A Hidden Absorbing Semi-Markov Model for Informatively Censored Temporal Data: Learning and Inference

Ahmed M. Alaa[†] AHMEDMALAA@UCLA.EDU
[†]*Electrical Engineering Department*
University of California, Los Angeles (UCLA)
Los Angeles, CA 90095-1594, USA

Mihaela van der Schaar[†,*] MIHAELA.VANDERSCHAAR@ENG.OX.AC.UK
[*]*Department of Engineering Science*
University of Oxford
Parks Road, Oxford OX1 3PJ, UK

Editor: xxxxxxxxxxxxx

Abstract

Modeling continuous-time physiological processes that manifest a patient's hidden (and evolving) clinical states is a key step in approaching many problems in healthcare. In this paper, we develop the *Hidden Absorbing Semi-Markov Model* (HASMM): a versatile probabilistic model that is capable of capturing the modern electronic health record (EHR) data. Unlike existing models, an HASMM accommodates irregularly sampled, temporally correlated, and informatively censored physiological data, and can describe non-stationary clinical state transitions. Learning an HASMM from the EHR data is achieved via a novel *forward-filtering backward-sampling* Monte-Carlo EM algorithm that exploits the knowledge of the end-point clinical outcomes (informative censoring) in the EHR data, and implements the E-step by sequentially sampling the patients' clinical states in the reverse-time direction while conditioning on the future states. Real-time inferences are drawn via an efficient message-passing algorithm that operates on a virtually constructed discrete-time *embedded Markov chain* that mirrors the patient's continuous-time state trajectory. We illustrate the operation of the proposed algorithms using synthetic data, and demonstrate the utility of the HASMM model in a critical care prognosis setting using a real-world dataset for patients admitted to Ronald Reagan UCLA Medical Center.

Keywords: Hidden Semi-Markov Models, Medical Informatics, Monte Carlo methods.

1. Introduction

Modeling the latent clinical states of a patient using evidential physiological data is a ubiquitous problem that arises in many healthcare settings, including disease progression modeling (Schulam and Saria (2015); Mould (2012); Wang et al. (2014); Jackson et al. (2003); Sweeting et al. (2010); Liu et al. (2015)) and critical care prognosis (Moreno et al. (2005); Matos et al. (2006); Yoon et al. (2016)). Accurate physiological modeling in these settings confers an *instrumental value* that manifests in the ability to provide early diagnosis, individualized treatments and timely interventions (e.g. early warning systems in critical care hospital wards (Yoon et al. (2016)), early diagnosis and drug recommendation for Scleroderma pa-

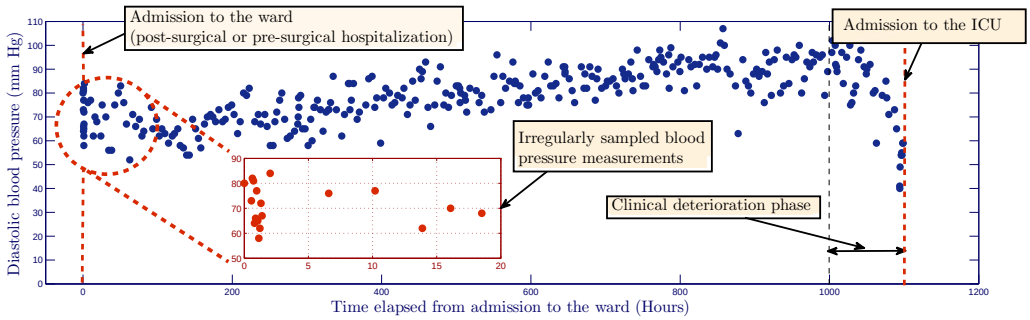

Figure 1: An episode of the diastolic blood pressure measurements (as recorded in the EHR) for a patient hospitalized in a regular ward for 50 days and then admitted to the ICU after the ward staff realized she is clinically deteriorating. Measurements are censored in accordance with the ICU admission time.

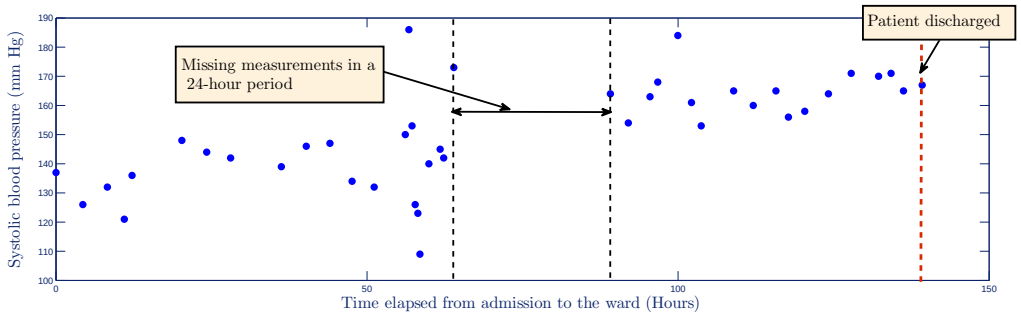

Figure 2: An episode of the systolic blood pressure measurements for a patient hospitalized in a regular ward for 6 days and then discharged home by the ward staff. Measurements are missing in a 24-hour period during the patient's stay in the ward.

tients (Varga et al. (2012)), early detection of a progressing breast cancer (Bartkova et al. (2005)), etc). Moreover, physiological modeling confers an *epistemic value* that manifests in the knowledge extracted from data about the progression and severity phases of a disease (Stelfox et al. (2012))), or the short-term dynamics of the physiological behavior of critically ill patients (Li-wei et al. (2013)). In this paper, we develop a versatile physiological model that fits a wide spectrum of healthcare settings, providing means for data-driven clinical diagnosis and prognosis that capitalize on the recent availability of data in the electronic health records (EHR)[1] (Gunter and Terry (2005)).

Modern EHRs comprise *episodic* data records for individual (anonymized) patients; every patient episode is a temporal sequence of clinical findings (e.g. visual field index for Glaucoma patients (Liu et al. (2015)), CD4 cell counts for HIV-infected patients (Guihenneuc-Jouyaux et al. (2000)), etc), lab test results (e.g. white cell blood count for post-operative

1. A recent data brief from the Office for National Coordinator (ONC) for healthcare technology shows that the adoption of EHR in US hospitals exhibited a spectacular increase from 9.4% in 2008, 27.6% in 2011, to 75.5% in 2014 (Charles et al. (2015)).

patients under immunosuppressive drugs (Cholette et al. (2012)), etc), or vital signs (e.g. blood pressure and O_2 saturation (Yoon et al. (2016))). The time span of these episodes may be as short as few days in short-term hospitalization episodes (e.g. patients with solid tumors, hematological malignancies or neutropenia who are hospitalized in a regular wards before or after a surgery (Kause et al. (2004); Hogan et al. (2012); Kirkland et al. (2013))), or as long as few years in longitudinal episodes (e.g. chronic obstructive pulmonary disease may evolve from a mild Stage I to a very severe Stage IV over a time span of 10 years (Pedersen et al. (2011); Wang et al. (2014))).

Hidden Markov Models (HMMs) and their variants have been widely deployed as a convenient machinery for modeling general dynamical systems with the latent states that manifest via noisy observation variables (Smyth (1994); Zhang et al. (2001); Giampieri et al. (2005); Genon-Catalot et al. (2000); Ghahramani and Jordan (1997)). Such models have achieved considerable success in various applications, such as topic modeling (Gruber et al. (2007)), speaker diarization (Fox et al. (2011)), and speech recognition (Rabiner (1989)). However, the nature of the clinical state estimation problem, together with the format of the modern EHR data pose the following set of serious challenges that confound classical HMM models:

1- Non-stationarity: Recently developed disease progression models, such those in (Wang et al. (2014)) and (Liu et al. (2015)), use conventional stationary Markov chain models for state transitions. In particular, they assume that state transition probabilities are independent of time. However, this assumption is seriously at odds with even casual observational studies which show that the probability of transiting from the current state to another state depends on the time spent in the current state (Lagakos et al. (1978); Huzurbazar (2004); Gillaizeau et al. (2015)). This effect, which violates the memorylessness assumptions adopted by continuous-time Markovian models, was verified in patients who underwent renal transplantation (Foucher et al. (2007, 2008)), patients who are HIV infected (Joly and Commenges (1999); Dessie (2014); Foucher et al. (2005)), and patients with chronic obstructive pulmonary disease (Bakal et al. (2014); Wang et al. (2014)).

2- Irregularly spaced observations: The times at which the clinical findings of a patient (vital signs or lab tests) are observed is controlled either by clinicians (in the case of hospitalized inpatients), or by the patient's visit times (in the case of a chronic disease follow up). The time interval between every two measurements may vary from one patient to another, and may also vary for the same patient within her episode. This is reflected in the structure of the episodes in the EHR records, as shown in Figure 1 and 2. Figure 1 depicts an actual diastolic blood pressure episode for a patient hospitalized in a regular ward for 1200 hours (50 days)[2]. The patient's stay in the ward was concluded with an admission to the ICU after the ward staff realized she was clinically deteriorating. As we can see, the blood pressure measurements in the first 20 hours were initially taken with a rate of 1 sample per hour, and then later the rate changed to 1 sample every 5 hours[3].

2. A detailed description for the data involved in this paper is provided in Section 4.
3. While Figure 1 illustrates a short-term episode for a critical care patient, similar effects are experienced in longitudinal episodes for patients with chronic disease (see Figure 4 in (Wang et al. (2014))).

An inference algorithm that runs in real-time for that patient must reason about her latent state while considering not only the blood pressure measurements, but also the times at which these samples where gathered. Thus, a direct application of a regular, discrete-time HMM (e.g. the models in (Murphy (2002); Rabiner (1989); Yu (2010); Matos et al. (2006); Guihenneuc-Jouyaux et al. (2000))) will not suffice for jointly describing the latent states and observations, and hence ensuring accurate inferences.

3- Discrete observations of a continuous-time phenomena: A patient's physiological signals and latent states evolve in continuous time; however, the observed physiological measurements are gathered at discrete time steps. The intervals between observed measurements can vary quite significantly; as we can see in Figure 2, the systolic blood pressure for a patient who stayed in a ward for 140 hours exhibits an entire day without measurements [4]. This means that the patient may encounter multiple hidden state transitions without any associated observed data. These effects will render more complicated learning and inference problems since the inference algorithms need to consider potential unobserved trajectories of state evolution between every two timestamps. This challenge, which has been ignored by the older literature (Jackson et al. (2003); Guihenneuc-Jouyaux et al. (2000)), was recently addressed in (Nodelman et al. (2012); Wang et al. (2014); Liu et al. (2015)), but only on the basis of memoryless Markov chain models for the hidden states, for which tractable inferences that rely on the solutions to Chapman-Kolmogorov equations can be executed. However, incorporating non-stationarity in state transitions (i.e. addressing challenge (1) in this list) would make the problem of reasoning about a continuous-time process through discrete observations much more complicated, which creates the demand for new machinery to handle inferences in such a setting.

4- Lack of supervision: The episodes in the EHR may be labeled with the aid of domain knowledge (e.g. the stages and symptoms of some chronic diseases, such as chronic kidney disease (Eddy and Neilson (2006)), are known to clinicians and may be provided in the EHR). However, in many cases, including the case of (post or pre-operative) critical care patients, we do not have access to any labels for the patients' states. Hence, unsupervised learning approaches need to be used for learning model parameters from EHR episodes. We focus in this paper on problems where no labeling or domain knowledge is provided for the states in the EHR episodes. While unsupervised learning of discrete-time HMMs has been extensively studied and is well understood (e.g. the Baum-Welch EM algorithm is predominant in such settings (Zhang et al. (2001); Yu (2010); Rabiner (1989))), the problem of unsupervised learning of continuous-time models for which both the patient's states and state transition times are hidden is far less understood, and indeed far more complicated.

5- Censored observations: Episodes in the EHR are usually terminated by an informative intervention or event, such as death, ICU admission, discharge, etc. This is known as *informative censoring* (Scharfstein and Robins (2002); Huang and Wolfe (2002); Link (1989)). Unlike classical HMM settings where training sets comprise fixed length HMM sequence instances, a typical EHR dataset would comprise a set of episodes with different

4. This may have resulted due to the patient undergoing a surgery or an intervention, or because the EHR recording system accidentally did not receive the data from the clinicians during that day.

durations, and the duration of each episodes is itself informative of the entire state evolution trajectory. Learning in such settings requires novel algorithms that can efficiently compute the likelihood of observing a set of episodes conditioned on their durations and terminating states, which is not possible using the classical Baum-Welch algorithm (Rabiner (1989)).

In order to address the challenges above, we develop a new model –which we call the *Hidden Absorbing Semi-Markov Model* (HASMM)– as a versatile generative model for a patient's (physiological) episode as recorded in the EHR. The HASMM captures non-stationary transitions for a patient's clinical state via a continuous-time semi-Markov model with explicitly specified state sojourn time distributions. Informative censoring is captured via absorbing states that designate clinical endpoint outcomes (e.g. cardiac arrest, mortality, recovery, etc); entering an absorbing state of an HASMM stimulates censoring events (e.g. clinical deterioration leads to an ICU admission which terminates the physiological observations for a monitored patient in a ward, etc). The HASMM is as a segment model that accounts for the temporal correlations among the observation variables that are generated by the same hidden state during its sojourn period (Ostendorf et al. (1996)). Moreover, the HASMM models the physiological data gathering process (i.e. follow up visits, vital sign gathering, lab tests, etc) as an arbitrary point process, and hence it can handle irregularly sampled observation variables. An elaborate comparison between the HASMM and existing models is provided in Section 5.

Real-time inference in the setting we consider is very challenging; while conventional discrete-time forward-backward inference algorithms assume that states change only at observation times (Murphy (2002); Yu (2010))), inferences in a continuous-time setting must take into account the time intervals between the irregularly sampled observation variables, and reason about the latent state trajectories between every two observed variables. To that end, we develop efficient diagnostic and prognostic HASMM inference algorithms that can estimate a patient's latent state, and predict her future state trajectory in real-time. The inference algorithms deal with an irregularly and arbitrarily sampled continuous-time state evolution process by constructing a virtual, discrete-time *embedded Markov chain* that fully describes the patient's state transitions at observation times, including potential intermediate transitions that can take place between the observation times. The embedded Markov chain is constructed in an offline stage by solving a system of *Volterra integral equations of the second kind* using the *successive approximation* method; the solution to this system of equations, which parallels the Chapman-Kolmogorov equations in ordinary Markov chains, describe the HASMM's and semi-Markovian state transitions as observed at arbitrarily selected discrete timestamps.

Offline learning of the HASMM model parameters from patients' episodes in an EHR is a daunting task. Since the HASMM is a continuous-time model, we cannot directly use the classical Baum-Welch EM algorithms for learning its parameters (Rabiner (1989)). Moreover, the semi-Markovianity of an HASMM yields an intractable integral in the E-step of the Expectation-Maximization (EM) formulation, and since the HASMM's state transitions are not captured by the conventional continuous-time Markov chain transition rate matrices, we cannot make use of the *Expm* and *Unif* methods that were used in (Hobolth and

Jensen (2011)), and more recently in (Liu et al. (2015)) for evaluating the integrals involved in the E-step of learning continuous-time HMMs. To address this challenge, we develop a novel *forward-filtering backward-sampling Monte Carlo EM* (FFBS-MCEM) algorithm that approximates the integral involved in the E-step by efficiently sampling the latent clinical trajectories conditioned on observations in the EHR by exploiting the informative censoring of the patients' episodes. The FFBS-MCEM algorithm samples the latent clinical states of every (offline) patient episode in the EHR as follows: it starts from the known clinical endpoints, and sequentially samples the patient's states by traversing in the reverse-time direction while conditioning on the future states, and then uses the sampled state trajectories to evaluate a Monte Carlo approximation for the E-step.

The rest of the paper is organized as follows. In Section 2, we present the HASMM model. Inference and learning algorithms are developed in Section 3. In Section 4, we conduct a set of experiments on synthetic data to illustrate the operation and performance of the proposed learning and inference algorithms, and we demonstrate the utility of the HASMM in the problem of critical care prognosis using a real-world dataset for patients admitted to Ronald Reagan UCLA Medical Center. Conclusions are drawn in Section 5.

2. The Hidden Absorbing Semi-Markov Model (HASMM)

In this section, we introduce the basic abstract structure of the continuous-time HASMM (Subsection 2.1), and then we propose the distributional specifications for the model's variables (Subsection 2.2).

2.1 Abstract Model

We start by describing the HASMM's hidden state evolution process, and then we describe the structure of its observable variables.

2.1.1 Hidden States

We consider a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{R}_+}, \mathbb{P})$, over which a continuous-time stochastic process $X(t)$ is defined on $t \in \mathbb{R}_+$. The process $X(t)$ corresponds to a temporal trajectory of the patient's hidden clinical states, which take on values from a finite *state-space* $\mathcal{X} = \{1, 2, \ldots, N\}$. Because the process $X(t)$ takes on only finitely many values, it can be decomposed in the form[5]

$$X(t) = \sum_n X_n \cdot \mathbf{1}_{\{\tau_n \le t < \tau_{n+1}\}}, \tag{1}$$

where $(X(t))_{t\in\mathbb{R}_+}$ is a càdlàg path (i.e. right-continuous with left limits), and the interval $[\tau_n, \tau_{n+1})$ is the time interval accommodating the n^{th} hidden state of the system, which takes on a value $X_n \in \mathcal{X}$. Every path $(X(t))_{t\in\mathbb{R}_+}$ on the stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{R}_+}, \mathbb{P})$ is a *semi-Markov path* (Janssen and De Dominicis (1984); Durrett (2010)), where the *sojourn time* of every state n, which we denote as $S_n = \tau_{n+1} - \tau_n$, is drawn from a *state-specific* distribution $v_j(S_n = s\,|\lambda_j) = d\mathbb{P}(S_n = s\,|X_n = j\,)$, with λ_j being a state-specific *duration*

5. By convention, we set $\tau_1 = 0$.

parameter associated with state $j \in \mathcal{X}$. Unlike ordinary time-homogeneous semi-Markov transitions, where the transition probabilities among states are assumed to be constant conditioned on there being a transition from the current state (Gillaizeau et al. (2015); Murphy (2002); Johnson and Willsky (2013); Yu (2010); Dewar et al. (2012); Guédon (2007)), our model accounts for *duration-dependent* semi-Markov transitions. In other words, the transition probability from one state to another depends on the time elapsed in the current state, i.e.

$$\mathbb{P}(X_{n+1} = j | X_n = i, S_n = s) = g_{ij}(s), \tag{2}$$

where $g_{ij} : \mathbb{R}_+ \to [0,1]$, $\forall i, j \in \mathcal{X}$ is a *transition function* for which $\frac{\partial g_{ij}(s)}{\partial s}$ is well defined, and $\sum_{j=1}^{N} g_{ij}(s) = 1, \forall s \in \mathbb{R}_+, i \in \mathcal{X}$.

Now consider the bi-variate (renewal) process $(X_n, S_n)_{n\in\mathbb{N}_+}$, which comprises the sequence of states and sojourn times. The semi-Markovian nature of $X(t)$ implies that $(X_n, S_n)_{n\in\mathbb{N}_+}$ satisfies the following condition on its transition probabilities

$$\begin{aligned}
\mathbb{P}(X_{n+1} = j, S_n \le s | \mathcal{F}_{\tau_n^-}) &= \mathbb{P}(X_{n+1} = j, S_n \le s | X_n = i) \\
&= \mathbb{P}(X_{n+1} = j | X_n = i, S_n \le s) \cdot \mathbb{P}(S_n \le s | X_n = i) \\
&= \mathbb{E}_{S_n}\left[g_{ij}(S_n) | S_n \le s\right] \cdot V_i(s|\lambda_i) \\
&= \bar{g}_{ij}(s) \cdot V_i(s|\lambda_i),
\end{aligned} \tag{3}$$

where $V_i(.)$ is the cumulative distribution function of state i's sojourn time, and $\bar{g}_{ij}(s)$ is the probability mass function that reflects the probability that a patient's next state being j given that she was at state i and her sojourn time in i is less than (or equal to) s. Based on (3), we define the *semi-Markov transition kernel* as a matrix-valued function $\mathbf{Q} : \mathbb{R}_+ \to [0,1]^{N\times N}$, with entries $\mathbf{Q}(s) = (Q_{ij}(s))_{i,j\in\mathcal{X}}$ that are given by

$$Q_{ij}(s) = \bar{g}_{ij}(s) \cdot V_i(s|\lambda_i). \tag{4}$$

The semi-Markov kernel $\mathbf{Q}$ describes the dynamics of $X(t)$ in continuous time, and will play an important role in constructing efficient inference algorithms in Subsection 3.1.

Since patients can start their observable episode at an arbitrary clinical state (i.e. we only observe the physiological measurements starting from the time when the patients are hospitalized or start taking clinical tests), then it follows that the initial state X_1 is random[6]. The initial state distribution is given by

$$\mathbf{p}^o = [p_1^o, p_2^o, \ldots, p_N^o]^T,$$

where $p_j^o = \mathbb{P}(X(0) = j)$, and $\sum_{j=1}^{N} p_j^o = 1$.

The hidden states reflect different levels of clinical risk (i.e. progression stage indexes of a chronic disease or phases of clinical deterioration (Sweeting et al. (2010);Chen and Zhou (2011))). In that sense, state 1 is regarded as the "least risky state", and state N is regarded as the "most risky state". We define and interpret states 1 and N as follows:

6. We do not consider left-censored observations in this model.

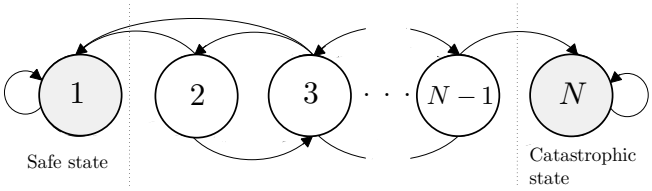

Figure 3: The Markov chain model for the HASMM.

- **State** 1 is denoted as the *safe state*, and represents the state at which the patient is at minimum (or no) risk (e.g. clinically stable post-operative patient, etc).

- **State** N is denoted as the *catastrophic state*, and represents the state at which the patient is at severe risk or encounters an adverse event (e.g. a very severe stage of a chronic disease (Bakal et al. (2014)), a cardiac or respiratory arrest (Subbe et al. (2001)), mortality (Knaus et al. (1991)), etc).

We assume that whenever the system enters either state 1 or state N, it remains there forever[7]. Therefore, we model states $\{1, N\}$ as *absorbing states*, whereas we model the remaining states in $\mathcal{X}/\{1, N\}$ as *transient states* that represent intermediate levels of risk. Following the assumptions in (Murphy (2002); Johnson and Willsky (2013)), we eliminate the self-transitions for all transient states by setting $g_{ii}(s) = 0, Q_{ii}(s) = 0, \forall s \in \mathbb{R}_+, i \in \mathcal{X}/\{1, N\}$, whereas we restrict the transitions from states 1 and N to self-transitions only, i.e. $g_{ii}(s) = 0, i \in \{1, N\}$. Figure 3 depict the Markov chain model for the sequence $\{X_n\}_{n\in\mathbb{N}_+}$.

We define $\mathcal{A}_1$ as the event that the path $(X(t))_{t\in\mathbb{R}_+}$ is absorbed in the safe state 1, i.e. $\mathcal{A}_1 = \{\lim_{t\to\infty} X(t) = 1\}$, and $\mathcal{A}_N$ as the event that $(X(t))_{t\in\mathbb{R}_+}$ is absorbed in the catastrophic state N, i.e. $\mathcal{A}_N = \{\lim_{t\to\infty} X(t) = N\}$. Since $(X(t))_{t\in\mathbb{R}_+}$ is an absorbing semi-Markov chain[8], we know that $\mathbb{P}(\mathcal{A}_1 \vee \mathcal{A}_N) = 1$, and since the events $\mathcal{A}_1$ and $\mathcal{A}_N$ are mutually exclusive, it follows that $\mathbb{P}(\mathcal{A}_N) = 1 - \mathbb{P}(\mathcal{A}_1)$. The quantity $\mathbb{P}(\mathcal{A}_N)$ describes a patient's prior risk of ending in the catastrophic state, whereas $\mathbb{P}(\mathcal{A}_N \,|\mathcal{F}_t)$ describes the patient's posterior risk of ending in the catastrophic state having observed its evolution history up to time t[9]. Define T_s as an $\mathcal{F}$-stopping time representing the absorption time of

7. The model can be easily extended to accommodate an arbitrary number of competing absorbing states.
8. We assume that the transition functions $g_{ij}(s)$ for any transient state i is non-zero for every s. Hence, it follows that $(X(t))_{t\in\mathbb{R}_+}$ is an absorbing semi-Markov chain since it has 2 absorbing states, each of which can be visited starting from any other state (Durrett (2010)).
9. In the clinical applications under consideration, transient states can be ordered by their respective relative risks of encountering event $\mathcal{A}_N$ in the subsequent transitions, i.e. in a 5-state chain, it is more likely for the patient to be absorbed in state 5 in the future when it is in state 4 than when it is in state 3. For instance, it is more likely for a patient's chronic obstructive pulmonary disease that is currently assessed to have a severity degree of GOLD1 (mild severity as defined in the GOLD standard Pedersen et al. (2011)) to progress (in the near future) to a severity degree of GOLD2 (moderate) rather than GOLD3 (severe).

the path $(X(t))_{t\in\mathbb{R}_+}$ in either state 1 or state N[10], i.e.

$$T_s = \inf\{t \in \mathbb{R}_+ : X(t) \in \{1, N\}\}.$$

Finally, we define K as the (random) number of state realizations in the sequence $\{X_n\}_{n=1}^K$ up to the stopping time T_s, which has to be concluded by either state 1 or N, e.g. when $\mathcal{X} = 4$, the sequences $\{1\}, \{4\}, \{2, 3, 2, 3, 4\}$, and $\{3, 2, 1\}$ are valid, random-length realizations of $\{X_n\}_{n=1}^K$, and each represents a certain state evolution trajectory for the patient.

2.1.2 Observations and Censoring

The path $(X(t))_{t\in\mathbb{R}_+}$ is unobservable; what is observable is a corresponding process $(Y(t))_{t\in\mathbb{R}_+}$ on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{R}_+}, \mathbb{P})$, the values of which are drawn from an *observation-space* $\mathcal{Y}$, and whose distributional properties are dependent on the latent states' path $(X(t))_{t\in\mathbb{R}_+}$. The observable process $(Y(t))_{t\in\mathbb{R}_+}$ can be put in the form

$$Y(t) = \sum_n Y_n(t) \cdot \mathbf{1}_{\{\tau_n \le t < \tau_{n+1}\}}, \tag{5}$$

where $(Y(t))_{t\in\mathbb{R}_+}$ is a càdlàg path, comprising a sequence of function-valued variables $\{Y_n(t)\}_{n=1}^K$, with $Y_n : [\tau_n, \tau_{n+1}) \to \mathcal{Y}$. Even though the path $(Y(t))_{t\in\mathbb{R}_+}$ is accessible (observable), only a sequence of irregularly spaced samples of it is observed over time, and is denoted by $\{Y(t_m)\}_{t_m\in\mathcal{T}}$, where $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$ is the set of observed measurements, and M is the total number of such measurements. We say that the process is censored if $M < \infty$; typical episodes in an EHR are censored: observations stop at some point of time due to a release from care, an ICU admission, mortality, etc.

The sampling times in $\mathcal{T}$ represent the times at which a patient with a chronic disease took clinical tests (i.e. time intervals in $\mathcal{T}$ spans years), or the times at which clinicians have gathered vital signs for a monitored critically ill patient in a hospital ward (i.e. time intervals in $\mathcal{T}$ span days or hours). We assume that the sampling times in $\mathcal{T}$ are drawn from a *point-process* $\Phi(\zeta) = \sum_{m\in\mathbb{N}_+} \delta_{t_m}$, which is defined on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{N}}, \mathbb{P})$, and with δ_t being the Dirac measure. The point process $\Phi(\zeta)$ is parametrized by an intensity parameter ζ, but is assumed to be independent of the latent states path[11]. Define $\mathcal{T}_n$ as the set of M_n samples that are gathered during the interval $[\tau_n, \tau_{n+1})$[12], i.e. $\mathcal{T}_n = \{t_m : t_m \in \mathcal{T}, t_m \in [\tau_n, \tau_{n+1})\}, M_n = |\mathcal{T}_n|$, and $\sum_n M_n = M$. Since $\mathcal{T}_n$ could possibly be empty ($\mathcal{T}_n = \emptyset$), some states can have no corresponding observations (i.e. an inpatient may exhibit a transition to a deteriorating state during the night without her blood pressure

10. Since the sequence $\{X_n\}_n$ is almost surely stopped (i.e. $T_s < \infty$ with probability 1), then the number of transitions that exhibited by $\{X_n\}_n$ until the absorption time T_s (i.e. number of jumps in $X(t)$) is almost surely finite.
11. This means that the sampling times are uninformative of the latent states, which makes the inference problem more challenging. The HASMM model can be easily extended to incorporate a state-dependent sampling process using a Cox process (Lando (1998)) or a Hawkes process (Hawkes and Oakes (1974)) to modulate the intensity parameter ζ.
12. Note that what is observed is a sequence of sampling times $\mathcal{T}$, the elements of which are not labeled by the corresponding state indexes, for that the states are latent, i.e. the sets $\mathcal{T}_n$ are latent.

being measured. Recall the illustration in Figure 2).

The paths $\{Y_n(t)\}_{n=1}^K$ are assumed to be conditionally independent given the hidden states' sequence $\{X_n\}_{n=1}^K$, and hence we have that

$$\{Y(t_m)\}_{t_m\in\mathcal{T}_n} \perp\!\!\!\perp \{Y(t_m)\}_{t_m\in\mathcal{T}_{n+1}} \,|\, X_n, X_{n+1}\,, \forall n \in \{1,2,\ldots,K-1\}.$$

The observed samples generated under every state X_n and sampled at the times in $\mathcal{T}_n$ are drawn from $\mathcal{Y}$ according to a distribution $\mathbb{P}(\{Y(t_m)\}_{t_m\in\mathcal{T}_n}|X_n=j,\Theta_j)$, where Θ_j is an *emission parameter* that controls the distributional properties of the observations generated under state j.

The number of observation samples is finite: the observed sequence is *censored* at some point of time, which we call the censoring time T_c, after which no more observation samples are available. Censoring reflects an external intervention/event that terminated the observation sequence, i.e. mortality event, intensive care unit (ICU) admission, etc. We assume that censoring is *informative* (Scharfstein and Robins (2002);Huang and Wolfe (2002);Link (1989)), i.e. the censoring time is correlated with the absorption time T_s, and T_s strictly precedes T_c (in an almost sure sense). That is, T_c is an $\mathcal{F}$-stopping time that is given by $T_c = T_s + S_K$, i.e. once the patient enters state 1 or state N, the observations stop after the patient's sojourn time in that state (i.e. observations stop after a time S_K from the entrance in the absorbing state). Therefore, the duration distributions $v_1(s|\lambda_1)$ and $v_N(s|\lambda_N)$ of states 1 and N are used to determine the censoring times conditioned on the chain $\{X_n\}_{n=1}^K$ being absorbed at time T_s.

Every sample from the HASMM is an episode comprising a random-length sequence of hidden states $\{X_n\}_{n=1}^K$, and a random-length sequence of observations $\{Y(t_m)\}_{m=1}^M$ together with the associated observation times. We only observe $\{Y(t_m)\}_{m=1}^M$; the latent states' path $X(t)$, the number of realized states K, the association between observations and states (i.e. the sets $\mathcal{T}_n$) are all unobserved, which makes the inference problem very challenging, but captures the realistic EHR data format and the associated inferential hurdles. In the next subsection, we specify the model's generative process and present an algorithm to generate episodic samples from an HASMM.

2.2 Model Specification and Generative Process

As have been discussed in Subsection 2.1, the hidden and observables variables of an HASMM can be listed as follows:

- **Hidden variables:** The hidden states sequence $\{X_n\}_{n=1}^K$ and the states' sojourn times $\{S_n\}_{n=1}^K$ (or equivalently, the transition times $\{\tau_n\}_{n=1}^K$).

- **Observable variables:** The observed episode $\{Y(t_m)\}_{m=1}^M$ and the associated sampling times $\mathcal{T}=\{t_m\}_{m=1}^M$.

The HASMM model parameters that generate both the hidden and observable variables are encompassed in the parameter set Γ, i.e.

$$\Gamma = \left(\underbrace{N}_{\text{State cardinality}}, \underbrace{\lambda = \{\lambda_j\}_{j=1}^{N}}_{\text{State duration}}, \underbrace{\mathbf{p}^o}_{\text{Initial states}}, \underbrace{\mathbf{Q} = \{Q_{ij}(s)\}_{i,j=1}^{N}}_{\text{Transitions}}, \underbrace{\mathbf{\Theta} = \{\Theta_j\}_{j=1}^{N}}_{\text{Emission}}, \underbrace{\zeta}_{\text{Sampling}} \right).$$

Since the point process $\Phi(\zeta)$ does not reveal any information about the latent states, and hence plays no role in inference, we will drop it from the parameter set Γ in the rest of the paper. In the following, we specify the distributional properties for both the hidden and observable variables.

2.2.1 Distributional specifications for the hidden variables

We model the state sojourn time of every state $i \in \mathcal{X}$ via a Gamma distribution. The selection of a Gamma distribution ensures that the generative process encompasses ordinary continuous-time Markov models for the path $(X(t))_{t\in\mathbb{R}_+}$, since the exponential distribution[13] is a special case of the Gamma distribution (Durrett (2010)). Thus, if the underlying physiology of the patient is naturally characterized by memoryless state transitions, this will be automatically learned from the data via the parameters of the Gamma distribution. The sojourn time distribution for state i is given by

$$v_i(s|\lambda_i = \{\lambda_{i,s}, \lambda_{i,r}\}) = \frac{1}{\Gamma(\lambda_{i,s})} \cdot \lambda_{i,r}^{\lambda_{i,s}} \cdot s^{\lambda_{i,s}} \cdot e^{-s\cdot\lambda_{i,r}}, s \geq 0,$$

where $\lambda_{i,s} > 0$ and $\lambda_{i,r} > 0$ are the shape and rate parameters of the Gamma distribution respectively.

Now we specify the structure of the transition kernel $\mathbf{Q}(s) = (Q_{ij}(s))_{i,j}, i, j \in \mathcal{X}$. Recall from (4) that the each element in the transition kernel matrix can be written as $\mathbb{E}_S\left[g_{ij}(S)|S \leq s\right] \cdot V_i(s|\lambda_i)$. Having specified the distribution $v_i(s|\lambda_i)$ as a Gamma distribution, it remains to specify the function $g_{ij}(s)$ in order to construct the elements of $\mathbf{Q}(s)$. The transition functions $(g_{ij}(s))_{i,j}$ are given by *Multinomial logistic* functions as follows

$$g_{ij}(s) = \frac{e^{(\eta_{ij}+\beta_{ij}\cdot s)}}{\sum_{k=1}^{N} e^{(\eta_{ik}+\beta_{ik}\cdot s)}}, \tag{6}$$

where $\eta_{ik}, \beta_{ik} \in \mathbb{R}_+, \eta_{ik} = -\infty, \forall i = k$. The parameters $(\eta_{ij})_{j=1}^{N}$ determine the baseline values for the transition probability mass out of state i, i.e. $g_{ij}(0)$, whereas the parameters β_{ij} controls the rate with which this transition probability mass changes over time. If $\beta_{ij} = 0$, then we have that $g_{ij}(s) = g_{ij}(0) = \frac{e^{\eta_{ij}}}{\sum_{k=1}^{N} e^{\eta_{ik}}}, \forall s \in \mathbb{R}_+$, i.e. the transition probability out of state i remains constant irrespective of the sojourn time in that state. If $\beta_{ij} > 0$, then $g_{ij}(s)$ changes monotonically over time, with a rate that is increasing in β_{ij}.

13. Note that a semi-Markov chain reduces to a Markov chain if the sojourn times are exponentially distributed.

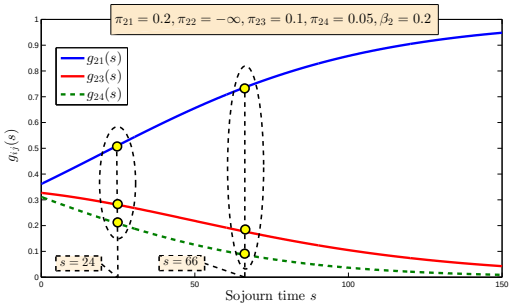

Figure 4: Exemplary transition functions $(g_{2j})_{j=1}^{4}$ for a 4-state HASMM.

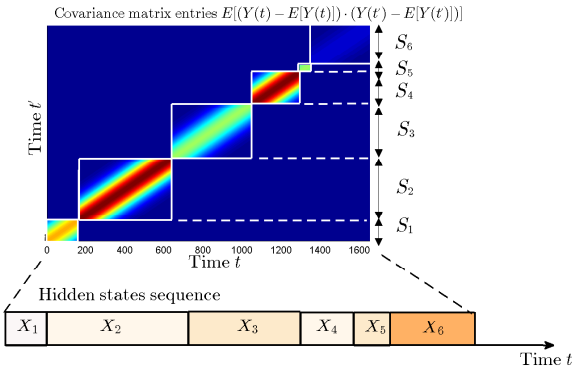

Figure 5: Depiction for the correlation structure of the observable variables for an underlying state sequence $\{X_n\}_{n=1}^{6}$.

The parametrization for $g_{ij}(s)$ in (6) captures the effect of the patient's sojourn time in a certain state on the uncertainty about her future states. The parameter β_{ij} biases the transitions to a specific state as more time elapses in the current state, i.e. the more time the patient spends in the current state, the less uncertain we are about her next state. To see how this effect materialize in the definition of $g_{ij}(s)$, we note that $\lim_{s\uparrow\infty} g_{ij}(s) = \mathbf{1}_{\{\eta_{ij}=\max_k \eta_{ik}\}}$. That is, when the sojourn time in state X_n becomes asymptotically large, i.e. $S_n \to \infty$, the uncertainty about the state X_{n+1} drops to zero[14]. Figure 4 depicts exemplary transition functions $(g_{ij}(s))_{i,j}$ for a 4-state HASMM. It can be seen that as the sojourn time increases ($\beta_{2j} = 0.2, \forall j$), the transition probabilities approaches a degenerate distribution that places a probability mass of 1 on a certain state (state 1 in this case).

2.2.2 Distributional specifications for the observable variables

As explained in Subsection 2.1, the observable process $Y(t)$ can be decomposed as $Y(t) = \sum_{n=1}^{K} Y_n(t) \cdot \mathbf{1}_{\{\tau_n \le t < \tau_{n+1}\}}$, where the paths $(Y_n(t))_{n=1}^{K}$ are conditionally independent given the state sequence $\{X_n\}_{n=1}^{K}$. Since observations are drawn from $Y(t)$ at arbitrarily, and irregularly spaced time instances $\mathcal{T}$, we have to model the distributional properties of $Y(t)$ in continuous time. Thus, we model every path $Y_n(t)$ defined over $[\tau_n, \tau_{n+1})$ as a segment drawn from a Gaussian Process (GP), with a parameter set Θ_i that depends on the corresponding latent state $X_n = i$ (Rasmussen (2006)). The GP associated with $X_n = i$ is parametrized by

14. Similar effects for the sojourn time on the transition probabilities has been demonstrated in the progression of breast cancer from healthy to preclinical states in (Taghipour et al. (2013)), where age (the main risk factor for breast cancer) was shown to affect the probability of progressing across the states of healthy to preclinical, clinical and death. These effects may be also prevailing in other diseases, or in critical care settings where the length of time during which a patient stays clinically stable may imply that the patient is more likely to transit to a more healthy state in the future. Through the HASMM model, we can recognize whether or not this effect is evident in the EHR data, i.e. whether the transition function reflects an underlying homogeneous (if $g_{ij}(s)$ is independent of s) or duration-dependent transitions by learning the parameter β_{ij}. Moreover, the parameter β_{ij} is defined per state; the HASMM model can capture scenarios where transitions are duration-independent from some states, but are duration-dependent from others.

a constant mean function $m_i(t) = m_i$ and a *squared-exponential* covariance kernel $k_i(t, t') = \sigma_i^2 \, e^{-\frac{1}{2\ell_i^2}||t-t'||^2}$; the GP parameters associated with state i are given by $\Theta_i = (m_i, \sigma_i, \ell_i)$, i.e. $Y_n(t)|X_n = i \sim \mathcal{GP}(\Theta_i)$. When $\mathcal{Y}$ is multidimensional, we adopt the *multitask GP* defined in (Bonilla et al. (2007)).

We note that the HASMM model is a *segment model* (Ostendorf et al. (1996); Murphy (2002); Yu (2010); Guédon (2007)), i.e. observation samples that are defined within the sojourn time of the same state are correlated, but observation samples in different states are independent. Figure 5 depicts the correlation structure of the observable variables in terms of the covariance matrix of a discrete version of $Y(t)$ generated under a specific hidden state sequence. We can see that conditioned on the hidden state sequence, the covariance matrix is a block diagonal matrix, where the sizes of the blocks are random and are determined by the hidden states' sojourn times.

The sampling times in $\mathcal{T}$ are generated by the point process $\Phi(\zeta)$, which for the sake of completeness of the model description, we specify as a Poisson process with an intensity parameter ζ. Note though that since we assume the sampling times are uninformative of the latent states path $X(t)$, the distributional specification of $\Phi(\zeta)$ is ancillary the inference and learning algorithms developed in Section 3.

2.2.3 Sampling episodes from an HASMM

We conclude this Section by presenting an Algorithm for sampling episodes from an HASMM with a parameter set Γ. Algorithm 1 (`GenerateHASMM(`Γ`)`)[15] samples a patient's episodes by first sampling an initial state from $\mathcal{X}$, and then sequentially samples sojourn times s from the Gamma distribution, and new states using the semi-Makrov kernel $\mathbf{Q}(s)$, until an absorbing state is drawn. Figure 6 depicts an episode that is sampled by Algorithm 1.

Algorithm 1 Sampling episodes from an HASMM

procedure GenerateHASMM(Γ)
  Input: HASMM model parameters $\Gamma = (N, \lambda, \mathbf{p}^o, \mathbf{Q}(s), \mathbf{\Theta}, \zeta)$
  Output: An episode $(\{X_n\}_{n=1}^K, \{\tau_n\}_{n=1}^K, \{Y(t_m)\}_{m=1}^M, \{t_m\}_{m=1}^M)$
  $\tau_1 \leftarrow 0$, $k \leftarrow 1$, $\mathcal{T} \sim \text{Poisson}(\zeta)$ ▷ Initializations
  $x_1 \sim \text{Multinomial}(p_1^o, p_2^o, \ldots, p_N^o)$ ▷ Sample an initial latent state
  $s_1 \sim \text{Gamma}(\lambda_{x_1,s}, \lambda_{x_1,r})$, $\tau_2 \leftarrow \tau_1 + s_1$
  $\mathcal{T}_1 = \{t \in \mathcal{T} : \tau_1 \leq t \leq \tau_2\}$
  while $x_k \notin \{1, N\}$ **do** ▷ Sample latent states until absorption
    $x_{k+1} \sim \text{Multinomial}(g_{x_k 1}(s_k), g_{x_k 2}(s_k), \ldots, g_{x_k N}(s_k))$
    $s_{k+1} \sim \text{Gamma}\left(\lambda_{x_{k+1},s}, \lambda_{x_{k+1},r}\right)$, $\tau_{k+2} \leftarrow \tau_{k+1} + s_{k+1}$
    $\mathcal{T}_{k+1} = \{t \in \mathcal{T} : \tau_{k+1} \leq t \leq \tau_{k+2}\}$
    $\{y(t_m)\}_{t_m \in \mathcal{T}_{k+1}} \sim \mathcal{GP}(\Theta_{x_{k+1}})$ ▷ Sample observations from a Gaussian Process
    $k \leftarrow k + 1$
  end while
  return $(\{x_n\}_{n=1}^K, \{\tau_n\}_{n=1}^K, \{y(t_m)\}_{m=1}^M, \{t_m\}_{m=1}^M)$
end procedure

15. Matlab codes are available at https://github.com/ahmedmalaa/JMLRHASMM.

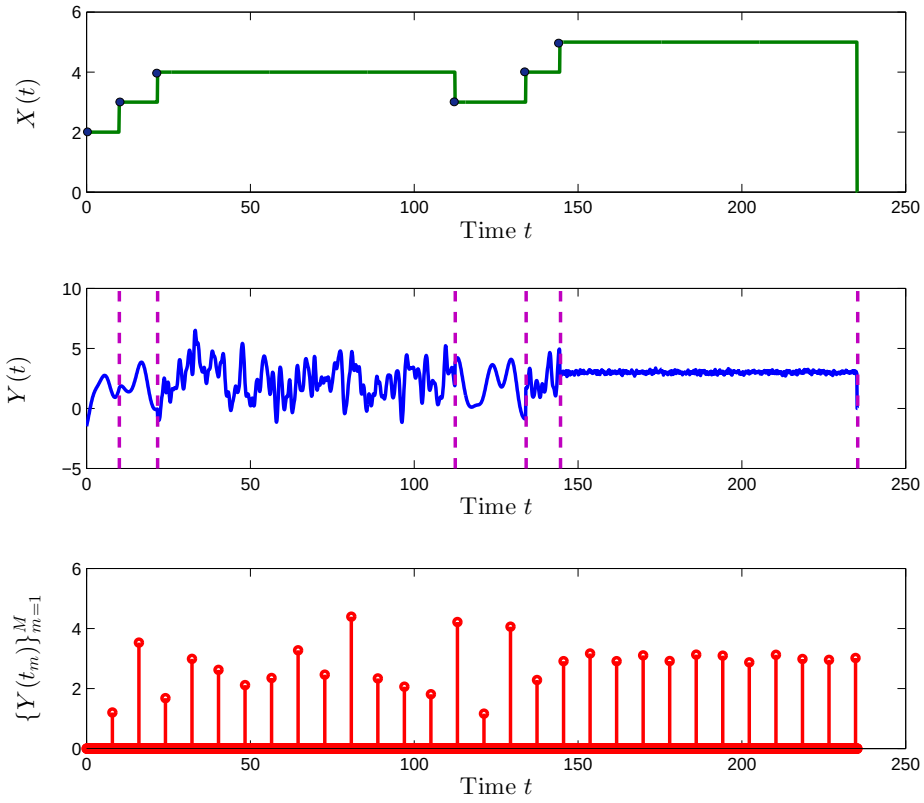

Figure 6: An episode generated by `GenerateHASMM(Γ)` with $N = 5$. The realized hidden state sequence (upper) is $\{2, 3, 4, 3, 4, 5\}$, and is absorbed in state 5. The patient's continuous time physiological signal $Y(t)$ (middle) is accessible by the clinicians, but only a set of discrete observations are gathered and observed $\{Y(t)\}_{m=1}^{M}$ (bottom).

3. Inference and Learning Algorithms

Our goal is to learn the HASMM model parameters that describes the patients' episodes from an offline EHR dataset $\mathcal{D}$ of previous patients, and then use the learned model to carry out diagnostic and prognostic inferences for new patients. A typical dataset $\mathcal{D}$ comprises D episodes, each of which contains only the patient's observable (temporal) variables (the bottom plot in Figure 6), i.e. the latent state trajectory $X(t)$ is hidden and the dataset is not labeled by the patient's state sequence. The format of the dataset $\mathcal{D}$ can be described as follows

$$\mathcal{D} = \left\{ \{y_m^d\}_{m=1}^{M^d}, \{t_m^d\}_{m=1}^{M^d}, T_c^d, l^d \right\}_{d=1}^{D},$$

where y_m^d is the m^{th} observed sample (e.g. m^{th} clinical finding, lab test, vital sign, etc) of the d^{th} patient, t_m^d is the time at which this sample was gathered, T_c^d is the censoring time for patient d's episode, and $l^d \in \{1, N\}$ is an endpoint label (e.g. mortality, ICU admission, etc), i.e. the state in which the patient's state trajectory is absorbed.

In Section 3.1, we develop online algorithms that carry out diagnostic and prognostic inferences for a monitored patient's episode in real-time. In particular, we are interested in the following inference tasks:

Inference tasks:

Given an ongoing realization of an episode $\{y(t_1), y(t_2), \ldots, y(t_m)\}$ at time t_m (before the censoring time T_c), and the HASMM model parameter Γ that has generated this realization (i.e. $\{y(t_1), y(t_2), \ldots, y(t_m)\}$ is sampled via the algorithm `GenerateHASMM(`Γ`)`), we aim at carrying out the following inference tasks:

- **Task 1 (Diagnosis):** Infer the patient's current clinical state, i.e. compute

$$\mathbb{P}(X(t_m) = j \,|\, Y(t_1) = y(t_1), \ldots, Y(t_m) = y(t_m), \Gamma).$$

- **Task 2 (Dynamic Survival Analysis):** Compute the patient's risk of absorption in the catastrophic state as a function of the future time horizon, i.e.

$$\mathbb{P}(X(t) = N \,|\, Y(t_1) = y(t_1), \ldots, Y(t_m) = y(t_m), \Gamma),\ t \geq t_m.$$

- **Task 3 (Prognostic Risk Scoring):** Compute the patient's risk of absorption in the catastrophic state, i.e.

$$\mathbb{P}(\mathcal{A}_N \,|\, Y(t_1) = y(t_1), \ldots, Y(t_m) = y(t_m), \Gamma).$$

In the rest of this Section, we drop the conditioning on Γ for notational brevity. Task 1 corresponds to disease severity estimation for patients with chronic disease, or clinical acuity assessment for critical care patients. Task 2 is concern with a patient's dynamic survival analysis; computing a hazard function describing the time to an adverse event. Traditionally, this type of analysis is done using the Cox proportional hazard model (Cox and Oakes (1984)), but was limited to regressing a single, static time-to-event hazard curve; Task 2 allows for dynamically estimating a patient's survival as more observable variables are gathered over time. Task 3 corresponds to risk scoring for future adverse events for patients who have been monitored for some period of time, i.e. the risk of developing a future preclinical or clinical breast cancer state (Gail and Mai (2010)), the risk of clinical deterioration for post-operative patients in wards (Rothman et al. (2013)), the risk of mortality for ICU patients (Knaus et al. (1985)), etc.

In order to implement the inference tasks above, we need first to learn the parameter set Γ that generates the patients' episodes using the offline dataset $\mathcal{D}$. The learning task, which we tackle in Section 3.3, can be described as follows.

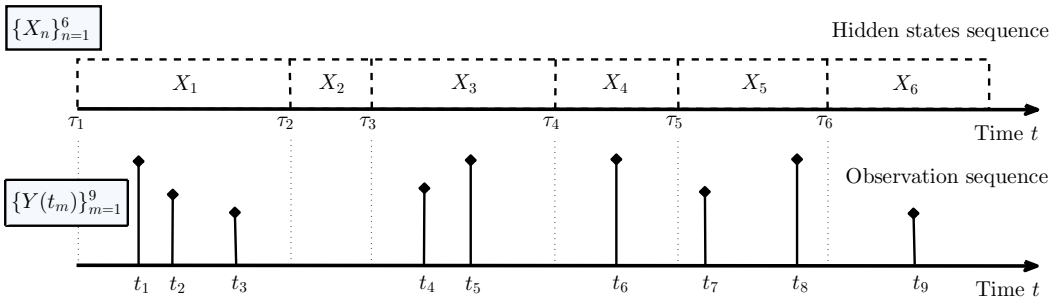

Figure 7: An exemplary HASMM episode with 6 hidden state realizations and 9 observed samples.

Learning task:

Given an offline dataset $\mathcal{D} = \left\{\{y_m^d\}_{m=1}^{M^d}, \{t_m^d\}_{m=1}^{M^d}, T_c^d, l^d\right\}_{d=1}^{D}$, the learning task retrieves the most likely instantiation of the HASMM model that could have generated the episodes in $\mathcal{D}$, i.e.

$$\Gamma^* = \arg\max_\Gamma \mathbb{P}(\mathcal{D} \,|\, \Gamma).$$

We start by developing inference algorithms that execute the inference tasks 1, 2 and 3 in the next Subsection.

3.1 The HASMM Inference Tasks

The inference tasks discussed in the previous Subsection are confronted with 3 main challenges –listed hereunder– that hinder the direct deployment of classical forward-backward message-passing routines.

1. In addition to the clinical states $\{X_n\}_{n=1}^K$ being unobserved, the transition times among the states, $\{\tau_n\}_{n=1}^K$, are also unobserved (i.e. we do not know the time at which the patient's state changed). Thus, unlike the discrete-time models in (Murphy (2002); Johnson and Willsky (2013); Yu (2010); Dewar et al. (2012); Guédon (2007)), in which we know that the underlying states switch sequentially in a (known) one-to-one correspondence with the observations, in an HASMM the association between states and observations is unknown. Figure 7 depicts an exemplary HASMM episode with 6 realized states and 9 observations samples; in this realization, the association between the observations $\{Y(t_1), Y(t_2), Y(t_3)\}$ and state X_1 is hidden. The importance of reasoning about the hidden transition times is magnified by the duration-dependence of the transition probabilities that govern the sequence $\{X_n\}_{n=1}^K$.

2. Since observations are made at random and arbitrary time instances, some transitions may not be associated with any evidential data. That is, as it is the case for state X_2 in Figure 7, there is no guarantee that for every state X_n, an observation is drawn during its occupancy, i.e. $[\tau_n, \tau_{n+1})$. In a practical setting, the inference algorithm should be able to reason about the state trajectories even in silence periods that come with no observations (recall the example in Figure 2 where observations of a critical care patient's systolic blood pressure stop for an entire day). Hence, one cannot directly

discretize the time variable and use the discrete-time HMM inference algorithms (e.g. the algorithms in (Rabiner (1989))) since in that case we would exhibit time steps that come with no associated observations, and with potential state transitions.

3. The HASMM model assumes that observations that belong to the same state are correlated (e.g. in Figure 7, each of the subset of observations $\{Y(t_1), Y(t_2), Y(t_3)\}$, $\{Y(t_4), Y(t_5)\}$ and $\{Y(t_7), Y(t_8)\}$ are not drawn independently conditioned on the latent state since they are sampled from a GP), thus we cannot use the variable-duration and explicit-duration HSMM inference algorithms in (Murphy (2002); Johnson and Willsky (2013); Yu (2010); Guédon (2007)), as those assume that all observations are conditionally independent given the latent states. Our model is closer to a segment-HSMM model (Yu (2010); Guédon (2007)), but with irregular spaced continuous-time observations and an underlying duration-dependent state evolution process, which requires a different construction of the forward messages.

In the following, we develop inference algorithms that deal with episodes generated from an HASMM and address the above challenges.

Inference Task 1: Diagnostic Inference via Forward Filtering

Given a realization of an episode $\{y(t_1), y(t_2), \ldots, y(t_m)\}$ at time t_m, the posterior probability of the patient's current clinical state $X(t_m)$ is given by

$$\begin{aligned}\mathbb{P}(X(t_m) = j \,|\, y(t_1), \ldots, y(t_m)) &= \frac{d\mathbb{P}(X(t_m) = j, y(t_1), \ldots, y(t_m))}{d\mathbb{P}(y(t_1), \ldots, y(t_m))} \\ &= \frac{d\mathbb{P}(X(t_m) = j, y(t_1), \ldots, y(t_m))}{\sum_{j=1}^{N} d\mathbb{P}(X(t_m) = j, y(t_1), \ldots, y(t_m))}. \end{aligned} \tag{7}$$

The above application of Bayes' rule implies that computing the joint probability density $d\mathbb{P}(X(t_m) = j, y(t_1), \ldots, y(t_m))$ suffices for computing the posterior probability of the patient's clinical states. Define $\alpha_m(j, w)$ as the forward message for the j^{th} state at the m^{th} observation time (i.e. t_m) with a lag w as follows

$$\alpha_m(j, w) = d\mathbb{P}(X(t_m) = j, t_m - t_{m-w+1} \leq S(t_m) \leq t_m - t_{m-w}, y(t_1), \ldots, y(t_m)),$$

where $S(t_m)$ is the time elapsed between the transition to the current state, i.e. $X(t_m) = j$, and the time instance t_m. That is, the forward message $\alpha_m(j, w)$ is simply the joint probability that the current state is j, that the associated observations are $(y(t_1), \ldots, y(t_m))$, and that the current state has lasted for the last w measurements. For notational brevity, denote the event $\{t_m - t_{m-w+1} \leq S(t_m) \leq t_m - t_{m-w}\}$ as $\Xi(m, w)$. Thus, $\alpha_m(j, w)$ can be written as

$$\alpha_m(j, w) = \sum_{i=1}^{N} \sum_{w'=1}^{m-w} d\mathbb{P}(X(t_m) = j, \Xi(m, w), X(t_{m-w}) = i, \Xi(m - w, w'), \{y(t_u)\}_{u=1}^{m}), \tag{8}$$

which can be decomposed using the conditional independence properties of the states, observable variables and sojourn times as follows

$$d\mathbb{P}(X(t_m) = j, \Xi(m, w), X(t_{m-w}) = i, \Xi(m - w, w'), \{y(t_u)\}_{u=1}^{m}) =$$

$$d\mathbb{P}(\{y(t_u)\}_{u=m-w+1}^{m} \,|\, X(t_m)=j) \times \underbrace{\mathbb{P}(X(t_m)=j \,|\, X(t_{m-w})=i, \Xi(m-w,w^{'}))}_{p_{ij}(t_m-t_{m-w},\Xi(m-w,w^{'}))} \times$$
$$\underbrace{d\mathbb{P}(\Xi(m,w) \,|\, X(t_m)=j)}_{V_j(t_m-t_{m-w}|\lambda_j)-V_j(t_m-t_{m-w+1}|\lambda_j)} \times \underbrace{d\mathbb{P}(X(t_{m-w})=i, \Xi(m-w,w^{'}), \{y(t_u)\}_{u=1}^{m-w})}_{\alpha_{m-w}(i,w')}. \quad (9)$$

The first term, $d\mathbb{P}(\{y(t_u)\}_{u=m-w+1}^{m} \,|\, X(t_m)=j)$, is the probability density of the observable variables in $\{y(t_u)\}_{u=m-w+1}^{m}$ conditioned on the hidden state being $X(t_m)=j$ and that the time instances $\{t_u\}_{u=m-w+1}^{m}$ reside in the sojourn time of $X(t_m)=j$. The second term, $p_{ij}(t_m-t_{m-w},\Xi(m-w,w^{'}))$, is the *interval transition probability*, i.e. the probability that the hidden state sequence transits to state j after a period of t_m-t_{m-w}, given that it's sojourn time in state $X(t_{m-w})=i$ at time t_m is at least t_m-t_{m-w+1}, and at most $t_m-t_{m-w-w'}$. The third term is the probability that the sojourn time in state $X(t_m)=j$ is between t_m-t_{m-w+1} and t_m-t_{m-w}, whereas the fourth term, $\alpha_{m-w}(i,w^{'})$, is the $(m-w)^{th}$ forward message with a lag of $w^{'}$. Thus, we can write the m^{th} forward message with a lag w as follows

$$\alpha_m(j,w) = d\mathbb{P}(\{y(t_u)\}_{u=m-w+1}^{m} \,|\, X(t_m)=j)\times$$

$$\sum_{i=1}^{N}\sum_{w'=1}^{m-w} p_{ij}(t_m-t_{m-w},\Xi(m-w,w^{'}))\cdot(V_j(t_m-t_{m-w}|\lambda_j)-V_j(t_m-t_{m-w+1}|\lambda_j))\cdot\alpha_{m-w}(i,w^{'}). \quad (10)$$

As we can see in (10), one can express $\alpha_m(j,w)$ using a recursive formula that makes use of the older forward messages $\{\alpha_{m-w}(i,w^{'})\}_{w=1}^{m}$, where $\alpha_o(i,w^{'})=0$, which allows for an efficient dynamic programming algorithm to infer the patient's clinical state in real-time; this is important in critical care settings where prompt risk assessments are crucial for timely clinical intervention.

The construction of the forward messages in (10) parallels the structure of forward message-passing in segment-HSMM (See Section 1.2 in (Murphy (2002)) and Section 4.2.2 in (Yu (2010))), but with the following differences. In (10), the time interval between every two observation samples is irregular, which reflects in the correlation between the observations in $\{y(t_u)\}_{u=m-w+1}^{m}$ (depends on the covariance kernel of the GP, and the probability of the current latent state's sojourn time being encompassing the most recent w samples, i.e. $(V_j(t_m-t_{m-w}|\lambda_j)-V_j(t_m-t_{m-w+1}|\lambda_j))$. However, the most challenging ingredient of the forward message is the interval transition probability $p_{ij}(t_m-t_{m-w},\Xi(m-w,w^{'}))$. This is because unlike the discrete-time HSMM models in (Murphy (2002); Yu (2010)), which exhibit transitions only at discrete time steps that are always accompanied with evidential observations, i.e. no hidden transitions can occur between observation samples, and the transitions among hidden states are duration-independent, in an HASMM, transitions can occur at arbitrary time instances, multiple transitions can occur between two observation samples, and transitions are duration-dependent.

In order to evaluate the term $p_{ij}(t_m-t_{m-w},\Xi(m-w,w^{'}))$, we construct a virtual (discrete-time) bi-variate *embedded Markov chain* $\{X(t_w),t_w\}_{w=1}^{m}$, the transition probabil-

ities of which are equal to the interval transition probabilities, i.e. $p_{ij}(\tau, \{s_1, s_2\})$ is the probability that the embedded Markov chain transits from state (i, t) to state $(j, t+\tau)$ given that state i in the original continuous-time semi-Markov chain has started at a time instance that lies between $t - s_2$ and $t - s_1$. In the recent work in (Liu et al. (2015)), a similar embedded Markov chain analysis was conducted for a CT-HMM, but for which the underlying state evolution process was assumed to be a duration-independent, ordinary Markov chain for which the expressions for $p_{ij}(t_m - t_{m-w}, \Xi(m-w, w^{'}))$ are readily available by virtue of the exponential distributions of the memoryless state sojourn times.

Recall that the semi-Markov kernel of the hidden state sequence $\{X_n\}_{n=1}^{K}$ is defined as $Q_{ij}(\tau) = \mathbb{P}(X_{n+1} = j, S_n \leq \tau | X_n = i)$, i.e. the probability that the sequence transits from state i to state j given that the sojourn time in i is less than or equal to τ. Now consider the interval transition probability $p_{ij}(\tau)$ for the semi-Markov path $(X(t))_{t \in \mathbb{R}_+}$ defined as $p_{ij}(\tau) = \mathbb{P}(X(t+\tau) = j \,|\, X(t) = i)$. From (Kulkarni (1996)), we know that the functions $(p_{ij}(\tau))_{i,j \in \mathcal{X}}$ solve the following system of integral equations

$$p_{ij}(\tau) = \delta_{ij} \cdot (1 - Q_i(\tau)) + \sum_{k=1}^{N} \int_0^{\tau} \frac{\partial Q_{ik}(u)}{\partial u} \cdot p_{kj}(\tau - u)\, du, \, \forall i, j \in \mathcal{X}, \tag{11}$$

where δ_{ij} is the Kronecker delta function, and $Q_i(\tau) = \sum_{k=1}^{N} Q_{ik}(\tau)$ (recall that $Q_{ii}(\tau) = 0$).

Since we are evaluating the functions $(p_{ij}(\tau, s))_{i,j}$ and not $(p_{ij}(\tau))_{i,j}$, we need to consider a *truncated semi-Markov kernel* $Q_{ij}(\tau, s)$ that conditions the interval transition probabilities on the elapsed time in state i being s. It can be easily shown that this "left-truncated" kernel is given by $Q_{ij}(\tau, s) = \frac{Q_{ij}(\tau) - Q_{ij}(s)}{1 - Q_{ij}(s)}$. Thus, by modifying the terms in (11) accordingly, the functions $(p_{ij}(\tau, s))_{i,j}$ can be obtained by solving the following system of integral equations

$$p_{ij}(\tau, s) = \delta_{ij} \cdot (1 - Q_i(\tau, s)) + \sum_{k=1}^{N} \int_s^{\tau} \frac{\partial Q_{ik}(u, s)}{\partial u} \cdot p_{kj}(\tau - u, s)\, du, \, \forall i, j \in \mathcal{X}, \tag{12}$$

where $Q_i(\tau, s) = \sum_{k=1}^{N} Q_{ik}(\tau, s)$. The term $\frac{\partial Q_{ij}(\tau, s)}{\partial \tau}$ can be easily computed by invoking the representation of the semi-Markov kernel provided in (4). From (4), we can write $Q_{ij}(\tau)$ as follows

$$\begin{aligned}
Q_{ij}(\tau) &= \mathbb{E}_S \left[g_{ij}(S) | S \leq \tau \right] \cdot V_i(\tau | \lambda_i) \\
&= \int_0^{\tau} g_{ij}(S) \cdot \frac{1 - V_i(S|\lambda_i)}{V_i(\tau|\lambda_i)}\, dS \cdot V_i(\tau|\lambda_i) \\
&= \int_0^{\tau} g_{ij}(S) \cdot (1 - V_i(S|\lambda_i))\, dS \\
&= \int_0^{\tau} \frac{e^{\pi_{ij}(1+\beta_i S)}}{\sum_{k=1}^{N} e^{\pi_{kj}(1+\beta_k S)}} \cdot (1 - V_i(S|\lambda_i))\, dS \\
&= \int_0^{\tau} \frac{e^{\pi_{ij}(1+\beta_i S)}}{\sum_{k=1}^{N} e^{\pi_{kj}(1+\beta_k S)}} \cdot \left(1 - \frac{\gamma(\lambda_{i,s}, \lambda_{i,r} S)}{\Gamma(\lambda_{i,s})}\right) dS,
\end{aligned} \tag{13}$$

where we have substituted for $V_i(S|\lambda_i)$ with the cumulative density function of a Gamma distributed random variable. The left-truncated semi-Markov kernel is then given by

$$Q_{ij}(\tau, s) = \frac{\int_s^\tau \frac{e^{\pi_{ij}(1+\beta_i S)}}{\sum_{k=1}^N e^{\pi_{kj}(1+\beta_k S)}} \cdot \left(1 - \frac{\gamma(\lambda_{i,s}, \lambda_{i,r} S)}{\Gamma(\lambda_{i,s})}\right) dS}{1 - \int_0^s \frac{e^{\pi_{ij}(1+\beta_i S)}}{\sum_{k=1}^N e^{\pi_{kj}(1+\beta_k S)}} \cdot \left(1 - \frac{\gamma(\lambda_{i,s}, \lambda_{i,r} S)}{\Gamma(\lambda_{i,s})}\right) dS}, \tag{14}$$

which can be easily evaluated numerically using a Riemann sum.

The system of equations in (12) is a non-homogeneous system of Volterra integral equations of the second kind (Polyanin and Manzhirov (2008)), which is analogous to the Chapman-Kolmogorov system of equations in ordinary Markov chains. Obtaining an analytic solution to (12) is a tedious problem: for a fixed s, we solve the system in (12) via the *successive approximation* method (Opial (1967)). That is, we initialize the interval transition probabilities by the corresponding entries in the semi-Markov kernel matrix $\mathbf{Q}$(s) as follows[16]

$$p_{ij}^{(o)}(\tau, s) = Q_{ij}(\tau, s),\ \forall i, j \in \mathcal{X},$$

and for the z^{th} iteration, we update the interval transition functions as follows

$$p_{ij}^{(z)}(\tau, s) = \delta_{ij} \cdot (1 - Q_i(\tau, s)) + \sum_{k=1}^{i} \int_s^\tau \frac{\partial Q_{ik}(u, s)}{\partial u} \cdot p_{kj}^{(z)}(\tau - u, s)\, du +$$

$$\sum_{k=i+1}^{N} \int_s^\tau \frac{\partial Q_{ik}(u, s)}{\partial u} \cdot p_{kj}^{(z-1)}(\tau - u, s)\, du,\ \forall i, j \in \mathcal{X},$$

i.e. we use the most recent interval transition function for updating the other functions in every iteration. Iterations stop when the criterion $\int_0^\infty |p_{ij}^{(z)}(\tau, s) - p_{ij}^{(z-1)}(\tau, s)| < \epsilon$, $\forall i, j \in \mathcal{X}$. By observing that the integral in (12) is a convolution integral, we have that

$$\int_s^\tau \frac{\partial Q_{ik}(u, s)}{\partial u} \cdot p_{kj}(\tau - u, s)\, du = \left(\frac{\partial Q_{ik}(., s)}{\partial u} \star p_{kj}(., s)\right)(\tau) - \left(\frac{\partial Q_{ik}(., s)}{\partial u} \star p_{kj}(., s)\right)(s), \tag{15}$$

where $\star$ is the convolution operator. Using the reformulation in (15), we can use a more efficient Fast Fourier Transform (FFT) algorithm to update the interval transition probabilities at each step instead of computing the convolution integral, which accelerates the computations.

It is important to note that we do not need to solve the system of equations in (12) during real-time inference. Instead, we create a look-up table of (discretized) transition function

16. This is a reasonable initialization since the entries of the semi-Markov kernel correspond to interval transition probabilities conditioned on there being no intermediate transitions on the way from state i to state j. For existence and uniqueness of the solutions to Volterra equations, please refer to (Polyanin and Manzhirov (2008)). For convergence of the successive approximation iterations, please refer to (Opial (1967)).

Algorithm 2 Constructing a look-up table of interval transition probabilities

procedure TransitionLookUp($\pi = \{\pi_{ij}\}_{i,j}$, $\beta = \{\beta_i\}_i$, $\lambda = \{\lambda_i\}_i$, ϵ)
  Input: Semi-Markov kernel parameters $\pi = \{\pi_{ij}\}_{i,j}, \beta = \{\beta_i\}_i$
  Output: A look-up table $(\tilde{p}_{ij}(a\Delta\tau, b\Delta s))_{i,j,a,b}$
  Set the values of A (number of steps for τ), B (steps for s), $\Delta\tau$, Δs (step sizes)
  for $a = 1$ to A, $b = 1$ to B **do**
    $Q^{\tau}_{ij}(a\Delta\tau) \leftarrow \sum_{x=1}^{a} \frac{e^{\pi_{ij}(1+\beta_i x\Delta\tau)}}{\sum_{k=1}^{N} e^{\pi_{ik}(1+\beta_i x\Delta\tau)}} \left(1 - \frac{\gamma(\lambda_{i,s},\lambda_{i,r}x\Delta\tau)}{\Gamma(\lambda_{i,s})}\right) \Delta\tau$
    $Q^{s}_{ij}(b\Delta s) \leftarrow \sum_{x=1}^{b} \frac{e^{\pi_{ij}(1+\beta_i x\Delta s)}}{\sum_{k=1}^{N} e^{\pi_{ik}(1+\beta_i x\Delta s)}} \left(1 - \frac{\gamma(\lambda_{i,s},\lambda_{i,r}x\Delta s)}{\Gamma(\lambda_{i,s})}\right) \Delta s$
    $Q_{ij}(a\Delta\tau, b\Delta s) \leftarrow \frac{Q^{\tau}_{ij}(a\Delta\tau) - Q^{s}_{ij}(b\Delta s)}{1 - Q^{s}_{ij}(b\Delta s)}$
  end for
  $e = \epsilon + 1$
  $z \leftarrow 1$
  $\tilde{p}^{(o)}_{ij}(a\Delta\tau, b\Delta s) \leftarrow Q_{ij}(a\Delta\tau, b\Delta s)$, $\forall a, b, i, j$.
  while $e > \epsilon$ **do**
    $ConvQ_{i,j,k}(a\Delta\tau, b\Delta s) \leftarrow \text{IFFT}\left(\text{FFT}\left(\text{diff}\left(Q_{ik}(a\Delta\tau, b\Delta s)\right)\right), \text{FFT}\left(\tilde{p}^{(z-1)}_{jk}(a\Delta\tau, b\Delta s)\right)\right)$,
    $\tilde{p}^{(z)}_{ij}(a\Delta\tau, b\Delta s) \leftarrow \delta_{ij}\, Q_{ij}(a\Delta\tau, b\Delta s) + \sum_{k=1}^{N} ConvQ_{i,j,k}(a\Delta\tau, b\Delta s)$
    $z \leftarrow z + 1$
    $e \leftarrow \max_{i,j,b}\left\{\sum_{a=1}^{A}\left|\tilde{p}^{(z)}_{ij}(a\Delta\tau, b\Delta s) - \tilde{p}^{(z-1)}_{ij}(a\Delta\tau, b\Delta s)\right|\right\}$
  end while
  return $(\tilde{p}_{ij}(a\Delta\tau, b\Delta s))_{i,j,a,b}$
end procedure

offline $(\tilde{p}_{ij}(a\Delta\tau, b\Delta s))_{i,j,a,b}$, and then we query this table when performing real-time inference for monitored patients. Hence, efficient and fast inferences can be provided for critical care patients for whom prompt diagnostic inferences are necessary for the efficacy of clinical interventions. Algorithm 2 shows a pseudocode for constructing a look-up table of interval transition probabilities, `TransitionLookUp`($\pi = \{\pi_{ij}\}_{i,j}$, $\beta = \{\beta_i\}_i$, $\lambda = \{\lambda_i\}_i$, ϵ), which takes as an input the parameters of the semi-Markov kernel, the states' sojourn times parameters, and a precision level ϵ (to terminate the successive approximation iterations), and outputs the interval transitions look-up table. In Algorithm 2, FFT and IFFT refer to the fast Fourier transform operation and its inverse, respectively, and "diff(.)" refers to a numerical differentiation operation.

Now that we have constructed the algorithm `TransitionLookUp` to compute the interval transition probabilities in the look-up table $(\tilde{p}_{ij}(a\Delta\tau, b\Delta s))_{i,j,a,b}$, we can implement a forward-filtering inference algorithm using dynamic programming (by virtue of the recursive formula in (10)). In particular, the posterior probability of the patient's current clinical state in terms of the forward messages can be written as

$$\mathbb{P}(X(t_m) = j \,|\, y(t_1), \ldots, y(t_m)) = \frac{\sum_{w=1}^{m} \alpha_m(j, w)}{\sum_{k=1}^{N}\sum_{w=1}^{m} \alpha_m(k, w)}. \tag{16}$$

Algorithm 3 Forward filter inference

procedure ForwardFilter(Γ, $\{y(t_w)\}_{w=1}^m$, ϵ)
  Input: Observed samples $\{y(t_w)\}_{w=1}^m$, HASMM parameters Γ, and precision ϵ
  Output: The posterior state distribution $\{\mathbb{P}(X(t_m)=j \mid \{y(t_w)\}_{w=1}^m)\}_{j=1}^N$
  $\tilde{p}_{ij}(a\Delta\tau, b\Delta s) \leftarrow$ TransitionLookUp($\pi=\{\pi_{ij}\}_{i,j}$, $\beta=\{\beta_i\}_i$, $\lambda=\{\lambda_i\}_i$, ϵ)
  $\alpha_1(j,1) = \mathbb{P}(y(t_1) \mid X(t_1)=j) \sum_{i=1}^N \tilde{p}_{ij}(t_1,0) \cdot p_i^o, \ \forall j \in \mathcal{X}$
  for $z = 2$ to m **do**
    for $w = 1$ to z **do**
      $a^*(z,w) = \arg\min_a |t_z - t_{z-w} - a\Delta\tau|$
      $b^*(z,w,w') = \arg\min_b \left|t_{z-w} - t_{z-w-w'+1} - b\Delta s\right|$
      $\alpha_z(j,w) = \mathbb{P}(\{y(t_u)\}_{u=z-w+1}^z \mid X(t_z)=j) \times$

$$\sum_{i=1}^{N} \sum_{w'=1}^{z-w} \tilde{p}_{ij}(a^*(z,w)\Delta\tau, b^*(z,w,w')\Delta s) \cdot (1 - V_j(t_z - t_{z-w+1}|\lambda_j)) \cdot \alpha_{z-w}(i,w')$$

    end for
  end for
  $\mathbb{P}(X(t_m)=j \mid \{y(t_u)\}_{u=1}^m) = \frac{\sum_{w=1}^m \alpha_m(j,w)}{\sum_{k=1}^N \sum_{w=1}^m \alpha_m(k,w)}$
  return $\{\mathbb{P}(X(t_m)=j \mid \{y(t_w)\}_{w=1}^m)\}_{j=1}^N$
end procedure

Algorithm 3, `ForwardFilter`, implements real-time inference of a patient's clinical state given a sequence of measurements $\{y(t_1), \ldots, y(t_m)\}$. In Algorithm 3, we invoke `TransitionLookUp` initially to construct the look-up table of transition probabilities, but in practice, the look-up table can be constructed in an offline stage once the HASMM parameter set Γ is known. The number of computations can be reduced by limiting the lags w for every forward message $\alpha_m(j,w)$ to the samples in $\mathcal{T}$ that reside in a period $t_m - T_{max}$, where T_{max} is derived from the Gamma distribution of the sojourn time (e.g. T_{max} can be selected such that $v_i(s \leq T_{max}|\lambda_i) > 90\%$). The complexity of `ForwardFilter` is similar to the conventional forward algorithms in (Rabiner (1989)).

Inference Task 2: Prognostic Dynamic Survival Analysis

In this task, we focus on inferring the patient's survival function given the sequence of observed variables $\{y(t_u)\}_{u=1}^m$, i.e. we compute the function

$$\begin{aligned}\bar{S}(t_m,\tau) &= \mathbb{P}(T_s > t_m + \tau \mid \{y(t_u)\}_{u=1}^m) \\ &= 1 - \mathbb{P}(X(t_m+\tau) = N \mid \{y(t_u)\}_{u=1}^m). \end{aligned} \tag{17}$$

That is, the patient's survival function, which is the probability that her clinical state is not absorbed in the catastrophic state N after a time period τ starting from t_m, is equivalent to the complement of the probability of that the hidden state process is absorbed in state N in less than a period of τ starting from time instance t_m.

Note that the survival function $\bar{S}(t_m, \tau)$ in (17) can be written as

$$\begin{aligned}\bar{S}(t_m, \tau) &= 1 - \mathbb{P}(X(t_m + \tau) = N \,|\, \{y(t_u)\}_{u=1}^m) \\ &= 1 - \sum_{j=1}^{N} \mathbb{P}(X(t_m + \tau) = N \,|\, X(t_m) = j) \cdot \mathbb{P}(X(t_m) = j \,|\, \{y(t_u)\}_{u=1}^m) \\ &= 1 - \sum_{j=1}^{N} p_{jN}(\tau, 0) \cdot \frac{\sum_{w=1}^m \alpha_m(j, w)}{\sum_{k=1}^N \sum_{w=1}^m \alpha_m(k, w)}. \end{aligned} \tag{18}$$

Using the procedures `TransitionLookUp` and `ForwardFilter`, we can update the patient's survival curve at every time instance t_m by plugging in the interval transition probabilities obtained from `TransitionLookUp`, together with the forward messages computed via `ForwardFilter` as follows

$$\bar{S}(t_m, a\Delta\tau) = 1 - \sum_{j=1}^{N} \tilde{p}_{jN}(a\Delta\tau, 0) \cdot \frac{\sum_{w=1}^m \alpha_m(j, w)}{\sum_{k=1}^N \sum_{w=1}^m \alpha_m(k, w)}. \tag{19}$$

Survival analysis plays an important role in guiding many clinical decisions, such as deciding the frequency of breast cancer screening (Taghipour et al. (2013)), predicting hospital readmission (Kansagara et al. (2011)), making discharge decisions for ICU or critically ill inpatients (Moreno et al. (2005)), and planning multi-stage interventions (Foucher et al. (2007)). The survival function in (19) is computed for an individual patient through her individual physiological trajectory $\{y(t_u)\}_{u=1}^m$, and hence can guide various survival-related clinical decisions for a monitored patient in an individualized manner.

Inference Task 3: Prognostic Risk Scoring

Prognostic risk scoring plays an important role in designing screening guidelines (Gail and Mai (2010)), acute care interventions (Knaus et al. (1985)) and surgical decisions (Foucher et al. (2007)). A risk score is an aggregate measure of the survival function $\bar{S}(t_m, \tau)$, i.e. it corresponds to the probability that the patients encounters an adverse event (abstracted as state N in our model) at any futuristic time step starting from time t_m. That is, the patient's risk score at time t_m can be formulated as

$$\begin{aligned} R(t_m) &= \mathbb{P}(\mathcal{A}_N \,|\, \{y(t_u)\}_{u=1}^m) \\ &= 1 - \mathbb{P}(X(\infty) = N \,|\, \{y(t_u)\}_{u=1}^m), \end{aligned} \tag{20}$$

which can be computed using the outputs of `TransitionLookUp` and `ForwardFilter` as follows

$$R(t_m) = \sum_{j=1}^{N} \tilde{p}_{jN}(A, 0) \cdot \frac{\sum_{w=1}^m \alpha_m(j, w)}{\sum_{k=1}^N \sum_{w=1}^m \alpha_m(k, w)}. \tag{21}$$

Therefore, the procedures `TransitionLookUp` and `ForwardFilter` suffice for executing all the diagnostic and prognostic inference tasks of inference. Performance of these algorithms is investigates in Section 4. In the next Subsection, we focus on the HASMM learning task.

3.2 The HASMM Learning Task

3.3 The HASMM Learning Task

In Section 3.1, we developed (diagnostic and prognostic) inference algorithms that can deal with patients in real-time assuming that the true HASMM parameter set Γ is known. In practice, the parameter set Γ is not known, and has to be learned from an offline EHR dataset $\mathcal{D}$ that comprises D episodes for previously hospitalized or monitored patients, i.e.

$$\mathcal{D} = \left\{ \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, \{t_m^{(d)}\}_{m=1}^{M^{(d)}}, T_c^{(d)}, l^{(d)} \right\}_{d=1}^{D}.$$

In this Section, we develop efficient algorithms that compute the Maximum Likelihood (ML) estimate of Γ given a dataset $\mathcal{D}$, defined as $\Gamma^* = \arg\max_\Gamma \Lambda(\mathcal{D} \,|\, \Gamma)$, where $\Lambda(\mathcal{D} \,|\, \Gamma) = \mathbb{P}(\mathcal{D} \,|\, \Gamma)$ is the likelihood of the dataset $\mathcal{D}$ given the parameter set Γ.

We focus on the challenging scenario when no domain knowledge or diagnostic assessments for the patient's latent states are provided in the dataset $\mathcal{D}$[17] (with the exception of the absorbing state which is declared by the variable $l^{(d)}$), i.e. the learning algorithm is *unsupervised*. For such a scenario, the main challenge in constructing the ML estimator Γ^* resides in the hiddenness of the patients' state trajectories in the training dataset $\mathcal{D}$; the dataset $\mathcal{D}$ contains only the sequence of observable variables, their respective observation times, the episodes censoring time and the state in which the trajectory was absorbed. If the patients' latent state trajectories $(X(t))_{t\in\mathbb{R}_+}$ were observed in $\mathcal{D}$, the ML estimation problem $\Gamma^* = \arg\max_\Gamma \mathbb{P}(\mathcal{D} \,|\, \Gamma)$. would have been straightforward; the hiddenness of $(X(t))_{t\in\mathbb{R}_+}$ entails the need for marginalizing over the space of all possible latent trajectories conditioned on the observed variables, which is a hard task even for conventional continuous-time HMM models (Liu et al. (2015); Nodelman et al. (2012); Leiva-Murillo et al. (2011); Metzner et al. (2007)). As we will show later in this Section, more complications are faced in an HASMM model due to the time-inhomogeneity and semi-Markovianity of state transition, and the segmental nature of the observation variables (i.e. temporal correlation between the observed variables).

In order to construct the ML estimator for Γ, we start by writing the complete likelihood, i.e. the likelihood of an HASMM with a parameter set Γ to generate both the hidden states trajectory $\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}}$ and the observable variables $\{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}$ of the d^{th} episode in the dataset $\mathcal{D}$ as follows

$$\mathbb{P}\left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\middle|\, \Gamma \right) =$$

17. For some problems, such as chronic kidney disease progression estimation (Eddy and Neilson (2006)), the EHR records may include some anchors or assessments to the latent states over time. A simpler version of the learning algorithm proposed in this Section can be used to deal with such datasets. In critical care settings, it is more common that the EHR records are not labeled with any clinical state assessments over time (Yoon et al. (2016)).

$$\mathbb{P}(X_1^{(d)}|\Gamma)\cdot\mathbb{P}(S_1^{(d)}|X_1^{(d)},\Gamma)\cdot\mathbb{P}(\{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)}\in\mathcal{T}_1^{(d)}}|X_1^{(d)},\Gamma)\times$$
$$\prod_{n=2}^{K^{(d)}}\mathbb{P}(X_n^{(d)}\,\Big|\,X_{n-1}^{(d)},S_{n-1}^{(d)},\Gamma)\cdot\mathbb{P}(S_n^{(d)}\,\Big|\,X_n^{(d)},\Gamma)\cdot\mathbb{P}(\{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)}\in\mathcal{T}_n^{(d)}}\,\Big|\,X_n^{(d)},\Gamma). \tag{22}$$

The factorization in (22) follows from the conditional independence properties of the HASMM variables. Since we cannot observe the latent states trajectory $\{X_n^{(d)},S_n^{(d)}\}_{n=1}^{K^{(d)}}$, the ML estimator deals with the expected likelihood $\Lambda(\mathcal{D}\,|\,\Gamma)$, which is evaluated by marginalizing the complete likelihood over the latent paths $(X(t))_{t\in\mathbb{R}_+}$, i.e.

$$\Lambda(\mathcal{D}\,|\,\Gamma)=\int\cdots\int\mathbb{P}\left(\left\{\{X_n^{(d)},S_n^{(d)}\}_{n=1}^{K^{(d)}},\{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}\right\}_{d=1}^{D}\,\Big|\,\Gamma\right)dX^{(1)}(t)\cdots dX^{(D)}(t). \tag{23}$$

Assuming that the episodes in $\mathcal{D}$ are independent, we can write (23) as

$$\begin{aligned}\Lambda(\mathcal{D}\,|\,\Gamma)&=\int\cdots\int\prod_{d=1}^{D}\mathbb{P}\left(\{X_n^{(d)},S_n^{(d)}\}_{n=1}^{K^{(d)}},\{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}\,\Big|\,\Gamma\right)dX^{(1)}(t)\cdots dX^{(D)}(t)\\&=\prod_{d=1}^{D}\int\mathbb{P}\left(\{X_n^{(d)},S_n^{(d)}\}_{n=1}^{K^{(d)}},\{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}\,\Big|\,\Gamma\right)dX^{(d)}(t),\end{aligned} \tag{24}$$

which can be further decomposed as

$$\begin{aligned}\Lambda(\mathcal{D}\,|\,\Gamma)=\prod_{d=1}^{D}\int&\mathbb{P}(X_1^{(d)}|\Gamma)\cdot\mathbb{P}(S_1^{(d)}|X_1^{(d)},\Gamma)\cdot\mathbb{P}(\{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)}\in\mathcal{T}_1^{(d)}}|X_1^{(d)},\Gamma)\times\\&\prod_{n=2}^{K^{(d)}}\mathbb{P}(X_n^{(d)}\,\Big|\,X_{n-1}^{(d)},S_{n-1}^{(d)},\Gamma)\cdot\mathbb{P}(S_n^{(d)}\,\Big|\,X_n^{(d)},\Gamma)\cdot\mathbb{P}(\{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)}\in\mathcal{T}_n^{(d)}}\,\Big|\,X_n^{(d)},\Gamma)\,dX^{(d)}(t).\end{aligned} \tag{25}$$

Finding the ML estimate Γ^* by direct maximization of $\Lambda(\mathcal{D}\,|\,\Gamma)$ is not possible due to the intractability of the integral in (25), i.e. $\Lambda(\mathcal{D}\,|\,\Gamma)$ has no analytic maximizer. The hardness of evaluating the expected likelihood $\Lambda(\mathcal{D}\,|\,\Gamma)$ follows from the fact that we need to average the complete likelihood over an infinite number of continuous paths. That is, for every episode d, both the number of states $K^{(d)}$ and the sets $\mathcal{T}_n^{(d)}$ are random; evaluating the integral in (25) requires enumerating a large number of possible associations between the observable variables and latent states, which renders the evaluation of $\Lambda(\mathcal{D}\,|\,\Gamma)$ intractable.

As it is the case for classical discrete and continuous-time HMMs, solving the maximization problem $\Gamma^*=\text{argmax}_\Gamma\Lambda(\mathcal{D}\,|\,\Gamma)$ can be approached using the Expectation-Maximization (EM) algorithm (Liu et al. (2015); Nodelman et al. (2012); Rabiner (1989)). The iterative EM algorithm starts with an initial guess Γ^o for the parameter set, and maximizes a proxy for the log-likelihood function in the p^{th} iteration through the following steps:

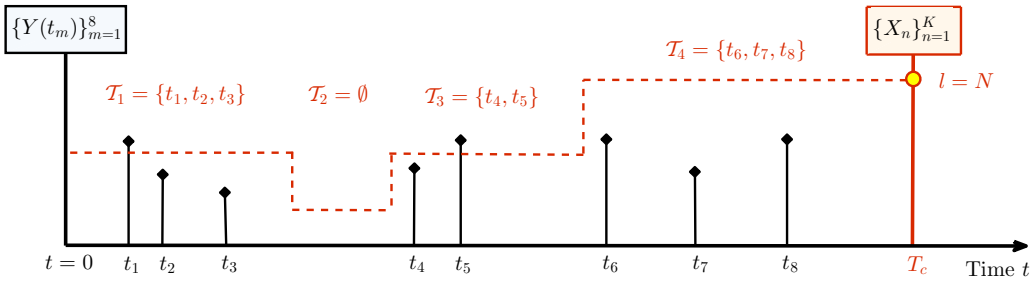

Figure 8: An episode that comprised 8 observable samples, censored at time T_c, and absorbed in state N (catastrophic state). The dashed state trajectory is a trajectory that could have generated the observables with a positive probability. Computing the proximal log-likelihood requires averaging over infinitely many paths that, as the depicted dashed path, could have generated the observables with a positive probability.

- **E-step:** $Q(\Gamma; \Gamma^{p-1}) =$

$$\sum_{d=1}^{D} \mathbb{E}\left[\log\left(\mathbb{P}\left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\middle|\, \Gamma\right)\right) \middle| \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, X^{(d)}(T_c^{(d)}) = l^{(d)}, \Gamma^{p-1}\right].$$

- **M-step:** $\Gamma^p = \arg\max_\Gamma Q(\Gamma; \Gamma^{p-1})$.

The E-step computes the proximal expected log-likelihood $Q(\Gamma; \Gamma^{p-1})$, which entails evaluating the following integral

$$Q(\Gamma; \Gamma^{p-1}) = \sum_{d=1}^{D} \int \log\left(\mathbb{P}\left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\middle|\, \Gamma\right)\right) \times$$
$$\mathbb{P}\left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}} \middle| \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, X^{(d)}(T_c^{(d)}) = l^{(d)}, \Gamma^{p-1}\right) dX^{(d)}(t). \tag{26}$$

That is, the proximal expected log-likelihood $Q(\Gamma; \Gamma^{p-1})$ is computed by marginalizing the likelihood of the observed samples of the d^{th} episodes $\{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}$ over all paths $(X(t))_{t\in\mathbb{R}_+}$ that are censored at time $T_c^{(d)}$ and absorbed in state $l^{(d)}$. Figure 8 depicts the procedure for computing $Q(\Gamma; \Gamma^{p-1})$: given the observed absorbing state $l^{(d)}$ and the censoring time $T_c^{(d)}$, we average the likelihood of the observed samples $\{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}$ over all the latent paths that could have been absorbed in $l^{(d)}$ and censored at $T_c^{(d)}$.

Direct adoption of the conventional Baum-Welch implementation (e.g. the implementation in (Rabiner (1989))) of the EM algorithm for an HASMM is not possible due to the intractability of the integral in the E-step. In fact, the Baum-Welch approach has been successful for discrete-time HMM models (e.g. HMM (Rabiner (1989)), HSMM (Murphy (2002)), EDHMM and VDHMM (Yu (2010)), etc); contrarily, previously investigated

continuous-time HMM models have constantly struggled with the implementation of the E-step due to the need for integrating over latent continuous paths (Liu et al. (2015); Nodelman et al. (2012)). Previous EM approaches for learning continuous-time HMMs were restricted to time-homogeneous Markovian state trajectories: for these models, the properties of the transition matrix make the computation of $Q(\Gamma;\Gamma^{p-1})$ boil down to computing the expected state durations and transition counts (e.g. see Equations (12) and (13) in (Liu et al. (2015))). Different approaches have been developed in the literature for computing these quantities: (Wang et al. (2014)) assumes that the transition rate matrix is diagonalizable, and hence utilize a closed-form estimator for the transition rates, whereas (Liu et al. (2015)) uses the *Expm* and *Unif* methods (originally developed in (Hobolth and Jensen (2011))) to evaluate the integrals of the transition matrix exponential. Unfortunately, none of these methods could be utilized for computing the proximal log-likelihood $Q(\Gamma;\Gamma^{p-1})$ of an HASMM due to the time-inhomogeneity and semi-Markovianity of the state trajectory (i.e. state-durations are not exponentially distributed as it is the case in (Liu et al. (2015); Nodelman et al. (2012); Hobolth and Jensen (2011); Wang et al. (2014))). Further complication is encountered by our model due to the segmental nature of observations; the observed samples are not conditionally independent given the latent states, which requires enumerating all possible memberships of the observation samples in their respective latent states in order to account for their correlations. Moreover, non of the previous works considered informative censoring, i.e. our training set does not comprise equal duration episdoes, but rather the censoring times and absorbing states convey information about the latent path. In the rest of this Section, we develop an efficient EM algorithm that can compute $Q(\Gamma;\Gamma^{p-1})$ by directly distilling information from the censoring events that are apparent in the episdoes in $\mathcal{D}$.

Since computing $Q(\Gamma;\Gamma^{p-1})$ does not admit a closed-form solution, we resort to a Monte Carlo approach for approximating the integral involved in the E-step (Caffo et al. (2005)). That is, in the p^{th} iteration of the EM algorithm, we draw G random trajectories

$$\left(\{X_n^{(d,p,g)}, S_n^{(d,p,g)}\}_{n=1}^{K^{(d,p,g)}}\right)_{g=1}^{G}$$

for every episode d, and use those trajectories to construct a Monte Carlo approximation for the proximal log-likelihood function. Sample trajectories are drawn from the joint posterior distribution of the latent states and sojourn times given the observable variables and censoring information, i.e. the g^{th} sample trajectory is drawn as follows

$$\{X_n^{(d,p,g)}, S_n^{(d,p,g)}\}_{n=1}^{K^{(d,p,g)}} \sim \mathbb{P}\left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, X^{(d)}(T_c^{(d)}) = l^{(d)}, \Gamma^{p-1}\right), \tag{27}$$

for $g \in \{1, \ldots, G\}$. Hence, the proximal log-likelihood $Q(\Gamma; \Gamma^{p-1})$ can be approximated as follows

$$
\begin{aligned}
Q(\Gamma; \Gamma^{p-1}) &= \sum_{d=1}^{D} \int \log \left(\mathbb{P} \left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\middle|\, \Gamma \right) \right) \times \\
&\mathbb{P} \left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, X^{(d)}(T_c^{(d)}) = l^{(d)}, \Gamma^{p-1} \right) dX^{(d)}(t) \\
&\approx \sum_{d=1}^{D} \frac{1}{G} \sum_{g=1}^{G} \log \left(\mathbb{P} \left(\{X_n^{(d,p,g)}, S_n^{(d,p,g)}\}_{n=1}^{K^{(d,p,g)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\middle|\, \Gamma \right) \right). \qquad (28)
\end{aligned}
$$

Convergence of $Q(\Gamma; \Gamma^{p-1})$ to its Monte Carlo estimate for a large sample size G follows from the law of large numbers.

Sampling trajectories from the posterior distribution

$$
\mathbb{P} \left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, X^{(d)}(T_c^{(d)}) = l^{(d)}, \Gamma^{p-1} \right),
$$

as specified in (27) is not a straight forward task, for that the sampler needs to jointly sample the states and their sojourn times taking into account the time-inhomogeneous transitions among states, and that the number of variables sampled (number of states) $K^{(d,p,g)}$ in each trajectory is itself random.

The availability of the censoring information (censoring time T_c and absorbing state $X^{(d)}(T_c) = l^{(d)}$) for every episode d in $\mathcal{D}$ stimulates the development of a *forward-filtering backward-sampling* algorithm that goes in the reverse-time direction and sequentially samples the latent states conditioned on future states (Godsill et al. (2012)). That is, unlike the generative process (described by the routine `GenerateHASMM`(Γ)) which uses the knowledge of the parameter set Γ to generate sample trajectories by drawing an initial state and then sequentially goes forward in time and sample future states until absorption, the inferential process naturally goes the other way around: it exploits informative censoring by starting from the knowledge of the final absorbing state and censoring time, and sequentially sampling a trajectory by traversing backwards in time and conditioning on the future. We start constructing our forward-filter backward-sampler by first formulating the posterior probability of a latent trajectory $\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}}$ (from which we sample the G trajectories

as shown in (27)) in the p^{th} iteration of the EM algorithm as follows

$$
\begin{aligned}
&\mathbb{P}\left(\{X_n^{(d)}, S_n^{(d)}\}_{n=1}^{K^{(d)}} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, X^{(d)}(T_c^{(d)}) = l^{(d)}, \Gamma^{p-1}\right) \\
&\overset{(a)}{=} \mathbb{P}\left(X_{K^{(d)}}^{(d)} = l^{(d)}, S_{K^{(d)}}^{(d)} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, T_c, \Gamma^{p-1}\right) \times \\
&\quad \prod_{n=1}^{K^{(d)}-1} \mathbb{P}\left(X_n^{(d)}, S_n^{(d)} \,\middle|\, \underbrace{X_{n+1}^{(d)}, \ldots, X_{K^{(d)}}^{(d)}, S_{n+1}^{(d)}, \ldots, S_{K^{(d)}}^{(d)}}_{\text{Future trajectory}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, T_c, \Gamma^{p-1}\right) \\
&\overset{(b)}{=} \mathbb{P}\left(X_{K^{(d)}}^{(d)} = l^{(d)} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, \Gamma^{p-1}\right) \times \mathbb{P}\left(S_{K^{(d)}}^{(d)} \,\middle|\, X_{K^{(d)}}^{(d)} = l^{(d)}, S_{K^{(d)}}^{(d)} \le T_c, \Gamma^{p-1}\right) \times \\
&\quad \prod_{n=1}^{K^{(d)}-1} \mathbb{P}\left(X_n^{(d)}, S_n^{(d)} \,\middle|\, X_{n+1}^{(d)}, \ldots, X_{K^{(d)}}^{(d)}, S_n^{(d)} \le \underbrace{T_c - \sum_{w=n+1}^{K^{(d)}} S_w^{(d)}}_{\text{Elapsed time in the episode}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, \Gamma^{p-1}\right) \\
&\overset{(c)}{=} \mathbb{P}\left(X_{K^{(d)}}^{(d)} = l^{(d)} \,\middle|\, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, \Gamma^{p-1}\right) \times \mathbb{P}\left(S_{K^{(d)}}^{(d)} \,\middle|\, X_{K^{(d)}}^{(d)} = l^{(d)}, S_{K^{(d)}}^{(d)} \le T_c, \Gamma^{p-1}\right) \times \\
&\quad \prod_{n=1}^{K^{(d)}-1} \mathbb{P}\left(X_n^{(d)}, S_n^{(d)} \,\middle|\, X_{n+1}^{(d)}, S_n^{(d)} \le T_c - \sum_{w=n+1}^{K^{(d)}} S_w^{(d)}, \underbrace{\{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)} \in \mathcal{T}/\cup_{v=n+1}^{K^{(d)}} \mathcal{T}_v}}_{\text{Observable variables up to state } n}, \Gamma^{p-1}\right).
\end{aligned}
\tag{29}
$$

Part (a) in (29) decomposes the likelihood of the latent trajectory (using the Markovian nature of the process) into factors in which the likelihood of every state n is conditioned on the future trajectory starting from n (i.e. the states X_{n+1} up to the absorbing states, together with their corresponding sojourn times). In part (c), a sufficient statistic for the distribution of the sojourn time of state n is the time elapsed in the episode up to state n, i.e. the duration of state n cannot exceed the difference between the censoring time T_c and the sojourn time of the futuristic trajectory that stems from state n. In part (b), we further reduce the terms in the future trajectory that are relevant to sampling the past trajectory: a sufficient statistic for state n is state $n+1$, and the observable variables that do not lie in the futuristic trajectory. Thus, the factorization in part (c) of (29) shows that the likelihood of the n^{th} state and sojourn time depends on the future trajectory only through the next state, i.e. state $n+1$, the time elapsed in the episode by the end of state n, and the observable variables up to state n. Using Baye's rule, we can further represent the factors

in part (c) of (29) in terms of familiar quantities that characterize the HASMM as follows

$$\mathbb{P}\left(X_n^{(d)}, S_n^{(d)} \middle| X_{n+1}^{(d)}, S_n^{(d)} \leq T_c - \sum_{w=n+1}^{K^{(d)}} S_w^{(d)}, \{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)} \in \mathcal{T}/\cup_{v=n+1}^{K^{(d)}} \mathcal{T}_v}, \Gamma^{p-1}\right)$$
$$\propto \underbrace{\mathbb{P}\left(X_n^{(d)} \middle| \{y^{(d)}(t_m^{(d)})\}_{t_m^{(d)} \in \mathcal{T}/\cup_{v=n+1}^{K^{(d)}} \mathcal{T}_v}, \Gamma^{p-1}\right)}_{\text{Forward message}} \times \underbrace{\mathbb{P}\left(X_{n+1}^{(d)} \middle| X_n^{(d)}, S_n^{(d)}, \Gamma^{p-1}\right)}_{\text{Transition function}} \times$$
$$\underbrace{\mathbb{P}\left(S_n^{(d)} \middle| X_n^{(d)}, S_n^{(d)} \leq T_c - \sum_{w=n+1}^{K^{(d)}} S_w^{(d)}, \Gamma^{p-1}\right)}_{\text{Truncated sojourn time distribution}}. \tag{30}$$

Thus, a sampler for the latent states trajectories can be constructed using the forward messages (which we can compute via the `ForwardFilter` routine using the p^{th} iteration's parameter set Γ^{p-1}), the p^{th} estimate of the HASMM's transition functions $(g_{ij}(s))_{i,j}$, and the p^{th} estimate of the sojourn time distributions (which we have specified to be the Gamma distribution). A compact representation for the factors in (30) is given by

$$\begin{aligned}
\alpha_{\bar{m}}^{d,p}(j) &= \mathbb{P}\left(X_n^{(d)} = j \middle| \{y^{(d)}(t_m^{(d)})\}_{m=1}^{\bar{m}}, \Gamma^{p-1}\right), j \in \mathcal{X}, \\
g_{ij}^p(s) &= \mathbb{P}\left(X_{n+1}^{(d)} = j \middle| X_n^{(d)} = i, S_n^{(d)} = s, \Gamma^{p-1}\right), i, j \in \mathcal{X}, \\
v_j(s|\lambda_j^{p-1}) &= \mathbb{P}\left(S_n^{(d)} = s \middle| X_n^{(d)} = j, \Gamma^{p-1}\right), j \in \mathcal{X},
\end{aligned} \tag{31}$$

where the truncated sojourn time distribution, which captures the sojourn time of a state conditioned on the time elapsed in the episode, is given by

$$\mathbb{P}\left(S_n^{(d)} = s \middle| X_n^{(d)}, S_n^{(d)} \leq \bar{s}, \Gamma^{p-1}\right) = \frac{v_j(s|\lambda_j^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_j(\bar{s}|\lambda_j^{p-1})}. \tag{32}$$

Given (30), (31) and (32), the sampler in (27) boils down to a sampler that operates sequentially in the reverse time direction by sampling from the posterior probability of every state n given the future trajectory of states that starts from state $n+1$, i.e. in the p^{th} iteration of the EM algorithm, state n in the g^{th} sample of episode d is sampled as follows

$$(X_n^{(d,p,g)} = i, S_n^{(d,p,g)} = s) \Big| X_{n+1}^{(d,p,g)} = j \sim \frac{\alpha_{\bar{m}}^p(i) \cdot g_{ij}^p(s) \cdot \frac{v_i(s|\lambda_i^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_i(\bar{s}|\lambda_i^{p-1})}}{\sum_{k=1}^N \alpha_{\bar{m}}^p(k) \cdot g_{kj}^p(s) \cdot \frac{v_k(s|\lambda_k^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_k(\bar{s}|\lambda_k^{p-1})}}, \tag{33}$$

where $\bar{s} = T_c - \sum_{w=1}^{K^{(d,p,g)}} S_w^{(d,p,g)}$ is the time elapsed in episode d by the end of state n, and $\bar{m} = \arg\max_m \left\{t_m^{(d)} : t_m^{(d)} \in \mathcal{T}/\bigcup_{v=n+1}^{K^{(d,g)}} \mathcal{T}_v\right\}$ is the index of the most recent observation sample that does not belong to the future trajectory. It is clear from (33) that sampling a trajectory requires first a *forward pass* on the episode in which the forward messages are computed via the forward filtering algorithm using the current EM estimated of Γ (i.e.

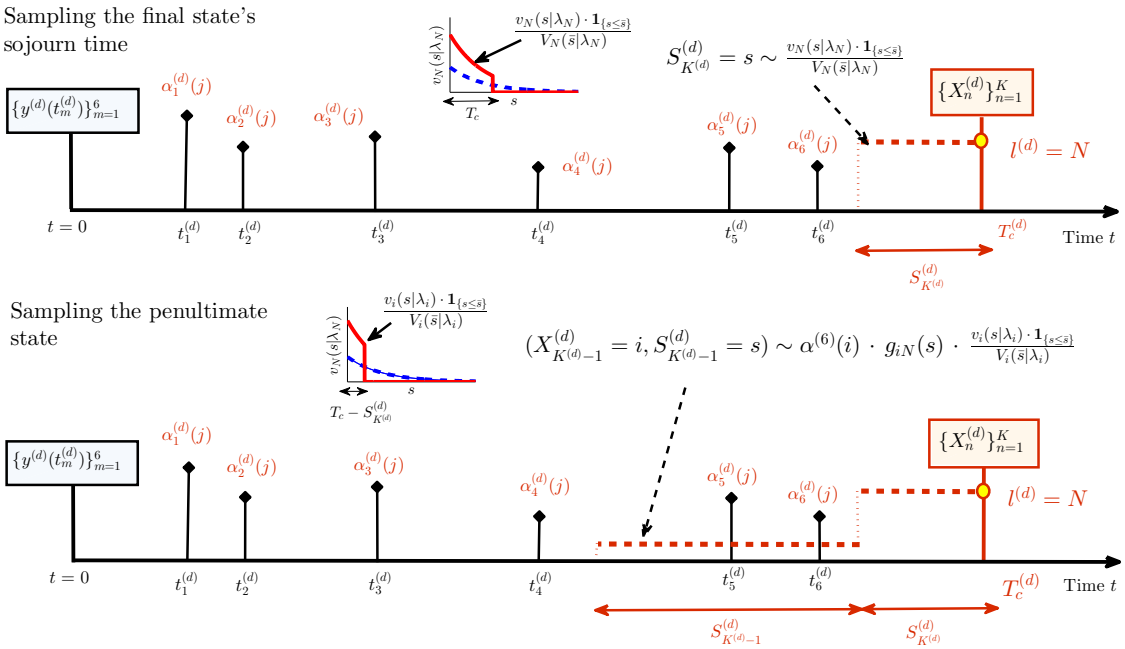

Figure 9: Depiction of the backward sampling pass for the last and penultimate states for an episode d. States are sampled in the reverse-time direction using the forward messages and the sampled future trajectory.

forward filtering), and then a *backward pass* is applied on the episode where starting from the censoring time, we sequentially sample states in the reverse time direction using the forward messages computed in the forward pass as described in (33).

Now that we have described the forward-filtering backward-sampling procedure for sampling the latent state trajectories conditioned on the episodes in the dataset $\mathcal{D}$, we provide a complete recipe for the Monte Carlo EM algorithm in terms of 4 main steps as follows

- **Step 1: The forward filtering pass**
  For every episode d in $\mathcal{D}$, compute the forward messages for all time instances $t_m^{(d)} \in \mathcal{T}^{(d)}$ using the current estimate for the parameter set Γ^{p-1}, i.e. invoke the routine $\texttt{ForwardFilter}(\Gamma^{p-1}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}}, \epsilon)$. The m^{th} forward message for episode d in the p^{th} iteration is denoted as $\alpha_m^p(j)$, $\forall j \in \mathcal{X}$.

- **Step 2: The backward sampling pass**
  Generate G sample trajectories for every episode d as follows:

  1. Given the absorbing state $l^{(d)}$, sample $S^{(d,p,g)}_{K^{(d,p,g)}}$ [18]

$$S^{(d)}_{K^{(d,p,g)}} \sim v_{l^{(d)}}(s|\lambda^{p-1}_{l^{(d)}}).$$

18. Note that $K^{(d,p,g)}$ is random and is not known ahead of time; however we will index the state variables with respect $K^{(d,p,g)}$ for simplicity of exposition.

2. Traverse backwards in time and sequentially sample states $K^{(d,p,g)}-1$, $K^{(d,p,g)}-2$,..., 1. The n^{th} state is sampled as follows

$$\bar{m} = \arg\max_m \left\{ t_m^{(d)} : t_m^{(d)} \in \mathcal{T} / \bigcup_{v=n+1}^{K^{(d,p,g)}} \mathcal{T}_v \right\}.$$

$$\bar{s} = T_c - \sum_{w=1}^{K^{(d,p,g)}} S_w^{(d,p,g)}$$

$$(X_n^{(d,p,g)} = i, S_n^{(d,p,g)} = s) \Big| X_{n+1}^{(d,p,g)} = j \sim \frac{\alpha_{\bar{m}}^p(i) \cdot g_{ij}^p(s) \cdot \frac{v_i(s|\lambda_i^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_i(\bar{s}|\lambda_i^{p-1})}}{\sum_{k=1}^N \alpha_{\bar{m}}^p(k) \cdot g_{kj}^p(s) \cdot \frac{v_k(s|\lambda_k^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_k(\bar{s}|\lambda_k^{p-1})}}.$$

The sequential sampling process above proceeds until $\bar{s}$ becomes sufficiently small, i.e. almost all the patient's episode duration is covered with a sampled latent state. Figure 9 provides a pictorial depiction for the process of sampling a single trajectory g using the procedure described above.

3. **Step 3: The E-step**
Compute the proximal log-likelihood function using the Monte Carlo approximation of average complete likelihood computed for all the sample trajectories generated in Setp 3:

$$Q(\Gamma; \Gamma^{p-1}) = \sum_{d=1}^{D} \frac{1}{G} \sum_{g=1}^{G} \log \left(\mathbb{P} \left(\{X_n^{(d,p,g)}, S_n^{(d,p,g)}\}_{n=1}^{K^{(d,p,g)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\middle|\, \Gamma \right) \right).$$

4. **Step 4: The M-step**
Update the HASMM parameters set $\Gamma^p = \arg\max_\Gamma Q(\Gamma; \Gamma^{p-1})$, and then go to Step 1.

Two obstacles hinder the direct application of the 4 steps listed above. First, sampling the bivariate random variable $(X_n^{(d,p,g)} = i, S_n^{(d,p,g)} = s)$ requires marginalizing over one of the two variables, which yields an intractable integral. Second, sampling G trajectories for every episode in every iteration of the EM algorithm can be computationally expensive. To overcome the first obstacle, we implement Step 2 via a Gibbs sampler, which operates as follows

If $K^{(d,p,g)} - n + 1 \leq K^{(d,p,g-1)}$:

$$X_n^{(d,p,g)} = i \Big| S_n^{(d,p,g-1)} = s, X_{n+1}^{(d,p,g)} = j \sim \frac{\alpha_{\bar{m}}^p(i) \cdot g_{ij}^p(s) \cdot \frac{v_i(s|\lambda_i^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_i(\bar{s}|\lambda_i^{p-1})}}{\sum_{k=1}^N \alpha_{\bar{m}}^p(k) \cdot g_{kj}^p(s) \cdot \frac{v_k(s|\lambda_k^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_k(\bar{s}|\lambda_k^{p-1})}}.$$

$$S_n^{(d,p,g)} = s \Big| X_n^{(d,p,g)} = i \sim \frac{v_i(s|\lambda_i^{p-1}) \cdot \mathbf{1}_{\{s \leq \bar{s}\}}}{V_i(\bar{s}|\lambda_i^{p-1})}$$

If $K^{(d,p,g)} - n + 1 > K^{(d,p,g-1)}$:

$$X_n^{(d,p,g)} = i \sim \alpha_{\bar{m}}^p(i)$$

$$S_n^{(d,p,g)} = s \,\Big|\, X_n^{(d,p,g)} = i, X_{n+1}^{(d,p,g)} = j \sim \frac{\alpha_{\bar{m}}^p(i) \cdot g_{ij}^p(s) \cdot \frac{v_i(s|\lambda_i^{p-1}) \cdot \mathbf{1}_{\{s \le \bar{s}\}}}{V_i(\bar{s}|\lambda_i^{p-1})}}{\sum_{k=1}^N \alpha_{\bar{m}}^p(k) \cdot g_{kj}^p(s) \cdot \frac{v_k(s|\lambda_k^{p-1}) \cdot \mathbf{1}_{\{s \le \bar{s}\}}}{V_k(\bar{s}|\lambda_k^{p-1})}}. \tag{34}$$

That is, if the state n in sample trajectory g has a counterpart in sample $g-1$ (i.e. the length of sample trajectory $g-1$ is large enough that it has more than n states), then we sample state n in trajectory g conditioned on the sojourn time of state n in trajectory $g-1$, and then sample the sojourn time of state n using the truncated sojourn time distribution. If state n has no counterpart in trajectory $g-1$, we first sample state n from the corresponding forward message, and then we sample the sojourn time conditioned on the state realization.

The second obstacle is solved by sampling G trajectories of the latent states only once, and then using these samples in all EM iterations, but with an adjustment for the computed proximal log-likelihood using importance weights (Booth and Hobert (1999)). That is, in the initial iteration, we generate the sample trajectories

$$\left(\{X_n^{(d,o,g)}, S_n^{(d,o,g)}\}_{n=1}^{K^{(d,o,g)}}\right)_{g=1}^G,$$

and then in the p^{th} EM iteration we implement the E-step as follows

$$Q(\Gamma; \Gamma^{p-1}) =$$

$$\sum_{d=1}^D \frac{1}{G} \sum_{g=1}^G \log\left(\mathbb{P}\left(\{X_n^{(d,o,g)}, S_n^{(d,o,g)}\}_{n=1}^{K^{(d,o,g)}}, \{y^{(d)}(t_m^{(d)})\}_{m=1}^{M^{(d)}} \,\Big|\, \Gamma\right)\right) \cdot \underbrace{\frac{\mathbb{P}\left(\{X_n^{(d,o,g)}, S_n^{(d,o,g)}\}_{n=1}^{K^{(d,o,g)}} \,\Big|\, \Gamma^{p-1}\right)}{\mathbb{P}\left(\{X_n^{(d,o,g)}, S_n^{(d,o,g)}\}_{n=1}^{K^{(d,o,g)}} \,\Big|\, \Gamma^o\right)}}_{\text{Importance weights}},$$

which makes it sufficient for the EM algorithm to rely on one sample of the latent trajectories in all its iterations, without the need to re-run the forward-filtering backward-sampling algorithm in each EM iteration.

4. Experiments

Experiments were conducted on data from a cohort of 6,321 patients who were hospitalized in a general medicine floor in a large academic medical center during the period between March 3^{rd} 2013, to February 4^{th} 2016. The patient population is heterogeneous with a wide variety of diagnoses. The patients in the cohort had a wide variety of diagnoses including septicemia, leukemia, hypertension, pneumonia, anemia, renal failure, heart failure, etc. Some of these patients underwent organ transplant surgeries or received chemotherapy which significantly ablate their immune system and leaves them at increased risk of clinical deterioration. Of the 6,321 patients, around 5% were admitted to the ICU; we handle the

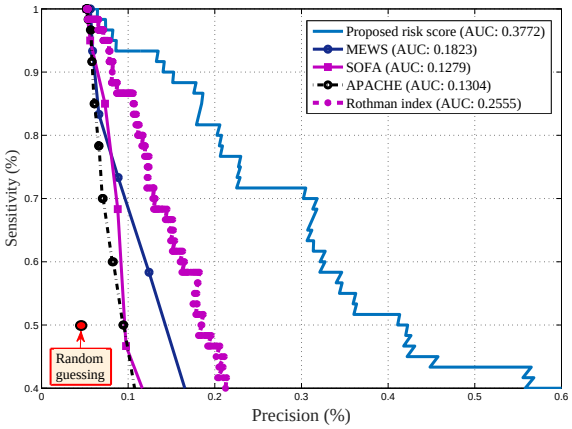

Figure 10: The ROC curve for the proposed risk score compared to state-of-the-art scores.

data imbalance by focusing on the sensitivity and precision measures for accuracy. The data associated with each patient involves 17 temporal physiological data streams that comprise vital signs (diastolic and systolic, blood pressure, Glasgow coma scale score, heart rate, respiratory rate, temperature, O_2 saturation, etc) and laboratory tests (white blood cell, Hemoglobin, Glucose, etc). The vital signs are sampled (approximately) once every 4 hours, whereas the laboratory tests are typically conducted every 24 hours. The length of the oatients' stay in the ward ranged from 4 hours to 2000 hours.

We first run our SMC-EM algorithm on the training set in order to find the HASMM parameter set Γ that best describes the observed episodes (physiological histories of the patients). We note that unlike the case of disease progression models where domain knowledge can inform the number of states, e.g. long-term stages of chronic disease progression (Liu et al. 2015), there is no domain knowledge on the nature of the clinical states for subacute care patients. Hence, we select the number of states via model selection; the Bayesian Information Criterion is used to select the least complex model that fits the observed episodes. The average sojourn time in each state was about 12 hours.

We validated the utility of the proposed risk score by evaluating the sensitivity, precision and timeliness of the early warning alarms prompted by our system as compared to the state-of-the-art risk scores currently deployed in hospital wards; namely MEWS and Rothman index, in addition to the APACHE and SOFA scores, which are normally used to predict mortality in the ICU, but have been recently validated for prognostication in wards (Yu et al. 2014). The Rothman index is the state-of-the-art risk scoring methodology in wards and is currently deployed in more than 70 hospitals in the US (Finaly 2014). Comparisons with baseline predictors including linear regression, random forest and LASSO were also conducted. The training set $\mathcal{D}$ comprises 5,130 episodes for patients who were admitted to the ward in the period between March 2013 and July 2015, whereas the remaining (most recently) admitted patients' episodes were used for testing.

Fig 2 demonstrates the ROC curves achieved by MEWS, Rothman index, APACHE, SOFA and the proposed risk score. It can be seen that the ROC curve achieved by the proposed score dominates those achieved by all other scores for all settings of precision and

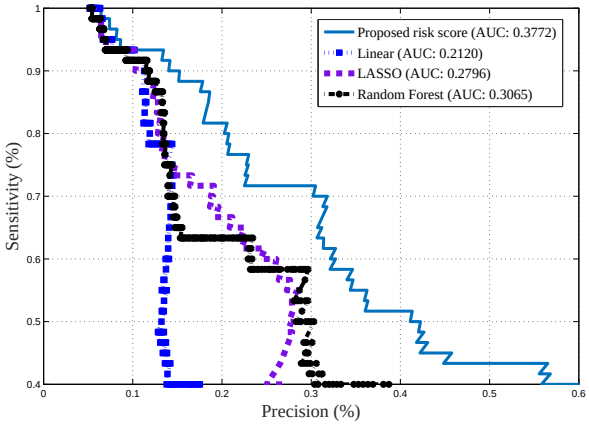

Figure 11: The ROC curve for the proposed risk score compared to baseline classifiers.

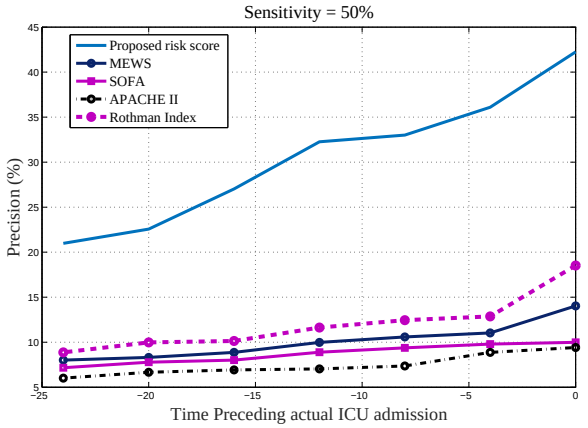

Figure 12: The timeliness of the proposed risk score compared to state-of-the-art scores.

sensitivity. The proposed risk score outperforms the Rothman index by many –perhaps all– measures. In particular, the proposed score offers a gain of 23.3% (p-value < 0.01) with respect to the AUC of the most competitive risk score, the Rothman index. Moreover, the proposed score provides significant improvements in precision at all sensitivity levels. For instance, for a sensitivity of 50%, the proposed risk score achieves a precision of 41% which is around 22% higher than that achieved by the Rothman index for the same sensitivity. This means that the proposed risk score can significantly reduce the rate of false ICU alarms in the subacute care wards, which would mitigate alarm fatigue and enhance a hospital's resource utilization. Fig 3 demonstrates the performance of the proposed risk score as compared to the baseline predictors; an AUC gain of 7% (p-value < 0.01) compared to random forest is reported.

To demonstrate the potential reductions of the false alarm rates that would result from improvements in the risk scoring precision achieved by our model, we list the number of false alarms per one true alarm for all the proposed risk model and the Rothman index at different levels of sensitivity in Table 1. As we can see, at a sensitivity of 50%, the proposed

risk score leads to only 0.84 false alarms for every 1 true alarm, whereas the Rothman index lead to 2.34 false alarms per true alarm.

The improvements achieved by our model can be attributed to: incorporating the entire physiological trajectory while assessing the current risk via forward-filtering (the Rothman index considers the only the latest physiological measurement), and accounting for temporal correlations and the cross correlations among the different vital signs and lab tests measurements via the multitask GP model (the Rothman index ignores those correlations and hence double-counts risk factors (see Eq. (1) in (Rothman et al. 2013)).

Table 1: Number of false alarm per one true alarm for different levels of sensitivity.

Sensitivity	**0.6**	**0.55**	**0.5**	**0.45**
Proposed risk score	1.14	0.96	0.84	0.8
Rothman index	3.11	2.67	2.34	2.01

Note that unlike state-of-the-art risk scores such as the Rothman index, which assigns high risk scores only to patients who appear to be in the absorbing clinical deterioration state (See Fig A1 in (Rothman et al. 2013)), our algorithm introduces foresightedness in the risk scoring methodology; it computes a patient's risk score taking into account the future trajectory of state evolution and not just the estimated acuity at the current moment, which provides significant gains in terms of the timeliness of its early warnings as compared to the other risk scores. This is illustrated in Fig 6 where we show the trade-off between the timeliness of an ICU admission alarm and its accuracy. It can be seen that for a sensitivity of 50% and precision of 32%, the proposed risk model can issue ICU alarms that are as early as 10 hours before the actual clinician's ICU transfer decision. Note that for the same prediction time and the same sensitivity, the Rothman index can only provide a precision of 12%. Therefore, the proposed risk model can provide the ward staff with a greater safety net for focusing their attention and delivering the care to the patients who are in real need in a timely manner, and allow them to plan for early ICU transfers that can boost the efficacy of consequent therapeutic interventions.

5. Conclusions

In this paper, we have developed a risk scoring and early warning system that can predict clinical deterioration for monitored patients on the wards, allowing for timely ICU admission and more efficient therapeutic interventions. The proposed risk scoring algorithm is based on a novel Hidden Absorbing Semi-Markov Model (HASMM) that relates a patient's evolving acuity to her observed physiology, and captures important aspects of the physiological data gathering process, such as the irregularly sampled physiological measurements and the informatively censored patients' episodes. We developed novel inference and learning algorithms that can use the EHR data to calibrate the HASMM model parameters and compute the patients' risks in real-time. Experiments conducted on a heterogeneous cohort of 6,321 patients show that the proposed risk score significantly outperforms the state-of-

the-art risk scoring technologies in terms of accuracy and timeliness, which translates into a significantly improved subacute care in hospital wards.

References

Jeffrey A Bakal, Finlay A McAlister, Wei Liu, and Justin A Ezekowitz. Heart failure re-admission: measuring the ever shortening gap between repeat heart failure hospitalizations. *PloS one*, 9(9):e106494, 2014.

Jirina Bartkova, Zuzana Hořejší, Karen Koed, Alwin Krämer, Frederic Tort, Karsten Zieger, Per Guldberg, Maxwell Sehested, Jahn M Nesland, Claudia Lukas, et al. Dna damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature*, 434 (7035):864–870, 2005.

Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2007.

James G Booth and James P Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, 1999.

Brian S Caffo, Wolfgang Jank, and Galin L Jones. Ascent-based monte carlo expectation–maximization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):235–251, 2005.

Dustin Charles, Meghan Gabriel, and JaWanna Henry. Electronic capabilities for patient engagement among us non-federal acute care hospitals: 2012-2014. *The Office of the National Coordinator for Health Information Technology*, 2015.

Baojiang Chen and Xiao-Hua Zhou. Non-homogeneous markov process models with informative observations with an application to alzheimer's disease. *Biometrical Journal*, 53 (3):444–463, 2011.

Jill M Cholette, Kelly F Henrichs, George M Alfieris, Karen S Powers, Richard Phipps, Sherry L Spinelli, Michael Swartz, Francisco Gensini, L Eugene Daugherty, Emily Nazarian, et al. Washing red blood cells and platelets transfused in cardiac surgery reduces post-operative inflammation and number of transfusions: Results of a prospective, randomized, controlled clinical trial. *Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, 13(3), 2012.

David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

Zelalem Getahun Dessie. Multi-state models of hiv/aids by homogeneous semi-markov process. *American Journal of Biostatistics*, 4(2):21, 2014.

Michael Dewar, Chris Wiggins, and Frank Wood. Inference in hidden markov models with explicit state duration distributions. *IEEE Signal Processing Letters*, 19(4):235–238, 2012.

Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

Allison A Eddy and Eric G Neilson. Chronic kidney disease progression. *Journal of the American Society of Nephrology*, 17(11):2964–2966, 2006.

Yohann Foucher, Eve Mathieu, Philippe Saint-Pierre, J Durand, and J Daures. A semi-markov model based on generalized weibull distribution with an illustration for hiv disease. *Biometrical journal*, 47(6):825, 2005.

Yohann Foucher, Magali Giral, Jean-Paul Soulillou, and Jean-Pierre Daures. A semi-markov model for multistate and interval-censored data with multiple terminal events. application in renal transplantation. *Statistics in medicine*, 26(30):5381–5393, 2007.

Yohann Foucher, M Giral, JP Soulillou, and JP Daures. A flexible semi-markov model for interval-censored data and goodness-of-fit testing. *Statistical methods in medical research*, 2008.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056, 2011.

Mitchell H Gail and Phuong L Mai. Comparing breast cancer risk assessment models. *Journal of the National Cancer Institute*, 102(10):665–668, 2010.

Valentine Genon-Catalot, Thierry Jeantheau, Catherine Larédo, et al. Stochastic volatility models as hidden markov models and statistical applications. *Bernoulli*, 6(6):1051–1079, 2000.

Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

Giacomo Giampieri, Mark Davis, and Martin Crowder. Analysis of default data using hidden markov models. *Quantitative Finance*, 5(1):27–34, 2005.

Florence Gillaizeau, Etienne Dantan, Magali Giral, and Yohann Foucher. A multistate additive relative survival semi-markov model. *Statistical methods in medical research*, page 0962280215586456, 2015.

Simon J Godsill, Arnaud Doucet, and Mike West. Monte carlo smoothing for nonlinear time series. *Journal of the american statistical association*, 2012.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. Hidden topic markov models. In *AISTATS*, volume 7, pages 163–170, 2007.

Yann Guédon. Exploring the state sequence space for hidden markov and semi-markov chains. *Computational Statistics & Data Analysis*, 51(5):2379–2409, 2007.

Chantal Guihenneuc-Jouyaux, Sylvia Richardson, and Ira M Longini. Modeling markers of disease progression by a hidden markov process: application to characterizing cd4 cell decline. *Biometrics*, 56(3):733–741, 2000.

Tracy D Gunter and Nicolas P Terry. The emergence of national electronic health record architectures in the united states and australia: models, costs, and questions. *Journal of medical Internet research*, 7(1):e3, 2005.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503, 1974.

Asger Hobolth and Jens Ledet Jensen. Summary statistics for endpoint-conditioned continuous-time markov chains. *Journal of Applied Probability*, pages 911–924, 2011.

Helen Hogan, Frances Healey, Graham Neale, Richard Thomson, Charles Vincent, and Nick Black. Preventable deaths due to problems in care in english acute hospitals: a retrospective case record review study. *BMJ quality & safety*, pages bmjqs–2012, 2012.

Xuelin Huang and Robert A Wolfe. A frailty model for informative censoring. *Biometrics*, 58(3):510–520, 2002.

Aparna V Huzurbazar. Multistate models, flowgraph models, and semi-markov processes. 2004.

Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

Jacques Janssen and R De Dominicis. Finite non-homogeneous semi-markov processes: Theoretical and computational aspects. *Insurance: Mathematics and Economics*, 3(3): 157–165, 1984.

Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-markov models. *Journal of Machine Learning Research*, 14(Feb):673–701, 2013.

Pierre Joly and Daniel Commenges. A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to aids. *Biometrics*, 55(3): 887–890, 1999.

Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.

Juliane Kause, Gary Smith, David Prytherch, Michael Parr, Arthas Flabouris, Ken Hillman, et al. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in australia and new zealand, and the united kingdomthe academia study. *Resuscitation*, 62(3):275–282, 2004.

Lisa L Kirkland, Michael Malinchoc, Megan OByrne, Joanne T Benson, Deanne T Kashiwagi, M Caroline Burton, Prathibha Varkey, and Timothy I Morgenthaler. A clinical deterioration prediction tool for internal medicine patients. *American Journal of Medical Quality*, 28(2):135–142, 2013.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

William A Knaus, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, and Anne Damiano. The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *Chest Journal*, 100(6):1619–1636, 1991.

Vidyadhar G Kulkarni. *Modeling and analysis of stochastic systems*. CRC Press, 1996.

Stephan W Lagakos, Charles J Sommer, and Marvin Zelen. Semi-markov models for partially censored data. *Biometrika*, 65(2):311–317, 1978.

David Lando. On cox processes and credit risky securities. *Review of Derivatives research*, 2(2-3):99–120, 1998.

Jose Leiva-Murillo, AA Rodrguez, and E Baca-Garca. Visualization and prediction of disease interactions with continuous-time hidden markov models. In *NIPS 2011 Workshop on Personalized Medicine*, 2011.

H Lehman Li-wei, Shamim Nemati, Ryan P Adams, George Moody, Atul Malhotra, and Roger G Mark. Tracking progression of patient state of health in critical care using inferred shared dynamics in physiological time series. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7072–7075. IEEE, 2013.

William A Link. A model for informative censoring. *Journal of the American Statistical Association*, 84(407):749–752, 1989.

Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, and James M Rehg. Efficient learning of continuous-time hidden markov models for disease progression. In *Advances in neural information processing systems*, pages 3600–3608, 2015.

Sergio Matos, Surinder S Birring, Ian D Pavord, and H Evans. Detection of cough signals in continuous audio recordings using hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53(6):1078–1083, 2006.

Philipp Metzner, Illia Horenko, and Christof Schütte. Generator estimation of markov jump processes based on incomplete observations nonequidistant in time. *Physical Review E*, 76(6):066702, 2007.

Rui P Moreno, Philipp GH Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, Jean-Roger Le Gall, et al. Saps 3from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive care medicine*, 31(10):1345–1355, 2005.

DR Mould. Models for disease progression: new approaches and uses. *Clinical Pharmacology & Therapeutics*, 92(1):125–131, 2012.

Kevin P Murphy. Hidden semi-markov models (hsmms). *unpublished notes*, 2, 2002.

Uri Nodelman, Christian R Shelton, and Daphne Koller. Expectation maximization and complex duration distributions for continuous time bayesian networks. *arXiv preprint arXiv:1207.1402*, 2012.

Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

Mari Ostendorf, Vassilios V Digalakis, and Owen A Kimball. From hmm's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on speech and audio processing*, 4(5):360–378, 1996.

Soren Erik Pedersen, Suzanne S Hurd, Robert F Lemanske, Allan Becker, Heather J Zar, Peter D Sly, Manuel Soto-Quiroz, Gary Wong, and Eric D Bateman. Global strategy for the diagnosis and management of asthma in children 5 years and younger. *Pediatric pulmonology*, 46(1):1–17, 2011.

Andrei D Polyanin and Alexander V Manzhirov. *Handbook of integral equations*. CRC press, 2008.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

Michael J Rothman, Steven I Rothman, and Joseph Beals. Development and validation of a continuous measure of patient condition using the electronic medical record. *Journal of biomedical informatics*, 46(5):837–848, 2013.

Daniel O Scharfstein and James M Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634, 2002.

Peter Schulam and Suchi Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.

Padhraic Smyth. Hidden markov models for fault detection in dynamic systems. *Pattern recognition*, 27(1):149–164, 1994.

Henry T Stelfox, Brenda R Hemmelgarn, Sean M Bagshaw, Song Gao, Christopher J Doig, Cheri Nijssen-Jordan, and Braden Manns. Intensive care unit bed availability and outcomes for hospitalized patients with sudden clinical deterioration. *Archives of internal medicine*, 172(6):467–474, 2012.

CP Subbe, M Kruger, P Rutherford, and L Gemmel. Validation of a modified early warning score in medical admissions. *Qjm*, 94(10):521–526, 2001.

MJ Sweeting, VT Farewell, and D De Angelis. Multi-state markov models for disease progression in the presence of informative examination times: An application to hepatitis c. *Statistics in medicine*, 29(11):1161–1174, 2010.

S Taghipour, D Banjevic, AB Miller, N Montgomery, AKS Jardine, and BJ Harvey. Parameter estimates for invasive breast cancer progression in the canadian national breast screening study. *British journal of cancer*, 108(3):542–548, 2013.

John Varga, Christopher P Denton, and Fredrick M Wigley. *Scleroderma: From pathogenesis to comprehensive management*. Springer Science & Business Media, 2012.

Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

J Yoon, A Alaa, S Hu, and M van der Schaar. Forecasticu: A prognostic decision support system for timely prediction of intensive care unit admission. pages 1680–1689, 2016.

Shun-Zheng Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010.

Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.