Computational Statistics and Data Analysis 220 (2026) 108359

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Dirichlet process multi-state mixture models

Rosario Barone [a,*], Andrea Tancredi [b]

[a] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milan, 20123, Italy
[b] Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Via del Castro Laurenziano 9, Rome, 00161, Italy

ARTICLE INFO

Keywords:
Bayesian nonparametrics
Clustering
Inhomogeneous Markov
Semi-Markov
Uniformization

ABSTRACT

A Bayesian nonparametric framework is introduced for modeling discretely observed trajectories of continuous-time multi-state processes. By employing Dirichlet Process Mixtures with Markov, inhomogeneous Markov, and semi-Markov kernels, the approach flexibly captures unobserved heterogeneity in the process dynamics. Crucially, the mixture structure induces a generalized form of non-Markovianity, as future state predictions depend on the entire observed history through component-specific weighting. This allows the model to capture complex temporal dependencies and memory effects beyond the scope of traditional multi-state models. The effectiveness of the methodology is demonstrated through simulation studies and an application to a real data set.

1. Introduction

Continuous-time multi-state models (CTMSM) represent processes that can be described as discrete states changing over time (Lawless, 2013). They are particularly valuable for analyzing longitudinal data, where individuals or entities are observed over time, and the modeling of their transitions between different states or conditions is of interest. The standard modeling approach relies on Markov processes. Alternative models that relax the Markov assumption are the semi-Markov models, where transitions between states may depend on the time elapsed since entry into the current state, and time-inhomogeneous Markov models, where transition rates depend on the time elapsed since the start of the process. All these models are widely used across various fields, including epidemiology, healthcare, economics, and engineering. Moreover, in many applications, the exact transition times between states are not observed; instead, only the states at arbitrary time points are recorded. These types of data are often referred to as panel data.

1.1. Frequentist and Bayesian approaches to multi-state model mixtures

In this article, we propose a Bayesian approach for estimating mixtures of CTMSM within the panel data framework. Mixture models allow for the heterogeneity among individuals by clustering them into a finite number of groups, each with a different set of model parameters. Failing to adjust for individual heterogeneity can lead to biases in estimating the parameters that control the process evolution, resulting in less accurate predictions, as recently demonstrated by Broomfield et al. (2024). Moreover, mixtures of Markov models provide a way to induce complex forms of non-Markovianity. In this framework, the conditional probability of future states given an observed history is expressed as a weighted average of the component-specific probabilities, each depending only on the last observed state, with the weights determined by the posterior probabilities that each mixture component receives from the entire history. In this way, mixtures of Markov models, or more general CTMSM, may capture temporal dependencies and

* Corresponding author.
E-mail address: rosario.barone@unicatt.it (R. Barone).

https://doi.org/10.1016/j.csda.2026.108359
Received 4 December 2024; Received in revised form 19 February 2026; Accepted 20 February 2026
Available online 24 February 2026

memory effects beyond the scope of traditional multi-state models. This perspective is consistent with Putter and Van Houwelingen (2015), who argue that latent components in multi-state models may be appropriately considered as a means to capture unobserved heterogeneity, but the primary modeling objective should be conditional prediction based on the observed history.

The idea of modelling state transitions using Markov chain mixture models dates back to the 1950s, with the introduction of the mover-stayer model (Blumen et al., 1955; Goodman, 1961). Since these seminal papers, numerous advancements in estimation methods and various model extensions have been proposed, from both frequentist and Bayesian perspectives. Focusing on recent years, Cook and Lawless (2018) considered a panel data framework, fitting a finite mixture of Markov CTMSM to the data by combining the expectation-maximization algorithm with the Fisher scoring algorithm proposed by Kalbfleisch and Lawless (1985) for estimating partially observed Markov CTMSM. Bayesian estimation of finite mixtures of discrete Markov multi-state models was first proposed by Pamminger and Frühwirth-Schnatter (2010) for clustering categorical time series. Bayesian estimation of mixtures of Markov CTMSM in a fully observed data framework was considered by Maystre et al. (2022), using a Dirichlet process mixture (DPM) (Lo, 1984; Escobar, 1994) to avoid selecting the number of mixture components. Finally, Luo et al. (2023) employed the DPM to account for heterogeneity in hidden Markov CTMSM within the panel data framework, comparing it with a reversible jump Markov chain Monte Carlo (MCMC) approach for clustering under finite mixture models with an unknown number of components. Similarly, Luo and Stephens (2021) explored mixtures of hidden Markov CTMSM, assuming that groups of trajectories may differ in the number of states in the underlying Markov model.

1.2. Relaxing the Markov assumption with panel data

Relaxing the Markov assumption within the panel data frameworks is challenging, particularly when accounting for heterogeneity. For continuous Markov processes, the Kolmogorov forward equations provide analytical solutions for transition probabilities, enabling likelihood computation via the matrix exponential. In contrast, for discretely observed semi-Markov CTMSM, a likelihood-based approach must rely on simplifying assumptions or complex numerical approximations. For instance, Kang and Lagakos (2006) assume constant transition intensities from at least one of the states. Titman and Sharples (2010) consider the tractable class of semi-Markov models represented by phase-type distributions for the sojourn time, while Titman (2014) uses phase-type distributions to approximate the likelihood of semi-Markov models with Weibull or Gamma sojourn time distributions. Alternatively, Armero et al. (2012) and Aastveit et al. (2023) restrict their analysis to acyclic multi-state models, excluding back transitions. Similar difficulties arise in time-inhomogeneous Markov models, where the Kolmogorov equations lack analytical solutions. A common approach is to assume piecewise time-homogeneity (Jackson, 2011). Several studies, including Titman (2011), Machado et al. (2021) and Kendall et al. (2024), have explored numerical solutions to the differential equations defining the transition probability matrix.

An alternative approach to panel data is through missing data techniques. For instance, Luo et al. (2023) and Luo and Stephens (2021) address partially observed data by reconstructing full process trajectories via rejection sampling, conditioned on consecutive observation pairs. Originally proposed by Hobolth and Stone (2009), this method was later adopted by Luo et al. (2021) to estimate hidden continuous-time Markov models with panel data. Trajectory reconstruction has been proposed not only for continuous-time Markov models. Barone and Tancredi (2022) embedded it within a Metropolis-within-Gibbs algorithm to estimate semi-Markov and non-homogeneous Markov multi-state models, using the distribution of a continuous Markov process–conditional on observed panel data–as the proposal. Instead of rejection sampling, they employed the uniformization algorithm, Hobolth and Stone (2009), which efficiently generates paths of continuous Markov processes within a fixed interval, conditioned on endpoints.

1.3. Bayesian nonparametric approach

In this paper, we extend Bayesian methods for estimating discretely observed CTMSM by introducing infinite mixtures of Markov, inhomogeneous Markov, and semi-Markov CTMSM, using a Dirichlet process prior (Ferguson, 1973) on the mixing measures. This leads to three DPM models, each with kernel densities tailored to the corresponding CTMSM class. Finite mixture models address heterogeneity by clustering individuals according to shared model parameters and, by relaxing restrictive parametric assumptions, offer a flexible framework for capturing complex and diverse data patterns. One of the main difficulties of this approach lies in the need to predefine the number of mixture components. Within the Bayesian framework, DPM models overcome this limitation by assuming an infinite number of potential components, allowing the data to determine the effective number of clusters while coherently quantifying uncertainty in the clustering structure through the posterior distribution over partitions. Moreover, they support sequential learning, as the predictive distribution updates consistently with the existing clustering structure when new data become available.

To adapt the DPM framework to discretely observed multi-state data, we incorporate a trajectory reconstruction method based on uniformization (Barone and Tancredi, 2022), enabling the DPM simulation algorithms to operate as if the data were fully observed.

The paper is organized as follows. Section 2 introduces the continuous-time multi-state models and the corresponding Dirichlet process mixture framework. Section 3 describes the MCMC algorithms for both fully and discretely observed data. Section 4 reports the results of a simulation study designed to evaluate the model's performance under different scenarios. Section 5 provides an empirical illustration using the Cardiac Allograft Vasculopathy dataset, and Section 6 presents the conclusions.

2. Multi-state models

Let us consider a continuous time process $Y(\cdot) = \{Y(t), t \geq 0\}$ with discrete state space $S = \{1, \dots, S\}$. A general definition of CTMSM may be given via its transition intensity function

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

which represents the instantaneous probability of a transition from state r to state s at time t when $\mathcal{F}_t$ is the entire history of the process at time t. Markov CTMSM have constant transition intensity functions, that is $q_{rs}(t, \mathcal{F}_t) = \gamma_{rs}$, with $\gamma_{rs} \geq 0$, $\gamma_{rr} = -\sum_{s \neq r} \gamma_{rs} = -\gamma_r$ and $\gamma_{rs} = 0 \; \forall s \neq r$ if r is an absorbing state. A formula for the transition probabilities for a Markov CTMSM can be obtained with the Chapman-Kolmogorov forward equations. Let Γ be the matrix with elements $\Gamma_{rs} = \gamma_{rs}$. Then $p_{rs}(t; \Gamma) = P\{Y(u+t) = s | Y(u) = r\}$ is the (r, s) element of the exponential matrix $\exp(t\Gamma) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \Gamma^k$.

Let $\{S_j, j \geq 0\}$ and $\{Z_j, j \geq 0\}$ represent respectively the sequences of states and jump times. In the homogeneous Markov case the state sequence $\{S_j, j \geq 0\}$ is a Markov chain with transition probabilities $p_{rs} = \gamma_{rs}/\gamma_r$, $s \neq r$ and the holding times $\{V_j = Z_j - Z_{j-1}, j \geq 1\}$ are independent exponential random variables with rates $\gamma_{s_{j-1}}$ depending on the departure state. Let $y = y(t)$ for $t \in [0, T]$ be an observed trajectory of the process. Moreover, let ℓ be the number of jumps observed in y. Assuming that the initial state s_0 is fixed, let $s = (s_1, \dots, s_\ell)$ be the state sequence in $(0, T]$ and $z = (z_1, \dots, z_\ell)$ be the jump times sequence. The Markov process density may be written as

$$\begin{aligned} p_M(y) = p_M(s, z) &= \left(\prod_{j=1}^{\ell} p_{s_{j-1} s_j} \gamma_{s_{j-1}} e^{-\gamma_{s_{j-1}}(z_j - z_{j-1})} \right) e^{-\gamma_{s_\ell}(T - z_\ell)} \\ &= \prod_{j=1}^{\ell} \gamma_{s_{j-1} s_j} e^{-\int_0^T \gamma_{y(t)} dt}. \end{aligned} \tag{1}$$

Suppose now to have n fully observed trajectories $\mathbf{y} = (y_1, \dots, y_n)$ where $y_i = y_i(t)$ for $t \in [0, T_i]$. Let $s_{i,1}, \dots, s_{i,\ell_i}$ be the state sequence for the trajectory y_i, $z_{i,1}, \dots, z_{i,\ell_i}$ be the jump time sequence, $N_{rs} = \sum_{i=1}^{n} \sum_{j=1}^{\ell_i} I(s_{i,j-1} = r, s_{i,j} = s)$ be the total number of transitions $r \to s$ across all the trajectories, $N_r = \sum_{s \neq r} N_{rs}$ the total number of completed holding times into the state r and $W_r = \sum_{i=1}^{n} \int_0^{T_i} I(y_i(t) = r) dt$ the total sojourn time for the state r across all the trajectories. The likelihood for the model parameters $\theta = (P, \gamma)$ where P is the $S \times (S-1)$ matrix with the transition probabilities p_{rs} and γ is the vector $(\gamma_1, \dots, \gamma_S)$ can be written as

$$L(\theta; \mathbf{y}) = L(P, \gamma; \mathbf{y}) = \prod_{rs} p_{rs}^{N_{rs}} \prod_{r} \gamma_r^{N_r} e^{-\gamma_r W_r}.$$

2.1. Inhomogeneous Markov processes

Assuming $q_{rs}(t, \mathcal{F}_t) = \gamma_{rs}(t)$, that is relaxing the assumption of time independence of the transition intensity functions, we define the time-inhomogeneous CTMSM. In this case the jump time Z_j depends on the last visited state S_{j-1} and its entry time Z_{j-1}. Let $\gamma_r(t)$ be the sum of the rate transition functions, i.e. $\gamma_r(t) = \sum_{s \neq r} \gamma_{rs}(t)$. The conditional density of $Z_j | S_{j-1}, Z_{j-1}$ is

$$p(z_j | z_{j-1}, s_{j-1}) = \gamma_{s_{j-1}}(z_j) e^{-\int_{z_{j-1}}^{z_j} \gamma_{s_{j-1}}(t) dt}, \tag{2}$$

while the conditional probability for $S_j | S_{j-1}, Z_j$ is

$$p(s_j | z_j, s_{j-1}) = \frac{\gamma_{s_{j-1}, s_j}(z_j)}{\gamma_{s_{j-1}}(z_j)}.$$

The density of a trajectory y can be generally written as

$$\begin{aligned} p_{IM}(y) = p_{IM}(s, z) &= \left(\prod_{j=1}^{\ell} \gamma_{s_{j-1} s_j}(z_j) e^{-\int_{z_{j-1}}^{z_j} \gamma_{s_{j-1}}(t) dt} \right) e^{-\int_{z_\ell}^{T} \gamma_{s_\ell}(t) dt} \\ &= \prod_{j=1}^{\ell} \gamma_{s_{j-1} s_j}(z_j) e^{-\int_0^T \gamma_{y(t)} dt}. \end{aligned} \tag{3}$$

A standard assumption for the inhomogeneous case is obtained by taking a Gompertz link function for the transition rates, i.e. $\gamma_{rs}(t) = p_{rs} e^{\beta_{0r} + \beta_{1r} t}$. In this case the conditional density of $Z_j | S_{j-1}, Z_{j-1}$ becomes

$$p(z_j | z_{j-1}, s_{j-1}) = e^{\beta_{0 s_{j-1}} + \beta_{1 s_{j-1}} z_j} e^{-\frac{\beta_{0 s_{j-1}}}{\beta_{1 s_{j-1}}} \left(e^{\beta_{1 s_{j-1}} z_j} - e^{\beta_{1 s_{j-1}} z_{j-1}} \right)},$$

and the conditional probability for $S_j | S_{j-1}, Z_j$ is $p(s_j | z_j, s_{j-1}) = p_{s_{j-1} s_j}$. Let $\beta = (\beta_{01}, \dots, \beta_{0S}, \beta_{11}, \dots, \beta_{1S})$ be the parameter vector of the coefficients of the Gompertz model and P be the matrix with the transition probabilities. In the case of n fully observed trajectories, the likelihood for the parameters $\theta = (P, \beta)$ of the Gompertz inhomogeneous model can be written by introducing the quantities

$$H_r = \sum_{i=1}^{n} \sum_{j : s_{i,j-1} = r}^{\ell_i} z_{i,j},$$

and

$$K_r(\beta_{0r},\beta_{1r}) = \sum_{i=1}^{n}\left\{\sum_{j\,:\,s_{i,j-1}=r}^{\ell_i}(e^{\beta_{1r}z_{i,j}} - e^{\beta_{1r}z_{i,j-1}}) + (e^{\beta_{1r}T_i} - e^{\beta_{1r}z_{i,\ell_i}})\mathbb{1}(s_{\ell_i}=r)\right\},$$

for $r = 1,\dots,S$. This way, the likelihood function can be written as

$$L(\theta;\mathbf{y}) = L(P,\beta;\mathbf{y}) = \prod_{rs} p_{rs}^{N_{rs}} \prod_{r} e^{\beta_{0r}N_r + \beta_{1r}H_r - e^{\beta_{0r}}K_r(\beta_{0r},\beta_{1r})/\beta_{1r}}.$$

2.2. Semi-Markov processes

Another class of continuous time multi-state models can be obtained by assuming that the transition intensity functions depend on the time spent in the current state. By taking

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

where T^* denotes the entry time in the last state assumed before time t, and

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

we define a semi-Markov CTMSM. The evolution of a semi-Markov process can be explained as follows. Let $F_r(u)$ represent the distribution with the hazard function $\sum_{l\neq r} q_{rl}(u)$ and define $p_{rs} = \int_0^\infty q_{rs}(u)(1-F_r(u))du$ and $F_{rs}(u) = \frac{1}{p_{rs}}\int_0^u q_{rs}(v)(1-F_r(v))dv$ for $s\neq r$. Then, the process $Y(t)$ is given by the state sequence produced by the Markov chain with transition probabilities p_{rs} and by the sojourn times, which depend on the transition states, generated independently using the distributions F_{rs}. Note also that we can obtain the transition intensity functions $q_{rs}(u)$ by specifying directly the transition probabilities p_{rs} and the conditional sojourn distributions F_{rs}. In fact we have $q_{rs}(u) = \frac{p_{rs}F'rs(u)}{1-F_r(u)}$, where $F_r(u) = \sum_{l\neq r} p_{rl}F_{rl}(u)$. Various parametric families can be suggested for $q_{rs}(u)$ or $F_{rs}(u)$. Assuming for example cause-specific hazards proportional to those of a distribution on $(0,\infty)$ with parameters depending only on the initial state, i.e. $q_{rs}(u) = p_{rs}q(u;\phi_r)$, the transition probabilities are p_{rs}, and the density of a trajectory y on the interval $[0,T]$ for a semi Markov process can be generally written as

$$p_{SM}(y) = p_{SM}(s,z) = \left(\prod_{i=1}^{\ell} p_{s_{i-1}s_i}q(z_i - z_{i-1};\phi_{s_{i-1}})e^{-\int_0^{z_i-z_{i-1}}q(u;\phi_{s_{i-1}})du}\right) e^{-\int_0^{T-z_\ell}q(u;\phi_{s_\ell})du}. \tag{4}$$

In the case of n fully observed trajectories with Weibull sojourn times depending only on the departure state, that is assuming $q_{rs}(u) = p_{rs}\gamma_r\alpha_r(\gamma_r u)^{\alpha_r-1}$ the likelihood function for the model parameter $\theta = (P,\gamma,\alpha)$ can be easily evaluated. In fact, adding to the notation previously introduced the quantities

$$U_r = \prod_{i=1}^{n}\prod_{j\,:\,s_{i,j-1}=r}^{\ell_i}(z_{i,j} - z_{i,j-1}),$$

and

$$W_r(\alpha_r) = \sum_{i=1}^{n}\left\{\sum_{j\,:\,s_{i,j-1}=r}^{\ell_i}(z_{i,j} - z_{i,j-1})^{\alpha_r} + (T_i - z_{\ell_i})^{\alpha_r}I(s_{\ell_i}=r)\right\},$$

for $r = 1,\dots,S$, the likelihood function is

$$L(\theta;\mathbf{y}) = L(P,\gamma,\alpha;\mathbf{y}) = \prod_{rs} p_{rs}^{N_{rs}} \prod_{r}(\alpha_r\gamma_r^{\alpha_r})^{N_r}U_r^{\alpha_r-1}e^{-\gamma_r^{\alpha_r}W_r(\alpha_r)}.$$

2.3. Dirichlet process mixtures of continuous time multi-state models

DPM models represent a very flexible tool for clustering and density estimation. They provide a way to model the uncertainty about the number of components in the mixture and their parameters, automatically determining the appropriate number of clusters given the data. This feature ensures the adequate flexibility of the model, allowing it to adapt to the characteristics of the observed individuals, resulting particularly useful in our context, where there can be partial observations of a continuous process and the identification of trends in subgroups of observations could be very complex.

More formally, let $p_\theta(s,z)$ be the probability density function of a CTMSM completely observed on the interval $[0,T]$, where $\theta\in\Theta$ represents the multi-dimensional parameter controlling the process. This parameter can be a rate matrix if a Markov kernel is used, or may represent a more complex parametric structure if an inhomogeneous Markov or semi-Markov kernel is considered for the

mixture model. Let G be a probability distribution defined on the parameter space Θ. We define the density function $p_G(s,z)$ of a mixture with multi-state model kernels $p_\theta(s,z)$ with respect to the mixing measure G as

$$p_G(s,z) = \int p_\theta(s,z) dG(\theta). \tag{5}$$

With a Dirichlet process (DP) prior on the mixing measure G with base distribution G_0 and concentration parameter M we get a CTMSM Dirichlet process mixture. Such a model can be also presented in a hierarchical form. In fact, let $Y_i(\cdot) = (s,z)_i$, for $i = 1,\dots,n$, be n fully observed paths on $[0,T_i]$ where the observation times T_i can be different across the paths. Then we may assume that

$$\begin{aligned} Y_i(\cdot)|\theta_i &\overset{ind}{\sim} p_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(MG_0). \end{aligned} \tag{6}$$

This definition holds for each of the presented models. By using a Markov, semi-Markov or time-inhomogeneous Markov density kernels, we get a different DPM model with the definition of the base measure G_0 depending on the chosen kernel. Notice that from a computational perspective the conjugacy between G_0 and the kernel density p_θ is particularly convenient. Unfortunately, this condition, as it will be explained below, holds only for the DPM of Markov CTMSM.

3. Markov chain Monte Carlo methods

The estimation of a DPM is typically performed using MCMC methods. Several MCMC strategies for DPMs have been developed, broadly classified into conditional and marginal approaches. Conditional methods approximate the infinite-dimensional process (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001, 2003), including retrospective (Papaspiliopoulos and Roberts, 2008) and slice algorithms (Walker, 2007; Kalli et al., 2011), later unified by Hastie et al. (2015). Marginal methods, which integrate out the random measure, were first proposed by Escobar (1994) and extended to non-conjugate cases by MacEachern and Müller (1998), with further developments by Neal (2000). Comparative studies (Papaspiliopoulos and Roberts, 2008; Kalli et al., 2011) show no clear dominance between the two approaches and for our framework we will use the marginal one.

3.1. Inference for DPM of fully observed CTMSM

Let $\mathbf{y} = (y_1,\dots,y_n)$ be the set of the n fully observed trajectories of a CTMSM defined on the state space S. Let $\boldsymbol{\theta} = (\theta_1,\dots,\theta_n)$ be the set with the n multi-dimensional parameters from which the n trajectories have been generated. For instance, in the homogeneous Markov case $\boldsymbol{\theta}$ is the set of n rate matrices $(\Gamma_1,\dots,\Gamma_n)$. Let θ_h^*, $h = 1,\dots,k$ denote the $k \le n$ unique elements of the set $\boldsymbol{\theta}$, that we indicate with Γ_h^* in the homogeneous Markov model. Let $\Psi_h = \{i : \theta_i = \theta_h^*\}$ be the set of labels comprising the hth cluster, for $h = 1,\dots,k$. Consider also the equivalent set of cluster membership indicators $\boldsymbol{\psi} = (\psi_1,\dots,\psi_n)$ where $\psi_i = h$ if $i \in \Psi_h$ indicates that the ith observation belongs to the hth cluster. Note that under the DPM model we have that

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

where $k_{\cdot}$ denotes the number of unique θ_ℓ among $\{\theta_1,\dots,\theta_i\}$ and $n_{i,j}$ denote the number of the jth of these unique values. The basic MCMC algorithm for conjugate DPM models performs a data clustering followed by the updating of the cluster parameters.

The clustering step is obtained by drawing the cluster indicator ψ_i conditionally on all the other allocation variables $\boldsymbol{\psi}_{-i} = (\psi_1,\dots,\psi_{i-1},\psi_{i+1},\dots,\psi_n)$. To perform this step we need to introduce the following quantities. Let k^- indicate the unique values of θ with exclusion of the ith observation, let $\mathbf{y}_h$ denote the trajectories arranged by clusters, i.e. the set of trajectories belonging to the hth cluster and $\mathbf{y}_h^- = \mathbf{y}_h \setminus \{(s,z)_i\}$ and let n_h^- denote the number of statistical units assigned to the hth cluster, excluding unit i if it belongs to it. Moreover, consider

$$p((s,z)_i|\psi_i = h, \mathbf{y}_h^-) = \int p_{\theta_h^*}((s,z)_i) d\pi(\theta_h^*|\mathbf{y}_h^-), \tag{7}$$

where $\pi(\theta_h^*|\mathbf{y}_h^-)$ is the posterior distribution of θ_h^* conditional on the paths of the hth cluster without the ith observation, and

$$p((s,z)_i) = \int p_{\theta_h^*}((s,z)_i) G_0(d\theta_h^*). \tag{8}$$

The posterior conditional probabilities $\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{y})$ are given by the following expression

$$\pi(\psi_i = h|\boldsymbol{\psi}_{-i}, \mathbf{y}) \propto \begin{cases} n_h^- p((s,z)_i|\psi_i = h, \mathbf{y}_h^-) & \text{for } h = 1,\dots,k^- \\ M p((s,z)_i) & \text{for } h = k^- + 1 \end{cases}. \tag{9}$$

The described framework naturally fits to the Markov case. By defining the centering measure G_0 as a product of independent Gamma(a_r, b_r) distributions for the rate parameter of each state r and independent Dirichlet(α_r) distributions, with $\alpha_r = (\alpha_{r1},\dots,\alpha_{rm})$, for the transitions probabilities from each state r, we can exploit the conjugacy between the defined prior distributions and the likelihood function. In fact, the conditional posterior distribution $\pi(\theta_h^*|\psi, \mathbf{y}) = \pi(\Gamma_h^*|\psi, \mathbf{y})$ is still a product of Gamma and Dirichlet

distributions. Specifically, let $W_{r,h}$ denote the sum of the holding times in state r for all trajectories lying in cluster h. Let $N_{r,h}$ represent the number of visits to state r and $N_{rm,h}$ denote the number of observed transitions from state r to state m for the trajectories in cluster h. Then we have

$$\pi(\Gamma_h^*|\psi,\mathbf{y}) = \prod_{r=1}^{S}\Gamma\left(\sum_{m\neq r}N_{rm,h}+\alpha_{rm}\right)\left(\prod_{m\neq r}\frac{{p^*}_{rm,h}^{N_{rm,h}+\alpha_{rm}-1}}{\Gamma\left(N_{rm,h}+\alpha_{rm}\right)}\right) \times \prod_{r=1}^{S}\frac{(b_r+W_{r,h})^{a_r+N_{r,h}}}{\Gamma(a_r+N_{r,h})}\left({\gamma^*}_{r,h}^{N_{r,h}+a_r-1}e^{-{\gamma^*}_{r,h}(W_{r,h}+b_r)}\right), \tag{10}$$

where $p^*_{rm,h}$ and $\gamma^*_{r,h}$ represent the transition probabilities and the rate parameters defining the transition rate matrix Γ^*_h. Note also that the integral defined in (7) is given by:

$$p((s,z)_i|\psi_i=h,\mathbf{y}_h^-) = \prod_{r=1}^{S}\left\{\frac{(b_r+W^-_{r,h})^{a_r+N^-_{r,h}}\Gamma\left(\Sigma_{s\neq r}(\alpha_{rm}+N^-_{rm,h})\right)}{\Gamma(a_r+N^-_{r,h})\prod_{s\neq r}\Gamma\left(\alpha_{rm}+N^-_{rm,h}\right)}\right\} \left\{\frac{\Gamma(a_r+N_{r,h})\prod_{s\neq r}\Gamma(\alpha_{rm}+N_{rm,h})}{(b_r+W_{r,h})^{a_r+N_{r,h}}\Gamma\left(\sum_{s\neq r}\left(\alpha_{rm}+N_{rm,h}\right)\right)}\right\}. \tag{11}$$

Analogously, the integral (8) becomes:

$$p((s,z)_i) = \prod_{r=1}^{S}\left\{\frac{b_r^{a_r}\Gamma\left(\sum_{s\neq r}\alpha_{rm}\right)}{\Gamma(a_r)\prod_{s\neq r}\Gamma\left(\alpha_{rm}\right)}\right\} \left\{\frac{\Gamma\left(a_r+N_r(y_i)\right)\prod_{s\neq r}\Gamma\left(\alpha_{rm}+N_{rm}(y_i)\right)}{\left(b_r+W_r(y_i)\right)^{a_r+N_r(y_i)}\Gamma\left(\sum_{s\neq r}\left(\alpha_{rm}+N_{rm}(y_i)\right)\right)}\right\}, \tag{12}$$

where $W_r(y_i)$ represents the total time spent by individual i in state r, while $N_r(y_i)$ represents the number of visits to state r, and $N_{rm}(y_i)$ represents the number of transitions from r to m.

To complete the MCMC simulation algorithm in the homogeneous Markov case, once the Gibbs type updating of the cluster indicators has been performed by drawing ψ_i from (9), for $i=1,\dots,n$, we need to update the components parameters θ^*_h for $h=1,\dots,k$ by the direct simulation from the distributions (10).

With the semi-Markov and in-homogeneous Markov kernels there is no conjugate base measure G_0 for the component parameters. In this case, to bypass the analytical evaluation of the integrals (7) and (8), MacEachern and Müller (1998) proposed the "no-gaps" sampler which updates the cluster indicators ψ_i conditionally on $\boldsymbol{\psi}_{-i}$ and the components parameters θ^*_h, $h=1,\dots,k$ requiring only simple likelihood calculations. In fact by considering $n-k$ auxiliary values $\theta^*_{k+1},\dots,\theta^*_n$ from the base measure G_0, a valid MCMC algorithm can be obtained by distinguishing the case where in the current partition the trajectory y_i belongs to a cluster with more than one element from the case where is the only member in a singleton cluster. In the former case ψ_i is drawn from

$$\pi(\psi_i=h|\boldsymbol{\psi}_{-i},\mathbf{y},\theta^*_h) \propto \begin{cases} n_h^- p_{\theta^*_h}((s,z)_i) & \text{for } h=1,\dots,k^- \\ \frac{M}{k^-+1}p_{\theta^*_{k^-+1}}((s,z)_i) & \text{for } h=k^-+1 \end{cases}, \tag{13}$$

while in the latter case ψ_i is updated with probability $1/k$ using the same probabilities as in (13) after a relabeling step that ensures the first k^- clusters are non empty.

In these non-conjugate cases, the updating of the component parameters generally cannot be obtained by the direct simulation from the posterior distributions $\pi(\theta^*_h|\psi,\mathbf{y})$. However, a Metropolis-Hastings step can be easily implemented with invariant distribution

$$\pi(\theta^*_h|\psi,\mathbf{y}) \propto G_0(\theta^*_h)\prod_{i=1:\psi_i=h}^{n}p_{\theta^*_h}((s,z)_i). \tag{14}$$

3.2. Inference for discretely observed CTMSM

We now suppose to have only partially observed data. Let $x_i=(x_{i,1},\dots,x_{i,m_i})$ be the sequence of states observed at times $t_{i,1},\dots,t_{i,m_i}$, for $i=1,\dots,n$. In order to fit a Dirichlet processes mixture of CTMSM also to these kind of data we propose to embed the algorithms outlined in Section 3.1 within a data augmentation step recovering the whole sample paths y_i, for $i=1,\dots,n$ until the last observed state. To simplify notation, when possible, we will omit subscripts for individuals.

3.2.1. Markov kernel

Firstly, let us consider the homogeneous Markov case. The reconstruction of a sample path y, given the discretely observed states x, can be achieved by using either straightforward rejection algorithms or uniformization-based methods (Hobolth and Stone, 2009).

In rejection algorithms, given the states r and s observed at times u and v, the algorithm draws $y(t)$ for $t \in [u, v]$ from a Markov CTMSM with starting state $y(u) = r$ and rate matrix Γ until the sample satisfies the transition constraint $y(v) = s$. In the panel data framework, the whole sample reconstruction is then obtained by drawing $y_i(t)$ for each interval $[t_{i,j}, t_{i,j+1}]$ for $j = 1, \dots, m_i$ and $i = 1, \dots, n$. Note, however, that despite the simplicity of this approach, embedding such algorithms into the MCMC machinery may be self-defeating due to the high rejection rates we may encounter when the interval length $t_{i,j+1} - t_{i,j}$ increases or when the transition matrix used for the simulation is not able to reproduce the observed data.

Contrary to the rejection samplers, the uniformization based algorithms are able to simulate directly a continuous time Markov process conditionally on the states observed at two interval endpoints. The key idea behind uniformization is that a continuous-time Markov process is equivalent to the process $Y^*(t)$ obtained by (i) simulating a Poisson process with rate $\mu = \max(\mathrm{diag}(\Gamma))$; (ii) generating a state sequence from a Markov chain with probability transition matrix $R = I + \frac{1}{\mu}\Gamma$; (iii) considering the points drawn from the Poisson processes as the jump points where the Markov chain transitions occur;(iv) removing the self transitions. Then, to simulate $y_i(t)$ on $[u, v]$ conditioned on both the interval end-points $y(u) = r$ and $y(v) = s$, we draw the number of change points $\tilde{N}$, considering the virtual ones, from a discrete random variable with distribution

$$P(\tilde{N} = n | y(u) = r, y(v) = s) \propto e^{-\mu(v-u)} \frac{(\mu(v-u))^n}{n!} \frac{R^n_{r,s}}{p_{rs}(v-u;\Gamma)}.$$

Conditional on $\tilde{N} = n$, $y(u) = r$ and $y(v) = s$ we generate the n transition times from a uniform distribution on $[u, v]$, while the state sequence can be drawn from a Markov chain with transition matrix R, starting state r and final state s. Finally, as in the unconditioned case, all the self-transitions are discarded.

The possibility to draw efficiently for each individual belonging to a given cluster the whole path $y_i(t) = (s, z)_i$ conditionally on the observed points x_i, that is to simulate directly $p_M(y_i|x_i, \Gamma_i)$, leads to the Gibbs sampler algorithm for discretely observed Markov CTMSM described in Algorithm 1.

Algorithm 1 Gibbs sampling for DPM of Markov multi-state models.

Path reconstruction:
for $i = 1$ to n **do**
  Draw $y_i \sim p_M(y_i \mid x_i, \Gamma^*_{\psi_i})$
end for

Clustering:
for $i = 1$ to n **do**
  Draw $\psi_i \sim \pi(\psi_i = h \mid \boldsymbol{\psi}_{-i}, \mathbf{y})$ ▷ Using (9), (11), and (12)
end for

Updating cluster parameters:
for $h = 1$ to k **do**
  Draw $\Gamma^*_h \sim \pi(\Gamma^*_h \mid \boldsymbol{\psi}, \mathbf{y})$ ▷ See (10)
end for

3.2.2. Time-inhomogeneous Markov and semi-Markov kernels

Recovering the entire trajectory y given the discretely observed states x in the semi-Markov and inhomogeneous Markov cases requires further computational effort. In the semi-Markov case, rejection algorithms can only be performed by drawing the complete path y until it agrees with the entire observed state sequence x. In fact the possibility to operate sequentially on disjoint sub-intervals $[t_j, t_{j+1}]$ is precluded by the dependence of the process on the entry time in the last state assumed before t_j. For the inhomogeneous case, it is possible to draw the sequence on t_j, t_{j+1} conditionally on the starting state x_j or to adapt the uniformization method to condition also to the end point x_{j+1}, as in the homogeneous case (Van Dijk, 1992). Recently, Barone and Tancredi (2022) proposed a Metropolis-Hastings algorithm with target distribution $p(y|x)$ that can be used both for semi-Markov and non-homogeneous Markov models. The proposal distribution $q(y)$ was the conditional density of a homogeneous or piece-wise homogeneous Markov model approximating the true distribution of $y|x$. This way, each proposal for y automatically matches the whole vector x; the acceptance probability for the proposed trajectory $y\prime$ is given by

$$A = \min\left\{1, \frac{p(y\prime|x)}{p(y|x)} \frac{p_M(y|x)}{p_M(y\prime|x)}\right\} = \min\left\{1, \frac{p(y\prime)}{p(y)} \frac{p_M(y)}{p_M(y\prime)}\right\}. \tag{15}$$

Notice that the Markovian trajectories conditioned on the observed points x can be simulated via the uniformization method described above and the Metropolis-Hastings step can be easily embedded in more complex MCMC algorithms.

The efficiency of the Metropolis-Hastings step will depend on the capacity of the proposal distribution $p_M(y|x)$ to mimic the target density $p(y|x)$. This can be obtained by tuning the parameters of the proposal density accordingly to the transition intensity functions of the data generating process. For example, consider a semi-Markov model with Weibull sojourn times given by the transition intensity functions $q_{rs}(u) = p_{rs}\gamma_r\alpha_r\left(\gamma_r u\right)^{\alpha_r - 1}$. In this case, we may draw the state sequence via the transition probabilities p_{rs} and for the rate parameters of the exponential holding time proposals we can take the inverse of the mean of the Weibull random variables, i.e. $\gamma_r/\Gamma(1 + \alpha_r)$ or directly the Weibull rate parameters γ_r. To ensure efficiency of the Metropolis-Hastings step

for inhomogeneous Markov models we may need to account for time dependence in the proposal distribution. To this aim, we could propose from a piecewise homogeneous Markov model with different rate matrices for each time interval $[t_j, t_{j+1}]$. For instance, assuming $q_{rs}(t) = p_{rs}\exp(\beta_{0r} + \beta_{1r}t)$ we may take as rate parameters for the proposed Markov process in the interval $[t_j, t_{j+1}]$ the quantities $exp(\beta_{0r} + \beta_{1r}t_j^*)$ where t_j^* is an internal point of the interval $[t_j, t_{j+1}]$. Alternatively, a less computationally expensive approach would be to propose the entire trajectory directly from a homogeneous Markov process with rates $\gamma_r = \exp\left(\beta_{0r} + \beta_{1r}t^*\right)$, where t^* is a point in the time interval $[0, T]$.

The possibility to perform a Metropolis-Hastings step providing the realization of the whole path $y_i(t) = (s_i, z_i)$ for each individual belonging to a given cluster conditionally on the observed points x_i, leads to the Metropolis-Within-Gibbs algorithm described in Algorithm 2.

Algorithm 2 Metropolis within Gibbs sampling for multi-state models.

Path reconstruction:
for $i = 1$ to n **do**
    Perform a Metropolis–Hastings step with acceptance probability given by Eq. (15), conditional on $\theta^*_{\psi_i}$
end for

Clustering:
for $i = 1$ to n **do**
    Draw ψ_i using the no-gaps algorithm from $\pi(\psi_i = h \mid \boldsymbol{\psi}_{-i}, \mathbf{y}, \theta^*_h)$ as defined in Eq. (13)
end for

Updating cluster parameters:
for $h = 1$ to k **do**
    Draw $\theta^*_h \sim \pi(\theta^*_h \mid \boldsymbol{\psi}, \mathbf{y})$ from Eq. (14)
end for

4. Simulation study

In this Section, we present a simulation study designed to evaluate the performance of the proposed method. We consider two simulation setups and two sample sizes, $n = 600$ and $n = 1200$, generating 100 datasets for each scenario. The objective is to assess the model's performance in two different situations: in the first case, we aim to evaluate its ability to correctly identify the exact number of subgroups in a clear setting with three well-separated components; in the second case, we analyze its performance in a more challenging context, where three components are present but two of them are very close to each other.

In the first setup, data are simulated from a mixture of three equally weighted components, each representing a continuous-time semi-Markov process. The state space includes four states, with the fourth being absorbing. Individuals are followed for up to 20 years, with observations recorded every six months. Each component is defined by distinct transition dynamics and Weibull-distributed sojourn times. The first cluster represents fast-progressing individuals, characterized by a shape parameter of 1.8 and a rate of 2 across all states. Transition probabilities are set to 0.8 for the next state and 0.1 for each of the others. The second cluster models intermediate progression, with shape 2.2 and rate 0.3. Transitions occur primarily toward the next state with probability 0.83, followed by 0.13 for the previous state and 0.04 for the remaining one–implying a decreasing likelihood of reaching higher states from state 1. The third cluster captures slower trajectories, with shape 2.3 and rate 0.1. Transition probabilities are 0.84 for the next state, 0.11 for the previous, and 0.05 for the remaining one. As in the intermediate case, transitions from state 1 become less likely as the target state increases from 2 to 4. The second setup differs from the first exclusively in the first component of the mixture, which represents individuals with slightly faster transitions compared to those in the second group. These are simulated using Weibull distributions characterized by a shape parameter of 1.2 and a rate of 1 across all states. Transition probabilities are set to 0.9 for the next state and 0.05 for each of the others. Fig. 1 shows the density functions of sojourn times in the two scenarios considered.

4.1. Posterior number of clusters

We evaluate the model's ability to recover the true number of mixture components by considering three values of the concentration parameter, $M \in \{0.05, 0.1, 1\}$. The choice of the parameter M is a critical aspect in any inferential approach based on DPM. Indeed, Miller and Harrison (2014) discussed the inconsistency of DPMs regarding the estimation of the number of components, whereas Frühwirth-Schnatter and Malsiner-Walli (2019) showed that a DPM with M centered on small values yields consistent estimates of the number of clusters. Figs. 2 and 3 report the distribution of the posterior modes of the number of observed components across 100 simulated datasets. In the first simulation setup (Fig. 2), the most frequently occurring posterior mode matches the true value ($k = 3$) for both $n = 600$ and $n = 1200$ when $M = 0.05$ and $M = 0.1$, although variability increases with sample size. For $M = 1$, the mode shifts slightly to $k = 4$ for $n = 600$ and to $k = 5$ for $n = 1200$, as theoretically expected. In fact, this behavior reflects both the tendency of larger M values to favor more components and the property of DPM to introduce additional clusters as the sample size increases. In the second setup (Fig. 3), where two components are closer together, performance improves as M increases for $n = 600$,

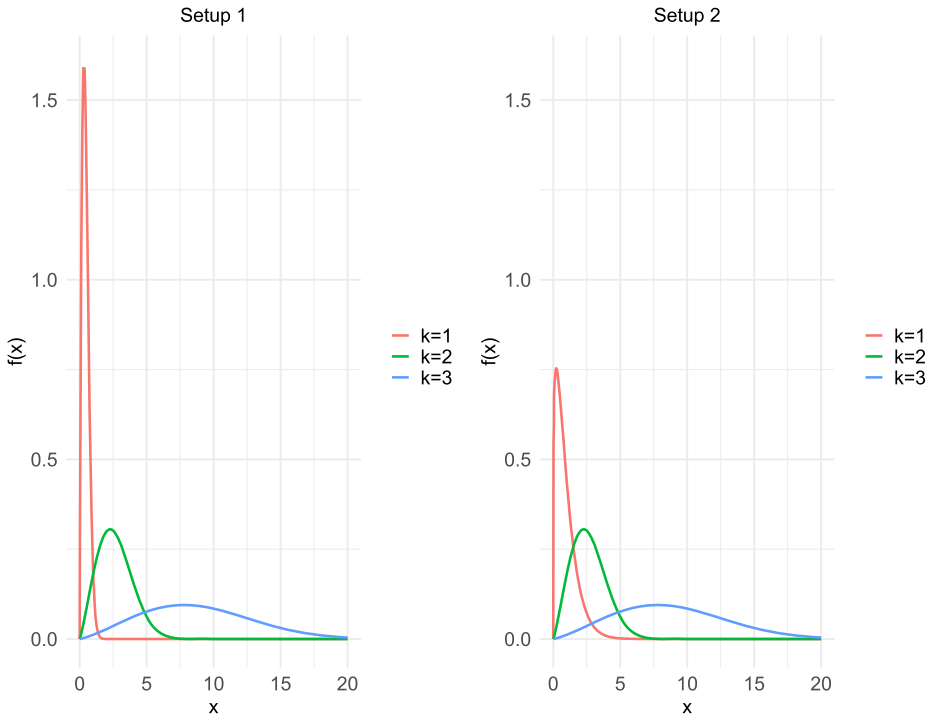

Fig. 1. Simulation Study. Theoretical distributions of sojourn times in the two simulation setups considered.

since a more flexible prior helps capture the underlying heterogeneity. Instead, for $n = 1200$, the best recovery of the true number of components occurs at $M = 0.1$. The results obtained for different values of M are also influenced by the relatively large sample sizes considered in our simulations. Since the expected number of clusters increases as $\mathbb{E}(K_n) \approx M \log(n)$ (Antoniak, 1974), larger values of n naturally call for smaller values of M in order to avoid an excessive proliferation of mixture components. In line with these considerations and supported by our simulation results, in the application presented in Section 5 we fix $M < 1$, which increases the concentration of observations within clusters, reduces the number of singleton components, and facilitates interpretability. Notice also that in the Bayesian nonparametric applications, M is often fixed (Guindani et al., 2014; Barone and Dalla Valle, 2023; Barone and Farcomeni, 2025), while alternative approaches place a prior on M and infer it from the data, as in the seminal DPM paper of Escobar and West (1995), trading interpretability for additional flexibility. See also Frühwirth-Schnatter and Malsiner-Walli (2019) and Ascolani et al. (2023) for recent theoretical developments in this direction.

4.2. Partition identification

We evaluate the accuracy of the clustering by comparing the representative partition produced by the model with the true data generating partition, using the Adjusted Rand Index (ARI). To derive a representative partition from the MCMC output, we apply the least-squares clustering method proposed by Dahl (2006). This approach relies on the posterior similarity matrix $C \in \mathbb{R}^{n \times n}$, where each entry C_{ij} denotes the proportion of MCMC iterations in which observations i and j are allocated to the same cluster:

$$C_{ij} = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}\{\psi_i^{(r)} = \psi_j^{(r)}\},$$

with R denoting the total number of MCMC iterations. Among all sampled partitions, the method selects the one that minimizes the squared distance to C:

$$\hat{c} = \arg\min_{c} \sum_{i<j} \left[\mathbb{1}\{c_i = c_j\} - C_{ij}\right]^2.$$

This procedure is robust to label switching and yields an interpretable and coherent summary of the posterior clustering structure. Table 1 reports the results for the average Adjusted Rand Index (ARI), computed as the mean of the ARI values obtained by comparing the representative partition estimated for each dataset with the corresponding true data-generating partition in both simulation setups. For the first setup, the results show excellent model performance at different values of the concentration parameter M and sample size n, indicating a strong agreement between the estimated and true partition, particularly for lower values of M. This suggests that the model effectively captures the underlying clustering structure, even in a context such as this, where latent heterogeneity

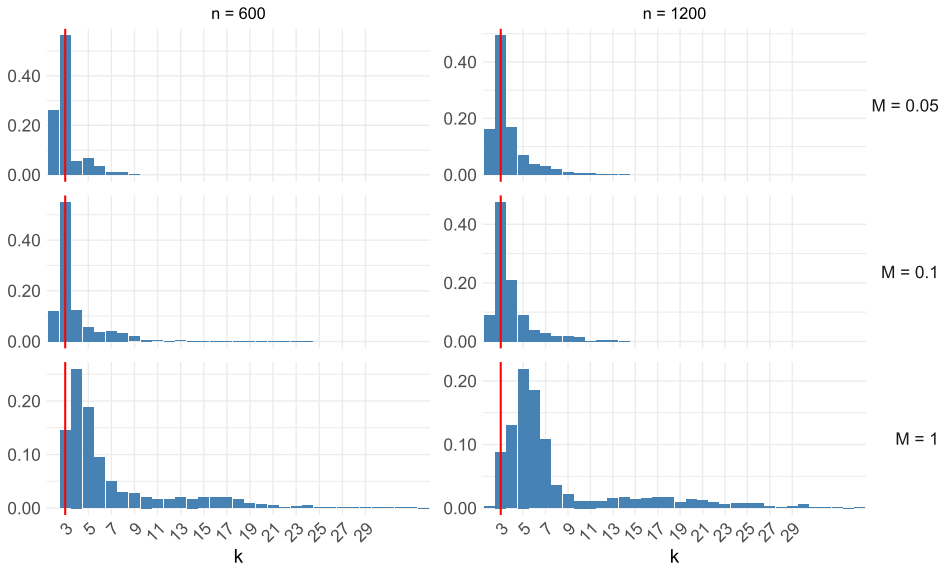

Fig. 2. Simulation study (setup 1). Distribution of posterior modes of the number of observed components across 100 simulated datasets for different values of M.

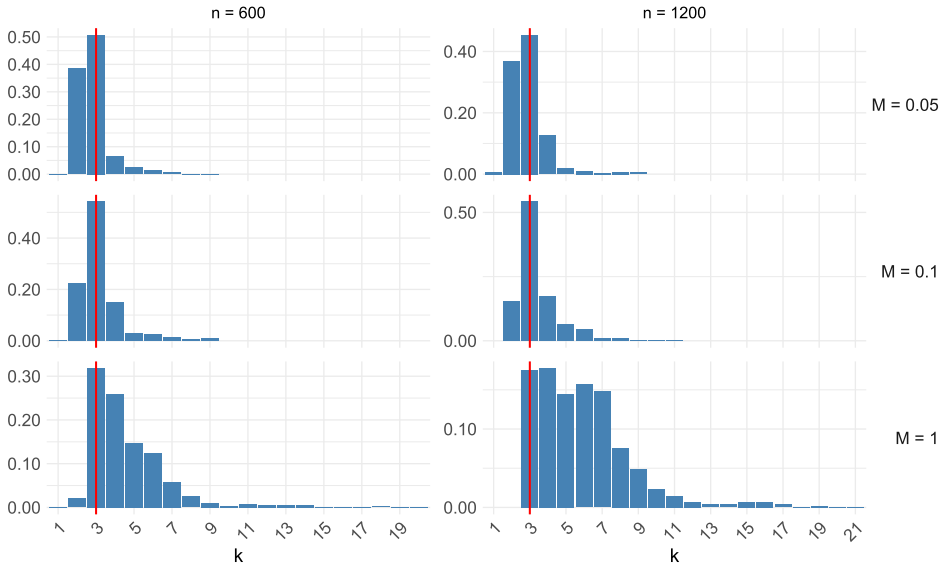

Fig. 3. Simulation study (setup 2). Distribution of posterior modes of the number of observed components across 100 simulated datasets for different values of M.

is especially difficult to detect. In the second setup, where the overall complexity increases due to the presence of two very similar mixture components, the model performance, as expected, slightly declines, but remains noteworthy. In particular, consistent with the findings reported in Section 4.1, for $n = 600$, the model with $M = 1$ recovers the true data generating partition with the best accuracy. However, for $n = 1200$, the best performance is achieved with $M = 0.1$, which provides an effective balance: it allows enough flexibility to detect a hidden cluster while maintaining the robustness required to handle a large sample size.

5. Cardiac allograft vasculopathy data

To analyze the performance of our models, we consider the CAV data set available in the `msm` package for R, see Jackson (2011). This data set has been used extensively in the literature on multi-state models to illustrate the fitting of new models or computational advancements (Sharples et al., 2003; Titman, 2011; Barone and Tancredi, 2022). The data document the progression of coronary

Table 1
Simulated data. Average Adjusted Rand Index across 100 replications.

setup 1	$n = 600$	$n = 1200$
$M = 0.05$	0.706	0.667
$M = 0.1$	0.708	0.737
$M = 1$	0.606	0.532
setup 2	$n = 600$	$n = 1200$
$M = 0.05$	0.535	0.525
$M = 0.1$	0.571	0.591
$M = 1$	0.582	0.566

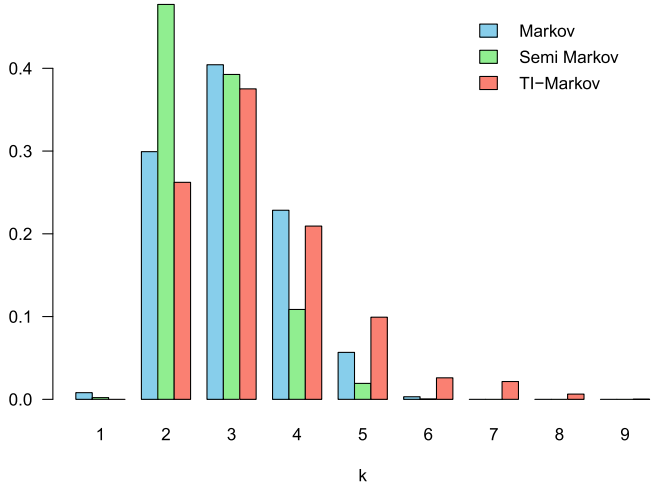

Fig. 4. CAV data. Posterior distributions of the number of observed mixture components k across MCMC iterations for the three fitted models.

allograft vasculopathy (CAV) in subjects who underwent heart transplant surgery. This disease leads to the deterioration of arterial walls and is a common cause of death post-transplant.

The data set provides the disease status (CAV-free (1), mild CAV (2), and moderate or severe CAV (3)) observed approximately every year after transplant for a set of 622 subjects. These subjects were followed up until their most recent visit if they were alive at the end of the observation period or until death (state (4)). Note that we use the version of the dataset provided in the `msm` R package. This version does not include information on the times at which patients stopped being under observation for survival follow-up. As a result, survival time may be underestimated, and some of the observed time-inhomogeneity may partially reflect bias induced by naïve censoring at the last recorded visit. These limitations should be taken into account when interpreting the results. The data include apparent transitions from higher to lower states, which are actually due to misclassification, as the deterioration of arterial walls is an irreversible process. We considered all the reverse transitions as remaining in the higher of the two states, and in the proposed models, we permit transitions only to adjacent states or to the death state. Note also that death times are observed exactly.

5.1. Results

We fitted to the CAV data the DP mixtures of Markov, time-inhomogeneous Markov, and semi-Markov models presented in the previous Sections. For all the three models, the DPM concentration parameter was fixed at $M = 0.25$ and the hyper-parameter vector for the Dirichlet distributions on the transition probabilities were set to 1. For each model, we ran 20,000 MCMC iterations of the MCMC sampler. Fig. 4 shows the posterior distributions of k_{obs}, i.e. the observed number of mixture components for the three models. For each model we may estimate the number of mixture components by the mode of the posterior distribution of k_{obs}. Tables 2 to 4 present the posterior estimates of the component parameters, conditioned on the estimated number of mixture components. Note that the cluster indices are denoted by ψ_i.

We first illustrate the results obtained by fitting the DPM of Markov model. We fix the hyper-parameters of the base gamma distributions to (0.1, 0.1). The model estimates three clusters, as indicated by the posterior distribution of k_{obs}. Posterior summaries conditioned on $k_{obs} = 3$ are reported in Table 2: from these estimates, we can observe a substantial difference among the three groups. In the first group, comprising approximately 94.6% of the individuals, we observe relatively low rates, with a slight increase in the departures from the second and third states, though they still remain at a low level. The posterior estimates of the transition probabilities suggest a smooth progression of the pathology, typically traversing all states before entering the absorbing state. The

Table 2
CAV data. Results for DPM of Markov models: posterior summaries conditional on the mode of the distribution of the observed clusters k.

	γ_1	γ_2	γ_3	p_{12}	p_{23}	p_{14}	p_{24}
$E(\cdot\|Y,\psi=1)$	0.128	0.292	0.278	0.787	0.808	0.213	0.192
$SD(\cdot\|Y,\psi=1)$	0.008	0.027	0.032	0.034	0.067	0.034	0.067
$q_{0.025}(\cdot\|Y,\psi=1)$	0.115	0.244	0.220	0.718	0.668	0.148	0.069
$q_{0.975}(\cdot\|Y,\psi=1)$	0.145	0.348	0.346	0.852	0.931	0.282	0.332
$E(\cdot\|Y,\psi=2)$	8.289	1.599	0.980	0.220	0.516	0.780	0.484
$SD(\cdot\|Y,\psi=2)$	5.194	2.505	1.281	0.201	0.285	0.201	0.285
$q_{0.025}(\cdot\|Y,\psi=2)$	0.169	0.006	0.027	0.009	0.030	0.208	0.024
$q_{0.975}(\cdot\|Y,\psi=2)$	19.438	8.609	4.215	0.792	0.976	0.991	0.970
$E(\cdot\|Y,\psi=3)$	7.858	1.565	1.056	0.267	0.496	0.733	0.504
$SD(\cdot\|Y,\psi=3)$	7.497	2.187	1.221	0.260	0.289	0.260	0.289
$q_{0.025}(\cdot\|Y,\psi=3)$	0.000	0.007	0.023	0.005	0.027	0.077	0.027
$q_{0.975}(\cdot\|Y,\psi=3)$	24.277	8.192	4.365	0.923	0.973	0.995	0.973

Table 3
CAV data. Results for DPM of time-inhomogeneous Markov models: posterior summaries conditional on the mode of the distribution of the observed clusters k.

	β_{01}	β_{02}	β_{03}	β_{11}	β_{12}	β_{13}	p_{12}	p_{23}	p_{14}	p_{24}
$E(\cdot\|Y,\psi=1)$	−2.405	−1.220	−1.955	0.097	−0.008	0.092	0.736	0.978	0.264	0.022
$SD(\cdot\|Y,\psi=1)$	0.140	0.197	0.237	0.027	0.031	0.026	0.040	0.045	0.040	0.045
$q_{0.025}(\cdot\|Y,\psi=1)$	−2.621	−1.580	−2.476	0.042	−0.061	0.043	0.668	0.921	0.173	0.001
$q_{0.975}(\cdot\|Y,\psi=1)$	−2.035	−0.866	−1.523	0.143	0.051	0.141	0.827	0.999	0.332	0.079
$E(\cdot\|Y,\psi=2)$	2.185	1.496	0.658	−2.027	−2.279	−0.185	0.601	0.453	0.399	0.547
$SD(\cdot\|Y,\psi=2)$	2.085	1.784	1.765	2.786	2.829	1.813	0.297	0.304	0.297	0.304
$q_{0.025}(\cdot\|Y,\psi=2)$	−4.060	−2.780	−2.888	−8.321	−7.541	−3.768	0.036	0.017	0.010	0.021
$q_{0.975}(\cdot\|Y,\psi=2)$	4.596	3.775	3.352	3.195	3.690	3.377	0.990	0.979	0.964	0.983
$E(\cdot\|Y,\psi=3)$	1.354	1.170	0.591	−1.035	−1.457	0.270	0.537	0.482	0.463	0.518
$SD(\cdot\|Y,\psi=3)$	2.528	1.804	1.632	2.533	2.663	1.883	0.316	0.302	0.316	0.302
$q_{0.025}(\cdot\|Y,\psi=3)$	−3.849	−3.030	−2.786	−5.692	−7.086	−3.202	0.018	0.019	0.017	0.020
$q_{0.975}(\cdot\|Y,\psi=3)$	4.607	3.805	3.254	3.470	3.550	4.319	0.983	0.980	0.982	0.981

second and third groups, composed of 3.2% and 2.2% of the sample, respectively, exhibit completely different behavior. Specifically, they show very high transition probabilities to the absorbing state from any state, especially from the first one, representing a subgroup of the population that, once affected by the pathology, faces sudden death. In particular, the second group exhibits a very high rate in the first state and a markedly elevated probability of transitioning from the first state to the absorbing state.

In the DPM of time-inhomogeneous Markov processes with transition intensity $q_{rs}(t) = p_{rs}\exp\left(\beta_{0r} + \beta_{1r}t\right)$ the prior for the regression coefficients is assumed to follow a Normal distribution centered at zero, with a standard deviation of 2. Table 3 shows the results. The first component (88.5% of the observations) shows negative intercepts and slightly positive temporal dependency parameters in the first and third states. We can associate the observations present in this cluster, as in the case of a Markovian process, with that fraction of the population that has a high probability of traversing all states before transitioning to the death state. In the second cluster (5.6% of the sample), we find a portion of the population exhibiting very high transition rates from each of the three states, that however decreases in a very strong way over time. There are high transition probabilities to the death state from each of the considered states. The last component, comprising 5.9% of the observed individuals, displays positive intercepts across all states, while temporal dependence parameters are negative in the first and second states, and positive in the third.

For the DPM of semi-Markov processes we assumed Weibull sojourn times as described in Section 2.2. The priors for the shape and rate parameters were assumed to follow a Gamma$(0.1, 0.1)$ distribution. For this model the mode of k_{obs} is two. Table 4 provides evidence of a slow progression group, comprising 92.88% of the observed individuals, represented by the first component and a fast progression group, accounting for 7.2% of the observed subjects, represented by the second component showing very high transition rates between states. For this model, the posterior summaries on the transition probabilities indicate that the first group is likely to slowly transition through all states, while the second group, unlike the results highlighted by the previous models, has a lower probability of entering the death state early, showing a very rapid progression through all states rather than a direct transition to the death state.

The results obtained from the implementation of the three mixture models are consistent with those reported by Barone and Tancredi (2022), presented in Table 5, who implemented the corresponding non-mixture specifications.

Finally, to assess the convergence of the proposed simulation algorithms, we conducted a comparative analysis using the DPM Markov case as a benchmark. Specifically, we compared the results obtained from the Gibbs sampler (Algorithm 1), which naturally fits the Markov framework by leveraging conjugacy with the base measure, with those produced by the no-gaps algorithm (Algorithm 2), applied under two alternative specifications: a DPM semi-Markov model with Weibull parameters fixed at 1, and a

Table 4
CAV data. Results for DPM of semi-Markov models: posterior summaries for the first and second clusters, conditional on the mode of the distribution of the observed clusters k.

	γ_1	γ_2	γ_3	α_1	α_2	α_3	p_{12}	p_{23}	p_{14}	p_{24}
$E(\cdot\|Y,\psi=1)$	0.132	0.271	0.273	1.332	0.961	1.026	0.737	0.883	0.263	0.117
$SD(\cdot\|Y,\psi=1)$	0.006	0.027	0.033	0.090	0.089	0.124	0.030	0.062	0.030	0.062
$q_{0.025}(\cdot\|Y,\psi=1)$	0.122	0.223	0.215	1.165	0.801	0.804	0.678	0.748	0.204	0.013
$q_{0.975}(\cdot\|Y,\psi=1)$	0.145	0.327	0.343	1.516	1.146	1.292	0.796	0.987	0.322	0.252
$E(\cdot\|Y,\psi=2)$	25.675	24.248	9.538	0.749	1.113	1.164	0.746	0.711	0.254	0.289
$SD(\cdot\|Y,\psi=2)$	15.114	17.240	7.863	0.301	1.334	2.124	0.265	0.313	0.265	0.313
$q_{0.025}(\cdot\|Y,\psi=2)$	2.440	0.177	0.118	0.323	0.194	0.184	0.063	0.041	0.002	0.002
$q_{0.975}(\cdot\|Y,\psi=2)$	61.970	62.065	28.296	1.540	4.210	5.512	0.998	0.998	0.937	0.959

Table 5
CAV data. Posterior summaries for standard Bayesian multi-state models as in Barone and Tancredi (2022).

Markov	p_{12}	p_{14}	p_{23}	p_{24}	γ_1	γ_2	γ_3			
$E(\cdot\|x)$	0.77	0.23	0.81	0.19	0.14	0.26	0.24			
$SD(\cdot\|x)$	0.03	0.03	0.06	0.06	0.01	0.02	0.03			
$q_{0.025}(\cdot\|x)$	0.71	0.18	0.69	0.08	0.12	0.22	0.19			
$q_{0.975}(\cdot\|x)$	0.82	0.29	0.91	0.30	0.15	0.30	0.29			
Semi Markov	p_{12}	p_{14}	p_{23}	p_{24}	γ_1	γ_2	γ_3	α_1	α_2	α_3
$E(\cdot\|x)$	0.71	0.29	0.85	0.15	0.14	0.28	0.28	0.94	0.92	0.99
$SD(\cdot\|x)$	0.03	0.03	0.08	0.08	0.01	0.03	0.04	0.05	0.08	0.13
$q_{0.025}(\cdot\|x)$	0.65	0.24	0.70	0.01	0.12	0.23	0.21	0.85	0.78	0.77
$q_{0.975}(\cdot\|x)$	0.76	0.35	0.99	0.30	0.15	0.33	0.37	1.04	1.10	1.27
Inohomogeneous	p_{12}	p_{14}	p_{23}	p_{24}	β_{0_1}	β_{1_1}	β_{0_2}	β_{1_2}	β_{0_3}	β_{1_3}
$E(\cdot\|x)$	0.67	0.33	0.98	0.02	−2.19	0.06	−1.24	−0.01	−2.00	0.10
$SD(\cdot\|x)$	0.03	0.03	0.02	0.02	0.09	0.02	0.18	0.03	0.23	0.02
$q_{0.025}(\cdot\|x)$	0.62	0.27	0.93	0.00	−2.36	0.02	−1.59	−0.07	−2.47	0.05
$q_{0.975}(\cdot\|x)$	0.73	0.38	1.00	0.07	−2.02	0.09	−0.88	0.04	−1.56	0.14

DPM non-homogeneous model with time regression parameters fixed at 0. In fact, both these models reduce to the standard Markov DPM case. The three algorithms yielded very similar results in terms of the posterior distribution of the number of clusters and model parameters, thereby confirming the reliability of our posterior sampling strategies. A comparison of the computational efficiency of the algorithms, however, falls outside the scope of this paper.

5.2. Model comparison

To perform model comparison, we evaluate the Watanabe-Akaike Information Criterion (WAIC), see Watanabe (2010) and Gelman et al. (2014), for each proposed multi-state DPM model and its corresponding non-mixture counterpart. Note that for a sample of independent observations $x_1, \dots, x_n$ from a model $p(x|\zeta)$ the WAIC is defined as:

$$\text{WAIC} = -2\sum_{i=1}^{n} \log\left(\mathbb{E}\left[p(x_i \mid \zeta)\right]\right) + 2\sum_{i=1}^{n} \text{Var}\left(\log p(x_i \mid \zeta)\right),$$

where the expectation and variance are computed with respect to the posterior distribution of the parameters. Hence, given the MCMC realizations of the model parameters ζ_r for $r = 1, \dots, R$, the WAIC is approximated by

$$\text{WAIC} \approx -2\sum_{i=1}^{n} \log\left(\frac{1}{R}\sum_{r=1}^{R} p(x_i \mid \zeta_r)\right) + 2\sum_{i=1}^{n}\left(\frac{1}{R}\sum_{r=1}^{R}\left(\log p(x_i \mid \zeta_r)\right)^2 - \left(\frac{1}{R}\sum_{r=1}^{R}\log p(x_i \mid \zeta_r)\right)^2\right).$$

To compute the WAIC for the proposed multi-state DPM models, we proceed as follows. For each partially observed trajectory $x_i = (x_{i_1}, \dots, x_{i,m_i})$, at each MCMC iteration we evaluate:

$$p(x_i \mid \zeta_r) = \sum_{l=1}^{k_r} \pi_{l,r}\, p(x_i \mid \theta^*_{l,r}), \tag{16}$$

where $\theta^*_{l,r}$, for $l = 1, \dots, k_r$, are the parameters of the k_r observed components drawn at iteration r, and $\pi_{l,r}$, for $l = 1, \dots, k_r$, are the mixture weights sampled from the posterior distribution of the truncated stick-breaking process associated with the DPM construction,

Table 6
CAV data. Model comparison using the Watanabe–Akaike Information Criterion (WAIC) for Markov, semi-Markov, and inhomogeneous Markov models, together with their DPM counterparts. The last two columns show the negative loglikelihood evaluated in the maximum a posteriori point and the corresponding penality term for the Akaike information criterion.

Model	$-2\sum_{i=1}^{n}\log E(p(x_i\|\zeta)$	$2\sum_{i=1}^{n} Var(log(p(x_i\|\zeta)))$	WAIC	$-2\sum_{i=1}^{n}\log p(x_i\|\hat{\zeta})$	2# of par.
Markov	3519.6	9.4	3529.1	3519.4	10
Semi-Markov	3517.3	17.7	3535.0	3516.9	16
TIM	3492.3	16.3	3508.5	3492.3	16
Markov-DPM	3456.6	23.7	3480.3	3454.2	34 (k = 3)
Semi-Markov DPM	3431.2	46.1	3477.3	3428.2	34 (k = 2)
TIM DPM	3424.7	29.1	3453.8	3422.2	52 (k = 3)

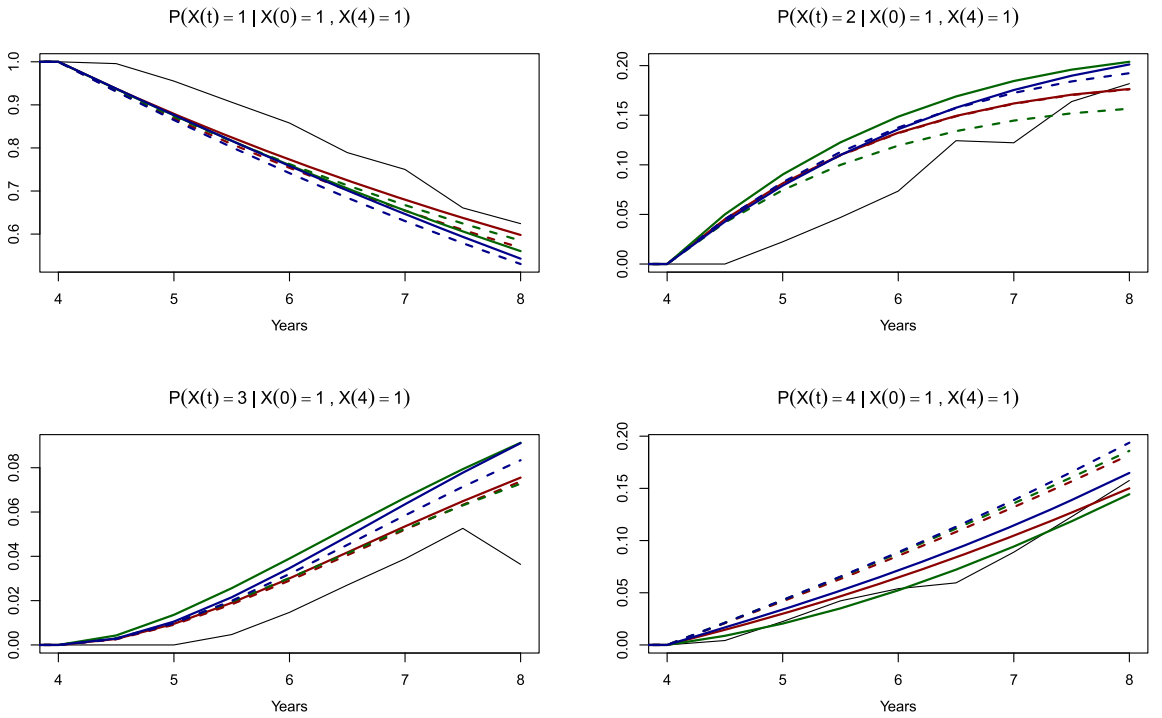

Fig. 5. CAV data. Conditional transition probabilities from state 1 at time 4, comparing DPM models (solid colored lines), non-mixture models (dashed colored lines), and empirical estimates (solid black lines). The red lines refer to the Markov specification, the green lines to the semi-Markov, and the blue lines to the inhomogeneous Markov. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

see Ishwaran and James (2001). These weights essentially correspond to the cluster proportions conditional on the cluster assignments observed at iteration r.

In the Markov case, the calculation of

$$p(x_i \mid \theta^*_{l,r}) = \prod_{j=2}^{m_i} p_{x_{i_j} x_{i_{j-1}}}(t_{i_j} - t_{i_{j-1}}; \Gamma^*_{l,r}),$$

can be efficiently performed using R packages that implement matrix exponentiation, such as `expm` or `msm` (Jackson, 2011). In contrast, for semi-Markov and non-homogeneous Markov models, the direct computation of $p(x_i \mid \theta^*_l)$ requires more sophisticated numerical approaches, which may significantly increase the computational burden. Nonetheless, thanks to the functions already present in the R packages `smms` (Aastveit et al., 2023) and `nhm` (Titman, 2011), this computation is feasible in our application.

Table 6 reports the WAIC values for each model specification. For each class of models, the DPM version achieves a lower WAIC compared to its parametric counterpart. Table 6 reports also the minimum of $-2\sum_{i=1}^{n}\log p(x_i \mid \zeta_r)$ observed across the MCMC iterations and the corresponding penalty under the standard Akaike information criterion considering the number of components fixed to the value k providing the maximum a posteriori estimate. Also following this hybrid approach the mixture models always outperform

the baseline counterpart. In addition, note that, in the fully Bayesian approach, the DPM of the time-inhomogeneous Markov model yields the lowest WAIC, suggesting that, for the data at hand, deviations from the standard Markovian assumptions present are best captured by this type of mixture model

As a final product of our data analysis, Fig. 5 shows the estimated probabilities $P(X(t) = j|X(s) = 1)$ for $s = 4$, $t \in [4, 8]$ and $j = 1, \ldots, 4$ obtained under the DPM models (solid colored lines) and the corresponding non-mixture counterparts (dashed colored lines). These estimates are also compared with the empirical ones (solid black lines) obtained considering all the patients that in the interval $[3.9, 4.1]$ were still in the state 1. The idea underlying the plot is to show the differences, among the proposed models, in the dynamic prediction of the future states, when we condition to an observed history up to a given time point. Similarly to the WAIC calculation, these estimated probabilities can be obtained by averaging across the MCMC iterations the quantities

$$P(X(t) = j|X(s) = 1, \zeta_r) = \sum_{l=1}^{k_r} P(X(t) = j|X(s) = 1, \theta^*_{l,r}) \frac{\pi_{l,r} P(X(s) = 1|\theta^*_r)}{\sum_{m=1}^{k_r} \pi_{m,r} P(X(s) = 1|\theta^*_r)}.$$

The final expression illustrates that, within the mixture model framework, the conditional probability of the future state at time t, given the most recently observed state at time s, is computed for each component and then weighted by the posterior probability of that component based on the observed history up to time s. As a result, components that better fit the observed data are given more weight in forecasting future states. Notably, in the selected scenario of Fig. 5 , this prediction mechanism leads to improved accuracy, especially for state 4, for which all three DPM models yield transition probability estimates that are closer to the empirical ones compared to the non-mixture models.

6. Conclusions

In this article, we have introduced a nonparametric Bayesian framework for continuous-time multi-state models observed at discrete time points. Specifically, we have proposed the use of Dirichlet Process Mixture models with Markov, semi-Markov, and time-inhomogeneous Markov kernels. This innovative approach offers the flexibility to capture new forms of unobserved heterogeneity. The effectiveness of our proposed method was demonstrated through its application to CAV data, a well-established dataset frequently used in the multi-state modeling literature. Specifically, the analysis revealed significant unobserved heterogeneity, indicating the presence of distinct subgroups within the population, each exhibiting different patterns of behavior over time. In particular, a predominant group of individuals follows a path of moderate disease progression, while a smaller subset of the population exhibits a notably higher probability of death in the early stages of the disease. This differentiation in disease progression underscores the utility of the proposed method in identifying and modeling the various trajectories individuals may experience, providing valuable insights for both clinical understanding and future interventions. The validity of the results was also supported by a model comparison conducted by the WAIC. For each kernel considered, the WAIC favors the CTMSM mixture model over the non-mixture counterpart, highlighting the suitability of the proposed models for analyzing this dataset.

In this work, we employed data augmentation techniques to reconstruct the underlying trajectories throughout our MCMC simulation algorithms. Note that in the Markov DPM case, the conjugacy between the base measure and the completely observed data likelihood allows for the derivation of the exact Gibbs sampler outlined in Section 3.2.1 and this is possible only under a data augmentation approach. Moreover, the use of a discrete data likelihood, except for the Markov case, is computationally demanding in a Bayesian approach since it requires solving integral or differential equations numerically at each MCMC iteration, see Titman (2011), Aastveit et al. (2023), Kendall et al. (2024). However, as highlighted by the WAIC evaluation, the calculations of the discrete data likelihood may be a necessary step even for a Bayesian analysis. Future research may further investigate the relative computational advantages and limitations of these two approaches.

Another promising direction for future research is the extension of the proposed approach to clustering panel data with censored or missing values. In this regard, Lin and Wang (2025) propose a modeling framework for longitudinal clinical data that accommodates mild outliers, censoring, and intermittent missingness, while Wang et al. (2025) present a flexible clustering approach based on parsimonious Gaussian mixtures with efficient estimation under the missing-at-random assumption. Together, these contributions offer valuable insights for future methodological developments.

Finally, following Barone and Tancredi (2022), we model the sojourn times in the semi-Markov case using a Weibull distribution which, generalizing the exponential distribution, reduces the semi-Markov CTMSM to the Markov process when the shape parameter equals one. However, the methodological structure of the proposed model and the implemented trajectory simulation algorithm are general enough to accommodate alternative specifications (e.g., log-normal, log-logistic), as illustrated by Klausch et al. (2023), who explored various distributions for transition times, many of which allow for conditional conjugacy.

Data availability

The CAV dataset is publicly available in the `msm` R package. The R Code is available at https://github.com/RosarioBarone.

Acknowledgments

This first author was partially supported by the Department of Excellence funds 2023–2027, Department of Economics and Finance, University of Rome “Tor Vergata” (UPB: DEF_ECC_2023–2027_Stefanucci; CUP: E83C23000320006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the funding institution.

References

Aastveit, M.E., Cunen, C., Hjort, N.L., 2023. A new framework for semi-Markovian parametric multi-state models with interval censoring. Stat. Methods Med. Res. 32 (6), 1100–1123.

Antoniak, C.E., 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. Ann. Stat. 2 (6),1152–1174.

Armero, C., Cabras, S., Castellanos, M.E., Perra, S., Quirós, A., Oruezábal, M.J., Sánchez-Rubio, J., 2012. Bayesian analysis of a disability model for lung cancer survival. Stat. Methods Med. Res. 25 (1), 336–351.

Ascolani, F., Lijoi, A., Rebaudo, G., Zanella, G., 2023. Clustering consistency with dirichlet process mixtures. Biometrika 110 (2), 551–558.

Barone, R., Dalla Valle, L., 2023. Bayesian nonparametric modeling of conditional multidimensional dependence structures. J. Comput. Graph. Stat. 32 (4), 1361–1370.

Barone, R., Farcomeni, A., 2025. Latent class multi-state quantile regression with a cure fraction: application to jail recidivism in the US. J. R. Stat. Soc. A Stat. Soc. 00 (0) 1-21.

Barone, R., Tancredi, A., 2022. Bayesian inference for discretely observed continuous time multi-state models. Stat. Med. 41 (19), 3789–3803.

Blumen, I., Kogan, M., McCarthy, P.J., 1955. The Industrial Mobility of Labor as a Probability Process. Cornell University.

Broomfield, J., Abrams, K.R., Freeman, S., Latimer, N., Rutherford, M.J., Crowther, M.J., Project HERCULES, t. C. I. N. R. G.i. , members, D. R. S.C., 2024. Modeling the multi-state natural history of rare diseases with heterogeneous individual patient data: a simulation study. Stat. Med. 43 (1), 184–200.

Cook, R.J., Lawless, J.F., 2018. Multistate Models for the Analysis of Life History Data. Chapman and Hall/CRC.

Dahl, D.B., 2006. Model-based clustering for expression data via a dirichlet process mixture model. Bayesian Inference for Gene Expression and Proteomics 4, 201–218.

Escobar, M.D., 1994. Estimating normal means with a dirichlet process prior. J. Am. Stat. Assoc. 89 (425), 268–277.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. 90 (430), 577–588.

Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. Ann. Stat. 1 (2), 209–230.

Frühwirth-Schnatter, S., Malsiner-Walli, G., 2019. From here to infinity: sparse finite versus dirichlet process mixtures in model-based clustering. Adv. Data Anal. Classif. 13, 33–64.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for bayesian models. Stat. Comput. 24, 997–1016.

Goodman, L.A., 1961. Statistical methods for the mover-stayer model. J. Am. Stat. Assoc. 56 (296), 841–868.

Guindani, M., Sepúlveda, N., Paulino, C.D., Müller, P., 2014. A bayesian semiparametric approach for the differential analysis of sequence counts data. J. R. Stat. Soc. C Appl. Stat. 63 (3), 385–404.

Hastie, D.I., Liverani, S., Richardson, S., 2015. Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. Stat. Comput. 25 (5), 1023–1037.

Hobolth, A., Stone, E.A., 2009. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. Ann. Appl. Stat. 3 (3), 1204–1231.

Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. 96 (453), 161–173.

Ishwaran, H., James, L.F., 2003. Generalized weighted chinese restaurant processes for species sampling mixture models. Stat. Sin. 13 (4), 1211–1235.

Ishwaran, H., Zarepour, M., 2000. Markov chain Monte Carlo in approximate dirichlet and beta two-parameter process hierarchical models. Biometrika 87 (2), 371–390.

Jackson, C.H., 2011. Multi-state models for panel data: the MSM package for R. J. Stat. Softw. 38 (8), 1–29.

Kalbfleisch, J.D., Lawless, J.F., 1985. The analysis of panel data under a Markov assumption. J. Am. Stat. Assoc. 80 (392), 863–871.

Kalli, M., Griffin, J.E., Walker, S.G., 2011. Slice sampling mixture models. Stat. Comput. 21 (1), 93–105.

Kang, M., Lagakos, S.W., 2006. Statistical methods for panel data from a semi-Markov process, with application to HPV. Biostatistics 8 (2), 252–264.

Kendall, E.B., Williams, J.P., Hermansen, G.H., Bois, F., Thanh, V.H., 2024. Beyond time-homogeneity for continuous-time multistate Markov models. J. Comput. Graph. Stat. 34 (2) , 668–682.

Klausch, T., Akwiwu, E.U., van de Wiel, M.A., Coupé, V. M.H., Berkhof, J., 2023. A Bayesian accelerated failure time model for interval censored three-state screening outcomes. Ann. Appl. Stat. 17 (2), 1285–1306.

Lawless, J., 2013. The design and analysis of life history studies. Stat. Med. 32 (13), 2155–2172.

Lin, T.-I., Wang, W.-L., 2025. Multivariate contaminated normal linear mixed models applied to alzheimer's disease study with censored and missing data. Stat. Methods Med. Res. 34 (3), 490–507.

Lo, A.Y., 1984. On a class of bayesian nonparametric estimates: I. density estimates. Ann. Stat. 12 (1), 351–357.

Luo, Y., Stephens, D.A., 2021. Bayesian inference for continuous-time hidden markov models with an unknown number of states. Stat. Comput. 31 (5), 57.

Luo, Y., Stephens, D.A., Buckeridge, D.L., 2023. Bayesian clustering for continuous-time hidden Markov models. Can. J. Stat. 51 (1), 134–156.

Luo, Y., Stephens, D.A., Verma, A., Buckeridge, D.L., 2021. Bayesian latent multi-state modeling for nonequidistant longitudinal electronic health records. Biometrics 77 (1), 78–90.

MacEachern, S.N., Müller, P., 1998. Estimating mixture of dirichlet process models. J. Comput. Graph. Stat. 7 (2), 223–238.

Machado, R. J.M., van den Hout, A., Marra, G., 2021. Penalised maximum likelihood estimation in multi-state models for interval-censored data. Comput. Stat. Data Anal. 153, 107057.

Maystre, L., Wu, T., Sanchis-Ojeda, R., Jebara, T., 2022. Multistate analysis with infinite mixtures of Markov chains. In: Uncertainty in Artificial Intelligence. PMLR, pp. 1350–1359.

Miller, J.W., Harrison, M.T., 2014. Inconsistency of Pitman-Yor process mixtures for the number of components. J. Mach. Learn. Res. 15 (1), 3333–3370.

Neal, R.M., 2000. Markov chain sampling methods for dirichlet process mixture models. J. Comput. Graph. Stat. 9 (2), 249–265.

Pamminger, C., Frühwirth-Schnatter, S., 2010. Model-based clustering of categorical time series. Bayesian Anal. 5 (2), 345–368.

Papaspiliopoulos, O., Roberts, G.O., 2008. Retrospective Markov chain Monte Carlo methods for dirichlet process hierarchical models. Biometrika 95 (1), 169–186.

Putter, H., Van Houwelingen, H.C., 2015. Frailties in multi-state models: are they identifiable? Do we need them? Stat. Methods Med. Res. 24 (6), 675–692.

Sharples, L.D., Jackson, C.H., Parameshwar, J., Wallwork, J., Large, S.R., 2003. Diagnostic accuracy of coronary angiography and risk factors for post–heart-transplant cardiac allograft vasculopathy. Transplantation 76 (4), 679–682.

Titman, A.C., 2011. Flexible nonhomogeneous markov models for panel observed data. Biometrics 67 (3), 780–787.

Titman, A.C., 2014. Estimating parametric semi-Markov models from panel data using phase-type approximations. Stat. Comput. 24 (2), 155–164.

Titman, A.C., Sharples, L.D., 2010. Semi-Markov models with phase-type Sojourn distributions. Biometrics 66 (3), 742–752.

Van Dijk, N.M., 1992. Uniformization for nonhomogeneous Markov chains. Oper. Res. Lett. 12 (5), 283–291.

Walker, S.G., 2007. Sampling the dirichlet mixture model with slices. Commun. Stat. Simul. Comput. 36 (1), 45–54.

Wang, W.-L., Lachos, V.H., Chen, Y.-C., Lin, T.-I., 2025. Flexible clustering via gaussian parsimonious mixture models with censored and missing values. Test 34 (2), 1–28.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.