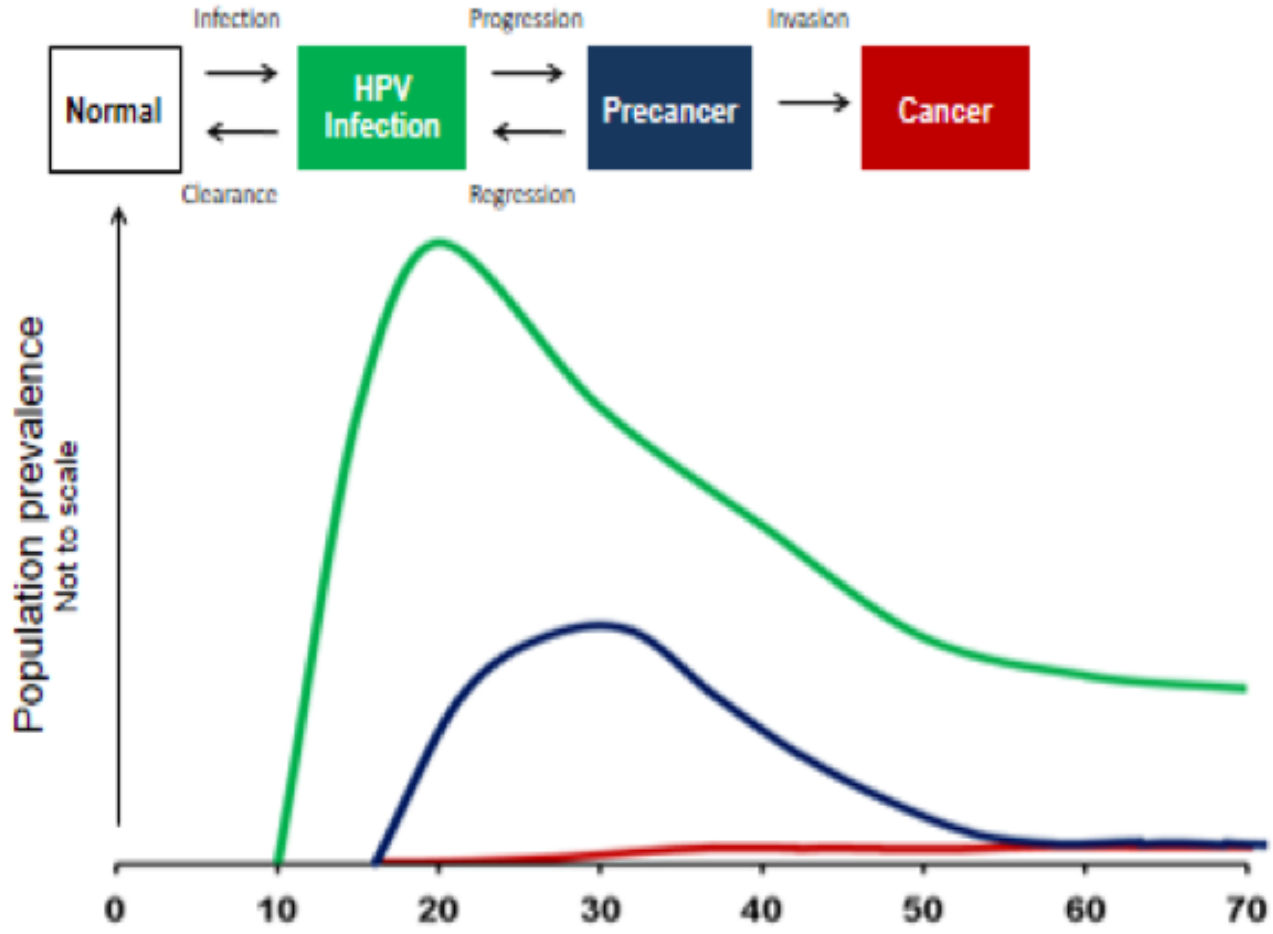
Modeling the Natural History of Cervical Cancer with Misclassification using Hidden Markov Models

Jordan Aron, Li Cheung, Hormuzd Katki, and Paul Albert

March 12, 2019

Big Picture

- HPV is closely related to cervical cancer



Big Picture

- Composite state space made up of three different test
- Incorporate misclassification (sensitivity and specificity)
- Subset of the population will always test negative
  - Monogomous relationships
  - Safe sex group

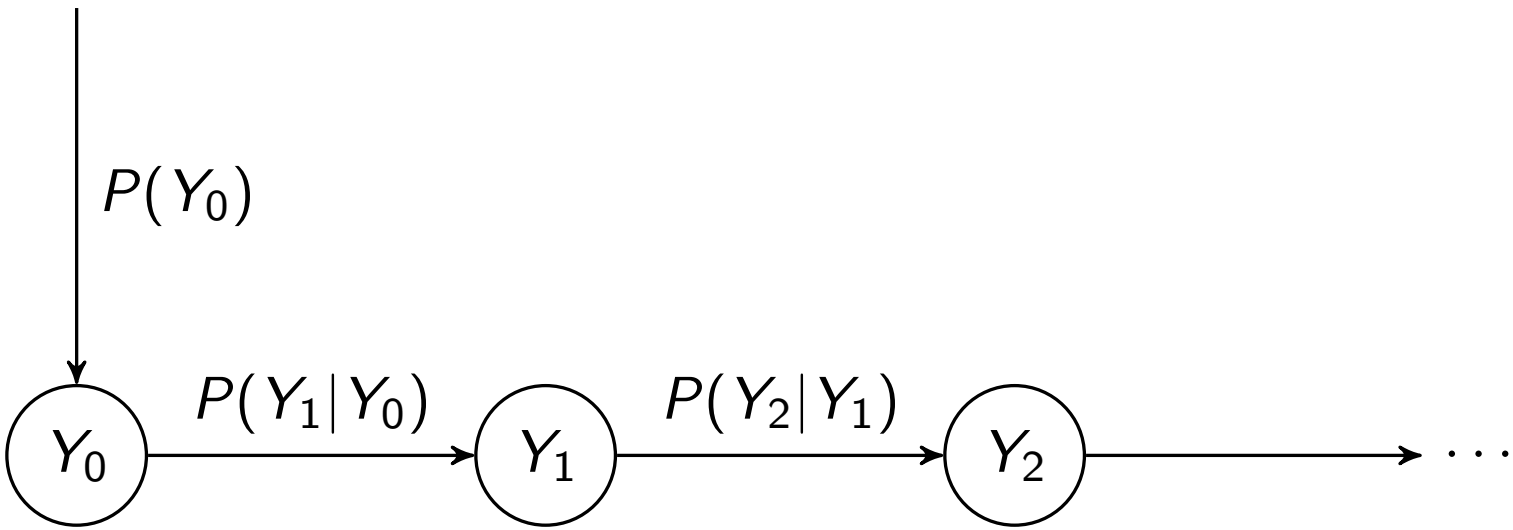
Method Overview

- First order hidden Markov model
- Mover-stayer component to allow for heterogeneous population
- EM algorithm

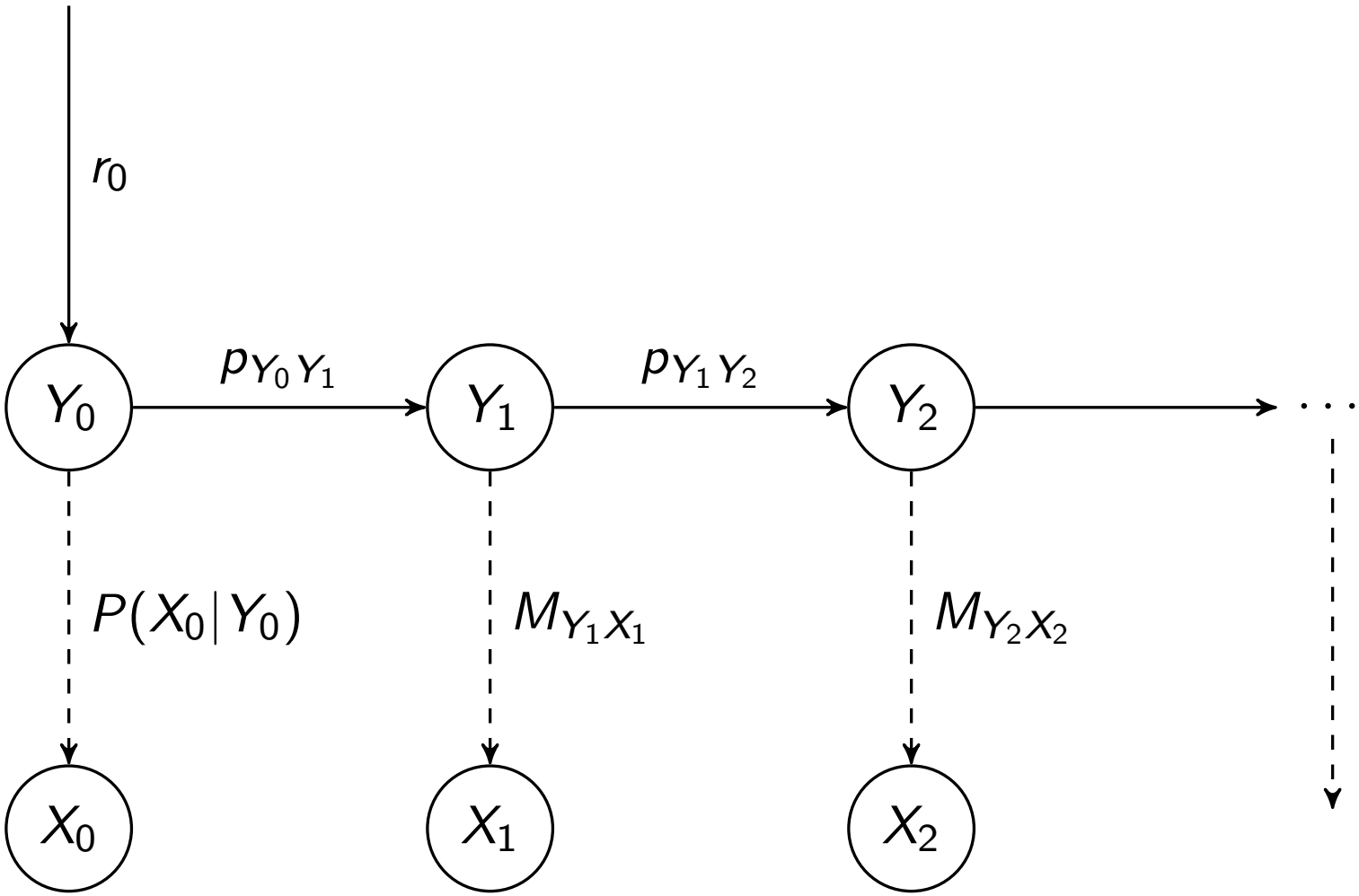
Data

- ALTS trial
  - Riskier women who tested for abnormal (ASCUS) cytology
  - Enrollment and follow visits up every 6 months for 2 years
- Kaiser Permanente northern California EMR data
  - Normal clinical practice
  - Differing intervals
- Three tests
  - HPV
  - Pap Smear (Cytology)
  - Punch biopsy (Histology)

Markov Chain



Hidden Markov Model



Notation

- $\mathbf{Y}_i = \{Y_{i1}, Y_{i2}, ..., Y_{in}\}$
  - First order Markov chain of latent states
  - Fully described by $P(Y_{i1})$ and $P(Y_{it}|Y_{it-1})$
- $\mathbf{X}_i = \{X_{i1}, X_{i2}, ..., X_{in}\}$
  - Vector of observable states
  - $P(X_{it}|\mathbf{Y}_i) = P(X_{it}|Y_{it})$ - Conditional independence assumption
- $r_g = P(Y_{i1} = g)$
- $p_{g\omega} = P(Y_{it} = \omega|Y_{it-1} = g)$
- $M_{g\omega} = P(X_{it} = \omega|Y_{it} = g)$

State Space

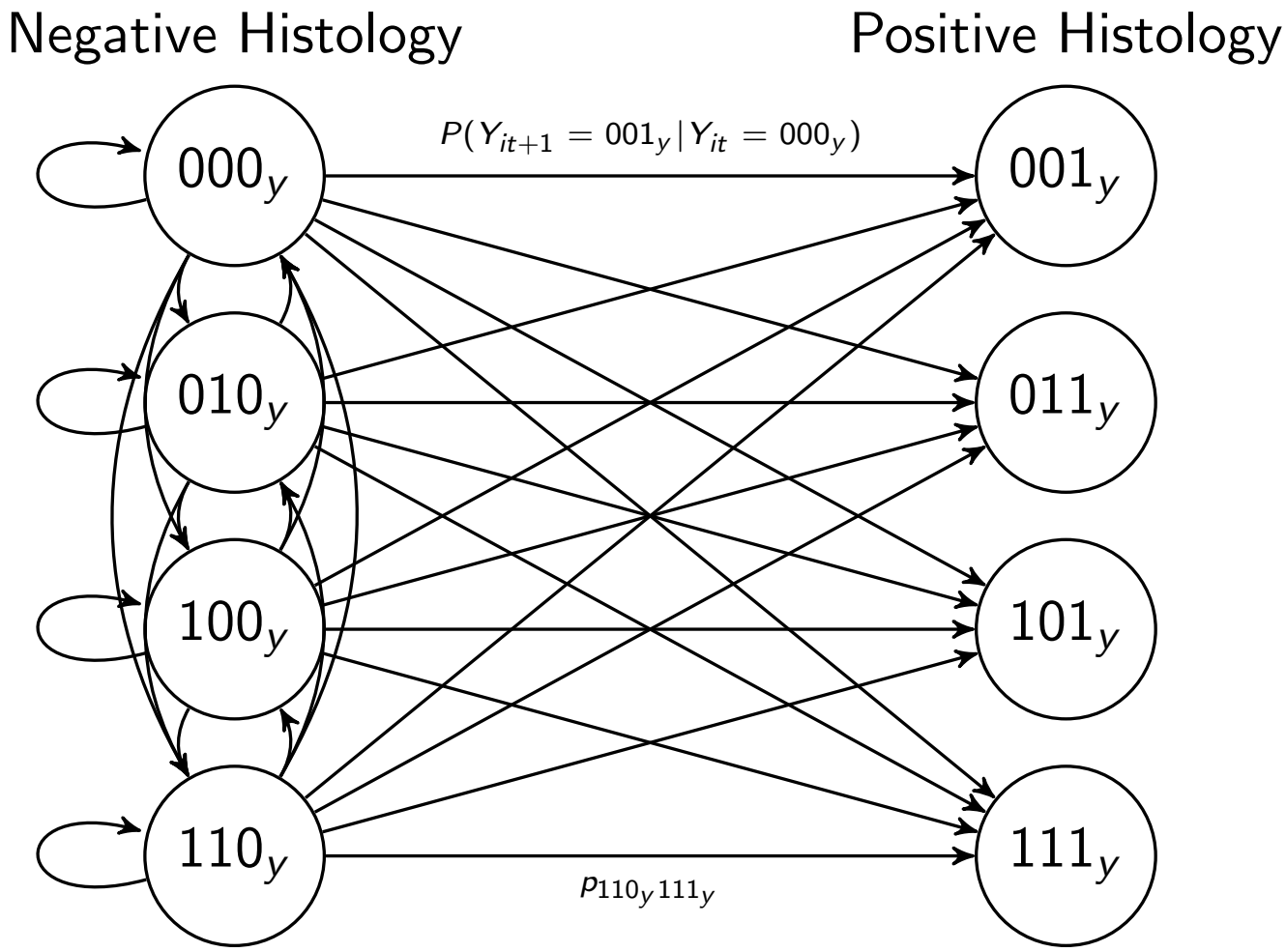
$$\text{HPV} = \begin{cases} 0 & \text{if negative} \\ 1 & \text{if positive} \end{cases}$$

$$\text{Cytology} = \begin{cases} 0 & \text{if normal} \\ 1 & \text{otherwise} \end{cases}$$

$$\text{Histology} = \begin{cases} 1 & \text{if precancer or cancer} \\ 0 & \text{otherwise} \end{cases}$$

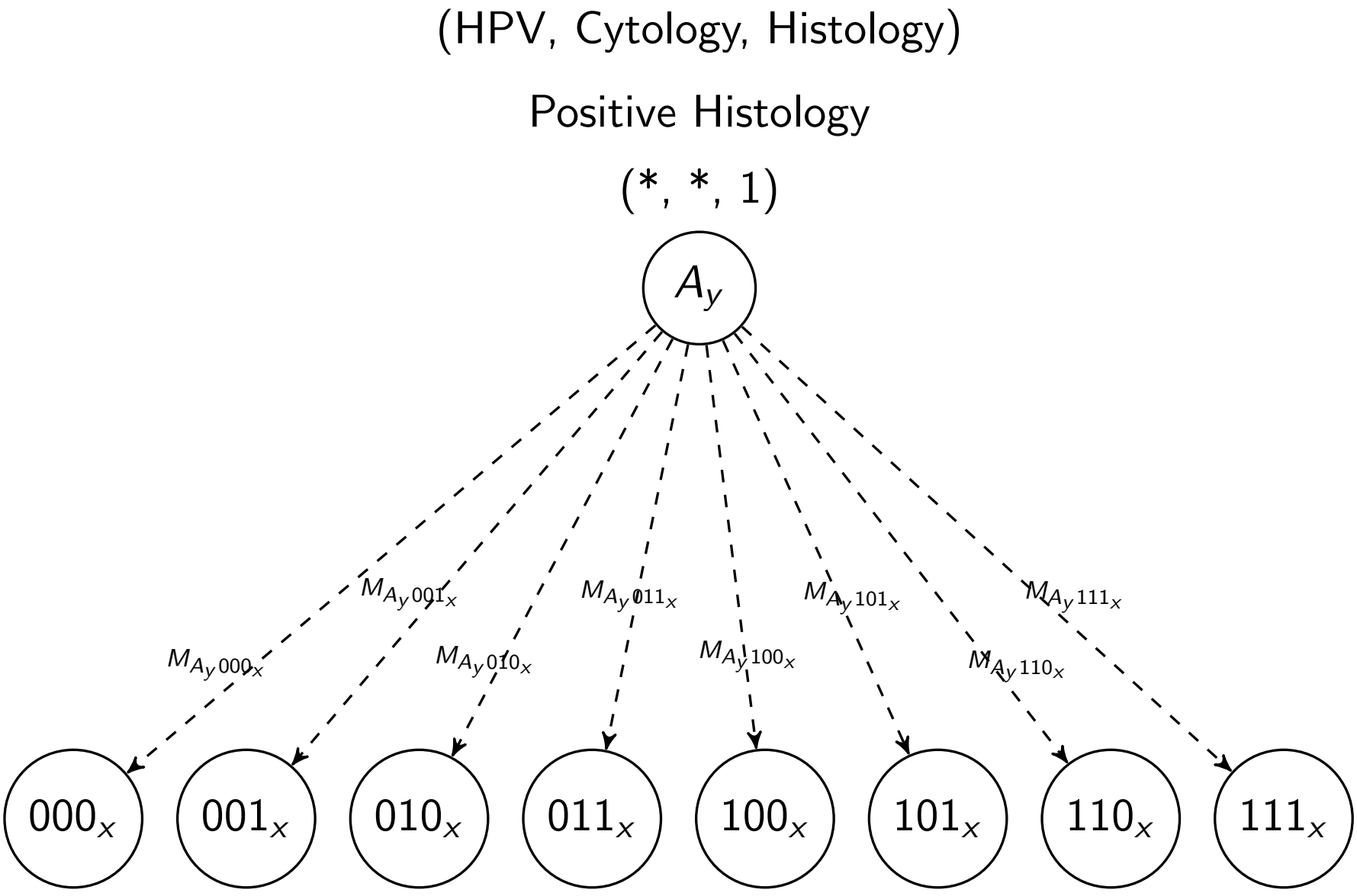
- State space is size $2 * 2 * 2 = 8$
- Each state is a tuple of (HPV, Cytology, Histology)
- The state space for $\mathbf{X}$ and $\mathbf{Y}$ are the same

Hidden Transition Structure

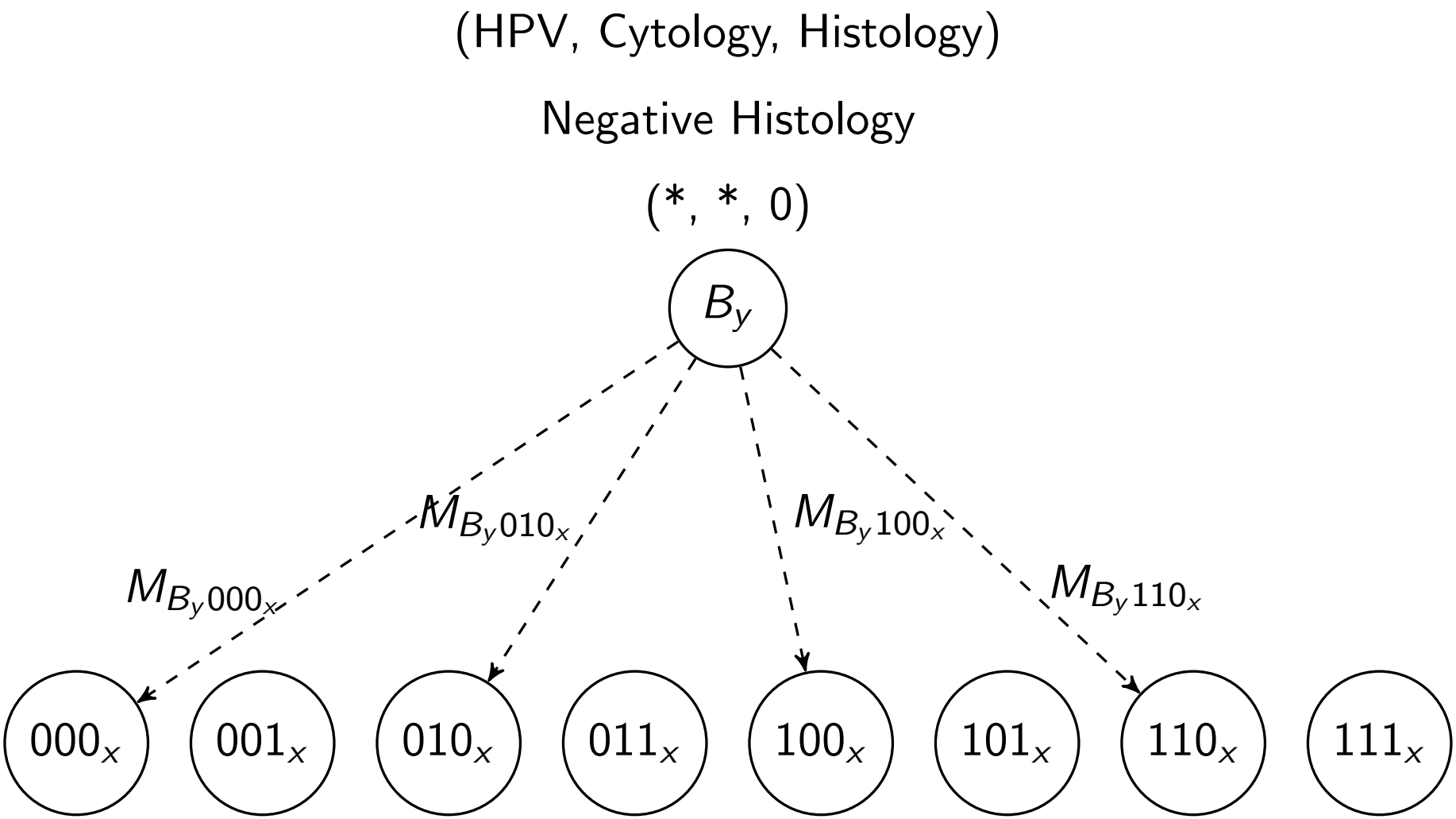


(HPV, Cytology, Histology)

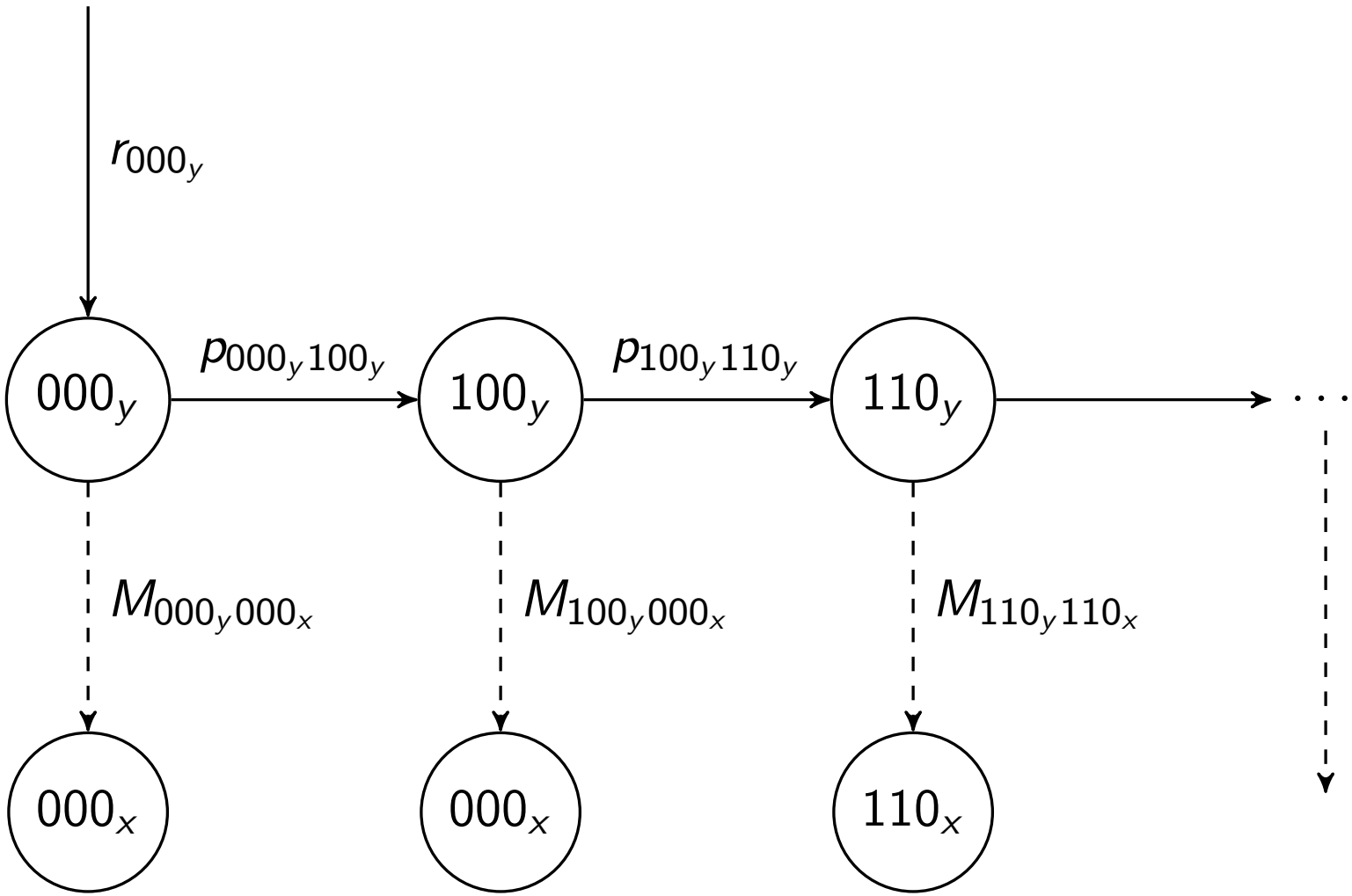
Classification Structure



Classification Structure



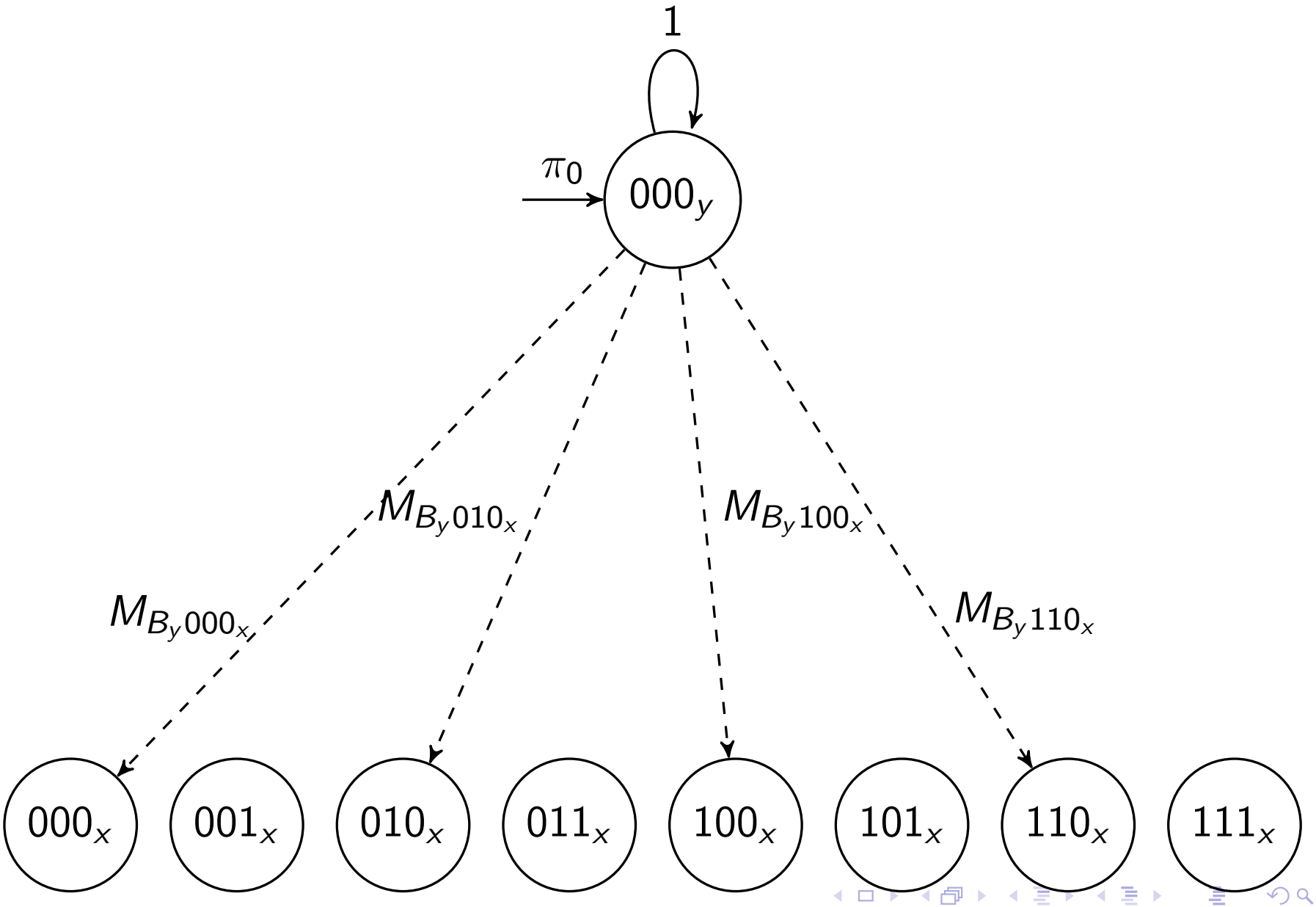
Hidden Markov Model



Mover Stayer

- Used to deal with heterogeneous populations
- Let Q_i be an indicator for whether individual i is a stayer
- Each individual has probability $\pi_0 = P(Q_i = 1)$ of being a stayer
  - Initial and all subsequent latent states are 000_y
  - Can still be misclassified

Stayer Model



HPV Persistence

- HPV persistence is a large factor
- Binary HPV states in a first order Markov chain cannot deal with persistence
- Expand the state space to deal with HPV persistence

$$\text{HPV} = \begin{cases} 0 & \text{if negative} \\ 1 & \text{if positive for one year} \\ 2 & \text{if positive for two years} \\ 3 & \text{if positive for three or more years} \end{cases}$$

Notation

- $\mathbf{X}_i = \{X_{i1}, ..., X_{in}\}$ = Vector of observable states
- $\mathbf{Y}_i = \{Y_{i1}, ..., Y_{in}\}$ = Vector of latent states
- Q_i is an indicator for if individual i is a stayer
- $r_g = P(Y_{i1} = g)$
- $p_{g\omega} = P(Y_{it} = \omega | Y_{it-1} = g)$
- $M_{g\omega} = P(X_{it} = \omega | Y_{it} = g)$
- $\pi_0 = P(Q_i = 1)$
- Where $Z_g(Y_{it}) = \begin{cases} 1 & \text{if } Y_{it} = g \\ 0 & \text{otherwise} \end{cases}$

Joint Distribution

$$
\begin{aligned}
& f_i(\mathbf{X}_i, \mathbf{Y}_i, Q_i) \\
& = \left[\pi_0 \prod_{t=1}^{n} P(X_{it} | Y_{it} = 000_y) \right]^{Q_i} && \text{Stayer} \\
& * \left[(1 - \pi_0) \prod_{g=000_y}^{k} P(Y_{i1} = g)^{Z_g(Y_{i1})} \right. && \text{Initial} \\
& * \prod_{t=1}^{n} \prod_{g=000_y}^{k} \prod_{\omega=000_y}^{k} P(Y_{it} = \omega | Y_{it-1} = g)^{Z_g(Y_{it-1}) Z_\omega(Y_{it})} && \text{Transition} \\
& * \left. \prod_{t=1}^{n} \prod_{g=000_y}^{k} P(X_{it} | Y_{it} = g)^{Z_g(Y_{it})} \right]^{1-Q_i} && \text{Classification}
\end{aligned}
$$

Likelihood Calculation

$$L(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) = \prod_{i=1}^{n} L(\mathbf{X}_i)$$

$$= \prod_{i=1}^{n} \sum_{l_1=000_y}^{k} \sum_{l_2=000_y}^{k} \cdots \sum_{l_n=000_y}^{k} \sum_{q=0}^{1} f_i(\mathbf{X}_i, (l_1, ..., l_n), q)$$

EM Algorithm

$$l^{(p)}(\Theta) = E[\log\ f_i(\mathbf{X}_i, \mathbf{Y}_i, Q_i)|\mathbf{X}_i] \qquad \text{E-Step}$$

$$\Theta^{(p+1)} = \max_{\Theta} l^{(p)}(\Theta) \qquad \text{M-Step}$$

- Where Θ is the vector of parameters (initial, transition, and classification probabilities)
- Iterate for $p = 1, 2, \ldots$ until convergence

Complete Data Log Likelihood

$$
\begin{aligned}
& E[\log\, f_i(\mathbf{X}_i, \mathbf{Y}_i, Q_i)|\mathbf{X}_i] \\
& = E[Q_i|\mathbf{X}_i](\log\, \pi_0 + \sum_{t=1}^{n} \log\, P(X_{it}|Y_{it} = 000_y)) && \text{Stayer} \\
& + (1 - E[Q_i|\mathbf{X}_i]) \Big[\log\, (1 - \pi_0) && \text{Mover} \\
& + \sum_{g=000_y}^{k} E[Z_g(Y_{i1})|\mathbf{X}_i] \log\, r_g && \text{Initial} \\
& + \sum_{t=2}^{n} \sum_{g=000_y}^{k} \sum_{\omega=000_y}^{k} E[Z_g(Y_{it-1}) Z_\omega(Y_{it})|\mathbf{X}_i] \log\, p_{g\omega} && \text{Transition} \\
& + \sum_{t=1}^{n} \sum_{g=000_y}^{k} \sum_{\omega=000_x}^{k} E[Z_g(Y_{it})|\mathbf{X}_i] Z_\omega(X_{it}) \log\, M_{gX_{it}} \Big] && \text{Classification}
\end{aligned}
$$

Forward Algorithm

$$\alpha_g(it) = \begin{cases} r_g M_{gx_{i1}} & \text{if } t = 1 \\ \sum_{\omega=1}^{k} \alpha_\omega(i, t-1) p_{\omega g} M_{gx_{it}} & \text{if } t > 1 \end{cases}$$

- $\alpha_g(it) = P(X_{i1}, X_{i2}, ..., X_{it}, Y_{it} = g | Q_i = 0)$
- Calculates the probability of being in latent state g at time t and the observations up to time t
- Need $\alpha(it-1)$ to calculate $\alpha_g(it)$

Backward Algorithm

$$\beta_g(it) = \begin{cases} 1 & \text{if } t = n \\ \sum_{\omega=0}^{k} p_{g\omega} M_{\omega x_{it+1}} \beta_\omega(it+1) & \text{if } t < n \end{cases}$$

- $\beta_g(it) = P(X_{it+1}, ..., X_{in} | Y_{it} = g, Q_i = 0)$
- Calculates the probability of the observations from time t+1 to n given being in latent state g at time t
- Need $\beta(it+1)$ to calculate $\beta_g(it)$

Forward-Backward Algorithm

- $\alpha_g(it) = P(X_{i1}, X_{i2}, ..., X_{it}, Y_{it} = g | Q_i = 0)$
- $\beta_g(it) = P(X_{it+1}, X_{it+2}, ..., X_{in} | Y_{it} = g, Q_i = 0)$
- $\alpha_g(it)$ and $\beta_g(it)$ are calculated independently

$$\begin{aligned}
&\alpha_g(it)\beta_g(it) \\
&= P(X_{i1}, ..., X_{it}, Y_{it} = g | Q_i = 0) * P(X_{it+1}, ..., X_{in} | Y_{it} = g, Q_i = 0) \\
&= P(\mathbf{X}_i, Y_{it} = g | Q_i = 0)
\end{aligned}$$

Calculating the Likelihood

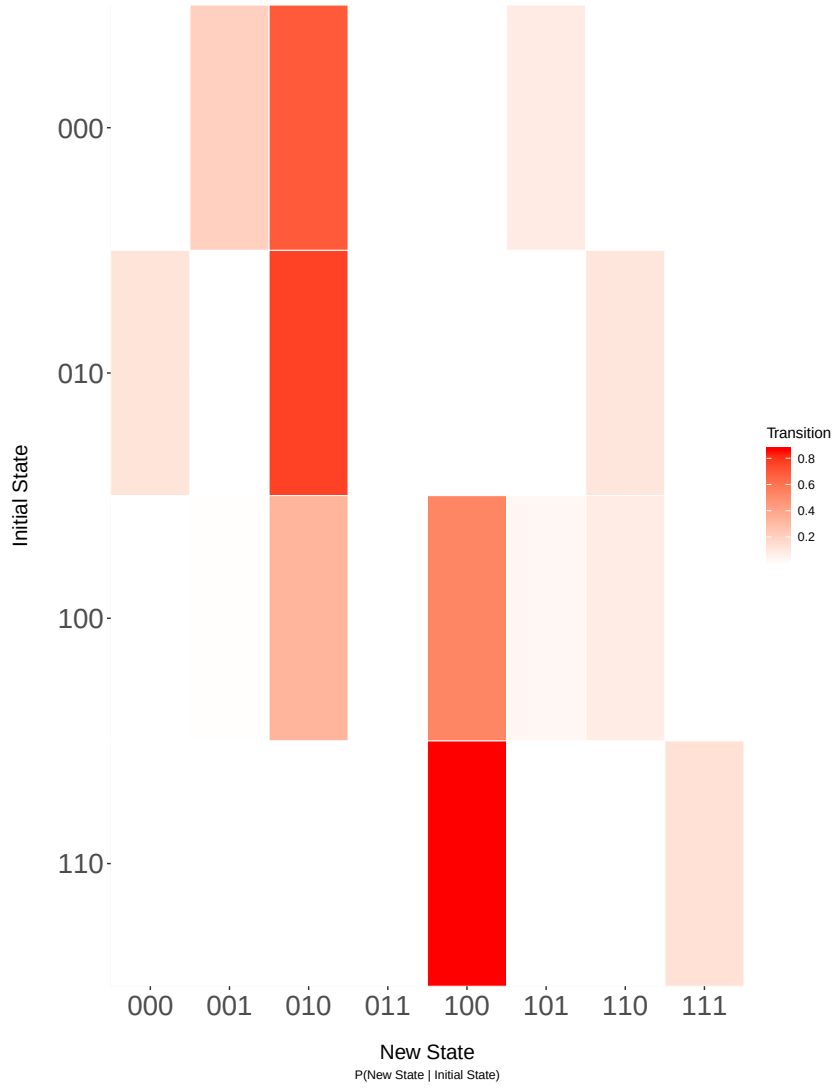
- $P(\mathbf{X}_i, Q_i = 0) = (1 - \pi_0) \sum_{g=000_y}^{k} \alpha_g(it)\beta_g(it)$
- $P(\mathbf{X}_i, Q_i = 1) = \pi_0 \prod_{t=1}^{n} P(X_{it}|Y_{it} = 000_y)$
- $P(\mathbf{X}_i) = P(\mathbf{X}_i, Q_i = 0) + P(\mathbf{X}_i, Q_i = 1)$

M-Step Calculations

- $E[Q_i|\mathbf{X}_i]$
- $E[Z_g(Y_{i1})|\mathbf{X}_i]$
- $E[Z_g(Y_{it-1})Z_\omega(Y_{it})|\mathbf{X}_i]$
- $E[Z_g(Y_{it})|\mathbf{X}_i]$

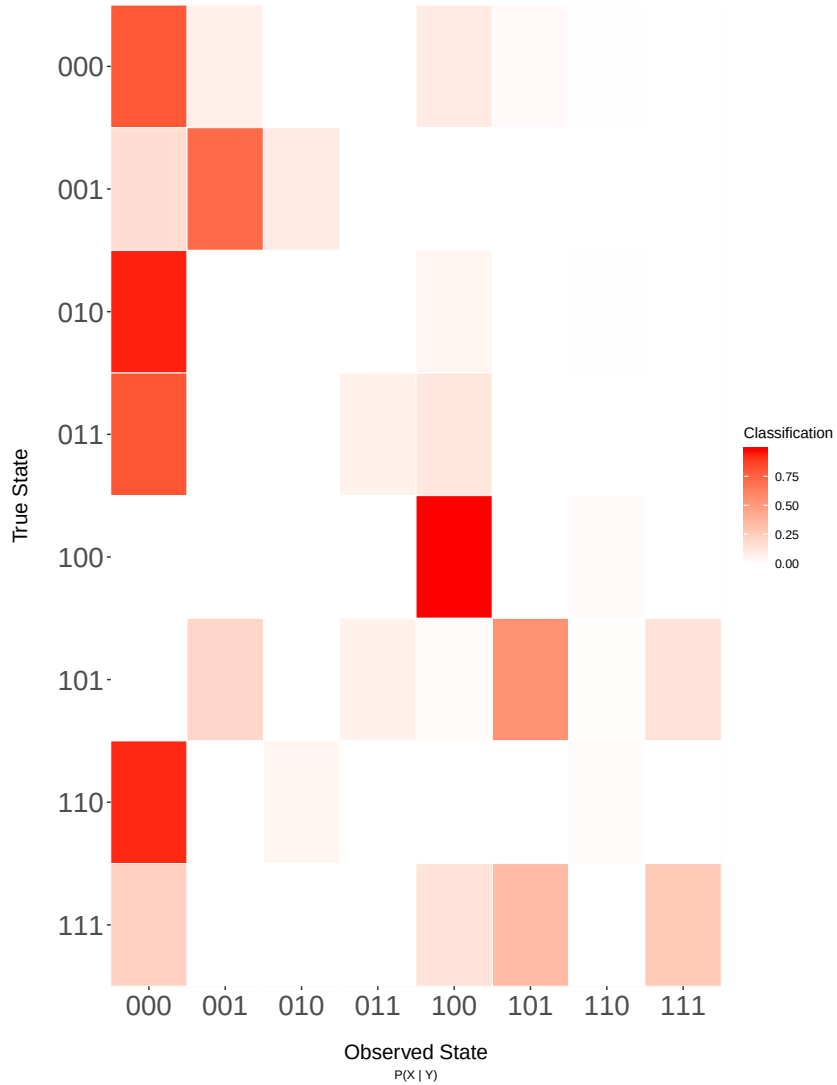
All can be calculated using the forward-backward algorithm

Transition Results



(HPV, Cytology, Histology)

Classification Results



(HPV, Cytology, Histology)

Partially Missing Data

- What happens when an individual is missing a singular test?
- $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{in})$
- $X^*_{it} = (X_{it_1}, X_{it_2}, ..., X_{it_m})$
  - Universe of all possible full observations
  - m is the maximum number of partial observations

		(HPV	Cytology	Histology)
Fully Observed :	X_{in}	=(0	1	1)
Partially Observed :	X^*_{it}	=(0	?	1)
	X_{it_1}	=(0	0	0)
	X_{it_2}	=(0	1	0)

Partially Missing Data

- $\mathbf{W}_i = (W_{i1}, ..., W_{it}, ..., W_{in})$
- $W_{it} = (W_{it_1}, W_{it_2}, ..., W_{it_m})$
  - Indicator variable for observed state if fully observed
- $W_{it_j} =$

$$\begin{cases} 1 & \text{if } X_{it_j} \text{ would be the fully observed state if } X_{it} \text{ was fully observed} \\ 0 & \text{Otherwise} \end{cases}$$

Joint Distribution with Partial Data

$$f_i(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{W}_i, Q_i)$$

$$= \left[\pi_0 \prod_{t=1}^{n} \prod_{j=1}^{m} P(X_{it_j} | Y_{it} = 000_y)^{W_{it_j}} \right]^{Q_i} \quad \text{Stayer}$$

$$* \left[(1 - \pi_0) \prod_{g=000_y}^{k} P(Y_{i1} = g)^{Z_g(Y_{i1})} \right. \quad \text{Initial}$$

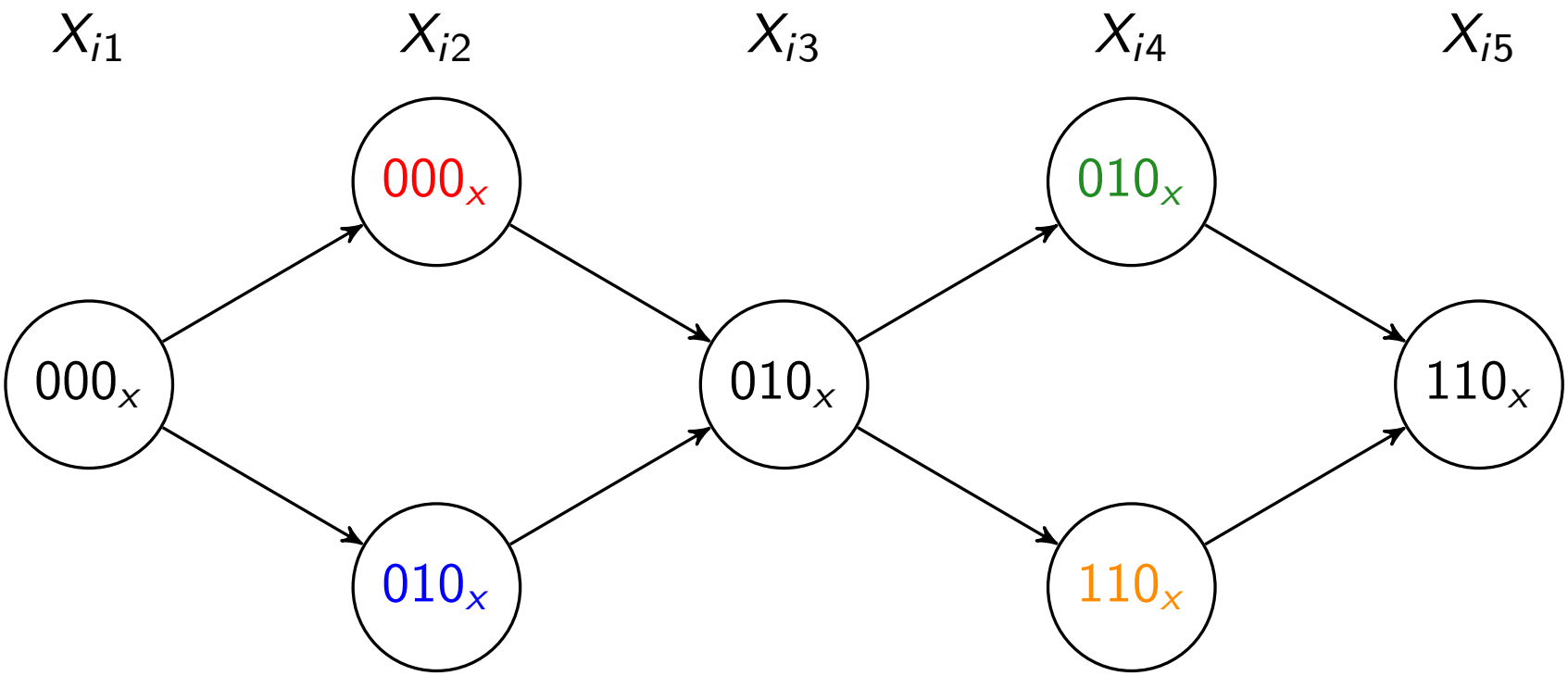
$$* \prod_{t=1}^{n} \prod_{g=000_y}^{k} \prod_{\omega=000_y}^{k} P(Y_{it} = \omega | Y_{it-1} = g)^{Z_g(Y_{it-1}) Z_\omega(Y_{it})} \quad \text{Transition}$$

$$\left. * \prod_{t=1}^{n} \prod_{g=000_y}^{k} \prod_{j=1}^{m} P(X_{it_j} | Y_{it} = g)^{Z_g(Y_{it}) W_{it_j}} \right]^{1-Q_i} \quad \text{Classification}$$

Complete Data Log Likelihood with Partial Data

$$
\begin{aligned}
& E[\log\, f_i(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{W}_i, Q_i)|\mathbf{X}_i] \\
& = E[Q_i|\mathbf{X}_i]\left[\log\, \pi_0 + \sum_{t=}^{n}\sum_{j=1}^{m} E[W_{it_j}|\mathbf{X}_i]\log\, M_{000_y X_{it_j}}\right] && \text{Stayer} \\
& + (1 - E[Q_i|\mathbf{X}_i])\left[\log(1-\pi_0)\right. && \text{Mover} \\
& + \sum_{g=000_y}^{k} E[Z_g(Y_{i1})|\mathbf{X}_i]\log\, r_g && \text{Initial} \\
& + \sum_{t=2}^{n}\sum_{g=000_y}^{k}\sum_{\omega=000_y}^{k} E[Z_g(Y_{it-1})Z_\omega(Y_{it})|\mathbf{X}_i]\log\, p_{g\omega} && \text{Transition} \\
& + \left.\sum_{t=1}^{n}\sum_{g=000_y}^{k}\sum_{\omega=000_x}^{k}\sum_{j=1}^{m} E[Z_g(Y_{it})W_{it_j}|\mathbf{X}_i]Z_\omega(X_{it})\log\, M_{gX_{it}}\right] && \text{Classification}
\end{aligned}
$$

Different Paths



$(000_x, 000_x, 010_x, 010_x, 110_x)$
$(000_x, 000_x, 010_x, 110_x, 110_x)$
$(000_x, 010_x, 010_x, 010_x, 110_x)$
$(000_x, 010_x, 010_x, 110_x, 110_x)$
(HPV, Cytology, Histology)

Forward Algorithm with Partial Data

$$\alpha_g(it)_{A_t} = \begin{cases} r_g M_g X_{i1_j} & \text{if } t = 1 \\ \sum_{\omega=1}^{k} \alpha_\omega(i, t-1)_{A_{t-1}} p_{\omega g} M_g X_{it_j} & \text{if } t \neq 1 \end{cases}$$

- $A_t = \{A_{t-1}, X_{it_j}\}$ and $A_0 = \{\}$
- $\alpha_g(it)_{A_t} = P(X_{i1} \in A_t, X_{i2} \in A_t, ..., X_{it} \in A_t, Y_{it} = g | Q_i = 0)$
- $\alpha_g(it) = \sum_{A_t} \alpha_g(it)_{A_t} = P(X_{i1}, X_{i2}, ..., X_{it}, Y_{it} = g | Q_i = 0)$

Backward Algorithm with Partial Data

$$\beta_g(it)_{B_t} = \begin{cases} 1 & \text{if } t = n \\ \sum_{\omega=0}^{k} p_{g\omega} M_{\omega X_{it+1_j}} \beta_\omega(t+1)_{B_{t+1}} & \text{if } t < n \end{cases}$$

- $B_t = \{X_{it+1_j}, B_{t+1}\}$ and $B_n = \{\}$
- $\beta_g(it)_{B_t} = P(X_{it+1} \in B_t, ..., X_{in} \in B_t | Y_{it} = g, Q_i = 0)$
- $\beta_g(it) = \sum_{B_t} \beta_g(it)_{B_t} = P(X_{it+1}, ..., X_{in} | Y_{it} = g, Q_i = 0)$

Next Steps

- Methods paper dealing with partial data
- Real life clinical application