

Hidden Markov Modeling

Nathaniel Osgood

University of Saskatchewan

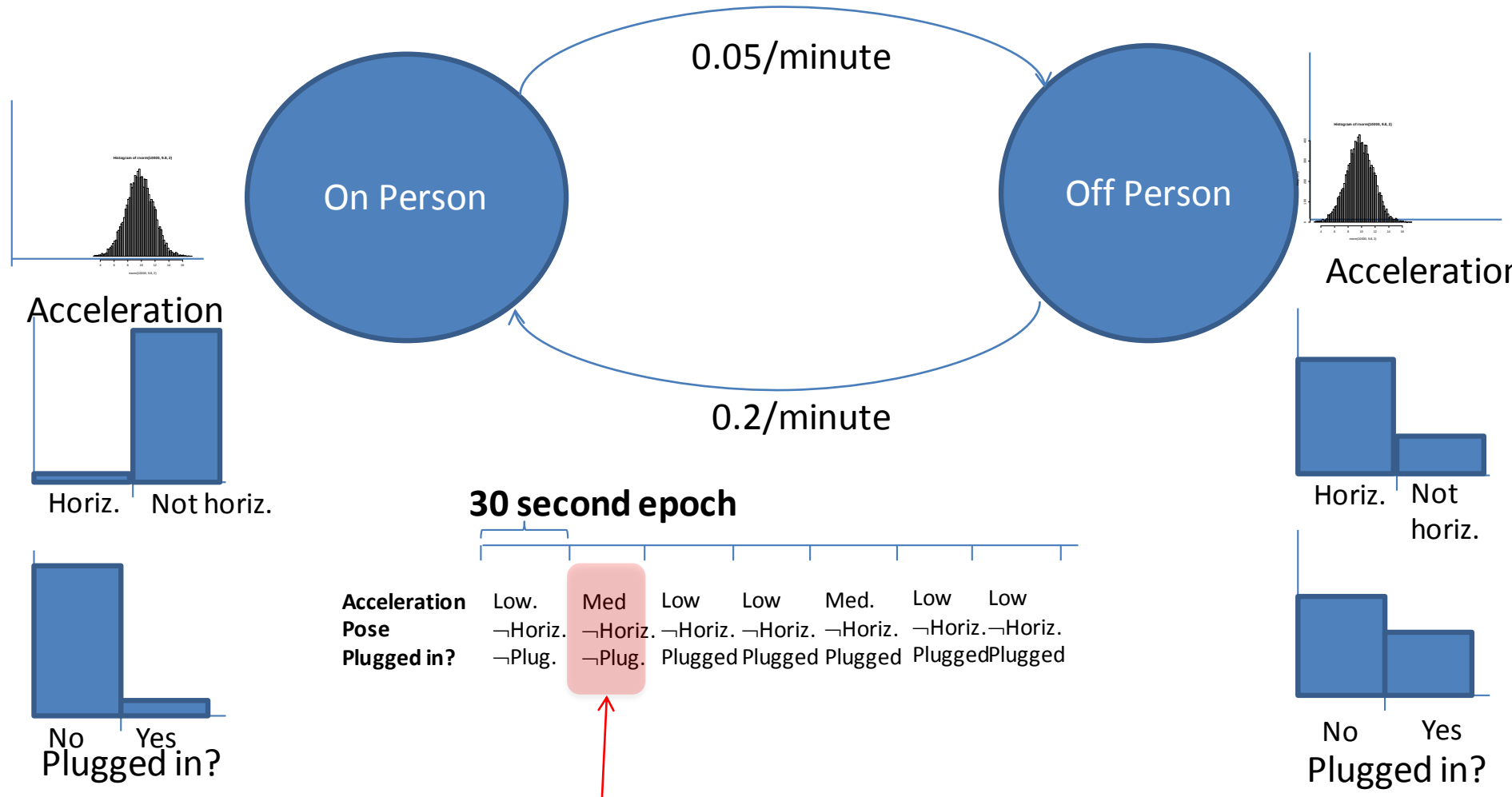
HMMs Depict Common Circumstance

- Some underlying system transitioning between a small set of (categorical/nominal) states
 - Often have some theory as to what the states might be
- ≥ 1 types of empirical observations over time
- Observations are ambiguous (e.g., noisy, sparse, and generally individually insufficient to conclusively identify the state)
 - In many cases: The distributions posited to obtain within particular states are likely to overlap strongly
- We'd like to infer with confidence “what’s going on” – i.e., in what state we are located

Research Questions Addressed

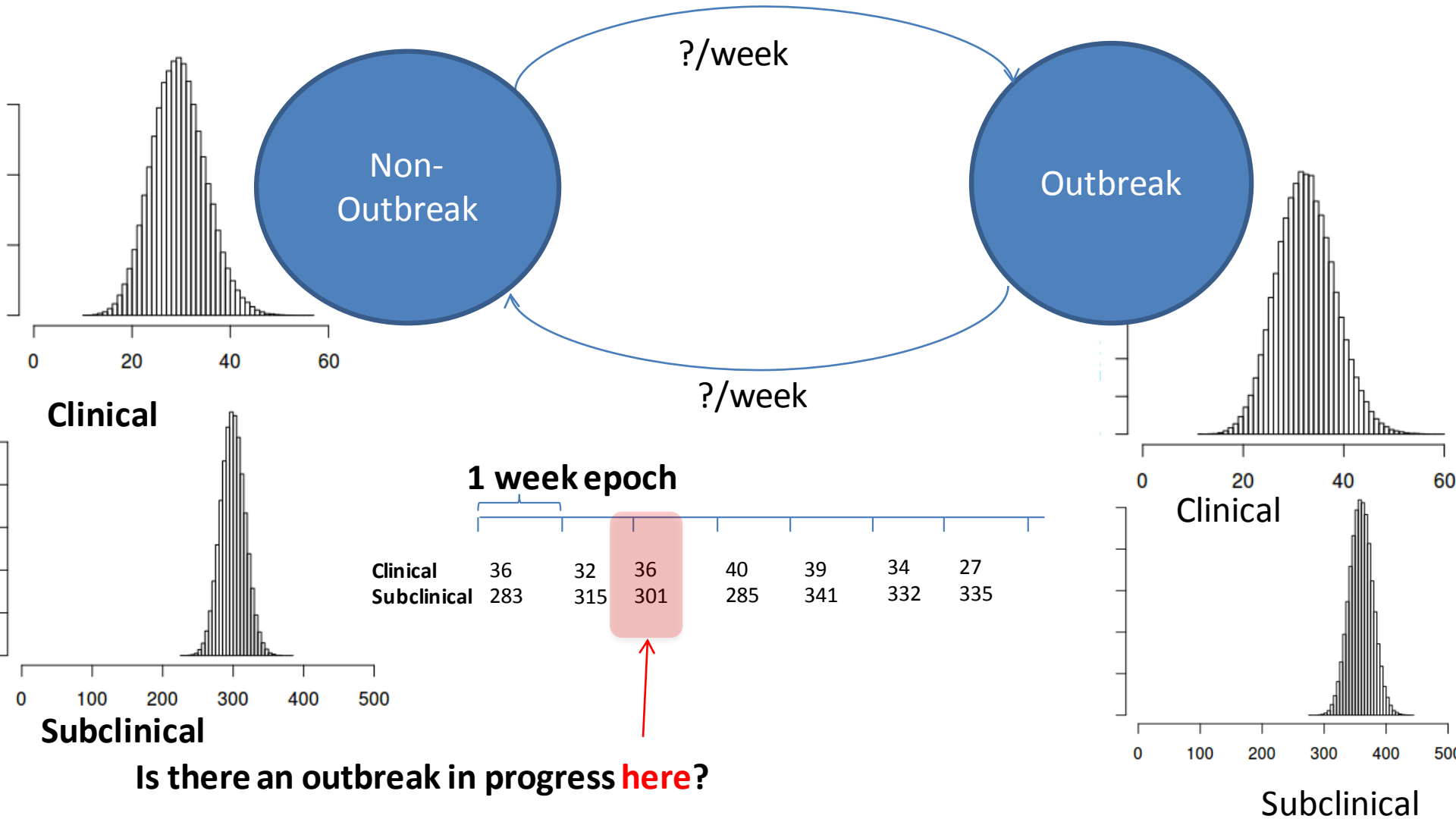
- What are the set of latent categorical states underlying the system?
- How often are transitions made between these states?
- What is the balance of time spent in different states?
- What categorical state am I in
 - at the current time?
 - at time t , looking back retrospectively?
- What is the most likely *sequence* of states that has transpired, in light of the data?

Example Machine Learning Model: Hidden Markov Models

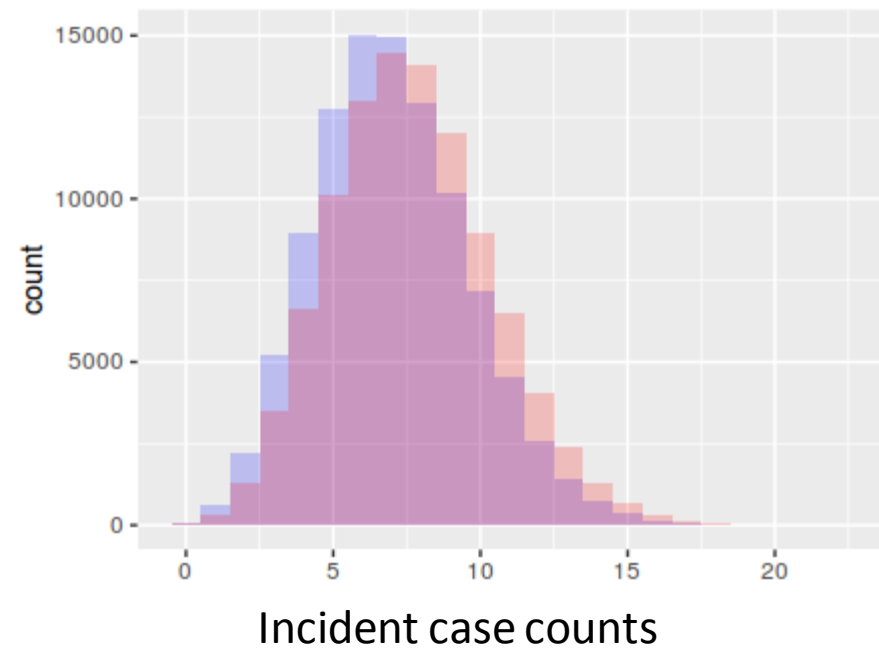


Is the phone on-person or off-person **here**?

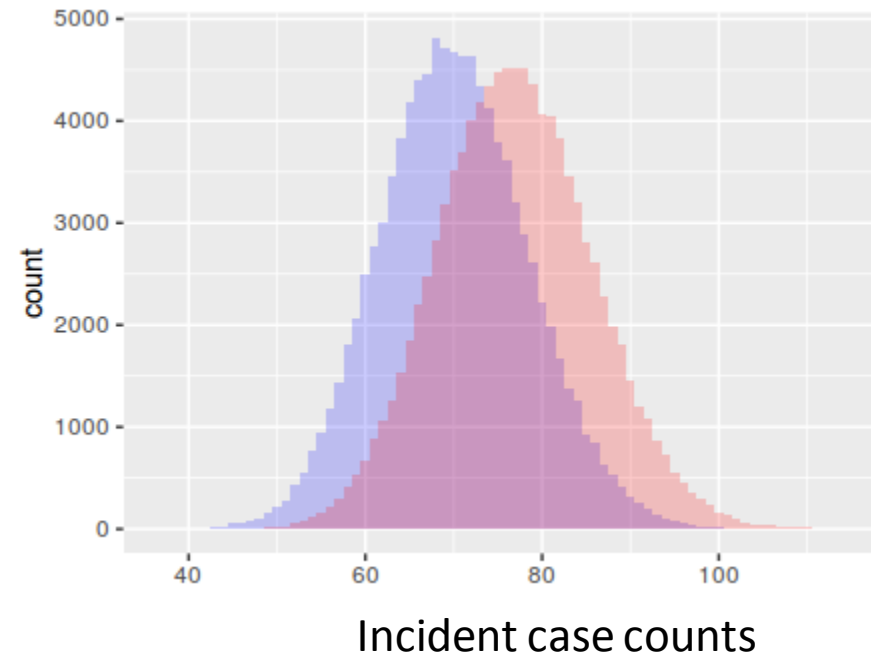
Example Machine Learning Model: Hidden Markov Models



Resolving Clinical and Subclinical States (Weekly)



Subclinical



Simplifications: Discrete States

- **Current state:** State of the system at any time as being one of a discrete set of possibilities
 - The current state changes over time
- The states are treated as
 - Mutually exclusive
 - Collectively exhaustive
- Examples
 - Contaminated vs. non-contaminated
 - Off-person vs. sitting vs. standing vs. lying down vs. active

Simplifications: Discrete **Time**

- Time is divided into equally sized “time steps”
- For each time step
 - Exactly 1 observation of each type
 - If observations sometimes do not occur, create a designated “no observation” category
 - If a given size of time step would make multiple observations likely, make time step smaller
- The system is in exactly 1 state $\Rightarrow$ Transitions occur instantaneously between time steps

Simplifications: Memoryless States

- The chance of departing each state in a given time step is treated as independent of how long that state has been occupied (residence time)
- If we want to capture non-memoryless states (e.g., where likelihood of transitioning rises with time), create successive series of states

Simplifications: Observations

- We assume that the probability of making a certain (set of) observations at a given time depends (only) on the state
 - Continuous observation (e.g., blood pressure, sleep duration): Probability distribution conditional on state
 - Discrete observation: e.g., occurrence of “screen turning on” or screen turning off event, occurrence of answering survey
- Conditional on being in a state, successive observations are held to be independent
 - i.e., each is independently drawn from the probability distribution for observations for that state

(Semi-) Formal Specification

- Set of possible states S
- Set of possible observations X
- Transition matrix (Γ): Probability of transitioning from one state to the next
 - Each row is specific to a state in which we are currently located
 - Each column is associated with a state to which we can head
 - A row gives probability of going from the associated state to each other state $\Rightarrow$ sums to 1
- T : Count of observations
- π : Probability of initially being in each state.
(Typically less critical to success)

Observation Probabilities $P(x|s)$

- s is a state
- x is an observation; for time $1 \leq i \leq T$, we write x_i
- $P(x|s)$ is the probability (density) of observing x in time step given that the HMM is currently in state s
- We generally specify this in one of two ways
 - Discrete possible outcomes: $|S| \times |X|$ “Emission matrix”
 - For each state, lists the probabilities of observing each possible observation while within a time step spent in that state
 - Continuous and some discrete possible outcomes:
Probability distribution function; variants:
 - Empirical (precise distribution, often smoothed); common in for supervised learning
 - Parametric distribution

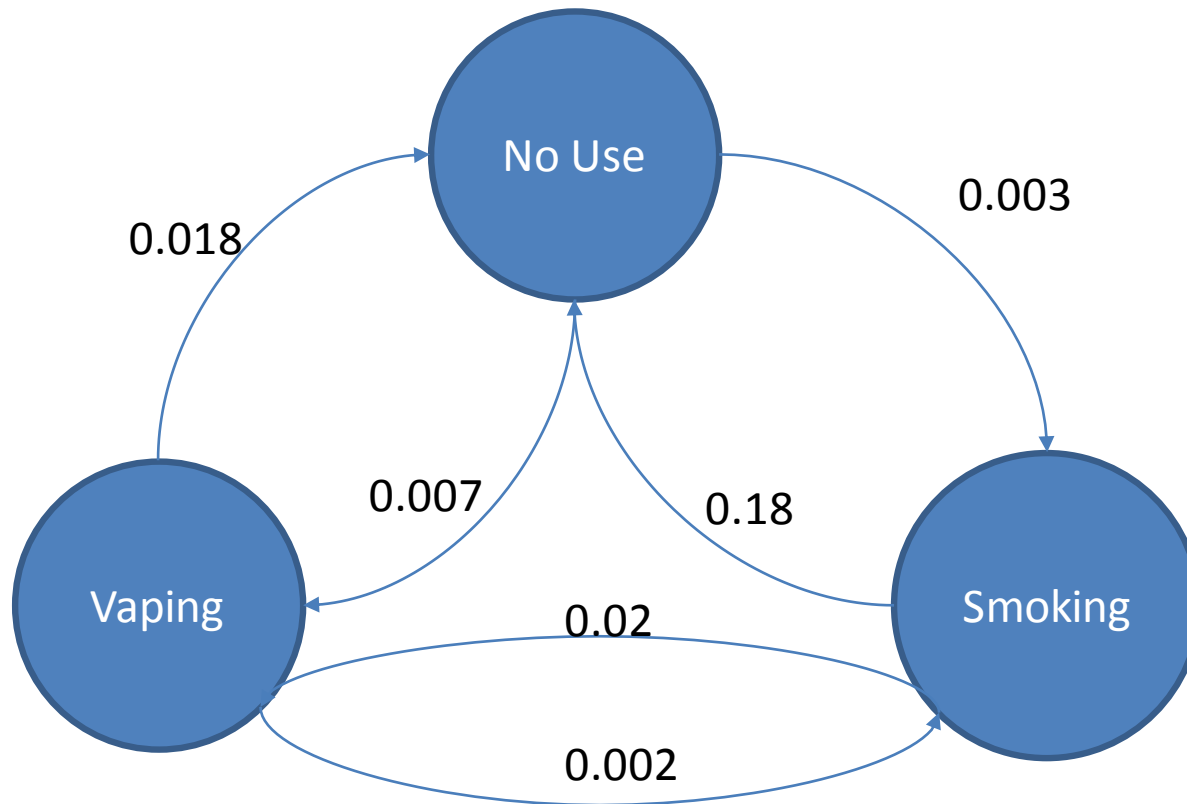
(Semi-) Formal Specification

- Transition matrix (Γ): Probability of transitioning from one state to the next
 - Each row is specific to a state in which we are currently located
 - Each column is associated with a state to which we can head
 - A row gives probability of going from the associated state

Example Transition Matrix

	No Use	Smoking	Vaping
No Use	0.99	0.003	0.007
Smoking	0.18	0.80	0.02
Vaping	0.018	0.002	0.98

Time step is 1 min

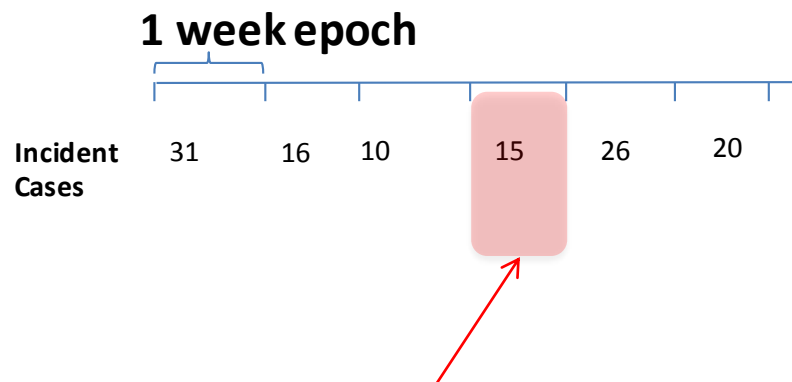


Modes of Use

- Iterate as required
 - Train on data with different parameter values
 - Cross-validate to test reliability of training
- Apply Hidden Markov Model

Example [Adapted from Zucchini, Ch. 2]

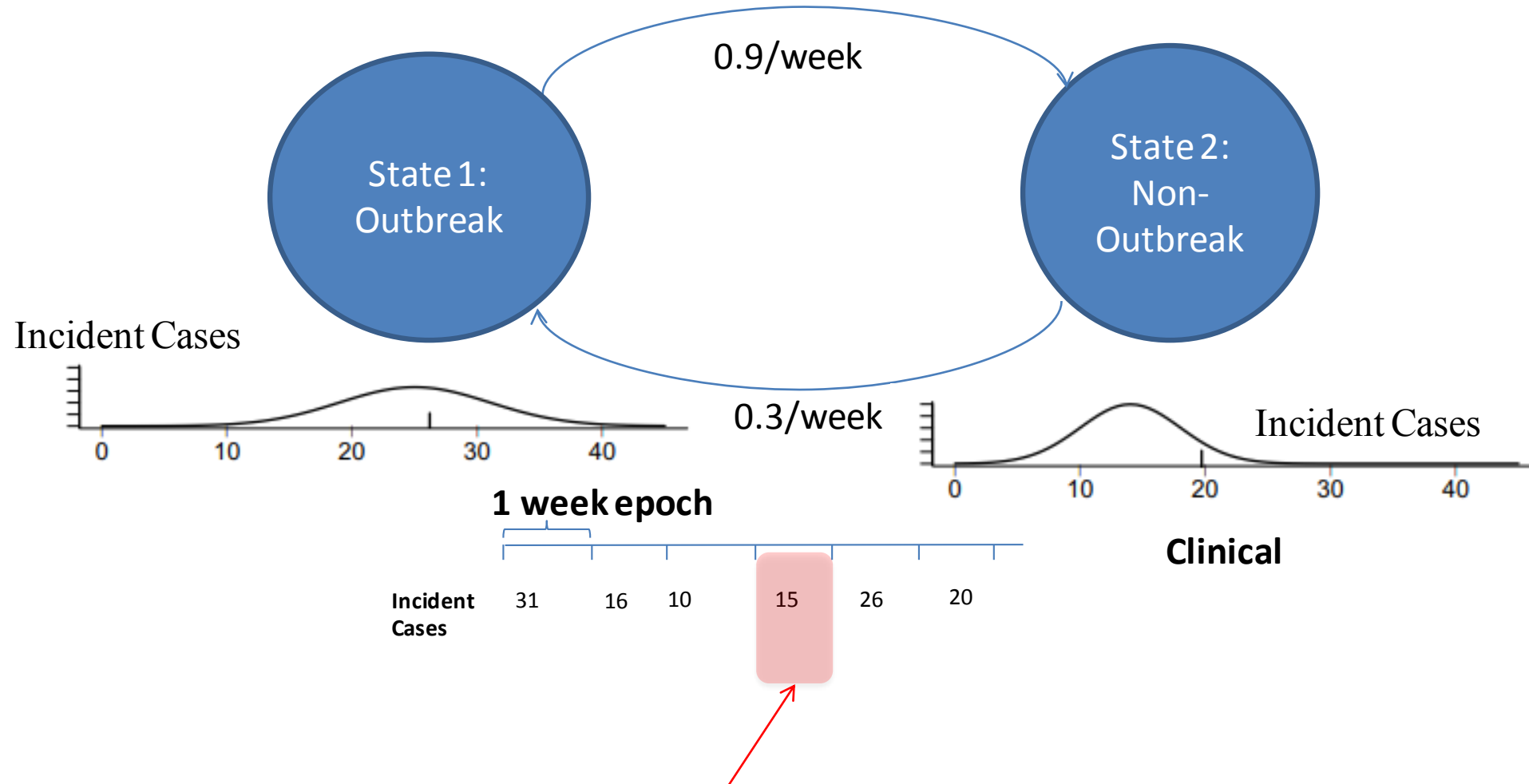
- We recognize that there are occurrences of outbreaks, and periods without outbreaks
- We know that periods with outbreaks tend to have higher incident case counts of reported illness



Is there an outbreak in progress **here**?

Example [Adapted from Zucchini, Ch. 2]

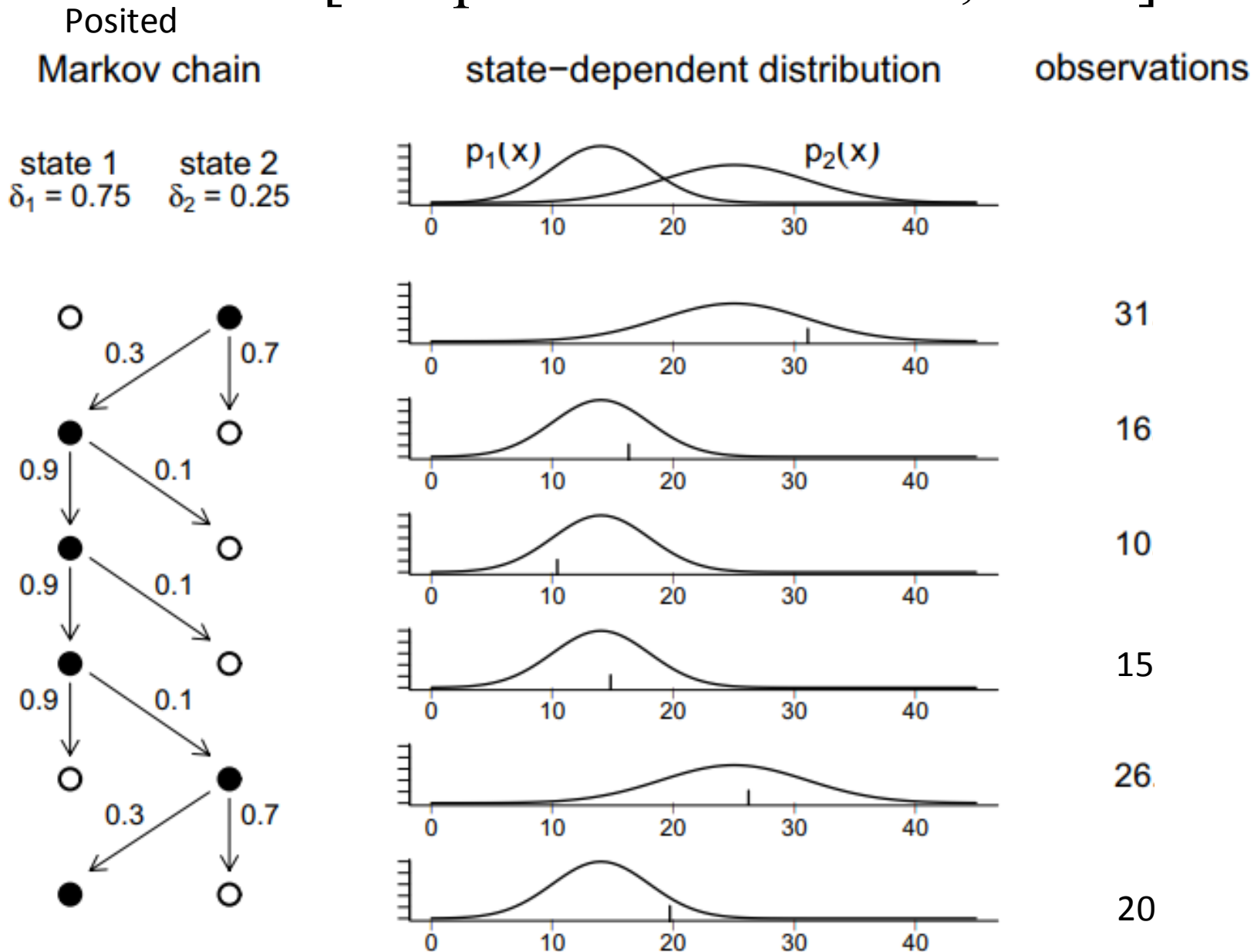
Our guessed characterization of underlying situation



Is there an outbreak in progress **here**?

Example

[Adapted from Zucchini, Ch. 2]



Example [from Zucchini, Ch. 2]

- How plausible is a given model in explaining the data?
- What is the most plausible model to explain the data?
- Given a plausible model, what is the likely underlying situation at a given point?

Key Algorithms

- Likelihood of observing sequence given HMM
- (Marginal) retrospective probability that was in each state for each time: **Forward-backwards algorithm**
 - This takes into account subsequent measurements
 - e.g., an observation shortly after t that suggests presence in a long-duration state $\Rightarrow$ supports inference that in that state at time t
- Single most likely sequence: **Viterbi Algorithm**

Modes of Training: **Supervised** learning

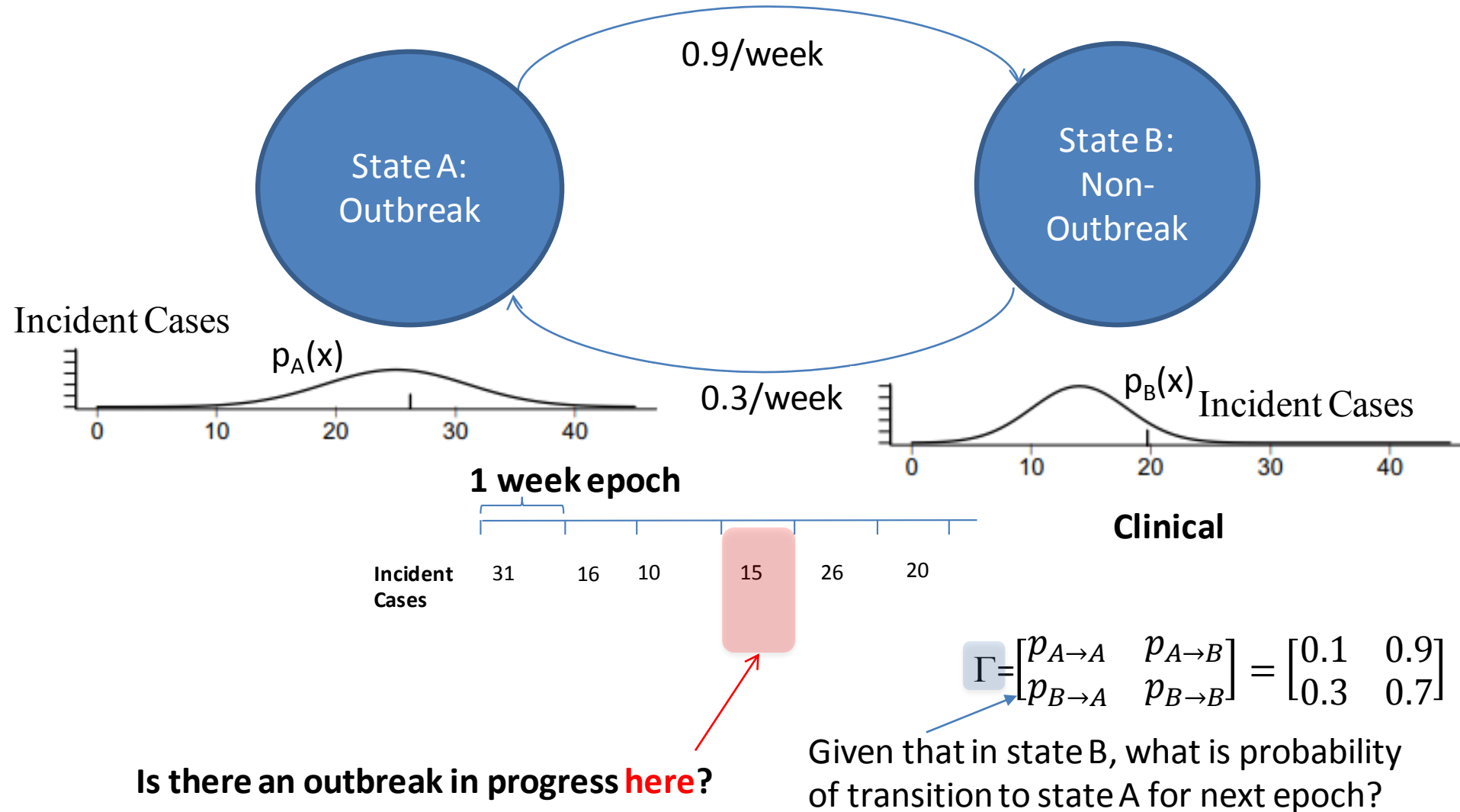
- Given:
 - Reliably labeled (“ground truthed”) data
 - Label classifying as to which state that in
 - Observations
 - HMM structure (e.g., count of states)
- Set HMM specifics based on labels together with observations, e.g.,
 - Compute transition probabilities given transitions observed in successive records
 - Fit or compute empirical probability distributions
- Then apply abduced HMM to other data

Modes of Training: **Unsupervised** learning

- Given: (Unlabeled) observations
- **Infer** HMM structure so as to best explain the patterns of the observed data
 - Cross-validation: Having trained on a subset of data (“training set”), we then evaluate on other data (“testing set”), and choose model that exhibits strongest such predictive powers
- Here, we are finding aspects of HMM structure that maximize some quantity
 - Likelihood of observing data (Maximum likelihood observation)
 - Entropy (Entropy maximization)
- Then apply abduced HMM to other data

Reminder Example [Zucchini, Ch. 2]

Our guessed characterization of underlying situation

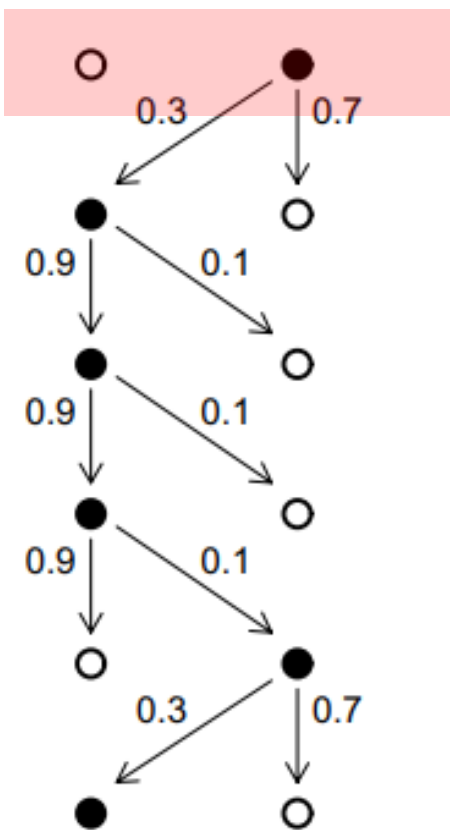


Reminder: Example

[Adapted from Zucchini, Ch. 2]

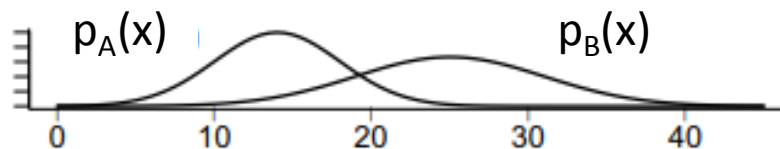
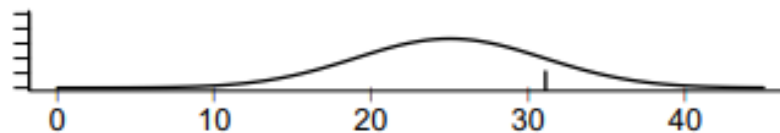
Posited
Markov chain

State A State B
 $\delta_1 = 0.75$ $\delta_2 = 0.25$

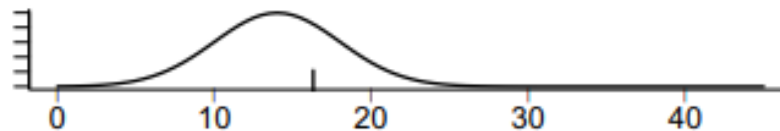


state-dependent distribution

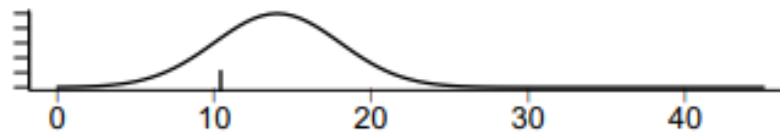
observations



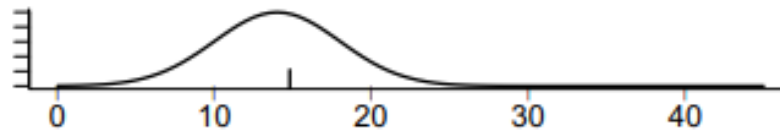
31



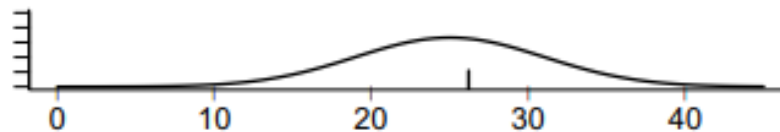
16



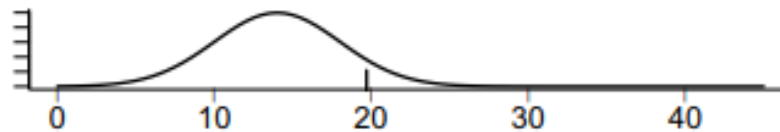
10



15



26

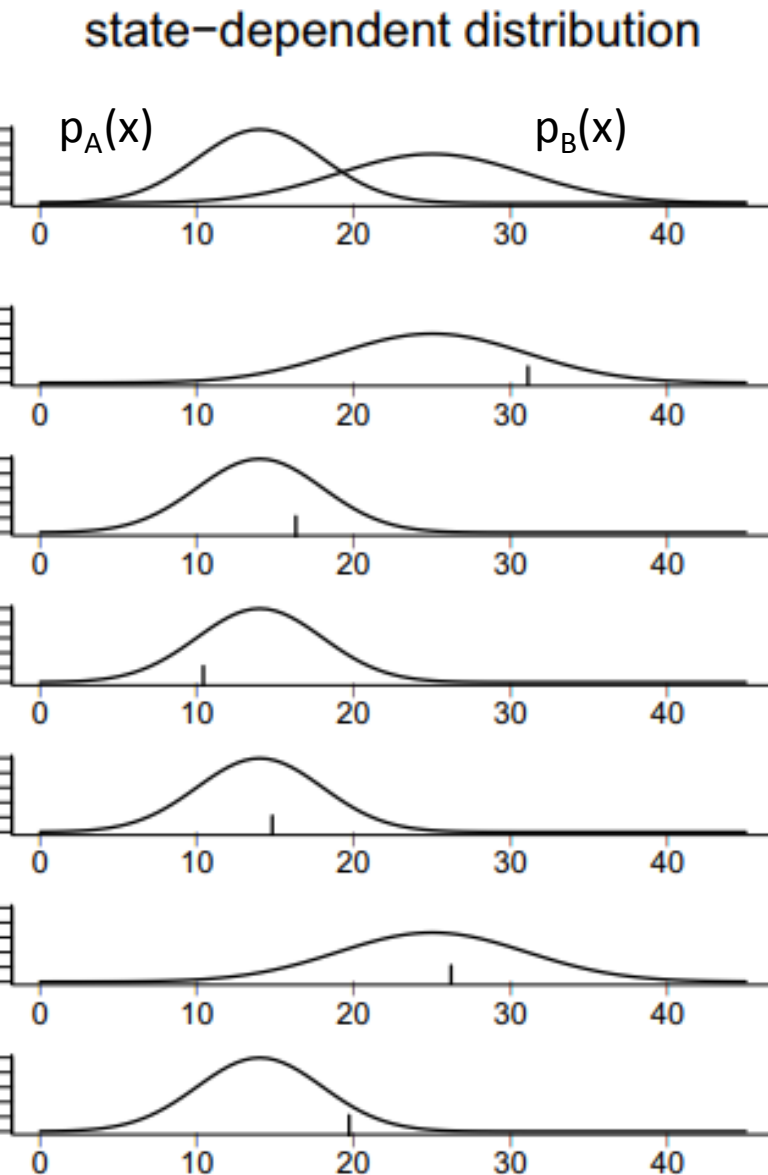


20

Example

[Adapted from Zucchini, Ch. 2]

Likelihood of
sequence
to this point
given in state



observations

$P(S)$

$$\begin{aligned}
 & \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} \\
 & \begin{bmatrix} 0.75 * P_{A \rightarrow A} + 0.25 * P_{B \rightarrow A} \\ 0.25 * P_{B \rightarrow B} + 0.75 * P_{A \rightarrow B} \end{bmatrix} \\
 & \alpha_1^A \rightarrow \begin{bmatrix} (0.75 * P_{A \rightarrow A} + 0.25 * P_{B \rightarrow A}) * P_A(31) \\ (0.25 * P_{B \rightarrow B} + 0.75 * P_{A \rightarrow B}) * P_B(31) \end{bmatrix} \\
 & \alpha_1^B \rightarrow \begin{bmatrix} (0.75 * P_{A \rightarrow A} + 0.25 * P_{B \rightarrow A}) * P_A(31) \\ (0.25 * P_{B \rightarrow B} + 0.75 * P_{A \rightarrow B}) * P_B(31) \end{bmatrix} \\
 & \alpha_2^A \rightarrow \begin{bmatrix} (\alpha_1^A * P_{A \rightarrow A} + \alpha_1^B * P_{B \rightarrow A}) * P_A(16) \\ (\alpha_1^B * P_{B \rightarrow B} + \alpha_1^A * P_{A \rightarrow B}) * P_B(16) \end{bmatrix} \\
 & \alpha_2^B \rightarrow \begin{bmatrix} (\alpha_1^A * P_{A \rightarrow A} + \alpha_1^B * P_{B \rightarrow A}) * P_A(16) \\ (\alpha_1^B * P_{B \rightarrow B} + \alpha_1^A * P_{A \rightarrow B}) * P_B(16) \end{bmatrix}
 \end{aligned}$$

π^T

$\pi^T \Gamma$

$\pi^T \Gamma P(x_1)$

$\pi^T \Gamma P(x_1) \Gamma P(x_2)$

10

15

26

20

Likelihood of Observing Sequence Given HMM

- Probability vector Probability of observing sequence $x_1 \dots x_T$ given that now in each state, often denoted α_T

$$\pi^T P(x_1) \Gamma P(x_2) \Gamma P(x_3) \dots \Gamma P(x_T) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} =$$

$$\pi P(x_1) \left(\prod_{i=2}^T \Gamma P(x_i) \right) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \alpha_T 1'$$

- π is vector of probability of being in different initial states
- $P(x_j)$ is a diagonal matrix with the probabilities of observing observation x_j in each state in successive diagonal elements
- Γ is the transition matrix that maps the current state to the next
- For a given T , this yields a value that constitutes the probability of having observed sequence $x_1 \dots x_T$ given the HMM

Forwards-Backwards Algorithm

- Probability vector α_t is the joint probability $P(x_1, x_2, \dots, x_t, S_t)$
 - This is termed the “forward probabilities” – relate to probability of being in a state in light of *earlier observations*

$$\alpha_t = \pi^T P(x_1) \Gamma P(x_2) \Gamma P(x_3) \dots \Gamma P(x_t) = \pi^T P(x_1) \left(\prod_{i=2}^t \Gamma P(x_i) \right)$$

- Probability vector β_t represents the conditional probability $P(x_{t+1}, x_{t+2}, \dots, x_T | S_t)$
 - This is termed the “backwards probabilities” – relate to probability of being in a state in light of *later observations*

$$\beta_t = \Gamma P(x_{t+1}) \Gamma P(x_{t+2}) \dots \Gamma P(x_T) 1' = \left(\prod_{i=t+1}^T \Gamma P(x_i) \right) 1'$$

- Marginal probability of being in state j at time t given all data before and after $= P(C_t = j | X^{(T)} = x^{(T)}) = \alpha_t * \beta_t / \text{likelihood of observing } x_1, x_2, \dots, x_T = \alpha_t * \beta_t / \alpha_T 1'$

Viterbi Algorithm

- Given HMM and observations, estimates **single most likely** sequence of (latent) states
- At a given time, the single most likely sequence may posit being in a state at that time that does not have the highest probabilities among all states
- Can be computed efficiently based on dynamic programming principles, based on two tables
 - $T_{\text{prob}}[i,j]$ of **probability** of most likely sequence so far between times 1 and j where posit that in state i at time j
 - $T_{\text{prev}}[i,j]$ of **identity of state time j-1** of most likely sequence so far between times 1 and j
- We take single most likely state from $T_{\text{prob}}[i,T]$ & work back

Focus of Training

- In different contexts, we will estimate different subsets of the HMM specification
- Most common: Transition matrix
- Common: Emission matrix, parameters of likelihood function

Caveats

- HMMs are generally not the right tool when there are
 - Continuous states
 - Memoryful states
- HMMs are at the boundary of systems science methods
 - Not non-linear $\Rightarrow$ We do not have a system where the “whole is greater than the sum of the parts”

Dynamic Models Add Further Power to HMMs

By Allowing For:

- Continuous states
- Continuous time
- Nonlinearity: Dependence of transition probabilities on state of the system
- ABMs:
 - Memoryful states
 - Richer Context (Networks, Geography, etc.)