Hierarchical Continuous-Time Hidden Markov Model for Cancer Screening Data

Rui Meng† and Herbert Lee†

†Department of Statistics, University of California, Santa Cruz

Introduction

Continuous-time hidden Markov models are an attractive approach for disease modeling because they are explainable and capable of handling both irregularly sampled, skewed and sparse data arising from real-world medical practice. 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 such as cancer. Moreover, the population at risk is not homogeneous either, since disease exposure and susceptibility can vary considerably. We model the heterogeneity of disease progression and regression using piece-wise constant intensity functions model the heterogeneity of risk in the population using a latent mixture structure. Different submodels under the mixture structure share the same latent states due to both clinical interpretation and model parsimony. We also consider flexible observational models dealing with model over-dispersion in real data. An efficient, scalable EM algorithm for inference is proposed with the theoretical guaranteed convergence property. We demonstrate our method's superior performance compared to other state-of-the-art methods using synthetic data and a real-world cervical cancer screening dataset from the Cancer Registry of Norway.

Objective

- Provide a better model predictive performance on cancer screening data.
- Provide a risk stratification approach.

Contributions

- Develop a latent mixture model to explain the frailty of the few women developing high-grade lesions and invasive cancer when exposed to the human papilloma virus.
- Promote parsimony in the mixture structure with great explanatory predictive power and solve the imbalance learning issue.
- Relieve the overdispersion issue in the observation model.

Model Notation

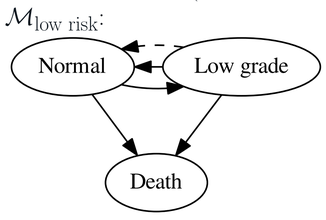
- Observations: $\boldsymbol{O} = \{O_n\}$.
- Number of screening tests: $\boldsymbol{E} = \{E_{ntk}\}$.
- Screening test results: $\boldsymbol{G} = \{G_{ntkl}\}$
- Covariates: $\boldsymbol{\theta} = \{\boldsymbol{\theta}_n\}$.
- Model parameters: $\boldsymbol{\psi} = \{\boldsymbol{\psi}_z\}$.
- Continuous-time hidden Markov models: $\{\mathcal{M}_z\}$
- Model indexes: $\boldsymbol{z} = \{z_n\}$.

Model

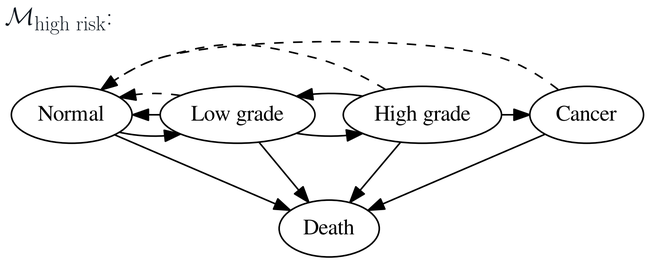
Hierarchical Model:

$$O_n|\boldsymbol{\psi}, z_n \sim \mathcal{M}_{z_n}(\boldsymbol{\psi}_{z_n}, \boldsymbol{\theta}_n)\,,$$
$$z_n|\boldsymbol{p} \sim \text{Cat}(\boldsymbol{p})\,.$$

Individual Models (Continuous-time hidden Markov model) [3]:
$\mathcal{M}_{\text{low risk}}$:



$\mathcal{M}_{\text{high risk}}$:



Observation Model (omitting the subscripts n and t):

$$E_k|s, \boldsymbol{\eta} \sim \text{Poisson}(\eta_{sk})\,,$$
$$\boldsymbol{G}_k|E_k, s, \tilde{\boldsymbol{\pi}} \sim \text{Multinomial}(E_k, \tilde{\boldsymbol{\pi}}_{sk})\,,$$
$$\tilde{\boldsymbol{\pi}}_{sk}|\tilde{\boldsymbol{\alpha}} \sim \text{Dir}(\tilde{\boldsymbol{\alpha}}_{sk})\,.$$

Inference

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

  $$p(z_n|O_n, \boldsymbol{\psi}^{(t-1)}) \propto \pi(z_n)p(O_n|z_n, \boldsymbol{\psi}^{(t-1)}) \sim \text{Cat}((\tilde{p}_{n1}, \ldots, \tilde{p}_{nZ})')\,,$$

  where $\tilde{p}_{nk} = \frac{p_k q_{nk}}{\sum_{z=1}^{Z} p_z q_{nz}}$ and $q_{ns} = p(O_n|z_n = s, \boldsymbol{\psi}_s^{(t-1)})$.

  - $p(O_n|z, \boldsymbol{\psi}_z)$ is accessible through the forward-filter backward-sample algorithm (FFBS), which is a sequential Monte Carlo approach first proposed in [2].

- Update the optimal state sequence $\boldsymbol{S}_n$ given corresponding observations O_n and model indicator z using the Viterbi algorithm [1] and we denote it as:

  $$\boldsymbol{S}_{nz}^{(t)} = \text{Viterbi}(O_n, \boldsymbol{\psi}_z^{t-1})\,.$$

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

  $$\boldsymbol{\psi}^{(t)} = \underset{\boldsymbol{\psi}}{\text{argmax}} \sum_{n=1}^{N} E_{z_n, \boldsymbol{S}_n}(\ell(\boldsymbol{\psi}|O_n, z_n, \boldsymbol{S}_n)|O_n, \boldsymbol{\psi}^{(t-1)})\,.$$

  - Instead of using the conditional posterior distribution for the expectation, we replace it by an approximate conditional posterior distribution

  $$q(z_n, \boldsymbol{S}_n|O_n, \psi^{(t-1)}) = p(z_n|O_n, \psi^{(t-1)})\mathbf{1}_{\boldsymbol{S}_{nz_n}^{(t)}}(\boldsymbol{S}_n)\,.$$

- Decompose parameters $\boldsymbol{\psi}$ into two parts, transition parameters $\boldsymbol{\psi}^{\text{tran}}$ and emission parameters $\boldsymbol{\psi}^{\text{emis}}$. Denote $\boldsymbol{\psi}_z^{\text{tran}}$ as transition parameters in the z^{th} model. We separately estimate $\boldsymbol{\psi}_z^{\text{tran}}$ and $\boldsymbol{\psi}^{\text{emis}}$ by

  $$\hat{\boldsymbol{\psi}}_z^{\text{tran}(t)} = \arg\max \sum_{n=1}^{N} p(z_n = z|O_n, \boldsymbol{\psi}^{(t-1)}) \log(\boldsymbol{S}_{nz}^{(t)}|z, \boldsymbol{\psi}_z^{\text{tran}})\,,$$
  $$\hat{\boldsymbol{\psi}}_z^{\text{emis}(t)} = \arg\max \sum_{n=1}^{N} p(z_n = z|O_n, \boldsymbol{\psi}^{(t-1)}) \log(O_n|\boldsymbol{S}_{nz}^{(t)}, \boldsymbol{\psi}_z^{\text{emis}})\,.$$

Simulation Study

Simulation Setting

We consider that two states, S_1 and S_2, two transition structures, A and B and a binary observation model.

- Structure A: CTIHMM with transition rates, $r_1 = r_2 = 0.1$ in the interval $(0, 5]$ while $r_1 = r_2 = 1$ in the interval $(5, 10]$.
- Structure B: CTIHMM with transition rates, $r_1 = r_2 = 1$ in the interval $(0, 5]$ while $r_1 = r_2 = 0.1$ in the interval $(5, 10]$.
- The observation model is developed by simple categorical distributions such that $p(O|S_1) \sim \text{Ber}(p_1 = 0.95)$ and $p(O|S_2) \sim \text{Ber}(p_2 = 0.05)$.
- We sample imbalanced time series via sampling 200 time series from structure A and 300 time series from structure B.
- All time series are assumed to start at state 1 and have 50 observations randomly located on the time interval $(0, 10)$.

Inference Results

We show both bootstrap mean and bootstrap standard error of model parameters.

Structure	$\hat{r}_1(t < 5)$	$\hat{r}_2(t < 5)$	$\hat{r}_1(t > 5)$	$\hat{r}_2(t > 5)$
A	0.07(0.01)	0.06(0.03)	1.08(0.12)	1.09(0.13)
B	0.85(0.11)	0.98(0.12)	0.12(0.02)	0.07(0.02)

- $\hat{p}_1 = 0.953(0.003)$ and $\hat{p}_2 = 0.041(0.004)$

Model Comparison

- We take 1000 balanced time series with 50 observations on each for testing, in which 500 of them are generated from structure A and the other 500 are generated from structure B for testing.
- We run different models to predict the last observations.

Method	ACC	AUC	F1	AP	P	R
LSTM (small)	0.802	0.854	0.805	0.740	0.794	0.816
LSTM (large)	0.783	0.856	0.786	0.720	0.776	0.796
stacked LSTM (small)	0.820	0.886	0.819	0.765	0.826	0.812
stacked LSTM (large)	0.766	0.858	0.767	0.704	0.765	0.768
GRU (small)	0.814	0.865	0.812	0.759	0.822	0.802
GRU (large)	0.802	0.867	0.803	0.742	0.801	0.804
HCTIHMM	**0.859**	**0.910**	**0.858**	**0.809**	**0.862**	**0.853**

- small: layer size 16; large: layer size 64.
- ACC: Accuracy; AUC: Area Under The Curve; F1: F1 score; AP: Average Precision; P: Precision; R: Recall.

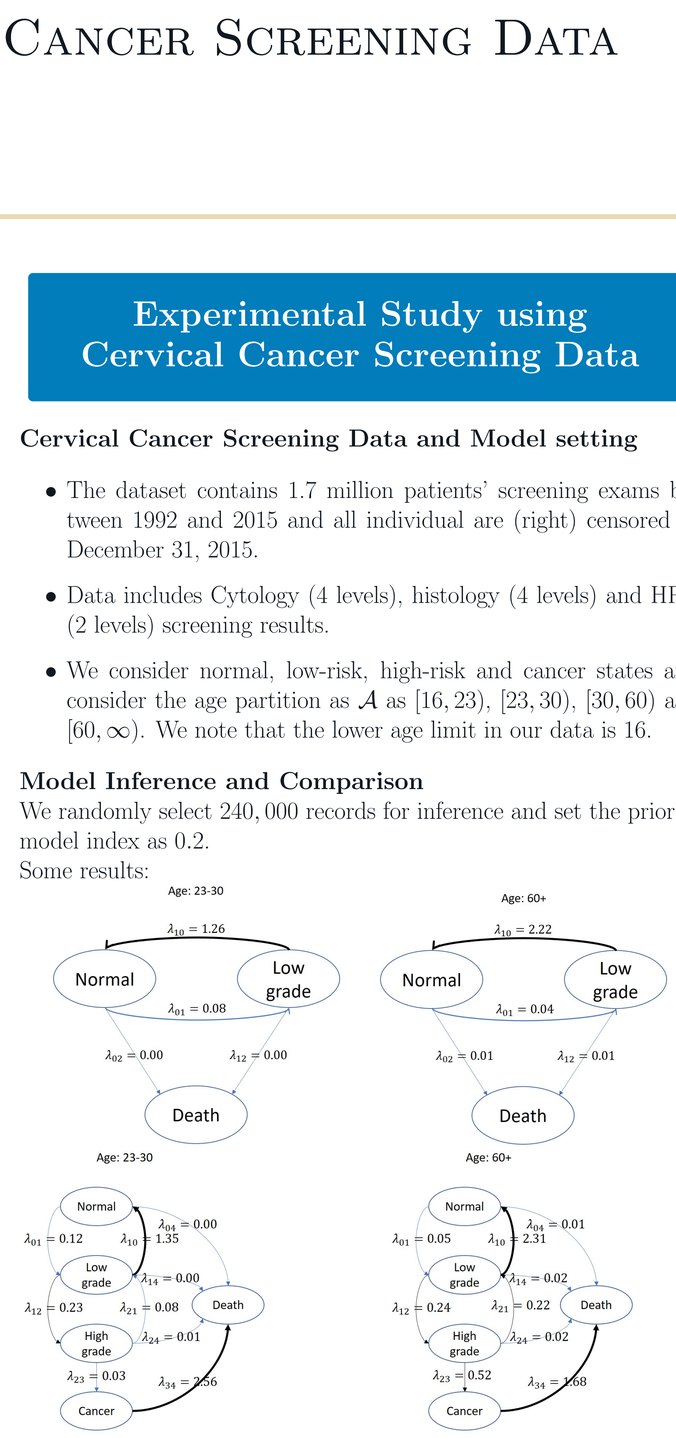
Experimental Study using Cervical Cancer Screening Data

Cervical Cancer Screening Data and Model setting

- The dataset contains 1.7 million patients' screening exams between 1992 and 2015 and all individual are (right) censored at December 31, 2015.
- Data includes Cytology (4 levels), histology (4 levels) and HPV (2 levels) screening results.
- We consider normal, low-risk, high-risk and cancer states and consider the age partition as $\mathcal{A}$ as $[16, 23)$, $[23, 30)$, $[30, 60)$ and $[60, \infty)$. We note that the lower age limit in our data is 16.

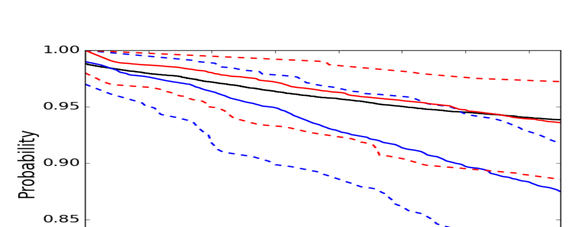
Model Inference and Comparison

We randomly select 240,000 records for inference and set the prior of model index as 0.2.
Some results:



Kaplan Curves from the hierarchical model and non-hierarchical model:

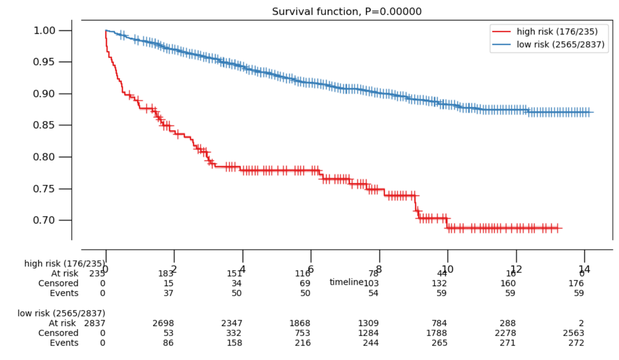
- Define failure as the first observation of a high-risk or cancer testresult directly following an initial normal or low-grade test result.
- Empirical Kaplan-Meier curve (black) and simulated Kaplan-Meiercurves, which are summarized using the 95% credible interval (dashedlines) and the median (solid lines), from the CTIHMM (blue) andHIHMM (red).



Risk Stratification

We random select 20,000 patients for risk stratification. We set an age threshold $t_0 = 35$ and take records before the threshold for risk stratification and take records after the threshold for model validation. Specifically, we plot empirical Kaplan Meier curve in different groups clustered by their risk level.

- To guarantee enough data for risk stratification, we require that woman should have 10 years under observation before the threshold t_0. In other word, woman's age at the first visit should less than $t_0 - 10$.
- To guarantee enough data for Kaplan Meier plotting, we require that woman have at least one visit after the threshold except for the censor visit.



References

[1] George David Forney. "The viterbi algorithm". In: *Proceedings of the IEEE* 61.3 (Mar. 1973), pp. 268–278. ISSN: 0018-9219. DOI: 10.1109/PROC.1973.9030.

[2] Genshiro Kitagawa. "Non-Gaussian State-Space Modeling of Nonstationary Time Series". In: *Journal of the American Statistical Association* 82.400 (1987), pp. 1032–1041. ISSN: 01621459. URL: http://www.jstor.org/stable/2289375.

[3] Braden C. Soper et al. "A hidden Markov model for population-level cervical cancer screening data". In: *Statistics in Medicine* n/a.n/a (). DOI: 10.1002/sim.8681. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8681. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8681.