Hidden Markov Models and Considerations for Digital Phenotyping Data

Patrick Emedom-Nnamdi

patrickemedom@hsph.harvard.edu

Overview

Hidden Markov Models (HMMs) are a flexible class of models for univariate and multivariate time series

- They assume that the distribution that generates an observation depends on the state of an *underlying and unobserved* Markov process
  - e.g. Psychological studies where we expect our responses to change due some underlying cognitive state that may unfold over time
- Allow us to conduct inference on unobservable state process
- HMMs are essentially suitable for settings where the observed data exhibits serial dependence
  - Other areas of applications include: ecology (e.g. animal behavior), environmental science (e.g. natural disasters and weather), financial data, and speech

Overview

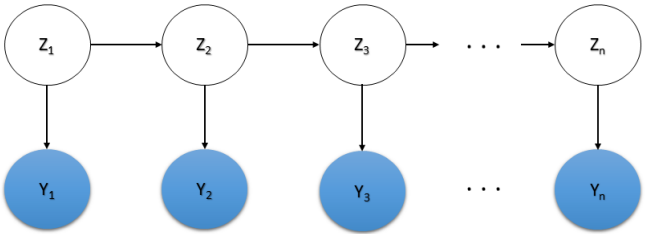
A **HMM** is a dependent mixture model comprised of two stochastic processes:

- Unobserved (or hidden) parameter process, $\{Z_t\}_{t=1}^n$, that follows a Markov chain, where $Z_t \in \{1, 2, \ldots, K\}$
  - Def. $(Z_1, \ldots, Z_n)$ is a Markov Chain if $Z_{t+1} \ (Z_1, \ldots, Z_{t-1}) \mid Z_t$
  - In words: "The future is conditionally independent of the past given the present."
- A state-dependent process, $\{Y_t\}_{t=1}^n$, where the distribution of our observed values depends only on the current state of underlying parameter process

Note: $t \in \{1, \ldots, n\}$ is an evenly-spaced, discrete unit of time

Overview

In other words, a HMM is a distribution, $p(y_1, \ldots, y_n, z_1, \ldots, z_n)$ that respects the following directed graph:



Hence, this reduces to following

$$p(y_1, \ldots, y_n, z_1, \ldots, z_n) = p(z_1)\, p(y_1 \mid z_1) \prod_{t=2}^{n} p(z_t \mid z_{t-1})\, p(y_t \mid z_t)$$

Modeling the joint distribution $\mathrm{p}(y_1, \ldots, y_n, z_1, \ldots, z_n)$

Initial Distribution:

- $\pi_k = p\left(Z_1 = k\right)$ and $\sum_k \pi_k = 1$
- π is a vector that represents the initial distribution of Z_1

Transition Probabilities:

- $T_{kj} = p\left(Z_t = j \mid Z_{t-1} = k\right)$ where $k, j = 1, \ldots, K$
- Thus, we have a $K \times K$ transition matrix T in where the (k,j)-th entry is T_{kj}
- $\sum_k T_{kj} = 1$ - each row of our transition matrix must sum to 1.

Emission Distributions:

- $\varepsilon_k\left(Y_t\right) = p\left(y_t \mid Z_t = k, \phi_k\right)$
- This is the distribution of our observed outcome. In general, this distribution may be either discrete (e.g. binomial, Poisson, multinomial), continuous (e.g. normal, Gamma), or multivariate

Example of a 2-State HMM

Suppose we have the following:

- **Number of hidden states**: $K = 2$. Hence, we have $Z_t \in \{1, 2\}$
- **Initial distribution**:

$$\pi = (0.5, 0.5)$$

- **Transition Matrix**:

$$T = \begin{bmatrix} .7 & .3 \\ .2 & .8 \end{bmatrix}$$

- **Emission Distribution**:

$$Y_t \mid Z_t = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right)$$

where $\mu = (-1, 1)$ and $\sigma = (1, 1)$

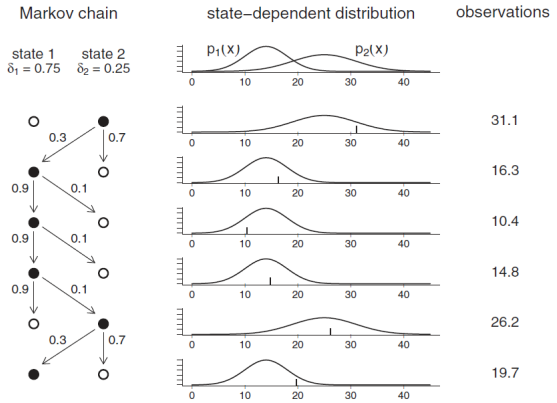


Figure 2.3 *Process generating the observations in a two-state HMM. The chain followed the path* $2, 1, 1, 1, 2, 1$, *as indicated on the left. The corresponding state-dependent distributions are shown in the middle. The observations are generated from the corresponding active distributions.*

Parameter Estimation via Maximum Likelihood

Suppose we observe a dataset $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$

- Our goal is to learn the HMM model parameters $\theta = \{\pi, T, \boldsymbol{\phi}\}$ via maximum likelihood
- Let $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ be the sequence of hidden states, where K is chosen a priori
- The observed data likelihood function $\mathcal{L}(\theta|\boldsymbol{Y})$ is obtained from the joint distribution $p(\boldsymbol{Y}, \boldsymbol{Z}|\theta)$ by marginalizing over all possible sequences the hidden states:

$$\mathcal{L}(\theta|\boldsymbol{Y}) = p(\boldsymbol{Y}|\theta) = \sum\nolimits_{\boldsymbol{Z}} p(\boldsymbol{Y}, \boldsymbol{Z}|\theta) \tag{1}$$

$$= \sum\nolimits_{\boldsymbol{Z}} p(Z_1|\boldsymbol{\pi}) \prod_{t=2}^{n} p(Z_t|Z_{t-1}, \boldsymbol{T}) \prod_{t=1}^{n} p(y_t|Z_t, \boldsymbol{\phi}) \tag{2}$$

$$= \sum\nolimits_{\boldsymbol{Z}} \prod_{k=1}^{K} \pi_k^{[Z_1=k]} \prod_{t=2}^{n} \left[\prod_{k=1}^{K} \prod_{j=1}^{K} T_{kj}^{[Z_t=j, Z_{t-1}=k]} \right] \prod_{t=1}^{n} \left[\prod_{k=1}^{K} p(y_t|Z_t = k, \phi_k) \right] \tag{3}$$

Maximum Likelihood via EM Algorithm

In general, direct maximization of $\mathcal{L}(\theta|\boldsymbol{Y})$ is difficult

- Involves summing over K^n sequences of hidden states (i.e. impractical for long sequences)

Expectation-Maximization (EM) algorithm is commonly used to efficient maximize the likelihood function in HMM

- **E-step** - Compute the posterior distribution of the latent states, $p(\boldsymbol{Z}|\boldsymbol{Y}, \theta^{\textbf{old}})$, and evaluate the expected complete data log-likelihood:

$$Q(\theta, \theta^{\textbf{old}}) = \mathbb{E}[\log p(\boldsymbol{Y}, \boldsymbol{Z}|\theta)\,|\boldsymbol{Y}, \theta^{\textbf{old}}] = \sum\nolimits_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}, \theta^{\textbf{old}}) \log p(\boldsymbol{Y}, \boldsymbol{Z}|\theta)$$

- **M-step** - Maximize $Q(\theta, \theta^{\textbf{old}})$ with respect to $\theta = \{\pi, T, \boldsymbol{\phi}\}$

Note: the E-step requires evaluating $p(Z_t|\boldsymbol{Y}, \theta^{\textbf{old}})$ and $p(Z_{t-1}, Z_t|\boldsymbol{Y}, \theta^{\textbf{old}})$, which can be efficiently computed using the *forward-backward algorithm*

What Now?

Important tasks that rely on $\hat{\theta} = \{\hat{\pi}, \hat{T}, \hat{\phi}\}$:

1. Decode the latent states via the Viterbi Algorithm
   - $\boldsymbol{Z} \in \text{argmax}_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{Y}, \hat{\theta})$
2. Compute the likelihood of HMMs via Forward Algorithm
   - Necessary for model selection via AIC and BIC (i.e. choosing the number of states)
3. Perform inference on the transition and emission model

Huang et al. (2018)

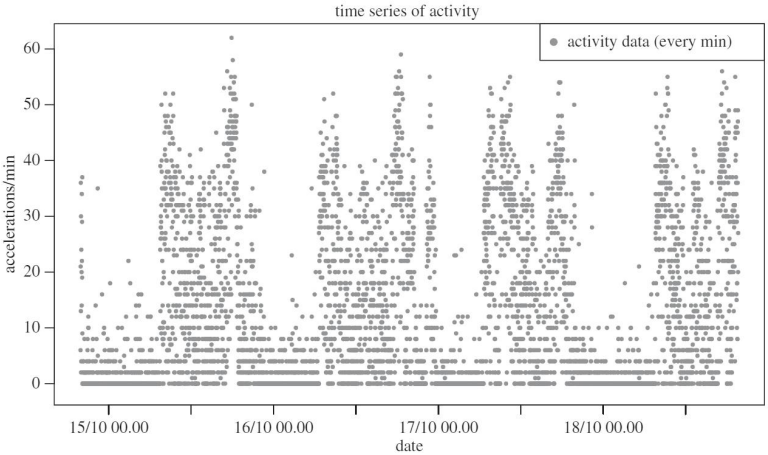


Figure 1. Example of raw accelerometer data: activity counts recorded per minute over 4 days with Move3 (Movisens GmbH, Germany) sensor with inbuilt accelerator ADXL345 (Analog Devices, MA, USA) fixed to the chest of a healthy individual (subject 16).

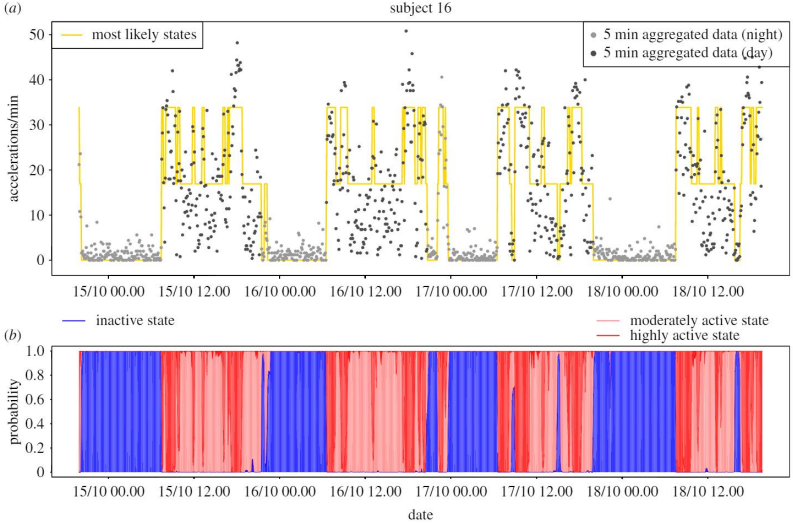


Figure 5. State estimation for example subject 16. (*a*) Time series of activity with yellow line indicating the mostly likely state using local decoding. (*b*) SP plot, i.e. cumulative plot of $P(S_t = j \mid Y^{(T)})$ for $j = 1$ (IA, blue), 2 (MA, light red), 3 (HA, dark red).

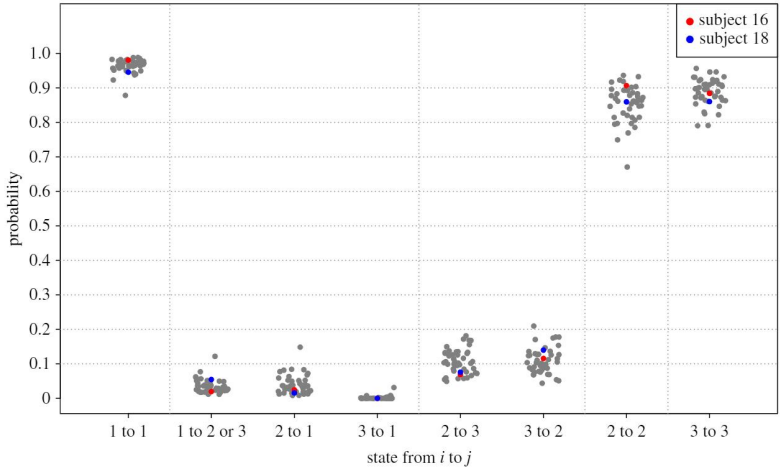


Figure 3. Estimated transition probabilities for 46 healthy individuals. The integers 1, 2 and 3 represent the inactive (IA), medium active (MA) and highly active (HA) states, respectively.

Considerations for Digital Phenotyping Data

- Patient heterogeneity
- Large (potentially high dimensional) feature space
- Irregular observation times
- Latent state space representation
- Computational efficiency

Accounting for Patient Heterogeneity

We assume that the observed data $\mathbf{Y}_i$ produced by each subject $i \in \{1, \ldots, N\}$ are independent, each with its own underlying sequence of hidden states

- This stems from the assumption that (1) each subject may transition between states according to their own internal process and (2) given a state, each subject's state dependent distribution may be different
- Hence, parameters of the HMM should vary across subjects (i.e. θ_i is now subject specific)

Hence, variability between the N subjects can be explained by

1. Covariate information
2. Inclusion of random effects

Covariate Information

Covariates can be included in either that transition probability matrix or in the state-dependent (emission) distributions to account for some variability

Transition Probabilities: Let $W_{kj} \in \mathbb{R}^p$ be a parameter vector and $C_i \in \mathbb{R}^p$ are baseline covariates for subject i. We can model the transition probability matrix using a multinomial logistic function

$$T_{kj,i}(W) = p\left(Z_{it} = j \mid Z_{i,t-1} = k, C_i\right) = \frac{\exp(\tau_{kj} + W_{kj}C_i)}{\sum_{h=1}^{K} \exp(\tau_{kh} + W_{kh}C_i)}$$

- Note that the condition $\sum_k T_{kj,i} = 1$ must not be violated. We must be careful not to overparameterize T_i
- Note that time varying covariates C_{it} can be included to produce a non-homogeneous HMM

Covariate information

Emission Distributions: Let $X_i \in \mathbb{R}^d$ be a subject-specific (or time-varying X_{it}) covariates and $\beta_k \in \mathbb{R}^d$ be state-dependent covariates

$$\varepsilon_k\left(Y_{it}\right) = p\left(y_{it} \mid Z_{it} = k, X_i, \beta_k\right)$$

- e.g. Let $Y_{it} \mid Z_{it} = k, X_i \sim \mathcal{N}\left(f_k(X_i), \sigma_k^2\right)$, where the conditional mean is some parametric function of the covariates
  - e.g. $f_k(X_i) = X_i\beta_k$

Estimation & Inference procedures under the inclusion of covariates in either the t.p.m or the emission distribution is nearly identical to traditional HMMs

- Modified EM algorithm that operates over N subjects
- Viterbi algorithm simply operates on each patient specific HMM, $\boldsymbol{Z_i} \in \text{argmax}_{\boldsymbol{Z_i}}\, P(\boldsymbol{Z_i}|\boldsymbol{Y_i}, \hat{\theta}_i)$

Random Effects

Random effects are parameters (generated from a random model) that allow the estimated effects of specific covariates (or the intercept) to vary across subjects

- They can also be included in the transition probability matrix or in the emission distribution
- Extremely useful in accounting for between-subject and within-subject variability

Random Effects

Model Structure I: Common t.p.m with random effects on emissions

Let Y_{it} follow a distribution from the exponential family conditional on the random effects, $\mathbf{u} \sim f(\mathbf{u}|\theta)$, the hidden states, $\mathbf{Z}$, and our model parameters θ:

$$f(y_{it}|Z_{it} = k, \mathbf{u}, \theta) = \exp\left\{\frac{(y_{it}\eta_{itk} - c(\eta_{itk}))}{a(\phi)} + d(y_{it}, \phi)\right\}$$

where, $\eta_{itk} = \tau_k + \mathbf{x}'_{it}\beta_k + \mathbf{w}'_{itk}\mathbf{u}$.

The likelihood for the model is specified as follows

$$\begin{aligned} L(\theta|\mathbf{y}) &= \int_{\mathbf{u}} \sum_{\mathbf{z}} f(\mathbf{y}|\mathbf{z}, \mathbf{u}, \theta) f(\mathbf{z}|\theta) f(\mathbf{u}|\theta) d\mathbf{u} \\ &= \int_{\mathbf{u}} \prod_{i=1}^{N} \Big\{ \sum_{z_i} \pi_{z_{i1}} f(y_{i1}|z_{i1}, \mathbf{u}, \theta) \prod_{t=2}^{n_i} P_{z_{i,t-1}, z_{it}} f(y_{it}|z_{it}, \mathbf{u}, \theta) \Big\} f(\mathbf{u}|\theta) d\mathbf{u}, \end{aligned}$$

where $\{P_{kl}\}$ is homogeneous transition probability matrix and $\{\pi_k\}$ is the vector initial probabilities; both quantities are common for all i.

Random Effects

Model Structure II: Random effects on t.p.m and emissions

Here the transition probabilities are modeled as

$$P(Z_{it} = l | Z_{i,t-1} = k, \mathbf{u}, \theta) = \frac{\exp(\tau^*_{kl} + \mathbf{x}^{*'}_{it}\beta^*_{kh} + \mathbf{w}^{*'}_{itkl}\mathbf{u})}{\sum_{h=1}^{k} \exp(\tau^*_{kh} + \mathbf{x}^{*'}_{it}\beta^*_{kh} + \mathbf{w}^{*'}_{itkh}\mathbf{u})}.$$

Furthermore, the likelihood under this model is follows as

$$L(\theta|\mathbf{y}) = \int_{\mathbf{u}} \prod_{i=1}^{N} \Big\{ \sum_{z_i} \pi_{i,z_{i1}} f(y_{i1}|z_{i1}, \mathbf{u}, \theta) \prod_{t=2}^{n_i} P_{i,z_{i,t-1},z_{it}} f(y_{it}|z_{it}, \mathbf{u}, \theta) \Big\} f(\mathbf{u}|\theta) d\mathbf{u},$$

Estimation & Inference:

- Random effects make traditional EM algorithm or direct maximization difficult to perform
- Altman proposes using direct maximization while employing both Guassian quadrature and quasi-Newton methods, or an Monte Carlo expectation-maximization algorithm (Altman 2007)

High Dimensional Data - Need for Parsimonious Models

When dealing with digital phenotyping data, its common to assume that the feature space of $\mathbf{X} \in \mathbb{R}^d$ is large or, even, high dimensional

- Determining which set of covariates belong either to the transition model or the emission model may require clinical guidance or an exhaustive search

Variable selection methods for both the transition model and emission model are need

- Select relevant features that affect between state-transitions
- Select relevant features that fully characterize the true emission model
  - e.g. introduce feature saliencies (Adams et al. 2016)
    - ⋆ parameters represent the probability that a feature is relevant by distinguishing between state-dependent and state-independent distributions
  - e.g. state-dependent variable selections methods (e.g. proximal gradient descent methods, such as LASSO on M-step of the EM)

Irregular Observation Times

HMMs assume that data are sampled regularly over **discrete intervals**. However, this assumptions is commonly violated in clinical and digital phenotyping studies

- Patient data are sampled irregularly over time
  - e.g. instances of informative missingness (i.e. missed surveys, zero reading when a subject's smartphone is turned off)

Irregular Observation Times

Continuous-time HMM is an HMM where both the transition model between states and the arrival of observations can occur over continuous time intervals

- Under this model, (1) hidden parameter process is unobserved and (2) the exact between-state transition times are also unobserved
- Rely on a transition rate matrix Q where each sojourn time in each state k is exponentially distributed
  - Inference focuses on (1) transition intensities between states and (2) mean sojourn time in each state
- Increased flexibility over discrete time HMMs
  - helps avoid the need of imputation for missing data
  - can be computationally costly for large datasets

Latent state space representation

Given an application, obtaining a discrete representation of states can be difficult

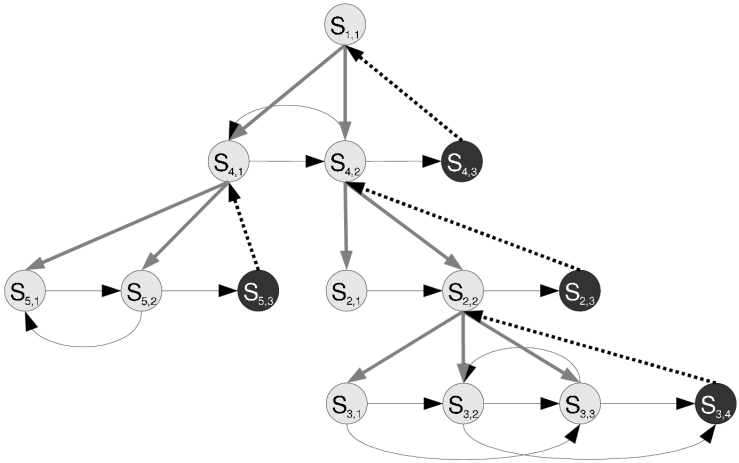
1. Choosing finite number of states may not be intuitive
2. Complex problem space results in a large number of states $\longrightarrow$ large parameter space

We can consider a **continuous-valued state process**, where the states are real-valued

- These are formally referred to as **State Space Models** (SSM) of which HMMs are a special case

Latent state space representation

Hierarchical HMMs



Computational Efficiency

- Estimating parameters and computing the likelihood of an HMM are computationally expensive tasks
- Typically digital phenotyping studies sample temporally dense (e.g. minutes, hours, days) data over long time periods (e.g. 6 months, 1 year, etc)
- Need for computationally efficient online algorithms that estimate model parameters as new data $\{\mathbf{Y}, \mathbf{X}\}_i^{\textbf{new}}$ is observed for each subject i

Thank you!