Supplementary Material for "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]

A. Prior Work on Interpretable Latent State Models

We highlight key differences between the present work and an interpretable, latent state model introduced by (Lian et al., 2014). In it, the model has one sequence of latent (K binary events) states (e.g., progression in a movie); each user experiences the same sequence of latent states but may react differently, resulting in a user-specific intensity function that produces a response to the latent process. In our model, on the other hand, the latent state process evolves independently from user to user. Thus the participants do not experience the same sequence of latent states. This is a key difference for our mobile health application.

In Lian et al. (2014), interpretability is achieved post-hoc and is not a built in feature of the model. Comparison of latent sources to movie scenes is performed and intuitive statements are made such as "One source relates to enhanced arousal intensity during a plane crash, key plot turning points, a climax, and a surprise denouement." We, on the other hand, will provide a systematic method of relating one latent state to particular emissions so that the states are easily interpreted, allowing us to make statements such as the 30-minute probability of lapse at high risk and low engagement is 74.4%.

B. Graphical Models for Section 2.3

The graphical model for the example discussed in Section 2.3 is given by $\{S_1 \rightarrow O_1,\, S_1 \rightarrow O_2,\, S_2 \rightarrow O_2\}$, where $x \rightarrow y$ denotes a directed arc from x to y.

[*]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.

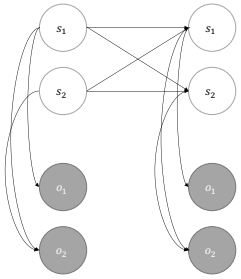

Figure 1. Graphical model for sec. 2.3. Recall that s_1 corresponds to stress and s_2 corresponds to craving. Then o_1 is conditionally independent of s_2, given s_1.

C. Details on Weighted Fisher Scoring

We have assumed complete factorization of the observation component; therefore, we only investigate the kth observation component. Let $O_k^{(n)}$ denote the kth observation component for the nth participant. Then the final term of the expected complete log-likelihood is given by

$$\sum_{n=1}^{N}\sum_{v=1}^{V_n}\mathbb{E}\left[\log p(O^{(n)}(t_v)=o_v \mid S(t_v)) \mid O_n[\mathbf{t}];\hat{Q},\hat{\Phi}\right].$$

Let $\gamma_v(s) = p\left(S(t_v)=s \mid O_n[\mathbf{t}];\hat{Q},\hat{\Phi}\right)$. Then the expected complete log-likelihood is equal to

$$\sum_{n=1}^{N}\sum_{v=1}^{V_n}\sum_{s}\gamma_v(s)\log p(O^{(n)}(t_v)=o_v \mid S(t_v)=s)$$
$$=\sum_{o}\sum_{s}b(s)\log p(O=o \mid S(t_v)=s).$$

where $\sum_o$ is a sum over the observed values of the observation process (e.g., if the process is binary then the sum is over $o=0$ and $o=1$) and $b(s)=\sum_v\sum_n\gamma_v(s)$ (i.e., the sum of all weights where the latent process is equal to the latent state $s \in S$.

Given weights $\{b(s)\}_{s\in S}$, we now show how to obtain the maximum-likelihood estimates via a *weighted* version of Fisher scoring. Let $\hat{\eta}_0$ denote the current estimate of the

linear predictor, and $\hat{\mu}_0$ the fitted value using the link function $\eta = g(\mu)$. Then define

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0$$

where the derivative of the link is evaluated at $\hat{\mu}_0$. Weights are defined by

$$W_0^{-1} = \bar{b}\left(\frac{d\eta}{d\mu}\right)_0^2 V_0$$

where V_0 is the variance function evaluated at $\hat{\mu}_0$, and $\bar{b} = (b(1), \ldots, b(S))$ is the vector of weights associated with the EM-algorithm. Then regress z_0 on the covariates with weight W_0 to give new estimates $\hat{\beta}$; from these form a new estimate $\hat{\eta}$. Repeat until changes are sufficiently small.

D. EM Convergence – Technical Details

Here we provide technical details related to Lemma 1 from the main body of the paper; we assume the event process is observed via self-report and therefore can be thought of as part of the observation process. When the event process is measured continuously, the below theory can easily be adjusted to include the necessary third term. We omit this for brevity as the conclusions are not changed.

We assume the observation schedule always starts with an observation at baseline (i.e, $t_1 = 0$). The next observation times t_i are (potentially stochastic) functions of the observed history H_i; therefore the observation schedule satisfies the restrictive conditional independence assumption. The study is of length ξ and therefore the observation schedule for each participant $\mathbf{t}$ is a random subset of the interval $[0, \xi]$.

We assume the probability of observing the sequence $\mathbf{o} = (o(t_1), \ldots, o(t_k))$ is given by:

$$\begin{aligned} p(\mathbf{X} = \mathbf{x}; \theta) &= \int_{\mathbf{s}} p(\mathbf{S} = \mathbf{s}, \mathbf{O} = \mathbf{o}; \theta) \\ &= \int_{\mathbf{s}} p(s(t_1)) \prod_{i=2}^{k} p(s(t_i) \mid s(t_{i-1}); \theta_z) \\ &\cdot \prod_{i=1}^{K} p(o(t_i) \mid s(t_i); \theta_o) p(t_i \mid H_i; \omega) \end{aligned}$$

where $\mathbf{S}$ is a latent Markov process, and $\theta = (\theta_s, \theta_x)$ parametrize the latent Markov process and the "measurement-error models" respectively. Note that the final term is independent of the latent process and therefore factors outside the summation; under the assumption of variational independence we see that the observation schedule component of the joint probability will not impact maximum likelihood estimation with respect to θ.

We observe N independent and identically distributed observation sequences, $\mathbf{o}_i = (o_i(t_{i,1}), \ldots, o_i(t_{i,k_i}))$ for $i = 1, \ldots, n$. Then the re-scaled log-likelihood of such a sample is given by

$$\begin{aligned} l_N(\theta) =& \frac{1}{N} \sum_{i=1}^{N} \log\left(\int_{\mathbf{s}} p(\mathbf{S} = \mathbf{s}, \mathbf{X}_i = \mathbf{x}_i; \theta)\right) \\ =& \frac{1}{N} \sum_{i=1}^{N} \log\left(\mathbb{E}_{\mathbf{S} \mid \mathbf{x}_i; \theta}\left[\frac{p(\mathbf{S} = \mathbf{s}, \mathbf{X}_i = \mathbf{x}_i; \theta')}{p(\mathbf{S} = \mathbf{s} \mid \mathbf{X}_i = \mathbf{x}_i; \theta')}\right]\right). \end{aligned}$$

By Jensen's inequality we have

$$\begin{aligned} l_N(\theta) \geq& \frac{1}{N} \mathbb{E}_{\mathbf{S} \mid \mathbf{x}_i; \theta'}\left[\sum_{i=1}^{N} \log\left(\frac{p(\mathbf{S} = \mathbf{s}, \mathbf{X}_i = \mathbf{x}_i; \theta')}{p(\mathbf{S} = \mathbf{s} \mid \mathbf{X}_i = \mathbf{x}_i; \theta')}\right)\right] \\ =& \frac{1}{N} \mathbb{E}_{\mathbf{S} \mid \mathbf{x}_i; \theta}\Bigg[\sum_{i=1}^{N} \log\left(p(\mathbf{S} = \mathbf{s}, \mathbf{X}_i = \mathbf{x}_i; \theta')\right) \\ & - \sum_{i=1}^{N} \log\left(p(\mathbf{S} = \mathbf{s} \mid \mathbf{X}_i = \mathbf{x}_i; \theta')\right)\Bigg] \\ =& Q_N(\theta \mid \theta') - H_N(\theta'). \end{aligned}$$

For a choice of θ', the E-step computes the function $\theta \to Q_N(\theta \mid \theta')$. The M-step then maximizes the Q-function for fixed θ':

$$M_N(\theta') = \arg\max_{\theta \in \Omega} Q_N(\theta \mid \theta')$$

For the HMM, we can decompose the Q-function as

$$Q_N(\theta \mid \theta') = Q_N^{(s)}(\theta_s \mid \theta') + Q_N^{(s)}(\theta_x \mid \theta').$$

That is, we can decompose the Q-function into a component only dependent on θ_s and one only dependent on θ_x. This implies the M-step is also decomposable and we can define

$$\begin{aligned} M_N^{(s)}(\theta') &= \arg\max_{\theta_s \in \Omega_s} Q_N^{(s)}(\theta_s \mid \theta') \quad \text{and} \\ M_N^{(x)}(\theta') &= \arg\max_{\theta_x \in \Omega_s} Q_N^{(x)}(\theta_x \mid \theta') \end{aligned}$$

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

$$\tilde{Q}(\theta \mid \theta') = \mathbb{E}[Q_N(\theta \mid \theta')] = \mathbb{E}\left[\mathbb{E}_{\mathbf{S} \mid \mathbf{X}, \theta'}\left[\log(p(\mathbf{S}, \mathbf{X}; \theta))\right]\right]$$

where the outside expectation is over the distribution of $\mathbf{o}$ (both the observation values and schedule). We call $\tilde{Q}$ the "population Q-function". The "population M-function" is defined as

$$\tilde{M}(\theta') = \arg\max_{\theta \in \Omega} \tilde{Q}(\theta \mid \theta').$$

D.1. Analysis of EM Algorithm

D.1.1. Population Level Algorithm

Let $\theta^\star$ denote the maximum likelihood estimate (MLE) of the population likelihood $\mathbb{E}[l_N(\theta)]$. Here we assume the MLE is *unique*. The MLE is a fixed point of the population M-function – i.e., it automatically satisfies

$$\theta^\star = \tilde{M}(\theta^\star).$$

We make the following conditions on the population Q-function.

Assumption 1 (Conditions on $\tilde{Q}$)**.** *We assume the following conditions on the population Q-function for $\theta \in \Omega$:*

1. *We assume $\tilde{Q}^{(s)}$ is λ_s-strongly concave:*

$$\tilde{Q}^{(s)}(\theta_s \mid \theta^\star) - \tilde{Q}^{(s)}(\theta'_s \mid \theta^\star) - \langle \nabla \tilde{Q}^{(s)}(\theta_s), \theta_s - \theta'_s \rangle \leq -\frac{\lambda_s}{2}\|\theta_s - \theta'_s\|_2^2$$

   for $\theta_s, \theta'_s \in \Omega_s$ and $\theta^\star \in \Omega$.

2. *We assume $\tilde{Q}^{(x)}$ is λ_x-strongly concave*

$$\tilde{Q}^{(x)}(\theta_x \mid \theta^\star) - \tilde{Q}^{(x)}(\theta'_x \mid \theta^\star) - \langle \nabla \tilde{Q}^{(x)}(\theta_x), \theta_x - \theta'_x \rangle \leq -\frac{\lambda_x}{2}\|\theta_x - \theta'_x\|_2^2$$

   for $\theta_x, \theta'_x \in \Omega_x$ and $\theta^\star \in \Omega$.

3. *For each $\theta_x \in \Omega_x$ and $\theta' \in \Omega$*

$$\begin{aligned}\|\nabla_x \tilde{Q}^{(x)}(\theta_x \mid (\theta_s^\star, \theta_x^\star)) - \nabla_x \tilde{Q}^{(x)}(\theta_x \mid (\theta'_s, \theta_x^\star))\|_2 &\leq L_{x,s}\|\theta_s^\star - \theta'_s\| \\ \|\nabla_x \tilde{Q}^{(x)}(\theta_x \mid (\theta_s^\star, \theta_x^\star)) - \tilde{Q}^{(x)}(\theta_x \mid (\theta_s^\star, \theta'_x))\|_2 &\leq L_{x,x}\|\theta_x^\star - \theta'_x\|\end{aligned}$$

4. *For each $\theta_s \in \Omega_s$ and $\theta' \in \Omega$*

$$\begin{aligned}\|\nabla_s \tilde{Q}^{(z)}(\theta_s \mid (\theta_s^\star, \theta_x^\star)) - \nabla_x \tilde{Q}^{(x)}(\theta_x \mid (\theta'_s, \theta_x^\star))\|_2 &\leq L_{x,s}\|\theta_s^\star - \theta'_s\| \\ \|\nabla_s \tilde{Q}^{(x)}(\theta_x \mid (\theta_s^\star, \theta_x^\star)) - \tilde{Q}^{(x)}(\theta_x \mid (\theta_s^\star, \theta'_x))\|_2 &\leq L_{x,x}\|\theta_x^\star - \theta'_x\|\end{aligned}$$

Lemma 1. *Under the above assumptions, for $\theta \in \Omega$ and pair (L, λ), then population M-function satisfies*

$$\|M(\theta) - \theta\| \leq \left(\frac{L}{\lambda}\right)\|\theta - \theta^\star\|.$$

Define $L = \max\{L_{s,x}, L_{s,s}\} + \max\{L_{x,s}, L_{x,x}\}$ and $\lambda = \min\{\lambda_s, \lambda_x\}$. Then for any point starting $\theta_0 \in \Omega$, the population EM-algorithm satisfies

$$\|\theta^{(t)} - \theta^\star\|_2 \leq \left(\frac{L}{\lambda}\right)^t \|\theta_0 - \theta^\star\|_2$$

for every $t = 1, 2, \ldots$.

Proof. By definition, the M-functions satisfy the following optimality condition

$$\langle \nabla Q^{(s)}(M^{(s)}(\theta^\star) \mid \theta^\star), M^{(s)}(\theta_s^\star) - M^{(s)}(\theta) \rangle \geq 0$$

and

$$\langle \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta), M^{(s)}(\theta_s^\star) - M^{(s)}(\theta) \rangle \leq 0.$$

Recall the optimal $\theta_s^\star$ is a fixed point of the $M^{(s)}$-operator (i.e., $M^{(s)}(\theta_s^\star) = \theta_s^\star$). Combining these inequalities yields

$$\begin{aligned}\langle \nabla Q^{(s)}(M^{(s)}(\theta^\star) \mid \theta^\star) - \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta), \\ \theta_s^\star - M^{(s)}(\theta) \rangle \geq 0.\end{aligned} \tag{1}$$

Recall λ_s-strong concavity, states that

$$\begin{aligned}Q^{(s)}(\theta_s \mid \theta^\star) - Q^{(s)}(\theta'_s \mid \theta^\star) - \langle \nabla Q^{(s)}(\theta_s \mid \theta^\star), \\ \theta_s - \theta'_s \rangle \geq \frac{\lambda_s}{2}\|\theta_s - \theta'_s\|_2^2.\end{aligned}$$

Switching places of θ_s and θ'_s and adding the resulting inequality to the above yields

$$\begin{aligned}\langle \nabla Q^{(s)}(\theta_s \mid \theta^\star) - \nabla Q^{(s)}(\theta'_s \mid \theta^\star), \\ \theta_s - \theta'_s \rangle \geq \lambda_s\|\theta_s - \theta'_s\|_2^2.\end{aligned} \tag{2}$$

Substitute $M^{(s)}(\theta) = \theta_s$ and $\theta_s^\star = M^{(s)}(\theta^\star) = \theta'_s$ into equation (2) yields

$$\begin{aligned}&\langle \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta^\star) - \nabla Q^{(s)}(M^{(s)}(\theta^\star) \mid \theta^\star), \\ &M^{(s)}(\theta) - M^{(s)}(\theta^\star) \rangle \geq \lambda_s\|M^{(s)}(\theta) - M^{(s)}(\theta^\star)\|_2^2 \\ &\qquad = \lambda_s\|M^{(s)}(\theta) - \theta_s^\star\|_2^2.\end{aligned}$$

Also, equation (1) implies

$$\begin{aligned}&\langle \nabla Q^{(s)}(M^{(s)}(\theta^\star) \mid \theta^\star) - \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta^\star), \\ &\theta_s^\star - M^{(s)}(\theta) \rangle \\ \geq\ &\langle \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta) - \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta^\star), \\ &\theta_s^\star - M^{(s)}(\theta) \rangle.\end{aligned}$$

Take the left-hand side of the above inequality. Then the second set of assumptions leads to

$$\begin{aligned}&\langle \nabla Q^{(s)}(M^{(s)}(\theta^\star) \mid \theta^\star) - \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta^\star), \\ &\theta_s^\star - M^{(s)}(\theta) \rangle \\ \leq &\|\nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta^\star) - \nabla Q^{(s)}(M^{(s)}(\theta) \mid \theta^\star)\|_2^2 \\ &\|\theta_s^\star - M^{(s)}(\theta)\|_2^2 \\ \leq &\left(L_{s,x} \cdot \|\theta_x^\star - \theta_x\|_2^2 + L_{s,s} \cdot \|\theta_s^\star - \theta_s\|_2^2\right) \\ &\|\theta_s^\star - M^{(s)}(\theta)\|_2 \\ \leq &\max\{L_{s,x}, L_{s,s}\}\left(\|\theta^\star - \theta\|_2^2\right)\|\theta_s^\star - M^{(s)}(\theta)\|_2\end{aligned}$$

where the second inequality is application of Cauchy-Schwarz, the third is due to item 3 in Assumptions 1. Combining the inequalities yields the equation

$$\lambda_s \|\theta_s^\star - M^{(s)}(\theta)\|_2 \leq \max\{L_{s,x}, L_{s,s}\}\|\theta^\star - \theta\|_2$$

The identical argument applied to the x-component yields

$$\lambda_x \|\theta_x^\star - M^{(x)}(\theta)\|_2 \leq \max\{L_{x,x}, L_{x,s}\}\|\theta^\star - \theta\|_2.$$

Combining these with $\lambda = \min(\lambda_s, \lambda_x)$ and $L = \max\{L_{s,x}, L_{s,s}\} + \max\{L_{x,x}, L_{x,s}\}$ yields

$$\|\theta^\star - M(\theta)\|_2 \leq \left(\frac{L}{\lambda}\right)\|\theta^\star - \theta\|_2$$

□

D.2. Finite Sample-size Statistical Guarantees for EM Algorithm

Definition 1 (Sample-to-population M-function gap)**.** *For a particular sample size $N \geq 0$ and constant $\delta \in (0,1)$, define $\epsilon^{(x)}(N,\delta)$ by:*

$$\sup_{\theta\in\Omega} p(\|M_N^{(x)}(\theta) - M^{(x)}(\theta)\|_2 \geq \epsilon^{(x)}(N,\delta)) \leq \delta$$

and $\epsilon^{(z)}(N,\delta)$ by:

$$\sup_{\theta\in\Omega} p(\|M_N^{(s)}(\theta) - M^{(s)}(\theta)\|_2 \geq \epsilon^{(s)}(N,\delta)) \leq \delta.$$

Lemma 2 below is the technical version of Lemma 1 in the main body of the paper; we end with a proof of this result.

Lemma 2. *The population M-function satisfies Assumptions 1 for all $\theta \in \Omega$. Then for sample size sufficiently large to ensure*

$$\epsilon_x(N,\delta) + \epsilon_s(N,\delta) \leq (1-\kappa)\, r - \kappa \max_{\theta_s\in\Omega_s} \|\theta_s - \theta_s^\star\|$$

Then for $\theta_0 \in \Omega$, with probability $1-2\delta$, the EM-algorithm satisfies

$$\|\hat\theta^{(t)} - \theta^\star\| \leq \kappa^t\|\theta_0 - \theta^\star\| + \frac{1}{1-\kappa}\left(\epsilon_x(N,\delta) + \epsilon_s(N,\delta)\right)$$

Proof. Define $\epsilon(N,\delta) = \epsilon_x(N,\delta) + \epsilon_x(N,\delta)$. With probability at least $1-2\delta$

$$\begin{aligned}\|\hat\theta^{(t+1)} - \theta^\star\| &= \|M_N(\hat\theta^{(t)}) - \theta^\star\| \\ &\leq \|M_N(\hat\theta^{(t)}) - \tilde M(\hat\theta^{(t)})\| + \|\tilde M(\hat\theta^{(t)}) - \theta^\star\| \\ &\leq \epsilon_x(N,\delta) + \epsilon_x(N,\delta) + \kappa\,\|\hat\theta^{(t)} - \theta^\star\|.\end{aligned}$$

Iterating on this we have

$$\begin{aligned}\|\hat\theta^{(t+1)} - \theta^\star\| &\leq \left[\sum_{s=0}^{t-1}\kappa^s\right]\epsilon(N,\delta) + \kappa^t\|\theta_0 - \theta^\star\| \\ &\leq \frac{1}{1-\kappa}\epsilon(N,\delta) + \kappa^t\|\theta_0 - \theta^\star\|.\end{aligned}$$

To complete the proof, note the sample size based inequality ensures that each iteration stays within the space Ω. □

E. Additional Details for the RSS Case Study

Here we present additional details regarding the subset of data analyzed from the recovery support studies on individuals with substance use disorders (SUDs). We start with a more detailed description of each component of the reduced observation $O(t) = (O_1(t), O_2(t), O_3(t), Y(t))$.

- $O_1(t)$: Ordinal response to the question "Rate the extent to which certain feelings help with/support your recovery." Original response was on a 0–7 scale. Based on studying the distribution of responses we collapsed into a three-level ordinal response:

  $$(0,1,2) \to 0,\ (3,4,5) \to 1,\ \text{ and }\ (6) \to 2.$$

  Here 2 is translates to current feelings *greatly* helping with/supporting your recovery. The fraction of observations per level is $(0.10, 0.26, 0.64)$. Figure 2 below presents histograms across participants of the fraction of EMAs when a participant responds with a particular rating.

- $O_2(t)$: a 3-level ordinal variable related to EMI usage by the following mapping:

  $$\text{None} \to 0,\ \ \text{1–3 times} \to 1,\ \ \text{and}\ \ \text{4+} \to 2$$

  The fraction of observations per level is $(0.76, 0.15, 0.09)$. Figure 3 below presents histograms across participants of the fraction of EMAs when a participant had a particular level of EMI usage.

- $O_3(t)$: an indicator of whether the participant kept all default answers to *all* questions within the self-report. The fraction of observations per level is $(0.61, 0.39)$. Figure 3 below presents a histogram (across participants) of the fraction of EMAs when the participant responded with all default answers.

- Y(t): the binary indicator of use of drugs/alcohol in past 30 minutes. The fraction of observations per level is $(0.94, 0.06)$ (i.e., 6% of all EMAs record drug/alcohol use within the). Figure 3 below presents a histogram (across participants) of the fraction of EMAs when the participant responded yes to having used drugs/alcohol within the past 30 minutes.

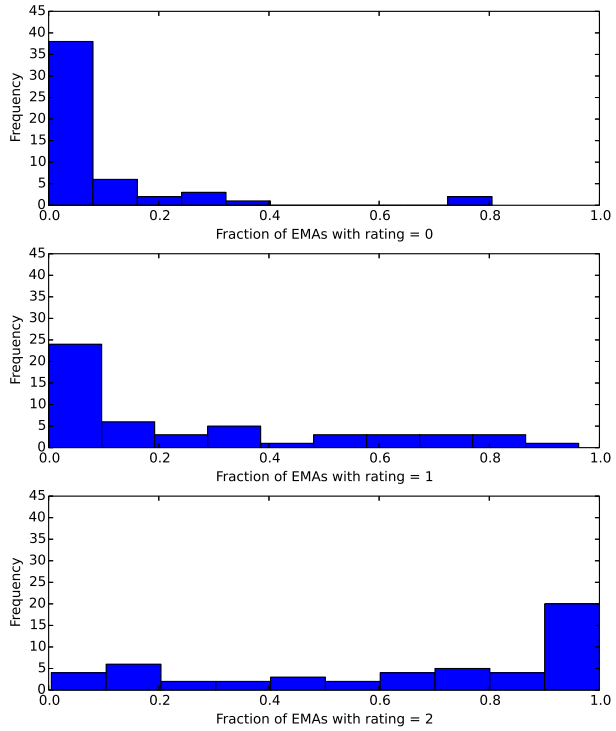

Figure 2. Histograms across participants related to $O_1(t)$

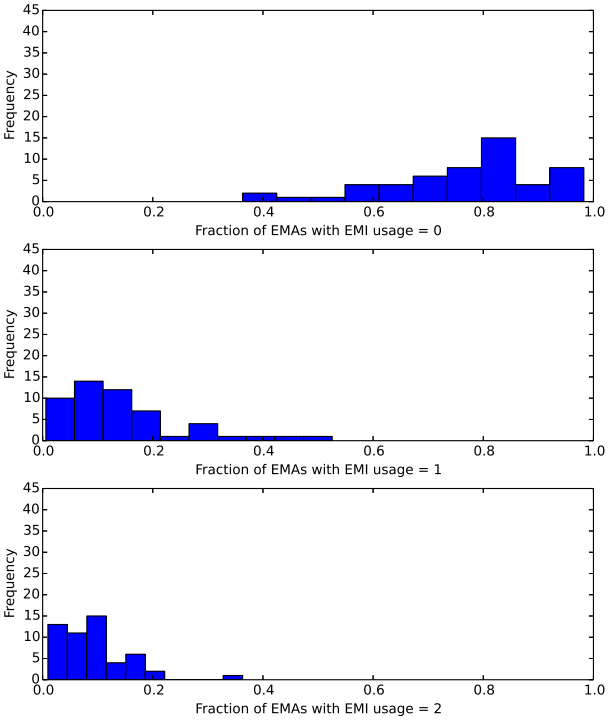

Figure 3. Histograms across participants related to $O_2(t)$

We assumed 2 latent binary sources–i.e., $S(t) = (S_1(t), S_2(t))$ with $S_i(t) \in \{0, 1\}$ for each $i \in \{1, 2\}$. We describe how we think about each and then show how this interpretation is achieved via link restriction.

- $S_1(t)$: We define this as *risk*. Here $S_1(t)$ is thought to be uniquely associated with O_1, the ordinal response on how current feelings help with support a participant's recovery.
- $S_2(t)$: We define this to be *engagement*. Here engagement is defined in a very specific manner: "thinking" through the EMA self-report and not simply filling a self-report with default answers. Here $S_2(t)$ is thought to be uniquely associated with $O_3(t)$.

We now specify the models for each observation component conditional on $S(t)$.

- Model for $O_2(t)$: Hierarchical model dependent on $S_2(t)$. If $S_2(t) = 1$ then the participant is not currently engaged, and therefore the response is independent of $S_1(t)$. In this case, the following proportional odds model defines the relationship:

$$\text{logit}(p(O_2(t) \leq j \mid S_1(t), S_2(t) = 1]) = \phi_j^{(1,1)}.$$

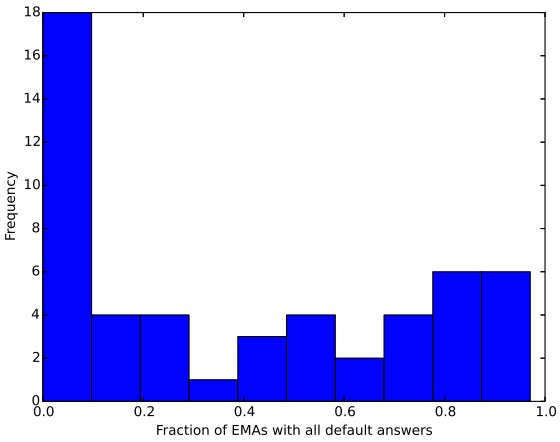

Figure 4. Histogram related to $O_3(t)$

If $S_2(t) = 1$ then the participant is currently engaged, and therefore the response depends on $S_1(t)$. In this case, the following proportional odds model defines the relationship:

$$\text{logit}(p(O_2(t) \leq j \mid S_1(t), S_2(t) = 0]) = \phi_j^{(1,0)} + S_1(t)\phi_1^{(1)}.$$

- Model for $O_2(t)$: Proportional odds model defined by

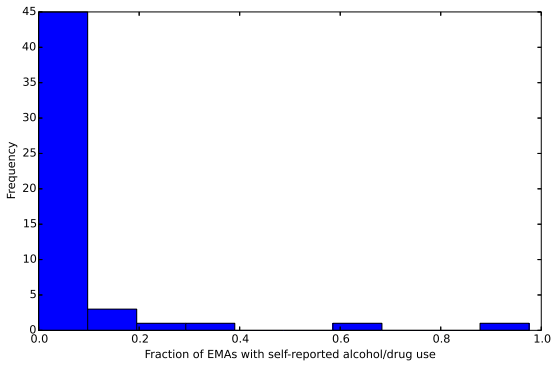

Figure 5. Histogram related to $Y(t)$

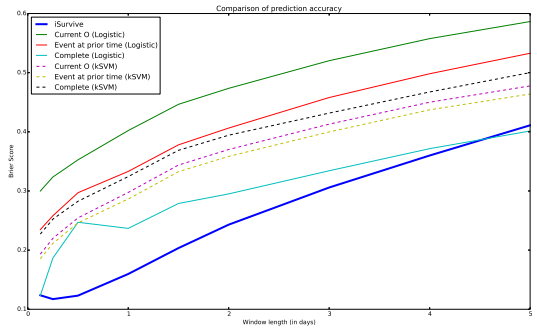

Figure 6. Cross-validated complete log-loss on recovery support services study for several discriminative models and iSurvive

the following relationship:

$$\text{logit}(p(O_2(t) \leq j \mid S(t)]) = \phi_j^{(2)} + S_1(t)\phi_1^{(2)} + S_2(t)\phi_2^{(2)}.$$

- Model for $O_3(t)$: GLM with logit link function (i.e., logistic regression) defined by the following relationship:

$$\text{logit}(\mathbb{E}[O_3(t) \mid S(t)]) = \phi_0^{(3)} + S_2(t)\phi_2^{(3)}.$$

- Model for $Y(t)$: GLM with logit link function (i.e., logistic regression) defined by the following relationship:

$$\text{logit}(\mathbb{E}[Y(t) \mid S(t)]) = \phi_0^{(4)} + S_1(t)\phi_1^{(4)} + S_2(t)\phi_2^{(4)}.$$

A log-linear model can also be fit with an "exposure" parameter of 30-minutes. This would allow for the user to make predictions for future self-reported drug/alcohol use over different windows. Since this was not necessary for the case study, we did not pursue this further.

F. Additional details on Synthetic Experiments

Table 1. Results from 10 runs of the synthetic experiments. We deleted a single extreme outlier from the $c = 2$ case.

c	Binary MAE	Count Weights Norm	Q rel. error
1	0.056±0.029	0.109±0.087	0.269±0.048
2	0.039±0.028	0.120±0.102	0.222±0.104
3	0.010±0.006	0.068±0.023	0.180±0.044

We assume the latent process has $p = 3$ sources: for ease of understanding we consider them to be stress, craving, and engagement. We generate a random Q matrix in the following way: we sample from a dirichlet with parameters $K = 8$ and $\alpha_1, \cdots, \alpha_K = 8$. We then zero out the diagonals, set them to be the negative of the sum of the remaining terms in the corresponding rows, and multiply the entire matrix by 5. For the observations, we have three ordinal ratings, each having a binary response. We also have count data to represent number of times EMIs (Ecological Momentary Interventions) are accessed via the application. The emission model for the ratings data is a GLM similar to equation (3) given by

$$\text{logit}\left(\mathbb{E}[O_k(t) \mid \mathbf{S}]\right) = \phi' S(t) \tag{3}$$

while for EMI the counts between two observations follow a homogeneous poisson process with mean rate parameter λ such that

$$\log(\lambda) = \phi' S(t) \tag{4}$$

We use logit emissions in order to see the effect of smart initialization, as the lack of strong concavity far from the true parameters can provide a challenge without it.

For the EMA, we use $\phi = (\phi_{baseline}, \phi_{stress}, \phi_{craving}, \phi_{engagement})$. Each EMA is only associated with one variable, and the true relationship between the baseline parameter and a latent variable weight is fixed. For the stress question this would be $\phi = c(1, -2, 0, 0)$, while for craving it would be $\phi = c(1, 0, -2, 0)$, where c is a scalar. The weights are the same magnitudes across questions. With this setup, c becomes a parameter that controls the emission noise. Larger values of c correspond to lower emission noise. Particularly, $c = 1, 2, 3$ correspond to probabilities 0.73, 0.88, and 0.95 for observing a 0. For the EMI we set the true weights to $\phi = c(0.4, 0.1, 0.2, 0.3)$. We set $c = 3$ for the true value, and $c = 0.25$ with the true relative proportions for the smart initialization. When not using

smart initialization, we set $c = 0$ and initialize Q randomly the same way we initialized the true Q.

Table 1 shows the results from running the algorithm with smart initialization 10 times for each of $c = 1, 2, 3$ with $N = 50$, $T = 252$ and 125 iterations. We deleted a single extreme outlier for $c = 2$, and found that the differences between the accuracy of $c = 1$ and $c = 2$ were not statistically significant at 95%, but the differences between $c = 3$ and either of $c = 1$ or $c = 2$ were for the binary case and count weights. The difference for Q in the latter case was just outside the 95% interval. For the terms shown, the binary MAE represents the mean absolute error between the true probability of observing a 0 and the learned probability across all states. The count weight norms represent the norm of the difference between generating and estimated parameters, and the Q relative error is the same as in (Liu et al., 2015).

References

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.