A Hierarchical Inhomogeneous Hidden Markov Model for Cancer Screening Modeling

Abstract

Continuous-Time Hidden Markov Models are an attractive approach for modeling clinical disease progression data because they are capable to handle both irregularly sampled and noisy data. Most applications in this context consider time-homogeneous models due to their relative computational simplicity. However, the time homogeneous assumption is too strong to accurately model the natural history of many diseases. Moreover, the population at risk is not homogeneous neither, since disease exposure and susceptibility can vary considerably. In this paper, we propose piece-wise stationary transition matrix to explain the heterogeneity in time. We propose a hierarchical structure for the heterogeneity in population, where prior information is considered to deal with unbalanced data. Moreover, an efficient, scalable EM algorithm is proposed for inference. We exemplify the feasibility and superiority of our model on a cervical cancer screening dataset from the Cancer Registry of Norway. Experiments show that our model outperforms state-of-the-art recurrent neural network models in terms of prediction accuracy and significantly outperforms a standard Continuous-Time Hidden Markov Model in generating Kaplan-Meier estimators.

1 Introduction

Population-based screening programs for identifying undiagnosed individuals have a long history in improving public health. Examples include screening programs for cancer (e.g., cervical, breast, colon), tuberculosis and fetal abnormalities. While the primary objective of such programs is to identify and treat undiagnosed individuals, these cancer screening programs and the population-level, longitudinal datasets associated with them, present many opportunities for the data-driven, computational sciences. In conjunction with modern analytic and computational techniques, such data have the potential to yield novel insights into the natural history of diseases as well as improving the effectiveness of the screening programs.

Hidden Markov Models (HMM) are a standard choice for disease progression modeling for at least three reasons. First, the underlying disease is represented as an unobserved, latent Markov process. Second, noisy measurements of the disease states are efficiently incorporated as conditional probability distributions in the emission mechanism. Third, any modeling assumptions for a particular application are easily incorporated into the transition probability matrix and emission mechanism.

However, standard HMMs assume that measurements are regularly sampled at discrete intervals which is often not the case in disease screening programs. Measurements are often irregularly sampled because patients come in for screenings at irregular intervals, even if regular screening tests are recommended. To deal with irregular sampling, Continuous-Time Hidden Markov Models (CTHMM) are often used since they easily handle samples taken at arbitrary time intervals. CTHMMs have been proposed in many applications such as networks [33], medicine [6], seismology [24] and finance [18]. Liu summarizes current inference methods for CTHMMs in [23] and proposes efficient EM-based learning approaches.

Preprint. Under review.

Because the natural history of many diseases depends heavily on the age of the individual, the time-homogeneous assumption is not valid. For this reason, time-inhomogeneous HMMs are more appropriate. Although such models have many appealing theoretical properties according to the Kolmogorov equations [35], parameter inference is intractable in most non-trivial cases. For this reason, many inference studies of continuous-time, time-inhomogeneous HMMs (CTIHMMs) in the medical domain depend on inefficient microsimulations [29, 27, 7]

Because of the computational issues, many previous HMM models of disease progression assume that the observations come from a homogeneous population. In large populations, this will typically not be the case. For example, in population-level screening data a large proportion of individuals have benign test results while only a small proportion have abnormal test results. Frailty models are proposed as a common methodology in epidemiological modeling [34].

To deal with these difficulties we introduce piece-wise constant transition intensity functions, which allow for tractable parameter inference yet are considerably more flexible in terms of time-inhomogeneity. We then propose a latent structure (i.e., frailty model) to capture unobserved population heterogeneity in terms of disease exposure and susceptibility. Specifically, we propose a new hierarchical hidden Markov model for disease progression in which patients are categorized into classes based on risk levels. Due to the expensive cost of the standard EM algorithm inference, we propose an efficient and scalable EM algorithm combining both soft and hard assignment in the E-step and an auto-differentiation based Limited-memory BFGS optimization method in the M-step.

We apply this model to cervical cancer screening data from the Cancer Registry of Norway. This is a true population-level dataset with over 1.7 million women and more than 10 million screening results. Based on the cervical cancer screening data, our model is exemplified to have a better predictive accuracy compared with state of the art recurrent neural network models, based on AUC (Area Under the Curve) under a binary classification framework. Moreover, our model is significantly better than a simple hidden Markov model by comparing model-generated Kaplan-Meier curves with observed Kaplan-Meier curves.

2 Related Work

Longitudinal observation data exist widely, especially in the healthcare area. [28] proposed multiple self-controlled case series to model the multiple drug exposure based on conditional poisson regression. [2] extends it from discrete time to continuous time using Hawkes process modeling. Moreover, [20] propose baseline regularization to leverage the diverse health profiles for adverse drug events. It is also a generalized linear model extended from [19].

For a screening test, the health status is of interest. It is crucial to consider a latent model of health status. The hidden Markov model is the state of the art approach. Most hidden Markov model variants consider only discrete time [14, 5, 30]. Continuous time hidden Markov models can handle data at any time stamp [10] and therefore are suitable for irregularly-sampled longitudinal data [3, 22, 32]. Furthermore, [23] summarizes and discusses learning approaches for continuous time hidden Markov models and proposes efficient EM-based learning approaches. Since screening processes significantly depend on a patient's age, our model is based on a CTIHMM, which is discussed in [29, 27, 7].

As a comparator, we consider recurrent neural networks, which are able to deal with variable-length time series and capture the temporal correlation. Variants of RNNs have been proposed to better balance memory needs and new features. [15, 8] propose the concept of a Long Short-Term Memory (LSTM) architecture, with variants utilized in handwritting recognition [11], language modeling [26], and video data [36]. [9] proposes the gated recurrent neural network (GRU) as another state of the art RNN model.

3 Model

We propose a hierarchical inhomogeneous HMM (HIHMM) to model disease progression. Here we assume all patients come from two risk categories: high disease exposure risk and low disease exposure risk. Each category has its own Markov model shown in Figure 1. Details are discussed in Section 5. This hierarchical structure of the HIHMM allows for an arbitrary number of latent frailty states, provided relevant Markov models can be ascribed to the disease progression associated with each.

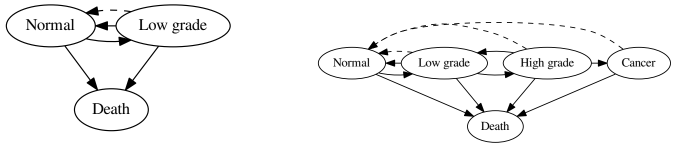

Figure 1: Transition structure of model $\mathcal{M}_0$ and $\mathcal{M}_1$. Solid lines denote the intensity transition while dashed lines denote that any state comes back to the normal state once treatment is completed.

3.1 Variables

Suppose there are N individuals in the screening population. Let individual n have T_n screening visits at ages $a_1, \ldots, a_{T_n}$. We assume Z categories are considered in the hierarchical frailty structure and introduce the following variables:

$$\begin{aligned} \text{Frailty State (hidden):} \quad & z_n \in \{1, \ldots, Z\} \\ \text{Disease States (hidden):} \quad & S_{nt} \in \{1, \ldots, M_{z_n}\} \\ \text{Number of screening tests (observable):} \quad & E_{ntk} \in \mathbb{N} \\ \text{Screening test results (observable):} \quad & G_{ntk} \in \mathbb{N}^{L_k} \end{aligned}$$

The underlying disease state of individual n is assumed to evolve according to a continuous-time, time-inhomogeneous Markov process assigned by its latent frailty class indicator z_n, where only screening results at specific time stamps with corresponding ages $a_1, \ldots a_{T_n}$ are observable. On the tth screening visit of individual n, S_{nt} refers to the latent disease state and the visit includes E_{ntk} tests of the kth test type and the corresponding results G_{ntk}, which is a L_k dimensional vector and the value on lth dimension refers to the number of the lth grade results.

3.2 Model of Disease Progression

As for the zth underlying Markovian disease process, it is parameterized by an $M_z \times M_z$ transition intensity matrix Q_z. For the simplicity of notation, we ignore the subscript z in the remainder of this section. The ijth element q_{ij} of Q satisfies $q_{ij} \geq 0$ for $i \neq j$ and $q_{ii} = -\sum_{i \neq j} q_{ij}$. The time spent in state i is exponentially distributed with rate $-q_{ii}$. Given that a transition occurs from state i, the probability of transitioning to state j is $\frac{q_{ij}}{q_i}$ where $q_i = \sum_{i \neq j} q_{ij}$. When Q is invariant for time t the model is homogeneous, otherwise the model is inhomegeneous.

3.2.1 Homogeneous Markov Model

For a homogeneous Markov process, we assume the initial state at t_0 is known, $p(S(t_1)) = 1$. We let $\boldsymbol{t}' = (t'_1, \ldots, t'_{T'})$ refer to the underlying transition timestamps and let $\boldsymbol{O} = (O_1, \ldots, O_T)$ denote observations at time $\boldsymbol{t} = (t_1, \ldots, t_T)$. Then the complete likelihood (CL) is given by

$$\begin{aligned} \text{CL} &= \prod_{i=1}^{T'} (q_{S(t'_i),S(t'_{i+1})} / q_{S(t'_i)}) q_{S(t'_i)} e^{-q_{S(t'_i)} \Delta_i} \prod_{j=1}^{T} p(O_j | S(t_j)) \\ &= \prod_{i=1}^{M} \left(e^{-q_i \tau_i} \prod_{j \neq i} q_{ij}^{n_{ij}} \right) \prod_{j=1}^{T} p(O_j | S(t_j)) \,, \qquad (1) \end{aligned}$$

where $\tilde{\Delta}_i = \tilde{t}_{i+1} - \tilde{t}_i$ and n_{ij} denotes the number of times the state changes from state i to state j during the whole process and τ_i denotes the duration that the process stays in state i. Since the underlying transition timestamps $\boldsymbol{t}'$ are not observable, the marginalized complete likelihood (MCL) is derived by marginalizing all $\boldsymbol{t}'$ as

$$\mathrm{MCL} = \prod_{i=1}^{T-1} P(\triangle_i)_{S(t_i),S(t_{i+1})} \prod_{j=1}^{T} p(O_j|S(t_j))\,,$$

where $\triangle_i = t_{i+1} - t_i$ and $P(\triangle_i) = e^{Q\triangle_i}$ is the transition probability matrix from time t_i to time t_{i+1}.

3.2.2 Inhomogeneous Markov Model

An inhomogeneous Markov process drops the time invariance assumption of Q by allowing Q to be a function of t. CL then becomes intractable, because the time spent in state i no longer follows an exponential distribution. An alternative approach is to consider the MCL. The only difference of the expression of MCLs between homogeneity and inhomogeneity is the computation of the transition matrix $P([t_i, t_{i+1}])$ from time t_i to time t_{i+1} for $i = 1, \dots, T-1$. For the inhomogeneous model, $P([t_i, t_{i+1}]) = \exp\{\int_{t_i}^{t_{i+1}} Q(t)dt\}$.

The transition intensity function $Q(t)$ can be modeled by any parametric model, but the computation of the matrix exponential $\exp\{\int_{t_i}^{t_{i+1}} Q(t)dt\}$ may be prohibitively expensive, even taking advantage of numerical computational methods. To ease this computational burden, we propose a piecewise constant transition intensity matrix Q, where each element q_{ij} is a piecewise constant function of time. Specifically, we partition time into I disjoint intervals covering the range of observable time. We then have a set of disjoint partitions $\mathcal{A} = \{A_i\}_{i=1}^{I}$. Each transition intensity function q_{ij} is a piecewise constant function via the defined partition $\mathcal{A}$, denoted by $q_{ij}(t) = \sum_{k=1}^{I} q_{ijk}\mathbf{1}_{A_k}(t)$, where $\mathbf{1}(\cdot)$ is a Kronecker delta function and $q_{ij\ell} \geq 0$. In this case, the inhomogeneous Markov process can be treated as a combination of several continuous-time homogeneous Markov processes, and the transition probability matrix Q can be computed as a product of transition probability matrices with respect to their corresponding partitions.

3.3 Hierarchical Model

Due to the significant population heterogeneity related to disease exposure risk, we propose a hierarchical model as follows. Let $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_Z)$ denote all model parameters and $\boldsymbol{\psi}_z$ be parameters for model z. Then the hierarchical model is given by

$$\begin{aligned} \boldsymbol{O}_n &\sim \mathcal{M}_{z_n}(\boldsymbol{\psi}_{z_n}, \boldsymbol{\theta}_n)\,, \\ z_n &\sim \mathrm{Cat}(\boldsymbol{p})\,, \end{aligned}$$

where $\boldsymbol{\theta}_n$ denotes all covariates for individual n. An informative prior of the model indicator z_n is proposed as a categorical distribution with hyper-parameters $\boldsymbol{p}$, which is used to provide expert knowledge of the model assignment. This prior contributes to reasonable model inference, especially when screening data are highly unbalanced in terms of latent class membership. Figure 1 shows the case where $Z = 2$ and index z_n has a Bernoulli prior with a parameter p, i.e., $z_n \sim \mathrm{Ber}(p)$.

4 Inference

Due to the latent characteristics of both model indices and patient states, the expectation maximization (EM) approach is considered. Moreover, considering the heterogeneity of our model, we marginalize the latent transition timestamps in our inference. We decompose the joint posterior distribution as $p(z_n, \boldsymbol{S}_n|-) = p(z_n|-)p(\boldsymbol{S}_n|z_n, -)$ and use soft assignment for $p(z_n|-)$ and hard assignment for $p(\boldsymbol{S}_n|z_n, -)$, because the computation of $p(\boldsymbol{S}_n|z_n, -)$ is prohibitively expensive. For simplicity, we ignore covariates $\boldsymbol{\theta}_n$ in the reminder of this section. Then recursive procedures are given as follows.

- Given previous estimates $\boldsymbol{\psi}^{(t-1)}$, compute the conditional posterior distribution of z_n:

  $$\begin{aligned} p(z_n|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) &\propto \mathrm{Cat}(z_n|\boldsymbol{p})p(\boldsymbol{O}_n|z_n, \boldsymbol{\psi}_{z_n}^{(t-1)}) \\ &\sim \mathrm{Cat}(\tilde{\boldsymbol{p}}_n)\,, \end{aligned} \tag{2}$$

  where $\tilde{p}_{nk} = \frac{p_k p(\boldsymbol{O}_n|z_n=k, \boldsymbol{\psi}_k^{(t-1)})}{\sum_{z=1}^{Z} p_z p(\boldsymbol{O}_n|z_n=z, \boldsymbol{\psi}_z^{(t-1)})}$ for $k = 1, \dots, Z$ and $p(\boldsymbol{O}_n|z, \boldsymbol{\psi}_z)$ is accessible through the forward-filter backward-sample algorithm (FFBS), which is a sequential Monte Carlo approach first proposed in [17].

- Update the optimal state sequence $\boldsymbol{S}_n$ given corresponding observations $\boldsymbol{O}_n$ and model indicator z using the Viterbi algorithm [13]:

$$\boldsymbol{S}_{nz}^{(t)} = \text{Viterbi}(\boldsymbol{O}_n, \boldsymbol{\psi}_z)\,. \tag{3}$$

- Maximize the expected marginal complete log-likelihood (EMCLL) with respect to $\boldsymbol{\psi}$ by

$$\begin{aligned}\boldsymbol{\psi}^{(t)} &= \arg\max_{\boldsymbol{\psi}} \sum_{n=1}^{N} E_{z_n,\boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \\ &= \arg\max_{\boldsymbol{\psi}} \sum_{n=1}^{N}\sum_{z=0}^{Z} p(z_n = z|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)})\left(\log p_z + \log p(\boldsymbol{S}_n|z, \boldsymbol{\psi}) + \log p(\boldsymbol{O}_n|\boldsymbol{S}_n, \boldsymbol{\psi})\right)\,.\end{aligned} \tag{4}$$

Since population screening datasets can contain millions of records. direct inference may be prohibitively expensive and more scalable approaches are necessary. We scale our EM algorithm by parallelizing the inference across observations using $\tilde{N}$ clusters, $\{C_n\}_{n=1}^{\tilde{N}}$, in three parts. We first compute the conditional posterior distribution of z_n in each cluster using (2). The time complexity for each cluster is $O(|C_n|ZM^2T)$ [16]. We then compute the optimal state sequences in each cluster C_n using (3) with the same time complexity $O(|C_n|ZM^2T)$ [1]. Finally, we compute the gradients in each cluster then reduce all local gradients to global gradients for the optimization in the M-step. In detail the EMCLL is rewritten as

$$\begin{aligned}&\textstyle\sum_{n=1}^{N} E_{z_n,\boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \\ = \;&\textstyle\sum_{\tilde{n}=1}^{\tilde{N}} \sum_{n\in C_{\tilde{n}}} E_{z_n,\boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}).\end{aligned}$$

Taking gradient on both sizes, we have

$$\begin{aligned}&\tfrac{\partial}{\partial\boldsymbol{\psi}} \textstyle\sum_{n=1}^{N} E_{z_n,\boldsymbol{s}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}) \\ = \;&\textstyle\sum_{\tilde{n}=1}^{\tilde{N}} \tfrac{\partial}{\partial\boldsymbol{\psi}} \textstyle\sum_{n\in C_{\tilde{n}}} E_{z_n,\boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|\boldsymbol{O}_n, z_n, \boldsymbol{S}_n)|\boldsymbol{O}_n, \boldsymbol{\psi}^{(t-1)}).\end{aligned}$$

Automatic differentiation (AD) [4] is utilized to compute the gradients in each cluster. Summing over all clusters, we get the gradient of the EMCLL. Using this gradient we adapt the Limited-Memory BFGS [21] algorithm to estimate $\boldsymbol{\psi}$. The complexity of each cluster in optimization is intractable, but it increase inference speed around $\tilde{N}$ times.

5 Experimental Results

The HIHMM is demonstrated on a true population-level cervical cancel screening test dataset from the Cancer Registry of Norway. Data used in the analyses will be available on request from the Cancer Registry of Norway, given legal basis according to the GDPR. This dataset contains 1.7 million patients' screening testing records. Each patient has a censored observation at the last time stamp t_c, denoted by O_c, which indicates whether the woman is dead or alive at time t_c. Each patient has treatment indices to show when and how many treatments occurred, and results of screening tests for each of cytology, histology and hpv. Cytology and histology have four levels of outcomes while hpv has two levels.

We set $Z = 2$ and the model processes are displayed in Figure 1. For model z, the initial state is modeled as $S_{z1}|a_1 \sim \text{Cat}(\boldsymbol{\pi}_z(\mathcal{A}, a_0))$ and $\boldsymbol{\pi}_{zi} \sim \text{Dir}(\boldsymbol{\alpha}_{zi})$, where a_1 denotes the age at the first screening test and $\mathcal{A}$ is a disjoint partition of observable ages, $\boldsymbol{\pi}_z(\mathcal{A}, a) = \boldsymbol{\pi}_{zi}$ if and only if $a \in \mathcal{A}_i$, and $\boldsymbol{\alpha}_{z\ell} \in \mathbb{R}^{+M_z}$.

The observations $\boldsymbol{O}$ have two levels: the number of screening tests $\boldsymbol{E}$ and the results of screening tests $\boldsymbol{G}$. Omitting the subscripts n and t, given state s, observations are modeled as

$$\begin{aligned}E_k &\sim \text{Poisson}(\eta_{sk})\,, \\ \boldsymbol{G}_k|E_k &\sim \text{Multinomial}(E_k, \tilde{\boldsymbol{\pi}}_{sk})\,, \\ \tilde{\boldsymbol{\pi}}_{sk} &\sim \text{Dir}(\tilde{\boldsymbol{\alpha}}_{sk})\,,\end{aligned}$$

where $\tilde{\boldsymbol{\alpha}}_{sk} \in \mathbb{R}^{+L_k}$ are hyper-parameters for observation model. The censored observations (dead or alive) are modeled by

$$p(O_c|S_T) = \begin{cases} P(t_T, t_c)_{S_T,\text{death}} & \text{if } O_c = \text{death}\,, \\ 1 - P(t_T, t_c)_{S_T,\text{death}} & \text{if } O_c \neq \text{death}\,. \end{cases}$$

We choose the age partition as $\mathcal{A}$ as $[0, 23)$, $[23, 30)$, $[30, 60)$ and $[60, \infty)$. More details are available in the supplementary material, including treatment modeling and inference in Appendix, age partition selection in Appendix, model training and optimization in Appendix.

In the proposed learning approach, we set the number of EM iterations at $N_{\text{EM}} = 100$, and in the Limited-memory BFGS (L-BFGS) approach we set the number of optimization iterations as $N_{\text{L-BFGS}} = 8$. Automatic differentiation is implemented using the autograd package [25] in Python.

5.1 Model Comparison

We randomly select 80000 patients' records for training and select another 20000 records for testing. The goal is to predict the status at the last visit. Specifically, if a patient has at least one result whose level is greater than 1, then the status is defined as high risk denoted as 1. Otherwise, the status is defined as low risk denoted as 0. Thus, the problem is defined as a binary classification problem.

The prediction procedure is defined as follows. After model training, let model parameter estimates be $\hat{\boldsymbol{\psi}}$. Given new patient historical records $\boldsymbol{O}^*$, compute the predictive distribution of model index $p(z^*|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}})$. Next, given model index z, compute the predictive distribution of the state at the second to last visit $p(S^*_{T-1}|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}})$ derived from FFBS. Then the predictive distribution of the state of the last visit is $p(S^*_T|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) \sum_z p(z^* = z|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) p(S^*_T|z, \boldsymbol{O}^*, \hat{\boldsymbol{\psi}})$ and the predictive distribution of screening test results is $p(\boldsymbol{G}^*_T|\boldsymbol{O}^*, \boldsymbol{E}^*_T, \hat{\boldsymbol{\psi}}) = \sum_s p(S^*_T = s|\boldsymbol{O}^*, \hat{\boldsymbol{\psi}}) p(\boldsymbol{G}^*_T|S^*_T = s, \boldsymbol{E}^*_T, \hat{\boldsymbol{\psi}})$. Finally, the predictive distribution of the last status is $G^* \sim \text{Ber}\,(p^*)$, where $p^* = p\left(\sum_{i=0}^{1} \sum_{j=2}^{3} \boldsymbol{G}^*_T[i, j] \geq= 1|\boldsymbol{O}^*, \boldsymbol{E}^*_T, \hat{\boldsymbol{\psi}}\right)$. and it is estimated by $\hat{G}^* = \begin{cases} 1 & p^* \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$.

For the comparator recurrent neural network models, each patient's record is modeled as one time-series and the features at each visit includes patient's age, patient's screening result and patient's treatment indicator. The screening result of patient n at the tth visit is $\vec{\boldsymbol{G}}_{n,t}$. The treatment indicator is equal to 1 if and only if the patient has accepted treatment. LSTM [8], stacked LSTM [12] and GRU [9] are implemented for model comparison. For stacked LSTM, two LSTMs are stacked. We summarize prediction results in Table 1. It shows our model outperforms state of the art methods overall on the set of criteria: Area Under the Curve (AUC), F1 value (F1), Average Precision (AP) and Recall (R). Furthermore, our HIHMM with prior setting $p = 0.001$ obtains the best overall prediction performance.

Table 1: Model prediction for the status of the last visit in terms of Accuracy (ACC), Area Under The Curve(AUC), F1, Average Precision (AP), Precision (P), Recall (R).

Method	Prior	ACC	AUC	F1	AP	P	R
LSTM		0.9920	0.8506	0.3496	0.1811	**0.7679**	0.2263
stacked LSTM		0.9914	0.8575	0.3108	0.1388	0.6393	0.2053
GRU		**0.9922**	0.8490	0.4322	0.2273	0.7108	0.3105
HIHMM	0.0001	0.9913	0.9136	0.5085	0.2649	0.5488	0.4737
HIHMM	0.001	0.9914	**0.9190**	**0.5210**	**0.2774**	0.5589	**0.4895**
HIHMM	0.01	0.9916	0.9154	0.5185	0.2757	0.5652	0.4789
HIHMM	0.1	0.9917	0.9129	0.5174	0.2758	0.5779	0.4684
HIHMM	0.2	0.9918	0.9087	0.4762	0.2426	0.6000	0.3947
HIHMM	0.3	0.9915	0.9081	0.3929	0.1837	0.6111	0.2895

5.2 Model Validation

Because we have expert opinion available, an informative prior is indeed preferable here, so we set the model index prior $p = 0.2$. We note that there is always a trade off between precision and recall, and $p = 0.2$ provides a reasonable balance based on model comparison results. We present two types of results on population-level data. First we present the MLEs for all model parameters along with bootstrapped standard deviations. Second we perform model validation using Kaplan-Meier estimators as suggested in [31].

We randomly divide all data into clusters such that each cluster has 100 individual observation sequences. Using a bootstrap technique, we randomly select 2400 clusters with replacement for model inference. We independently repeat the same inference on different selections 5 times. The mean and standard deviation of all parameter estimates are discussed in Appendix.

For model validation we randomly select 2400 clusters of data in which each cluster has 100 individual sequences of observations. We implement both the HIHMM and the CTIHMM for the same dataset. We follow the method proposed in [31] that utilizes Kaplan-Meier estimators to validate continuous-time HMMs. Kaplan-Meier estimators are defined according to the definition of a failure, or time-to-event. In multi-state models different failures can be defined depending on which features of the model and data are of interest. Here we define failure as the first observation of a high-risk or cancer test result directly following an initial normal or low-grade test result. Accurately predicting this time-to-event is of practical importance because clinical intervention is only possible in the high-grade state. Treating patients at this stage is critical to preventing precancerous lesions from progressing to cervical cancer.

The empirical Kaplan-Meier estimator is defined as $\hat{S}(t) = \prod_{i:t_i \leq t} (1 - d_i/n_i)$, where t_i is a time when at least one failure is observed, d_i is the number of failures that occurred at time t_i, and n_i is the number of individuals known to have survived up to time t_i. We randomly choose 24000 records to generate an empirical Kaplan-Meier estimator according to our definition of a failure. We generate Kaplan-Meier estimates by simulating 100 sequences from both the CTIHMM and HIHMM and repeat this 100 times. Figure 2 shows the empirical Kaplan-Meier curve in black, simulated Kaplan-Meier curves from the CTIHMM in blue, and simulated Kaplan-Meier curves from the HIHMM. As for the simulated Kaplan-Meier curves, solid lines denote the median curve and dashed lines denote the 95% credible intervals based on the 100 replications. The results show that the empirical Kaplan-Meier curve is always near the median and within the 95% credible intervals generated by the HIHMM. This is not the case with the CTIHMM. In this sense the HIHMM outperforms the CTIHMM in an important clinical metric.

On the other hand, the HIHMM has a relatively high Kaplan-Meier estimate at time 0 because the informative prior $p = 0.2$ is relatively small. This has the effect of driving simulated patients to more likely be in the low-risk model $\mathcal{M}_l$ at the initial time. Moreover, these patients are more likely to stay at the normal state for longer. However, the trend of the median curve from the HIHMM more closely tracks that of the empirical Kaplan-Meier curve, compared with the trend of the median curve from the CTIHMM. This suggests that the HIHMM models disease progression better than the CTIHMM. The Kaplan-Meier curves simulated from the CTIHMM are always underestimated.

6 Conclusion and Discussion

One of the possible applications of the HIHMM in the context of population-based screening programs is risk stratification of the population. The latent random variable z_n is an indicator of belonging to a frail class in the population. Given the learned model parameters $\boldsymbol{\psi}$ it is possible to compute the posterior probability of belonging to the frailty class for individual women. In other words, given an observed sequence of test results $\boldsymbol{O}_n$ and model parameters $\boldsymbol{\psi}$, the posterior predictive distribution $p(z_n|\psi, \boldsymbol{O}_n)$ is of interest. This parameter gives a measure of the likelihood of an individual to be at risk of developing cervical cancer conditioned on their observed test results. Such information could be used to more efficiently screen a population by avoiding the over screening of women at low-risk and the under screening of women at high-risk.

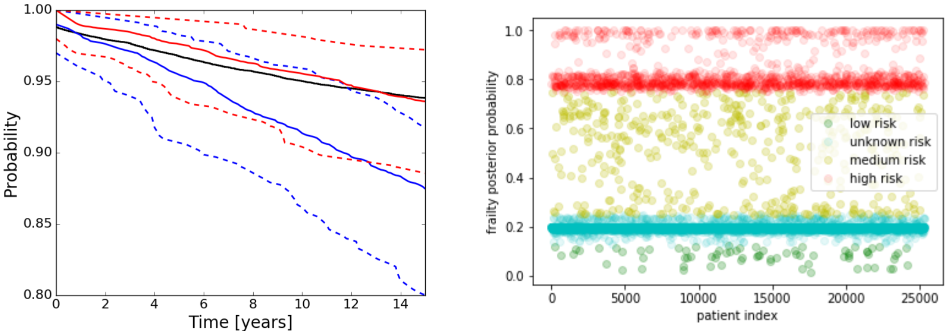

Figure 2: Left panel: Empirical Kaplan-Meier curve (black) and simulated Kaplan-Meier curves, which are summarized using the 95% credible interval (dashed lines) and the median (solid lines), from the CTIHMM (blue) and HIHMM (red). Right panel: Posterior probabilities of belonging to the frailty class for each individual from a test set. Risk stratification is possible by thresholding the probabilities. Threshold probabilities in this example are $(0, 0.125, 0.25, 0.75, 1)$. Color indicates falling between two probability thresholds.

Examples of these posterior probabilities are shown in Fig 2. For illustration purposes, we have chosen risk thresholds of $\{0.125, 0.25, 0.75\}$ with the following interpretation.

$$\begin{aligned}
0 \leq p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) < 0.125 &\implies \text{low-risk} \\
0.125 \leq p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) < 0.25 &\implies \text{unknown risk} \\
0.25 \leq p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) < 0.75 &\implies \text{medium-risk} \\
0.75 < p(z_n|\boldsymbol{\psi}, \boldsymbol{O}_n) \leq 1 &\implies \text{high-risk}
\end{aligned}$$

Two main clusters are apparent in the data corresponding to unknown risk and high risk. The unknown risk cluster is those patients close to the prior probability of 20%. These patients lack sufficient observations to make an informed decision about their risk profile. This suggests these patients should be followed up with the standard screening protocol. The high risk cluster is those patients who are more likely to be in a high-grade state. This suggests these patients may require immediate follow up. The two smaller clusters of low risk and medium risk are comprised of patients that may require decreased or increased screening frequencies, respectively, relative to the standard screening protocol.

In summary, this paper has made the following contributions:

- We make CTIHMM inference possible for population-level datasets by using piece-wise constant intensity functions and deriving a scalable inference algorithm.
- We put a hierarchical structure over the CTIHMM to explain population heterogeneity in terms of frailty, resulting in our HIHMM.
- We utilize prior distributions in the model to achieve more accurate estimates when dealing with imbalanced data.
- We perform full model inference and prediction on subset of a cancer screening dataset and show that our model outperforms comparators on the prediction task.
- We perform full model inference on a cancer screening dataset and show that modeling population heterogeneity improves performance in terms of Kaplan-Meier estimators.
- We illustrate how the model may be used to better inform public health professionals by providing a risk stratification mechanism.

References

[1] Arturs Backurs and Christos Tzamos. Improving Viterbi is Hard: Better Runtimes Imply Faster Clique Algorithms. *arXiv e-prints*, art. arXiv:1607.04229, Jul 2016.

[2] Yujia Bao, Zhaobin Kuang, Peggy Peissig, David Page, and Rebecca Willett. Hawkes process modeling of adverse drug reactions with longitudinal observational data. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 177–190, Boston, Massachusetts, 18–19 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v68/bao17a.html`.

[3] N Bartolomeo, P. Trerotoli, and G. Serio. Progression of liver cirrhosis to hcc: an application of hidden markov model. *BMC Med Research Method*, 11, 2011.

[4] Atilim Gunes Baydin, Barak Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18:1–43, 04 2018.

[5] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002.

[6] Alexandre Bureau, Stephen Shiboski, and James P. Hughes. Applications of continuous time hidden markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, 22(3):441–462, 2003. doi: 10.1002/sim.1270. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1270`.

[7] K Canfell, R Barnabas, Patnick J, and Beral V. The predicted effect of changes in cervical screening practice in the uk: results from a modelling study. *British journal of cancer*, 91(3): 530–536, 2004.

[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, art. arXiv:1406.1078, Jun 2014.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv e-prints*, art. arXiv:1412.3555, Dec 2014.

[10] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. 1965.

[11] P. Doetsch, M. Kozielski, and H. Ney. Fast and robust training of recurrent neural networks for offline handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 279–284, Sep. 2014. doi: 10.1109/ICFHR.2014.54.

[12] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. *arXiv e-prints*, art. arXiv:1505.08075, May 2015.

[13] G. D. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, March 1973. ISSN 0018-9219. doi: 10.1109/PROC.1973.9030.

[14] Jurgen V. Gael, Yee W. Teh, and Zoubin Ghahramani. The infinite factorial hidden markov model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1697–1704. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3518-the-infinite-factorial-hidden-markov-model.pdf`.

[15] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[16] Wael Khreich, Eric Granger, Ali Miri, and Robert Sabourin. On the memory complexity of the forward–backward algorithm. *Pattern Recognition Letters*, 31(2):91 – 99, 2010. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2009.09.023. URL `http://www.sciencedirect.com/science/article/pii/S0167865509002578`.

[17] Genshiro Kitagawa. Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, 1987. ISSN 01621459. URL `http://www.jstor.org/stable/2289375`.

[18] Vikram Krishnamurthy, Elisabeth Leoff, and Jörn Sass. Filterbased stochastic volatility in continuous-time hidden markov models. *Econometrics and Statistics*, 6:1 – 21, 2018. ISSN 2452-3062. doi: https://doi.org/10.1016/j.ecosta.2016.10.007. URL `http://www.sciencedirect.com/science/article/pii/S2452306216300144`. STATISTICS OF EXTREMES AND APPLICATIONS.

[19] Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. Baseline regularization for computational drug repositioning with longitudinal observational data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2521–2528. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL `http://dl.acm.org/citation.cfm?id=3060832.3060974`.

[20] Zhaobin Kuang, Peggy Peissig, Vitor Santos Costa, Richard Maclin, and David Page. Pharmacovigilance via baseline regularization with large-scale longitudinal observational data. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, 2017. doi: 10.1145/3097983.3097998.

[21] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(1-3):503–528, August 1989. ISSN 0025-5610. doi: 10.1007/BF01589116. URL `https://doi.org/10.1007/BF01589116`.

[22] Yu-Ying Liu, Hiroshi Ishikawa, Mei Chen, Gadi Wollstein, Schuman Joel S., and James M. Rehg. Longitudinal modeling of glaucoma progression us-ing 2-dimensional continuous-time hidden markov model. *Med Image Comput Assist Interv.*, 2013.

[23] 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 C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3600–3608. Curran Associates, Inc., 2015.

[24] Shaochuan Lu. A continuous-time hmm approach to modeling the magnitude-frequency distribution of earthquakes. *Journal of Applied Statistics*, 44(1):71–88, 2017. doi: 10.1080/02664763.2016.1161736. URL `https://doi.org/10.1080/02664763.2016.1161736`.

[25] Dougal Maclaurin. *Modeling, Inference and Optimization with Compposable Differentiable Procedures*. PhD dissertation, Harvard University, 2016.

[26] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. *arXiv e-prints*, art. arXiv:1708.02182, Aug 2017.

[27] Evan R. Myers, Douglas C. McCrory, Kavita Nanda, Lori Bastian, and David B. Matchar. Mathematical Model for the Natural History of Human Papillomavirus Infection and Cervical Carcinogenesis. *American Journal of Epidemiology*, 151(12):1158–1171, 06 2000. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a010166. URL `https://dx.doi.org/10.1093/oxfordjournals.aje.a010166`.

[28] Shawn E. Simpson, David Madigan, Ivan Zorych, Martijn J. Schuemie, Patrick B. Ryan, and Marc A. Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902, 2013. doi: 10.1111/biom.12078. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12078`.

[29] Frank A. Sonnenberg and Beck J. Robert. Markov models in medical decision making: A practical guide. *Med Decis Making*, 13:322–338, 1993.

[30] Cem Subakan, Johannes Traa, and Paris Smaragdis. Spectral learning of mixture of hidden markov models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2249–2257. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5518-spectral-learning-of-mixture-of-hidden-markov-models.pdf`.

[31] Andrew C. Titman and Linda D. Sharples. A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*, 27(12):2177–2195, May 2008. ISSN 02776715, 10970258. doi: 10.1002/sim.3033. URL `http://doi.wiley.com/10.1002/sim.3033`.

[32] 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*, KDD '14, pages 85–94, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623754. URL `http://doi.acm.org/10.1145/2623330.2623754`.

[33] Wei Wei, Bing Wang, and Don Towsley. Continuous-time hidden markov models for network performance evaluation. *Performance Evaluation*, 49(1):129 – 146, 2002. ISSN 0166-5316. doi: https://doi.org/10.1016/S0166-5316(02)00122-0. URL `http://www.sciencedirect.com/science/article/pii/S0166531602001220`. Performance 2002.

[34] Amy MF Yen, Tony HH Chen, Stephen W Duffy, and Chih-Dao Chen. Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. *Statistical Methods in Medical Research*, 19(5):529–546, 2010. doi: 10.1177/0962280209359862. URL `https://doi.org/10.1177/0962280209359862`. PMID: 20488838.

[35] A.I. Zeifman and Dean L. Isaacson. On strong ergodicity for nonhomogeneous continuous-time markov chains. *Stochastic Processes and their Applications*, 50(2):263 – 273, 1994. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(94)90123-6. URL `http://www.sciencedirect.com/science/article/pii/0304414994901236`.

[36] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video Summarization with Long Short-term Memory. *arXiv e-prints*, art. arXiv:1605.08110, May 2016.