MODELING AMYOTROPHIC LATERAL SCLEROSIS (ALS) PROGRESSION WITH THE CONTINUOUS-TIME HIDDEN MARKOV MODEL (CT-HMM)

A PREPRINT

Michael Colomb
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
colomb2@illinois.edu

Brian Luc
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
luc2@illinois.edu

Ophir Sneh
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
osneh2@illinois.edu

Sam Cheung
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
skcheun2@illinois.edu

January 12, 2019

ABSTRACT

Disease progression modeling is a data-driven approach to gain insight into the mechanics of a disease and to provide reliable prognostics to people afflicted with the condition. However, developing accurate progression models for amyotrophic lateral sclerosis (ALS, Lou Gehrig's disease, motor neurone disease) has proven difficult due to its rarity and the heterogeneity of patient trajectories. We hypothesize that a continuous-time hidden Markov model (CT-HMM) can efficiently and accurately model the evolution of ALS by training on observations of the symptomatic manifestation of the disease over time in patients who participated in longitudinal studies. This proposed method is motivated in part by the inherent ability of the CT-HMM to handle the inconsistent and missing observations that are common in data gathered through medical studies. After we trained and tested our model on the PRO-ACT dataset of ALSFRS scores, our results corroborated the hypothesis. We were able to capture 66.7% of the change in patients' final ALSFRS score between their second to last and last observation. Qualitatively, our generative model was able to *hallucinate* new, realistic patient trajectories that appear consistent with *a priori* statistics regarding survival analysis.

***K*eywords** Medical informatics · Disease progression modeling · Knowledge discovery · Amyotrophic lateral sclerosis · ALS · Continuous-time hidden Markov model · Graphical modeling · Markov jump process

1 Introduction

Traditionally, research into diagnostics and disease mechanics has relied on medical practitioners' experiential intuition to formulate hypotheses which can then be tested through clinical studies. However, recent work has shown that data-driven approaches can successfully improve early stage detection and reveal non-obvious patterns and structure in diseases. Not only do these advances have immediate implications on improving survival rates of diseases like cancer, but these methods can also provide further insight into the diseases' underlying mechanics, which can then provide medical researchers with both more insight into the disease for data-driven hypothesis generation and more insight into study design and patient stratification, leading to the reduced time and cost of these trials [1].

Amyotrophic lateral sclerosis (ALS, Lou Gehrig's disease, motor neurone disease) is a rare, yet deadly neurodegenerative disease that is not well understood. There currently exists no known cure or biomarker for early stage detection. At the time of diagnosis, patients usually report inexplicable and sudden difficultly doing simple physical tasks like walking, writing, or swallowing. Unfortunately, these symptoms only worsen and spread to other areas of the body as a patient's motor cortex degrades. The patients usually pass away from related complications and a loss of ability to independently breathe within two to four years. Prognosis is difficult, due to the heterogeneity of outcomes; while some patients progress very quickly, about 10% of patients survive longer than ten years. This makes end-of-life planning difficult and imposes additional stress on the patient and their family, while also increasing the difficulty of patient selection for clinical trials [2].

We hope to address this issue by developing a model for ALS progression. If we can provide reliable predictions of future symptoms that take advantage of the patient's already-observed progression, we will reduce some additional stress associated with end-of-life planning. On top of this, developing a highly interpretable model would provide ALS researchers with insight into the heterogeneous progression patterns of the disease, which could hopefully help lead to novel hypotheses and improved patient selection for clinical studies.

The hidden Markov model (HMM) is a generative model for sequential data that assumes that observations are sampled conditionally from some latent sequence of states. HMMs have been used to model disease progression, where the unobserved states of the model correspond to the stages of the disease, e.g. the stages of cancer [3]. At each patient observation, we assume that the patient is in a stage that explains their symptoms well. Then, instead of modeling a highly complex stochastic process, we can reduce the problem to modeling the process that describes the stage progression. Once we know their likely stage progression, we can *generate*, or predict, their future symptoms.

However, standard HMMs generally rely on the assumption of uniform time intervals between observations; if an HMM is trained on disease progression data where patient observations are performed once every thirty days, then the HMM will only be capable of predicting state transition probabilities for time intervals that are multiples of thirty days. This constraint also makes it unclear how to handle missing patient observations. *Ad hoc* approaches, like symptom interpolation between observations or dataset trimming, can introduce unwanted bias in the model or invalidate large proportions of available data. To relax the constraint of uniform intervals, we propose using a continuous-time HMM (CT-HMM) [4]. In our framework, CT-HMMs model state transitions not as hard probabilities, but as *jumps* from one state to another in real-time. Instead of parameterizing transition probabilities, we parameterize the exponential rates of these *jumps*. Because of this, CT-HMMs for disease progression modeling are capable of generating interpretable, probabilistic prognoses at any point in the future and are more robust to the missing or chronologically noisy observations that are common in medical records.

2 Data and Preprocessing

Our observed data was based on the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT). This dataset contained information about over 8500 ALS patients. Despite the PRO-ACT dataset containing many different sections, we only used data from the revised ALS Functional Rating Scale (ALSFRS-R) assessments. The ALSFRS is a survey sent to ALS patients with 10 questions about functional ability to complete basic tasks like speech, cutting food, handwriting, and others. The survey is meant to serve as a representation of the ALS's effects on patients, and is collected many times throughout the progression of the disease. The response to each question on the survey marks the patient's ability to complete a task, and is collected on a 0-4 discrete scale, with 0 being difficult or low functional ability and 4 being completely normal. The revised version (ALSFRS-R) of the survey simply splits up a question from the original version about general respitory ability into 3 more specific questions about Dyspnea, Orthopnea, and Respitory Insufficiency, all of which also exist on the same 4 point scale.

2.1 Preprocessing

The ALSFRS-R Dataset we used contained survey information from 6844 ALS patients. Note that not all patients in PRO-ACT participated in the ALSFRS-R dataset. Before learning a CTHMM for this data, we had to clean it. First, only some patients took the ALSFRS-R while other took the original ALSFRS survey. To simplify things, we decided to mainly used data from the ALSFRS questions. For the patients who took ALSFRS-R, we simply averaged the Dyspnea, Orthopnea, and Respiratory Insufficiency and used it as the patients' general respiratory ability score. Furthermore, certain patients had blank surveys for certain time steps, so we simply dropped those observations. We ended up with about 5900 unique anonymized subjects with about 60,000 total observations. Each observation for each patient contains the time since the start of the trials. For CTHMM, however, we were more interested in the time between observations, so we added an attribute with the time since the last observation to each row.

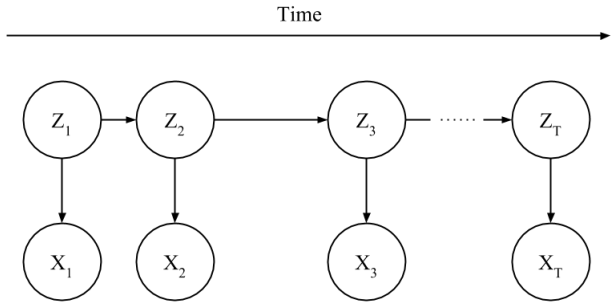

Figure 1: Graphical representation of the CT-HMM. Note the inconsistent intervals of observations.

Table 1: Notation

S	:= the set of K possible states
d	:= the dimensionality of our observed data ($d = 10$ in our application)
T	:= the length of the observed sequence
Δ_t	:= the time interval between the t'th and $t+1$'th observation
$\mathcal{X}$	:= the sequence of observed data, where $x_t \in \mathbb{R}^d$
$\mathcal{Z}$	:= the latent sequence, where $z_t \in S$
$\pi_{0\,K\times 1}$	:= initial probability matrix, where $\pi_{0i} = \mathbb{P}(z_1 = s_i)$
$A_{K\times K}$	:= transition probability matrix, where $A_{i,j} = \mathbb{P}(z_t = s_j \mid z_{t-1} = s_i)$
$R_{K\times K}$	:= rate matrix

3 Methods

3.1 Our Model

Discrete-time HMMs are generative models that explain sequential observations as emissions from a sequence of hidden states. In our specific case, the observed data is the survey information of a patient at a given time step, and the hidden states could be some transformation of the underlying disease progression. The key here is the Markov assumption: each emission is conditioned only on the current hidden state, and each hidden state is conditioned only on the previous hidden state. This allows us to tractably model complex dependencies across emissions relying only on the single dependency of the hidden states. HMMs are parameterized by the probability of transition $\mathbb{P}(z_t = s_j \mid z_{t-1} = s_i)$ from one hidden state s_i to another s_j, the probability $\mathbb{P}(x_t \mid z_t = s_i)$ of one hidden state s_i emitting an observation x_t, and the initial probabilities of hidden states $\mathbb{P}(z_1)$. Intuitively, when the underlying Markov chain that describes the state transitions is not conditioned on observations, we can model its evolution after each time step as:

$$\pi_{t+1} = A^T \, \pi_t \tag{1}$$

where π_0 is our initial distribution. Note that this suggests the constraint that A is a stochastic matrix:

$$\forall i \in [K], \quad \sum_{j=1}^{K} A_{i,j} = 1 \tag{2}$$

However, notice that this does not inherently model irregularly timed observations – there's nothing that differentiates 2 successive observations occurring 3 days apart, and another 2 observations occurring 40 days apart.

Continuous-time hidden Markov models address this problem by conditioning transition probabilities on the amount of time that has passed $\mathbb{P}(z_{t+1} = s_j \mid z_t = s_i, \Delta_t)$. CT-HMMs model a continuous latent stochastic process from which discrete emissions are observed. CT-HMMs better model irregular observations by employing a transition rate matrix, where roughly each value in the matrix describes a transition rate between two hidden states. The probabilities

of transition follow a exponential distribution and allow us to view the state transition probabilities after an arbitrary time via the matrix exponential:

$$\pi_{t+1} = e^{\Delta_t * R^T} \pi_t \tag{3}$$

where R is not a stochastic matrix, but $\exp(R^T)$ must be. This constraint implies:

$$\forall i \in [K], \quad R_{i,i} = -\sum_{\substack{j \in [K] \\ j \neq i}} R_{i,j} \tag{4}$$

$$\forall i, j \in [K] \quad \text{s.t.} \quad i \neq j, \quad R_{i,j} \geq 0 \tag{5}$$

This is more fitting for the ALSFRS data, since the time between surveys for each patient varies. Intuitively, patients visit their doctors at different intervals and report their functional status at different times. However, this brings on multiple challenges. Not only are the hidden states unobserved, but the transition times at which the hidden states are changing are also unobserved. Furthermore, for any two successive observations, CT-HMMs can actually change hidden states multiple times before the observed emission.

We initially considered modeling our emission probabilities with a categorical distribution, but decided to instead use a Gaussian partly due to the data preprocessing resulting in averaged values for the replaced respiratory question in ALSFRS, but also because using a Gaussian preserves the ordinality of the survey responses. Thus, our observations conditioned on state follow a multivariate Gaussian:

$$\mathbb{P}(x_t \mid z_t = s_i) \sim \mathcal{N}(\mu_{s_i}, \Sigma_{s_i}) \tag{6}$$

3.2 Parameters and Training

For the CT-HMM, we needed to learn the transition rate matrix, the mean for the emission probabilities, and the initial probability of each state. In an effort to simplify our model and the parameter space, we decided to use a fixed diagonal covariance matrix when learning emission distributions. Intuitively, having a fixed and diagonal covariance restricts our parameters to only model spheroid multivariate Gaussians (their PDFs extend in all dimensions equally). We could extend our model by learning variance while keeping the diagonal constraint to allow for axis-aligned stretching of the Gaussian. Finally, we could remove all constraints on the covariance, allowing the Gaussian to fit arbitrarily aligned clusters.

We use the Baum-Welch algorithm, an EM-based optimization technique, to learn posterior estimates of our model parameters. We used the forward-backward algorithm to dynamically compute the model likelihood and the posterior probabilities of latent state sequences. The forward recurrence is defined as:

$$\alpha_i(t) = \mathbb{P}(\mathcal{X}_{1:t}, z_t = s_i)$$

where:

$$\alpha_i(1) = \pi_0 \, \mathbb{P}(x_1 | z_1 = s_i)$$

$$\alpha_i(t+1) = \mathbb{P}(x_{t+1} | z_{t+1} = s_i) \sum_{j=1}^{K} \alpha_j(t) \left[e^{\Delta_t * R^T} \right]_{j,i}$$

and the backward recurrence is defined as:

$$\beta_i(t) = \mathbb{P}(\mathcal{X}_{t+1:T} | z_t = s_i)$$

where:

$$\beta_i(T) = 1$$

$$\beta_i(t) = \sum_{j=1}^{K} \beta_j(t+1) \left[e^{\Delta_t * R^T} \right]_{i,j} \mathbb{P}(x_{t+1} | z_{t+1} = s_j)$$

We use the efficient rate matrix updates from Liu et al. [4] which leverages existing literature for continuous-time Markov chain (CTMC) models (refer to Figure 2). These rate updates rely on calculating the expected time spent in specific states and the expected number of transitions between each of the states (these are the expectations that the

algorithm is calculating). Since ALS is a irreversible, degenerative disease, we restricted our transition rate matrix to be a forward matrix to save on computation. For the majority of the updates and computation, we operated in log space to avoid underflow and numerical errors. Furthermore, we corrected some known bias in our dataset by fixing some of the emission means. We fixed the means of the first state μ_{s_1} to 4, to correct for the the lack of healthy observations in our dataset despite knowing that these patients would have started as healthy, and we fixed the means of the last hidden state μ_{s_K} to be 0 to correct for the bias generated from patients pulling of out of trials near death.

Algorithm 2 The Expm Algorithm for Computing End-State Conditioned Statistics

for each state i in S **do**
  for $\Delta = 1$ **to** r **do**
    $D_i = \frac{(e^{t_\Delta A})_{(1:n),(n+1):(2n)}}{P_{kl}(t_\Delta)}$, where $A = \begin{bmatrix} Q & I(i,i) \\ 0 & Q \end{bmatrix}$
    $\mathbb{E}[\tau_i|O,T,Q] += \sum_{(k,l)\in L} C(\Delta,k,l)(D_i)_{k,l}$
  end for
end for
for each edge (i,j) in L **do**
  for $\Delta = 1$ **to** r **do**
    $N_{ij} = \frac{q_{ij}(e^{t_\Delta A})_{(1:n),(n+1):(2n)}}{P_{kl}(t_\Delta)}$, where $A = \begin{bmatrix} Q & I(i,j) \\ 0 & Q \end{bmatrix}$
    $\mathbb{E}[n_{ij}|O,T,Q] += \sum_{(k,l)\in L} C(\Delta,k,l)(N_{ij})_{k,l}$
  end for
end for

Figure 2: The transition rate matrix update from Liu et al. [4]. Note some notational differences. Most notably, they use Q, where we use R. The C values can be calculated using the results of the forward-backward algorithm.

4 Results

We used a dataset of 59259 observations and 6514 patients to train and test our model. Our final model had 12 hidden states, and was trained for 100 epochs. The following paragraphs contain our results. To note, in the future, we definitely want to explore training many more times with different parameter initializations. Training HMMs is notoriously difficult, and the problem is inherently non-convex, so models can easily be trapped in local optima. We found that our models seemed to converge within the first few epochs of training.

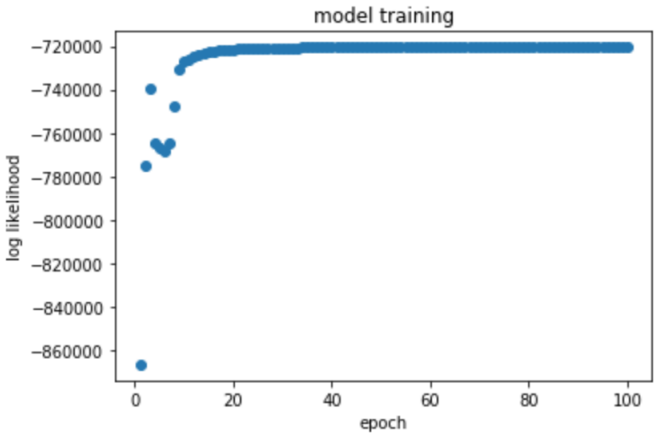

Figure 3: Our Log Likelihood graph during training. The model seemed to converge within the first 25 epochs. We aren't entirely sure why the model performance isn't monotonic. In theory, alternating updates in EM have been proven via Jensen's inequality to increase model performance.

Next, our initial hidden state probabilities. Note that the probabilities actually aren't monotonic. It seems that the majority of our patients are most likely to "start" in state 2.

State	0	1	2	3	4	5	6	7	8	9	10	11
Probability	0.24	0.18	0.39	0.02	0.02	0.1	0.008	0.01	0.006	0.001	0.0009	0

	0	1	2	3	4	5	6	7	8	9
0	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
1	3.78	3.83	3.86	3.77	3.83	3.08	3.32	2.14	1.12	3.78
2	3.58	3.67	3.75	2.91	2.62	2.67	3.54	3.55	3.49	3.76
3	1.10	1.73	2.15	3.64	3.58	3.54	3.80	3.64	3.43	3.46
4	2.16	2.65	2.75	3.00	2.56	2.40	2.93	2.37	1.24	3.13
5	3.66	3.80	3.81	2.89	2.46	2.32	2.97	2.21	1.01	3.60
6	3.45	3.70	3.60	2.95	2.11	1.29	1.58	1.17	0.15	3.31
7	3.25	3.48	3.56	1.36	0.99	1.37	2.39	2.77	1.88	3.53
8	3.20	3.59	3.50	0.68	0.46	0.45	1.03	1.38	0.25	3.17
9	1.41	2.25	1.95	0.53	0.27	0.34	0.81	1.24	0.21	2.82
10	0.44	0.98	1.06	2.80	1.87	1.77	2.37	2.22	1.02	2.80
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 4: The means for each observation given each state. The x-axis represents the 10 questions in the ALSFRS survey, and the y-axis represents the 12 states. We hardcoded the means for the first state to be 4 and the last state to be 0, making the assumption that patients (around 59000 observations) are initially in perfect health and are deceased at the end.

Next, we show our initial transition matrix, and our transition matrix after 120 days. To get this from our transition rate matrix, we simply use the matrix exponential:

$$P(z_t|z_{t-1}) = e^{\Delta_t * R^T}$$

where R is the transition rate matrix, Δ_t is the number of days since our last observation.

	0	1	2	3	4	5	6	7	8	9	10	11
0	0.96	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.96	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.96	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.96	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.97	0.01	0.01	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.97	0.01	0.01	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.01	0.01	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.01	0.01	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.01	0.01	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.01	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.01
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 5: Transition matrix after 0 days. Note that patients are extremely likely to stay in the same state after 0 days. Note that the transition matrix is upper triangular. Since ALS is degenerative, we can save on computation and only update the forward probabilities.

	0	1	2	3	4	5	6	7	8	9	10	11
0	0.0	0.01	0.02	0.02	0.03	0.04	0.06	0.08	0.10	0.14	0.19	0.30
1	0.0	0.01	0.01	0.02	0.03	0.04	0.06	0.08	0.10	0.14	0.20	0.30
2	0.0	0.00	0.01	0.02	0.03	0.04	0.06	0.08	0.11	0.14	0.20	0.31
3	0.0	0.00	0.00	0.01	0.03	0.04	0.06	0.08	0.11	0.15	0.21	0.32
4	0.0	0.00	0.00	0.00	0.02	0.04	0.05	0.08	0.11	0.15	0.22	0.34
5	0.0	0.00	0.00	0.00	0.00	0.03	0.05	0.08	0.11	0.16	0.23	0.36
6	0.0	0.00	0.00	0.00	0.00	0.00	0.04	0.07	0.11	0.16	0.24	0.39
7	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.10	0.16	0.25	0.42
8	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.17	0.27	0.47
9	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.30	0.54
10	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.67
11	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 6: Transition matrix after 120 days. Note that the values are very bottom heavy – after 120 days, patients are fairly likely to have transitioned multiple states. Note that the transition matrix is upper triangular. Since ALS is degenerative, we can save on computation and only update the forward probabilities.

To evaluate our model, we tested on a subset of 20% of the data. Our testing method, by Professor El-Kebir's suggestion, was to try and predict the final observed emissions (answers to the survey questions) for each patient, given their previous observed emissions. We did this by taking a weighted sum of our predicted final emissions over the distribution of our predicted final states. To quantify our errors, took the percentage difference between the total ALSFRS score of our final predicted emissions and our final observed emissions. We had a total of 33.3% error, meaning we were able to capture 66.7% of the change in the patients' final ALSFRS score between their second to last and last observed observation. To provide some visual clarity, we have also included an example comparison of the distributions between our final predicted emissions and our final observed emissions in Figure 7.

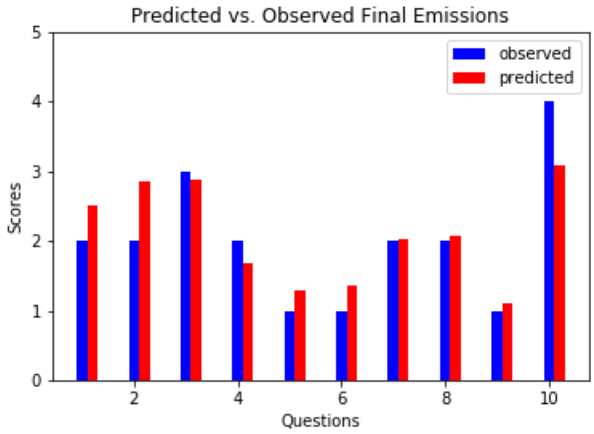

Figure 7: Plot of the distribution of final observed emissions (answers to the survey questions) vs. the distribution of our model's predicted final emissions for a single patient.

5 Conclusion and Future Work

We have shown a valid way to model ALS progression using Continuous Time Hidden Markov Models. Using the ALSFRS data from the PRO-ACT dataset, we trained a CTHMM with 12 hidden states efficiently using the rate matrix updates from Liu et al. [4]. We wrote our own code for the Baum-Welch algorithm, including the Forward, Backward, and updates. We also wrote Viterbi for inference.

In the future, there is more that can be done to attempt to improve our model as well as it's findings. For example, we assumed for simplicity that our Multivariate Gaussian model for emission probabilities had a constant and diagonal covariance. Relaxing this constraint and learning the covariance matrix (with or without the additional diagonal constraint) as an additional model parameter could improve the results of our model, at the cost of increasing it's complexity. Furthermore, there is room to experiment with different types of emission probability models entirely. The emission probabilities do not have to be modeled after a Multivariate Gaussian, but could instead be modeled as a Categorical Distribution, or another distribution of choice.

Apart from our emission probabilities, the hyperparameter for the number of hidden states could also be tweaked further in search of an optimal number. Furthermore, since HMMs are difficult to train and are an inherently non-convex problem, they are incredibly likely to converge to local optima. There are many more initializations of the rate matrix which can be performed in order to alleviate this fact and attempt to converge as close to a global optima as possible.

References

[1] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. *KDD'14*, 2014.

[2] Lewis P. Rowland and Neil A. Shneider. Amyotrophic lateral sclerosis. *New England Journal of Medicine; 344:1688-1700*, 2001.

[3] Miguel Simoes. A hidden markov model for cancer progression.

[4] 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.