iSurvive: An Interpretable, Event-time Prediction Model for mHealth

Walter H. Dempsey [*1] **Alexander Moreno** [*2] **Christy K. Scott** [3] **Michael L. Dennis** [3] **David H. Gustafson** [4]
Susan A. Murphy [1] **James M. Rehg** [2]

Abstract

An important mobile health (mHealth) task is the use of multimodal data, such as sensor streams and self-report, to construct interpretable time-to-event predictions of, for example, lapse to alcohol or illicit drug use. Interpretability of the prediction model is important for acceptance and adoption by domain scientists, enabling model outputs and parameters to inform theory and guide intervention design. Temporal latent state models are therefore attractive, and so we adopt the continuous time hidden Markov model (CT-HMM) due to its ability to describe irregular arrival times of event data. Standard CT-HMMs, however, are not specialized for predicting the *time* to a future event, the key variable for mHealth interventions. Also, standard emission models lack a sufficiently rich structure to describe multimodal data and incorporate domain knowledge. We present iSurvive, an extension of classical survival analysis to a CT-HMM. We present a parameter learning method for GLM emissions and survival model fitting, and present promising results on both synthetic data and an mHealth drug use dataset.

1. Introduction

In the emerging field of mobile health (mHealth) an important problem is the use of collected multimodal data – e.g., sensor streams along with self-report – to make time-varying predictions of events like lapse (Chih et al., 2014). Using latent state models for prediction is an attractive choice for three reasons: (1) States can be made interpretable by representing behavioral constructs such as stress and craving; (2) Emission models can handle noisy measurements; and (3) Parameters can capture domain knowledge. Moreover, an interpretable model can represent a theoretical relationship, such as the hypothesized link between increased stress and risk of smoking lapse, in a form which supports learning from data, simulation and visualization, and hypothesis testing. Such models can be a tool for data-driven design and testing of theoretical models by domain scientists (Nahum-Shani et al., 2015). Further, in the case of small sample sizes, the incorporation of domain knowledge may be critical for good performance. In such cases, superior performance relative to an alternative "black box" model can provide additional evidence for the correctness of a behavioral theory.

Discrete time hidden Markov models (DT-HMMs) are a standard tool for regularly-sampled sensor data, but many important datatypes, such as EMAs or detected periods of high stress, take the form of event data with irregular arrival times. Fortunately, recent work (Wang et al., 2014; Rao & Teh, 2013) makes it feasible to use continuous time HMMs (CT-HMMs) to model irregularly-sampled data. This paper builds on our prior work on efficient parameter learning algorithms for CT-HMMs (Liu et al., 2015).

In order to utilize CT-HMMs for mobile health interventions, however, two limitations with existing models must be addressed: (1) A mechanism is needed for predicting the *time* to future events; and (2) Emission models must go beyond the standard Gaussian and multinomial observations to embrace general multimodal data models. The first limitation can be addressed via classical methods for time-to-event prediction from survival analysis. Prior work in joint survival and longitudinal analysis has focused on shared-random effects models (Rizopoulos, 2012) or latent class models (Proust-Lima et al., 2014) with Gaussian emissions. Survival analysis has not been previously used in an interpretable hidden state setting with multimodal data. We develop a method for using the states of a CT-HMM as interpretable, time-varying covariates in a survival model. In prior work, Lian et al. (2014) developed an interpretable latent event process model; interpretability, however, was achieved post hoc rather than being a built-in feature of the model as is the case here. We describe additional differences between this prior work and our current approach in the supplementary material.

[*]Equal contribution [1]University of Michigan [2]Georgia Institute of Technology [3]Lighthouse Institute [4]University of Wisconsin-Madison. Correspondence to: Alexander Moreno <alexander.f.moreno@gatech.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

We address the second limitation, the treatment of multimodal observations, by using a factorized GLM emission model. Rather than simply "stacking" multimodal observations into a single vector, GLMs allow the specification of a different link function for each type of observation data, such as ordinal, count, and continuous data. Due to this choice, the M-step of EM does not have a closed form solution, and the temporal dependencies are different from standard GLM training. We derive an iterative M-step update approach to solve this problem, and provide convergence guarantees to match. We believe we are the first to use interpretable latent variables from a CT-HMM as covariates in a survival model with use of GLM emissions to handle multimodal data types in a CT-HMM. Our method allows us to make precise intuitive statements like "under high risk and low engagement, the probability of lapse is 74.4%." Our publicly-released software (`http://cbi.gatech.edu/Survival-HMM`) will enable the mHealth and data science communities to benefit from these new modeling capabilities. We show promising results both in simulation and on a real-world mHealth recovery support services dataset from individuals with substance use disorders.

2. Model Description

Our latent variable-based survival model, which we term iSurvive, has three components: (1) continuous-time hidden Markov process, (2) GLM emission models, and (3) event process. We assume there are N participants. For each participant, t is the time in hours since the start of the study. The study-window length ξ is pre-specified and fixed for all participants.

2.1. Continuous-Time HMM

A *continuous-time hidden Markov model* (Liu et al., 2015) is a continuous-time latent Markov process where state transitions and observations can occur at arbitrary times. There are two sources of hidden information: the states and their transition times. The estimation problem involves three sets of parameters: (1) an emission model $p(o|s)$, relating observations o to the latent state s; (2) a transition rate matrix Q that captures the exponentially distributed transition rates between states; and (3) an initial state distribution.

Let $\mathbf{S}$ be a vector Markov process of length p with vector representation $S(t) = [S_1(t), \ldots, S_p(t)]$ in which each $S_i(t)$ takes discrete values in $\{0, 1, \ldots, \ell_i\}$. Note that this can always be reduced to an equivalent representation with a single discrete state $\tilde{S}(t)$ with cardinality $\tilde{p} = \prod_{i=1}^{p}(\ell_i + 1)$. We will alternate between these representations as needed.

A $\tilde{p} \times \tilde{p}$-dimensional transition rate matrix Q governs transitions for the latent Markov process $\tilde{\mathbf{S}}$. The negative diagonal element $-Q_{ii}$ is the rate at which the process leaves state $i \in [\tilde{p}]$, assumed exponentially distributed with parameter $q_i = -Q_{ii}$. The equation $\sum_{j \neq i} Q_{ij} = -Q_{ii}$ must hold. If the latent process is currently in state i then at a transition time from state i, the probability of transitioning to state j is Q_{ij}/q_i. Suppose we observe the latent state transition times $(t'_1 = 0, t'_2, \ldots, t'_{V-1}, t'_V = \xi)$ and corresponding states $(\tilde{s}(t'_0), \ldots, \tilde{s}(t'_V))$. From these we deduce the sufficient statistics: (1) the number of transitions between states $\{n_{ij}\}_{i,j=1}^{\tilde{p}}$ and (2) the total length of time spent in each state $\{\tau_i\}_{i=1}^{\tilde{p}}$. The probability of this progression for the latent process is

$$\prod_{v=1}^{V-1} q_{\tilde{s}(t_v)} \exp(-q_{\tilde{s}(t_v)} \cdot (t_{v+1} - t_v)) \cdot Q_{\tilde{s}(t_v), \tilde{s}(t_{v+1})} / q_{\tilde{s}(t_v)}.$$

The CT-HMM also includes an observation process $\mathbf{O} = \{O(t)\}_{t \in [0,\xi]}$, which is only observed at observation times. Let $\mathbf{t} = (t_1, \ldots, t_V)$ denote the observation schedule; note this is a random subset of $[0, \xi]$. At each observation time t_i we observe the vector $O(t_i)$. We assume $O(t) \perp\!\!\!\perp (\mathbf{O}, \mathbf{S}) \mid S(t)$. Let $O[\mathbf{t}]$ denote the vector of observation values at the observation schedule $(O(t_1), \ldots, O(t_V))$.

Consider a participant who has already had k observation times $\mathbf{t}^{(k)}$. In this paper, we make the following conditional independence assumption:

$$t_{k+1} \perp\!\!\!\perp (\mathbf{O}, \mathbf{S}) \mid (\mathbf{t}^{(k)}, O[\mathbf{t}^{(k)}]) \tag{1}$$

In other words, the conditional distribution of the random interval $t_{k+1} - t_k$ only depends on the observed history. Under both above conditional independence assumptions, the joint probability of the latent process $\mathbf{S}$ and observation sequence $O[\mathbf{t}] = (O(t_1), \ldots, O(t_V))$ is equal to

$$\prod_{i=1}^{\tilde{p}} \left[\prod_{j \neq i} Q_{ij}^{n_{ij}} \right] e^{-q_i \tau_i} \prod_{v=1}^{V} p(O(t_v) \mid \tilde{s}(t_v)) p(t_v \mid H_v) \tag{2}$$

where H_v is the observed history $(\mathbf{t}^{(v-1)}, O[\mathbf{t}^{v-1}])$.

Each observation $O(t_k)$ is a multimodal vector. In an mHealth application, for example, we may observe several self-reported ordinal ratings (e.g. EMA) along with the number of times the mobile app has been used recently. Thus, we have both ordinal and count data. Our observation model summarizes each component of the observation vector in terms of the p latent sources.

We assume conditional independence of the M observation components given the latent process, leading to the emission factorization:

$$p(O(t) \mid S(t)) = \prod_{m=1}^{M} p(O_m(t) \mid S(t)).$$

This factorization simplifies the specification of the GLM for each observation component.

2.2. Emissions and Generalized Linear Models

A generalized linear model (GLM) is a flexible generalization of ordinary linear regression where the error distributions need not be Gaussian. A GLM has three components (McCullagh & Nelder, 1989; Agresti, 2015): (1) a response variable y following a dispersion exponential family with distribution $p(y \mid \eta, d(\tau))$ with natural parameter η and dispersion term $d(\tau)$. (2) a linear predictor $\phi' s$, where ϕ is a vector of weights and s our input vector; and (3) a link function g or activation function g^{-1}.

Suppose the p latent sources are binary (i.e., $l_i = 1$ for each i). For the j-th observation process, $\mathbf{O}_j = \{O_j(t)\}_{t\in[0,\xi]}$, the conditional mean given the latent process at time t equals $\mathbb{E}[O_j(t) \mid S(t)] = g^{-1}(\phi_0 + \phi' S(t))$. That is, the conditional expectation is equal to the activation function applied to a linear combination of the current values of the p latent sources plus a constant. Note this is an assumption of parsimony as the number of parameters in the fully nonparametric model would be 2^p.

2.3. Interpretability via Link Restriction

Much of behavioral science theory concerns latent states such as stress, craving, engagement and risk. We aim for our model outputs to be interpretable by the clinician, an important feature necessary for both acceptance and adoption by domain scientists. We achieve this through the following assumption: some of the variables collected are noisy measures of only *one* latent state and not the others.

For example, suppose $\mathbf{S}$ has latent binary sources $S_1(t)$ and $S_2(t)$ representing stress and craving, respectively. Further let $O_1(t)$ be a binary observation dependent *only* on stress, and $O_2(t)$ be a second observation dependent on both stress and craving (see Appendix B for the associated graphical model). Then taking the logit link function, the form of the conditional mean is

$$\text{logit}\left(\mathbb{E}[O_1(t) \mid \mathbf{S}]\right) = \phi_{\text{baseline}} + \phi_{\text{stress}} \cdot S_1(t). \quad (3)$$

Therefore $O_1(t)$ is conditionally independent of $S_2(t)$ given $S_1(t)$. That is, given information about the user's stress, the observation value does not depend on craving. We call such conditional independence assumptions *link restrictions*. iSurvive achieves interpretability via link restriction: for each latent source $\{S_i(t)\}_{t\in[0,\xi]}$ there exists at least one observation process $\{O_j(t)\}_{t\in[0,\xi]}$ such that $O_j(t)$ is a noisy measure of *only* $S_i(t)$. In an EMA context, this can be achieved with questions such as "Are you currently experiencing stress?" which target a single latent state construct. These direct observations enforce interpretability and allow us to incorporate additional more complex observation processes that can provide improved accuracy over self-report. Using a survival model then allows us to make intuitive statements such as "The probability of lapse within the next 30 minutes when the participant is stressed but not craving is 70%."

2.4. Event Process

We now build a model relating the interpretable, latent process $\mathbf{S}$ to the event process $\mathbf{Y} = \{Y(t)\}_{t\in[0,\xi]}$ of interest; this is a binary process where $Y(t) = 1$ implies an event occurs at time t. In our case study, for example, the event of interest is alcohol or drug use at a particular moment t. Survival analysis provides the appropriate tools for modeling the intensity function – the instantaneous rate of occurrence of the event – given the latent process. Let $N(t, t+s]$ be the number of events in the window $(t, t+s]$. Then the intensity function at time t is defined:

$$h(t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \mathrm{P}(N(t, t+\Delta t] > 0 \mid \mathbf{S}) \quad (4)$$

For this paper we consider the proportional hazards model which expresses the hazard as

$$h(t \mid \mathbf{S}) = h_0(t) \exp\left(\sum_{i=1}^{p} \phi_i S_i(t)\right) \quad (5)$$

where $h_0(t)$ is the baseline hazard function. In this paper we consider a constant baseline hazard; moreover, we presuppose the intensity only depends on the latent process at the current time t. The proposed model is an interpretable Cox process (Cox & Isham, 1980; Taylor et al., 2013) – a generalization of a Poisson process in which the intensity function is itself a stochastic process. Cox processes have found success in event prediction problems with respect to complex health data (Ranganath et al., 2015). Here the latent process is interpretable and therefore helps in answering both the prediction problem and the sequential decision making problem of interest. The Cox process assumes "lapses" are conditionally independent given the latent process. Such a model is appropriate when lapse is only a function of the latent behavioral constructs.

In mHealth applications, the event process can be measured either via sensors (Sarker et al., 2016; Hossain et al., 2014) (i.e., continuous monitoring) or self-report (i.e., intermittent monitoring). Alternatively, scientists may schedule observation times at which the participant is asked if they have used drugs within a prior window of time. In the case study for this work, participants were asked if they used either alcohol or drugs within the past 30 minutes. This can be modeled as

$$\mathrm{P}(N(t-\Delta, t] > 0) = 1 - \exp\left(-\int_{t-\Delta}^{t} h(s)ds\right) \quad (6)$$

Here we assume the window length Δ (i.e., 30 minutes) is short enough so that (1) the latent process is likely to be constant within the window; then given the latent process, the chance of no use (exponentiated term in eq. 6) can be well-approximated by $\exp(-\Delta \cdot (\phi_0 + \phi' S(t)))$, where $\phi_0 + \phi' S(t)$ is the hazard. We focus on the latter case with a similar discretized approximation for the remainder of the paper. For more on survival analysis see (Aalen et al., 2008; Cook & Lawless, 2007).

3. Parameter Estimation for iSurvive

Here we present an expectation-maximization algorithm for parameter estimation of iSurvive. Our development uses the context of our case study, and so we assume that the event process and observation processes are measured via self-report (EMA) following the same observation schedule (i.e., $Y(t) \subset O(t)$). It is straightforward to apply our approach to other use cases.

One property of self-report data is that participants may not respond at a scheduled observation time, and instead may decline to provide data. Let $\mathbf{M} = \{M(t)\}_{t\in[0,\xi]}$ be a binary process representing missing data. Suppose an observation is scheduled at time t_v; we write $M(t_v) = 1$ if the participant declines to provide information (i.e., the observation $O(t_v)$ is missing). Consider a participant who already has k observations at times $\mathbf{t}^{(k)}$. Then define $\mathbf{t}_0^{(k)}$ to be the set of observation times at which we *do* observe the observation (i.e., $\{t_i \in \mathbf{t}^{(k)} \text{ s.t. } M(t_i) = 0\}$).

We make the following conditional independence assumptions regarding the missing data indicator process:

$$M(t_{k+1}) \perp\!\!\!\perp \mathbf{S} \mid O[\mathbf{t}_0^{(k)}], M[\mathbf{t}^{(k)}], \mathbf{t}^{(k)}. \tag{7}$$

That is, the missing data indicator at t^{k+1} is independent of the latent process, given the observed history (i.e., observation and missing data processes at prior observation times of t^k). This assumption plus variational independence (i.e., no parameter sharing across components of the joint density) imply that likelihood estimation can ignore the missing data process. Note that observed missing data indicators can still be used in the survival and emission models. For example, missing data may be an indicator of future risk; assumption (7) only states that missing data may only depend on the latent process through the observed history.

A plausible alternative to assumption (7) is conditional independence given the observed history *and* the latent process at the observation time,

$$M(t_{k+1}) \perp\!\!\!\perp \mathbf{S} \mid S(t_{k+1}), O[\mathbf{t}_0^{(k)}], M[\mathbf{t}^{(k)}], \mathbf{t}^{(k)}. \tag{8}$$

Under this assumption, likelihood estimation cannot ignore the missing data process; however, the emission model becomes hierarchical and so can be readily handled within the iSurvive framework.

As the preceding discussion illustrates, the iSurvive framework is sufficiently flexible to describe a wide range of experiment designs and modeling assumptions. In the case study in Section 6, behavioral scientists identified through participant interviews that the primary cause for missed appointment was exogenous shocks to their schedule.

3.1. EM Method

For ease of presentation, we present the EM algorithm based on equation (2) under the conditional independence assumptions for observations and missing data given by equations (1) and (7) respectively. Excluding the initial state distribution, the *expected* complete log-likelihood is given by

$$\begin{aligned} L(Q,\Phi) = \sum_{i=1}^{P} \Bigg[\sum_{j\neq i} \log(q_{ij}) \mathbb{E}\left[n_{ij} \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \right] \Bigg] \\ - q_i \mathbb{E}\left[\tau_i \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \right] \\ + \sum_{v=1}^{V} \mathbb{E}\left[\log p(O(t_v) \mid S(t_v)) \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \right] \end{aligned}$$

where $(\hat{Q}^{(l)}, \hat{\Phi}^{(l)})$ are the parameter estimates from the l-th iteration. Under variational independence of the emission models and the latent Markov process, the maximization step can be done separately for the latent variable parameters and emission models.

We begin by describing the EM-steps for the transition matrix. The M-step for the transition matrix yields the following $(l+1)$ iteration estimate:

$$\hat{Q}_{ij}^{(l+1)} = \frac{\mathbb{E}\left[n_{ij} \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \right]}{\mathbb{E}\left[\tau_i \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \right]} \tag{9}$$

for $i \neq j$ and $\hat{Q}_{ii} = -\sum_{j\neq i} \hat{Q}_{ij}$.

The main challenge here is in the E-step, which (Liu et al., 2015) solved by breaking up the expectations into terms per observation, and terms conditioned on the possible state transitions between observations. Let $\zeta(v, s, s')$ denote the transition probability $p(S(t_v) = s, S(t_{v+1}) = s' \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)})$. Then

$$\begin{aligned} &\mathbb{E}\big[n_{ij} \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \big] \\ &= \sum_{v=1}^{V-1} \sum_{s,s'=1}^{S} \zeta(v,s,s') \times \mathbb{E}\left[n_{ij} \mid S(t_v) = s, S(t_{v+1}) = s', \hat{Q}^{(l)} \right] \\ &\mathbb{E}\big[\tau_i \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)} \big] \\ &= \sum_{v=1}^{V-1} \sum_{s,s'=1}^{S} \zeta(v,s,s') \times \mathbb{E}\left[\tau_i \mid S(t_v) = s, S(t_{v+1}) = s', \hat{Q}^{(l)} \right]. \end{aligned}$$

Liu et al. (2015) adapt methods from the continuous-time Markov chain (CTMC) literature to compute the end-state

conditioned expectations, and develop an equivalent inhomogeneous discrete-time hidden Markov model to calculate the pairwise beliefs $\zeta(v, s, s')$.

The density for the kth observation process $p(O_k(t) \mid S(t) = s; \phi)$ can be rewritten in the following exponential dispersion family form:

$$p(O_k(t) = o \mid S(t) = s) = h(o, \tau) \exp\left(\frac{\eta_s T(o) - A(\eta_s)}{d(\tau)}\right).$$

Define $\gamma_{v,s} = p(S(t_v) = s \mid O, \hat{Q}^{(l)}, \hat{\Phi}^{(l)})$. Then the corresponding component of the expected complete log-likelihood (ECLL) is given by

$$\sum_{v=1}^{V}\sum_{s=1}^{S}\gamma_{v,s}\left[\log h(O(t_v), \tau) + \frac{\eta_s T(O(t_v)) - A(\eta_s)}{d(\tau)}\right].$$

Maximizing the ECLL, which is done after maximization of the transition terms, does not have a closed form solution, and Fisher scoring or Newton's method must be used. However, Fisher scoring for GLMs assumes an unweighted log-likelihood, which is not the case for this objective because of the $\gamma_{v,s}$ terms. Appendix C extends the Fisher scoring method to the weighted setting above. A similar learning procedure was derived in (Escola et al., 2011), but focused on incorporating covariates into an HMM rather than using GLM emissions.

Algorithm 1 Forward-backward + Weighted Fisher scoring estimation procedure

Input: N participants, observation processes $\{\mathbf{O}^{(i)}\}_{i=1}^{N}$, observation times $\{\mathbf{t}^{(i)}\}_{i=1}^{N}$
Output: rate matrix $\hat{Q}$, emission parameters $\hat{\Phi}$
Smart initialization: $(\hat{Q}^{(0)}, \hat{\Phi}^{(0)})$
Set $l = 0$
repeat
  Use forward-backward algorithm to compute $\zeta(v, s, s')$ and $\gamma(v, s)$ for $v = 0, \ldots, V$ and $s, s' = 1, \ldots, S$.
  Compute $\mathbb{E}\left[\tau_i \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)}\right]$, and $\mathbb{E}\left[n_{ij} \mid O[\mathbf{t}], \hat{Q}^{(l)}, \hat{\Phi}^{(l)}\right]$.
  Compute $\hat{Q}^{(l+1)}$ via equation (9).
  Compute $\hat{\Phi}^{(l+1)}$ via weighted Fisher scoring with weights $\{\gamma(v, s)\}$.
until log-likelihood converges

3.2. Parameter Initialization and Convergence

Algorithm 1 presents the EM-algorithm for iSurvive using a combination of the forward-backward and weighted Fisher scoring procedures. The algorithm requires initial estimates $(\hat{Q}^{(0)}, \hat{\Phi}^{(0)})$, and these initial values will effect its convergence properties. We want to ensure that EM converges to intuitively reasonable estimates so that we can interpret the resulting model parameters in the context of relevant behavioral theory.

Prior theoretical work has provided some convergence guarantees for standard EM algorithms (Balakrishnan et al., 2014; Wang et al., 2015). To guarantee convergence, in each case strong concavity and first-order stability conditions are required along with reasonable parameter initialization. Here we provide a similar guarantee for EM parameter estimation for the case of CT-HMMs with factorized GLM emissions and conditional independence on the observation schedule. Our development is based on the assumption of strong concavity, which seems likely to hold based on prior work (Kakade et al., 2010) on almost strong concavity of exponential families. Our supporting simulation results in Section 5 provide additional evidence. A proof of strong concavity remains for future work.

For iSurvive, the Q-function of the EM-algorithm is decomposable into three components for the latent process, the observation model, and the event process. Decomposability allows us to discuss the assumption of strong concavity separately for each component; in particular, it allows us to isolate the issue of strong concavity for generalized linear models and investigate this independently of the other model components. Further, since the Q-function is decomposable, so is the M-step.

Under the assumption that the participant trajectories are independent and identically distributed, the law of large numbers ensures that as the sample size N increases, the sample-based Q-function approaches its expectation:

$$\begin{aligned}\tilde{Q}(\theta \mid \theta') &= \mathbb{E}\left[Q_N(\theta \mid \theta')\right] \\ &= \mathbb{E}\left[\mathbb{E}_{\mathbf{S} \mid O[\mathbf{t}], \mathbf{Y}, \theta'}[\log(p(\mathbf{S}, O[\mathbf{t}], \mathbf{Y}; \theta))]\right]\end{aligned}$$

The population M-function can then be defined as $\tilde{M}(\theta') = \arg\max_{\theta \in \Omega} \tilde{Q}(\theta \mid \theta')$. When applying EM, θ' corresponds to $\theta^{(t-1)}$, the parameters of the previous iteration. The population M-function is also decomposable; We now present Lemma 1 which provides convergence guarantees and motivates our approach to parameter initialization.

Lemma 1. *If each component of the population M-function satisfies strong concavity and first-order stability conditions for parameters $\theta \in \Omega$, then for sample size N sufficiently large the EM-algorithm satisfies*

$$\|\hat{\theta}^{(t)} - \theta^{\star}\| \leq \kappa^t \|\theta^{\star} - \hat{\theta}^{(0)}\| + \frac{1}{1-\kappa}\left(\psi(N, \delta)\right) \quad (10)$$

where θ^* is the MLE, $\hat{\theta}^{(t)}$ is the tth EM-iteration estimates, $\kappa = L/\lambda$, L is a measure of first-order stability, λ a measure of strong concavity, and $\psi(N, \delta)$ is the sum of high-probability lower bounds on the distance between $M_N(\theta)$ and $\tilde{M}(\theta)$ for each component.

See Appendix D for additional technical details. Lemma 1 motivates a practical problem: how do we choose the initial parameter estimates? iSurvive leverages GLMs for connecting the latent process to each component of the observable process. Kakade et al. (2010) shows (almost) strong concavity of exponential families. In particular, their Theorem 3.4 quantifies the fact that "exponential families behave in a strongly concave manner *only* in a (sufficiently small) neighborhood of $\theta^\star$. These findings combined with equation (10) highlight the importance of appropriately choosing $\hat{\theta}^{(0)}$. A good initialization will ensure convergence to high quality parameter estimates.

We propose a *smart initialization* strategy which leverages domain expertise. First, we treat rate matrix and emission initializations separately. For the rate matrix, we obtain an initial estimate of the hidden state sequences by assuming that the direct observations of the states resulting from link restriction (see Section 2.3) are noise-free. This allows us to estimate an initial Q corresponding to a CTMC over the observed states using the method of (Metzner et al., 2007). For the emission parameters, we hand-design an initial emission model by drawing from behavioral theory to connect the latent states to the observations. For example, we can ascertain the likelihood that a participant is experiencing stress if they answer positively to a question about being stressed. We can then choose GLM weight parameters accordingly. Note that this initialization method provides an indirect test of of our behavioral theories: if they are correct, they should lead to good predictive accuracy.

4. Prediction and Validation

iSurvive is designed to be an interpretable event prediction model. Here we show how to take parameter estimates $(\hat{Q}, \hat{\Phi})$ and answer the question "Given observed data $O[\mathbf{t}]$, what is the probability that a lapse will *not* occur at any of a future set of observation times $(t'_1, \ldots, t'_k)$?". Note we allow for multiple events to occur and thus are interested in whether any event occurs in the window or not. This is the primary question of interest when analyzing the recovery support services data in Section 6. In this case, prior data includes self-reported alcohol or drug use at observation times (i.e., $Y[\mathbf{t}] \subset O[\mathbf{t}]$). As we are interested in predicting self-reported lapse, we do not require the full generality of equation (6), as we only need to compute the chance of no use within a small prior window (i.e., a discretized approximation). At time t'_1 we can use the forward algorithm to compute the posterior distribution of the latent process at time t_1 conditional on the observed data. We can then compute the probability of *not* observing a lapse at time t'_1. To do this, we use a logistic emission model for $p(Y(t_i) = 0 \mid S(t_i) = s)$. This is equivalent to the discretized approximation described in Section 2.4 for the event process. We iterate on this procedure to compute a sequence of conditional probabilities of *not* lapsing (conditional on observed history and the fact that the participant has yet to lapse at any future scheduled observation time); multiplying them together yields the prediction of interest. Algorithm 2 presents pseudo-code.

We use the brier score (Blanche et al., 2015; van Houwelingen & Putter, 2011) and log-loss to define prediction accuracy. Use of the log-likelihood is inappropriate as it measures overall fit, which is not the main quantity of interest. We choose a time t and window-length Δ. We then produce the probability of no lapses at any of the scheduled observations times within the interval $(t, t + \Delta)$. Let $\pi_n(t, \Delta)$ denote this prediction for the nth participant. We compute an indicator of whether no events occur at these scheduled observation times within the window for the chosen participant $I_n(t, \Delta)$. The brier score is the average squared difference in these quantities $(\pi_n(t, \Delta) - I_n(t, \Delta))^2$.

The complete brier score and log-loss are then

$$BS(\Delta) \propto \sum_{n=1}^{N} \sum_{j=1}^{m_n} (\pi_n(t_j, \Delta) - I_n(t_j, \Delta))^2$$

$$LL(\Delta) \propto -\sum_{n=1}^{N} \sum_{j=1}^{m_n} \Big[I_n(t_j, \Delta) \log(1 - \pi_n(t_j, \Delta)) + (1 - I_n(t_j, \Delta)) \log(\pi_n(t_j, \Delta)) \Big]$$

respectively where $\{t_j\}_{j=1}^{m_n}$ are a set of chosen, participant specific times. The Brier score and log-loss are two ways to verify the accuracy of a probability forecast, the former ranges from 0 (completely accurate) to 1 (wholly inaccurate) while the latter from 0 to ∞.

For each experiment that follows, we perform a cross-validation based assessment of prediction accuracy. We randomly partition the N participants into groups of size K. Suppose $N/K = M$ and we label each partition uniquely $m = 1, \ldots, M$, then the Brier score and log-loss for the mth run is denoted $BS_m(\Delta)$ and $LL_m(\Delta)$ respectively. The cross-validated complete brier score and log-loss is then given by $\sum_m BS_m(\Delta)$ and $\sum_m LL_m(\Delta)$ respectively. We perform this calculation for various choices of Δ to observe performance over a range of window-lengths.

5. Synthetic Experiments

Our experiments are focused on the setting of self-reported alcohol or illicit drug use at scheduled observation times, which is the basis for our case study in Section 6.

We start with a synthetic experiment aimed at illustrating

Algorithm 2 Event prediction algorithm

Input: Rate matrix Q, emission parameters Φ, prior observations $O[\mathbf{t}]$; current time t_0 and future observation times $(t_1, \ldots, t_k)$.
Output: $\prod_{i=1}^{k} \psi(t_i)$ for $i = 1, \ldots, k$.
Compute $\alpha_t(s) \leftarrow p(S(t) = s \mid O[\mathbf{t}]; Q, \Phi)$
for $k' \in \{1, \ldots, k\}$ **do**
  Compute $P^{(i)} \leftarrow \exp(Q(t_i - t_{i-1}))$
  for $s \in \{1, \ldots, S\}$ **do**
    Set $\alpha_{t_i}(s) \leftarrow \left[\sum_{s'} \gamma_{t_{i-1}}(s') P^{(i)}_{s's}\right]$
    Set $\gamma_{t_i}(s) \leftarrow \alpha_{t_i}(s) p(Y(t_i) = 0 \mid S(t_i) = s)$
  end for
  Set $\psi(t_i) \leftarrow \sum_{s \in S} \gamma_{t_i}(s)$
  Normalize $\gamma_{t_i}(s) \leftarrow \gamma_{t_i}(s) / \sum_s \gamma_{t_i}(s)$
end for

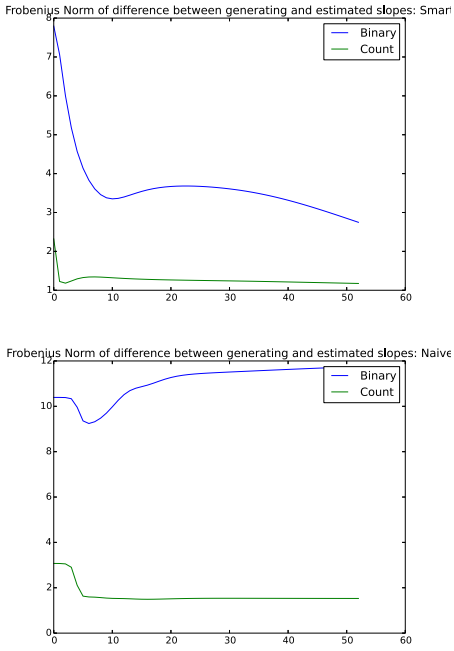

Figure 1. Convergence to the generating parameters (50 iterations) measured via Frobenius (l^2 in count case) norm a) with smart initialization b) without it. With smart initialization we recover good estimates of the generating parameters, while without it we do not even get closer to them in the binary case. The count difference norms are exponentiated to make visual comparison easier.

the importance of good initialization. Due to our desire for interpretability, we are concerned with how close the parameter estimates are to the true parameter values. We now briefly describe the synthetic experiment; see Appendix F for details. The latent process has three binary sources. We generate a random transition matrix, Q. We generate three ordinal ratings, each taking binary values. Each ratings question is associated with only one latent source (i.e., one binary question per latent source). We also generate count data to represent the number of times a user has used the app in the past 30 minutes. All emissions models are generalized linear models; for the ordinal ratings we assume a logit link while for the count data we assume a logarithmic link. For each, we are interested in observing the effect of smart initialization.

Figure 1 shows convergence (50 iterations) of the emission coefficients to the generating parameters a) with and b) without smart initialization. With smart initialization, convergence is very good, and without it, convergence is very poor; in fact, the estimated logistic parameters do not even converge toward the generating parameters.

6. Recovery Support Services – a Case Study

Here we analyze a set of recovery support studies on individuals with substance use disorders (SUDs). In particular, we analyze two pilot studies – a 5-week study of adults ($N = 23$) and a 6-week study of adolescents ($N = 29$) – where participants have recently been discharged from outpatient, intensive outpatient, or residential treatment. Each study was done using a modified version of the Addiction Comprehensive Health Enhancement Support System (ACHESS) (Gustafson et al., 2014). Participants of each pilot study have met criteria for substance use disorder in the year prior to the original treatment intake and have used alcohol or other drugs in the 90 days prior to the original treatment; prompts occur at six random times by the mobile phone per day. At each prompt, self-report data is collected concerning the prior 30 minutes exposure to internal and external protective and risk factors. Ratings were given to how each factor aids in their recovery or makes them want to use drugs/alcohol. Self-report also included questions regarding physical pain, illness or withdrawal from drugs/alcohol, level of craving for drugs/alcohol, exposure to drugs/alcohol, and resistance to drugs/alcohol. Participants were given smartphones enabled with 24/7 access to a range of drug abuse and HIV ecological momentary interventions (Dennis et al., 2015; Scott et al., 2017). Participants were asked to self-report use of drugs/alcohol in the prior 30 minutes. Sensor data was collected including information on when EMIs (Ecological momentary interventions; interventions delivered in real time) were accessed and the amount of time engaged with the EMI. Such a rich dataset of complex longitudinal data (e.g., simultaneously measuring EMI usage, drug use, self-reported ratings) is becoming increasingly common as the field of mobile health grows. We are interested in using the collected complex longitudinal data in event prediction – in particular the probability of *any* substance use within a future

window of time.

Prompts occur at random times so the observation schedule automatically satisfies the sequential conditional independence assumption. Prompts cannot occur from midnight to 6AM every morning; since participants are likely asleep at these times, we compute time since recruitment after removing these "sleeping windows". Consider a prompt at time t; we consider the reduced observation $O(t) = (O_1(t), O_2(t), O_3(t), Y(t))$. This is a 4-dimensional vector where $O_1(t)$ is a 3-level ordinal response to a question on how one's current feelings helps with/supports recovery, $O_2(t)$ is a 3-level ordinal variable related to EMI usage, $O_3(t)$ is a binary variable indicating whether the participant kept all default answers in the self-report, and $Y(t)$ is the binary event process indicating use of drugs/alcohol. We assume the latent process $\mathbf{S} = \{(S_1(t), S_2(t))\}_{t\in[0,\xi]}$ is comprised of two binary sources; $S_1(t)$ represents level of engagement; $S_2(t)$ represents level of risk. In this paper engagement is defined in terms of active engagement in self-report and is therefore connected to the indicator $O_3(t)$.

We now specify the models for each observation component conditional on $S(t)$. For the event variable $Y(t)$, we assume a logit model where the mean is a linear, additive model in terms of $S_1(t)$ and $S_2(t)$. For the engagement variable $O_3(t)$, we assume a logit model where the mean is a linear, additive model in terms of *only* $S_1(t)$. For the EMI usage variable $O_2(t)$, we assume a proportional odds model where the linear predictor is additive in terms of $S_1(t)$ and $S_2(t)$. The risk variable $O_1(t)$ is a mixture depending on engagement. Given the participant is not currently engaged, the responses are not related to latent risk; given the participant is currently engaged, we assume a proportional odds model where the linear predictor is additive only in terms of $S_2(t)$. Appendix E provides further details on the observations, latent states, and models.

We fit several alternative discriminative models aiming to predict future events given a fixed window length Δ. We fit both logistic regressions and kernel SVMs where the response is an indicator of use in a future window with a particular choice of features from the history. The first feature set is simply the current observation values; the second adds an additional covariate indicating whether an event occurred at the current observation time; the third adds an additional covariate indicating whether an event occurred in the prior twenty-four hours; the fourth adds an additional covariate indicating whether an event occurred in the prior week. Finally, we add as a covariate the number of scheduled observation times in the future window. For each discriminative model and iSurvive, we compute the cross-validated complete brier score and complete log-loss. Figure 2 shows that iSurvive outperforms the discriminative models in terms of the cross-validated Brier score for each $\Delta \leq 5$ days. Figure 8 in Appendix E shows iSurvive also outperforms the alternative discriminative models in terms of the cross-validated log-loss.

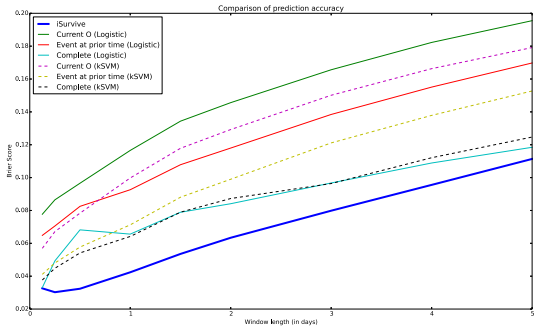

Figure 2. Cross-validated complete Brier score on recovery support services study for several discriminative models and iSurvive

Our analysis via iSurvive yields the interpretable finding that 30-minute probability of lapse is highest for individuals in the latent states of High risk and Low engagement (74.4%), decreases for High risk and High engagement (21.5%), decreases more for Low risk and Low Engagement (1.1%), and is negligible for Low risk and High engagement (0.1%). The finding is intuitive for our behavioral scientist collaborators and can be used to help decide what types of interventions to provide and when to provide them.

7. Conclusion and Future Work

In this paper we introduce iSurvive, an interpretable, event-time prediction model for mHealth. By using a continuous-time hidden Markov model and a factorized GLM emission model with link restriction, we can summarize our observations in terms of interpretable latent variables. We then use these in a survival model to predict event times. iSurvive is designed with an interest toward treatment policies; by having interpretable latent states, we hope to leverage iSurvive in optimizing the delivery of mobile health interventions as future work.

Acknowledgements

We thank the reviewers for their useful comments and the members of the Statistical Reinforcement Learning Lab at the University of Michigan for valuable discussions. We acknowledge support from the National Institutes of Health under BD2K grant U54EB020404 and grants R01EY13178-15, R01AA023187, P50 DA039838, and R01HL125440. This work was also supported by NIDA R01 DA035879 and additional grants from NIAAA, NIDA and NIMH.

References

Aalen, O., Borgan, O., and Gjessing, H. *Survival and event history analysis: a process point of view*. New York. Springer-Verlag, 2008.

Agresti, A. *Foundations of Linear and Generalized Linear Models*. Wiley, New York, first edition, 2015.

Balakrishnan, S., Wainwright, M.J., and Yu, B. Statistical guarantees for the em algorithm: From population to sample-based analysis. 2014.

Blanche, P., Proust-Lima, C., L. Loubère, C. Berr, Dartiguesand, J.F., and Jacqmin-Gadda, H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102–113, 2015.

Chih, M., Patton, T., McTavish, F., Isham, A., Judkins-Fisher, C. L., Atwood, A. K., and Gustafson, D. H. Predictive modeling of addiction lapses in a mobile health application. *Journal of Substance Abuse Treatment*, 46 (1):29–35, 2014.

Cook, R. and Lawless, J. *The statistical analysis of recurrent events*. Springer, 2007.

Cox, D.R. and Isham, V. *Point Processes*. Chapman & Hall, London, 1 edition, 1980.

Dennis, M.L., Scott, C. K., Funk, R. R., and Nicholson, L. A pilot study to examine the feasibility and potential effectiveness of using smartphones to provide recovery support for adolescents. *Substance abuse*, 36(4):486–492, 2015.

Escola, Sean, Fontanini, Alfredo, Katz, Don, and Paninski, Liam. Hidden markov models for the stimulus-response relationships of multistate neural systems. *Neural computation*, 23(5):1071–1132, 2011.

Gustafson, D.H., McTavish, F. M., Chih, M. Y., Atwood, A. K., Johnson, R. A., Boyle, M. G., M.S., Levy, Driscoll, H., Chisholm, S.M., Dillenburg, L., and Isham, A. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry*, 71(5):566–572, 2014.

Hossain, Syed Monowar, Ali, Amin Ahsan, Rahman, Md. Mahbubur, Ertin, Emre, Epstein, David, Kennedy, Ashley, Preston, Kenzie, Umbricht, Annie, Chen, Yixin, and Kumar, Santosh. Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, IPSN '14, pp. 71–82, 2014.

Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. *AISTATS*, pp. 381–388, 2010.

Lian, W., Rao, V.A., Eriksson, B., and Carin, L. Modeling correlated arrival events with latent semi-markov processes. *In Proceedings of the 29th International Conference on Machine Learning*, 2014.

Liu, Y., Song, L., Li, F., Li, S., and Rehg, J. Efficient continuous-time hidden markov model for disease modeling. *In Proceedings for Advances in Neural Information Processing Systems (NIPS)*, 2015.

McCullagh, P. and Nelder, J.A. *Generalized Linear Models*. Springer, 2nd edition, 1989.

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

Nahum-Shani, Inbal, Hekler, Eric B., and Spruijt-Metz, Donna. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, 34:1209–1219, 2015.

Proust-Lima, C., Sne, M., Taylor, J.M., and Jacqmin-Gadda, H. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23:74–90, 2014.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. The survival filter – joint survival analysis with a latent time series. *Uncertainty in Artificial Intelligence*, 2015.

Rao, Vinayak and Teh, Yee Whye. Fast mcmc sampling for markov jump processes and extensions. *Journal of Machine Learning Research*, 14(1):3295–3320, 2013.

Rizopoulos, D. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press, 2012.

Sarker, Hillol, Tyburski, Matthew, Rahman, Md Mahbubur, Hovsepian, Karen, Sharmin, Moushumi, Epstein, David H., Preston, Kenzie L., Furr-Holden, C. Debra, Milam, Adam, Nahum-Shani, Inbal, al'Absi, Mustafa, and Kumar, Santosh. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 4489–4501. ACM, 2016.

Scott, C. K., Dennis, M. L., Gustafson, D., and Johnson, K. A pilot study of the feasibility and potential effectiveness of using smartphones to provide recovery support. *Drug & Alcohol Dependence*, 171:e185, 2017.

Taylor, Benjamin M., Rowlingson, Barry, Moraga, Paula, and Diggle, Peter J. Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

van Houwelingen, Hans and Putter, Hein. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2011.

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

Wang, Zhaoran, Gu, Quanquan, Ning, Yang, and Liu, Han. High dimensional em algorithm: Statistical optimization and asymptotic normality. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pp. 2521–2529, 2015.