplots of the mixture of Gaussian copulas data and density estimation via both the models with sample size n=500 and link function (3.30).



Bibliography

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

Fentaw Abegaz, Irène Gijbels, and Noël Veraverbeke. Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis*, 110:43–73, 2012.

Hilary Aralis and Ron Brookmeyer. A stochastic estimation procedure for intermittently-observed semi-markov multistate models with back transitions. *Statistical Methods in Medical Research*, 28(3):770–787, 2019.

C Armero, S Cabras, ME Castellanos, S Perra, A Quirós, MJ Oruezábal, and J Sánchez-Rubio. Bayesian analysis of a disability model for lung cancer survival. *Statistical Methods in Medical Research*, 25:336–351, 2012.

M Concepcion Ausin and Hedibert F Lopes. Time-varying joint distribution through copulas. *Computational Statistics & Data Analysis*, 54(11):2383–2399, 2010.

Tim Bedford and Roger M Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1-4):245–268, 2001.

David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

Mogens Bladt and Michael Sørensen. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B*, 67(3):395–410, 2005.

Mogens Bladt and Michael Sørensen. Efficient estimation of transition rates between credit ratings from observations at discrete time points. *Quantitative Finance*, 9(2):147–160, 2009.

Christopher A Bush and Steven N MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

Hervé Cardot, Guillaume Lecuelle, Pascal Schlich, and Michel Visalli. Estimating finite mixtures of semi-markov chains: an application to the segmentation of temporal sensory data. *arXiv preprint arXiv:1806.04420*, 2018.

Luciana Dalla Valle, Fabrizio Leisen, and Luca Rossini. Bayesian nonparametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):523–548, 2018.

Anthony Christopher Davison. *Statistical Models*. Cambridge University Press, 2003.

Bianca L De Stavola. Testing departures from time homogeneity in multistate Markov processes. *Journal of the Royal Statistical Society: Series C*, 37: 242–250, 1988.

Ralph dos Santos Silva and Hedibert Freitas Lopes. Copula, marginal distributions and model selection: a bayesian note. *Statistics and Computing*, 18(3):313–320, 2008.

Michael D Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430): 577–588, 1995.

Michael David Escobar. Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. 1990.

Paul Fearnhead and Chris Sherlock. An exact Gibbs sampler for the Markov-modulated Poisson process. *Journal of the Royal Statistical Society: Series B*, 68(5):767–784, 2006.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Thomas S Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pages 287–302. Elsevier, 1983.

Denis Fougère and Thierry Kamionka. Bayesian inference for the mover–stayer model in continuous time with an application to labour market transition data. *Journal of Applied Econometrics*, 18(6):697–723, 2003.

Sylvia Frühwirth-Schnatter and Sylvia Kaufmann. Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1): 78–89, 2008.

Halina Frydman. Estimation in the mixture of markov chains moving with different speeds. *Journal of the American Statistical Association*, 100(471): 1046–1053, 2005.

Christian Genest and Louis-Paul Rivest. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association*, 88(423):1034–1043, 1993.

RC Gentleman, JF Lawless, JC Lindsey, and P Yan. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13(8):805–821, 1994.

Clara Grazian and Brunero Liseo. Approximate bayesian inference in semi-parametric copula models. *Bayesian Analysis*, 12(4):991–1016, 2017.

Asger Hobolth and Eric A Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics*, 3(3):1204–1231, 2009.

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

David Huard, Guillaume Évin, and Anne-Catherine Favre. Bayesian copula selection. *Computational Statistics & Data Analysis*, 51(2):809–822, 2006.

Francesca Ieva, Christopher H Jackson, and Linda D Sharples. Multi-state modelling of repeated hospitalisation and death in patients with heart failure: the use of large administrative databases in clinical epidemiology. *Statistical methods in medical research*, 26(3):1350–1372, 2017.

Arne Jensen. Markoff chains as an aid in the study of markoff processes. *Scandinavian Actuarial Journal*, 1953(sup1):87–91, 1953.

Harry Joe. Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, pages 120–141, 1996.

Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.

Xavier Joutard, Alain Paraponaris, Luis Sagaon Teyssier, and Bruno Ventelou. Continuous-time markov model for transitions between employment and non-employment: the impact of a cancer diagnosis. *Annals of Economics and Statistics/ANNALES D'ÉCONOMIE ET DE STATISTIQUE*, pages 239–265, 2012.

JD Kalbfleisch and Jerald Franklin Lawless. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80 (392):863–871, 1985.

Minhee Kang and Stephen W Lagakos. Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics*, 8(2): 252–264, 2007.

Christian Kronwald. *Credit rating and the impact on capital structure*. GRIN Verlag, 2010.

Dorota Kurowicka and Roger M Cooke. *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons, 2006.

Jerald Lawless. The design and analysis of life history studies. *Statistics in Medicine*, 32(13):2155–2172, 2013.

T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

Albert Y Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357, 1984.

Steven N MacEachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2): 223–238, 1998.

Aleksey Min and Claudia Czado. Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8(4): 511–546, 2010.

Peter Müller, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. *Bayesian nonparametric data analysis*, volume 18. Springer, 2015.

James Robert Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

Aidan O'Keeffe, Brian Tom, and Vernon Farewell. A case-study in the clinical epidemiology of psoriatic arthritis: multistate models and causal arguments. *Journal of the Royal Statistical Society: Series C*, 60(5):675–699, 2011.

Christoph Pamminger and Sylvia Frühwirth-Schnatter. Model-based clustering of categorical time series. *Bayesian Analysis*, 5(2):345–368, 2010.

M Pfeuffer, L Mostel, and M Fischer. An extended likelihood framework for modelling discretely observed credit rating transitions. *Quantitative Finance*, 19(1):93–104, 2018.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Arkady Shemyakin and Alexander Kniazev. *Introduction to Bayesian estimation and copula models of dependence*. Wiley Online Library, 2017.

A Sklar. Fonctions dé repartition à n dimension et leurs marges. *Université Paris*, 8(3.2):1–3, 1959.

Michael S Smith and Mohamad A Khaled. Estimation of copula models with discrete margins via bayesian data augmentation. *Journal of the American Statistical Association*, 107(497):290–303, 2012.

Peter X-K Song, Yanqin Fan, and John D Kalbfleisch. Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, 100 (472):1145–1158, 2005.

Andrea Tancredi. Approximate bayesian inference for discretely observed continuous-time multi-state models. *Biometrics*, 75(3):966–977, 2019.

Andrew C Titman. Estimating parametric semi-Markov models from panel data using phase-type approximations. *Statistics and Computing*, 24(2): 155–164, 2014.

Andrew C Titman and Linda D Sharples. Semi-Markov models with phase-type sojourn distributions. *Biometrics*, 66(3):742–752, 2010.

Juan Wu, Xue Wang, and Stephen G Walker. Bayesian nonparametric inference for a multivariate copula function. *Methodology and Computing in Applied Probability*, 16(3):747–763, 2014.

Juan Wu, Xue Wang, and Stephen G Walker. Bayesian nonparametric estimation of a copula. *Journal of Statistical Computation and Simulation*, 85(1):103–116, 2015.