# Machine Learning Methods for Neural Data Analysis

## More Hidden Markov Models

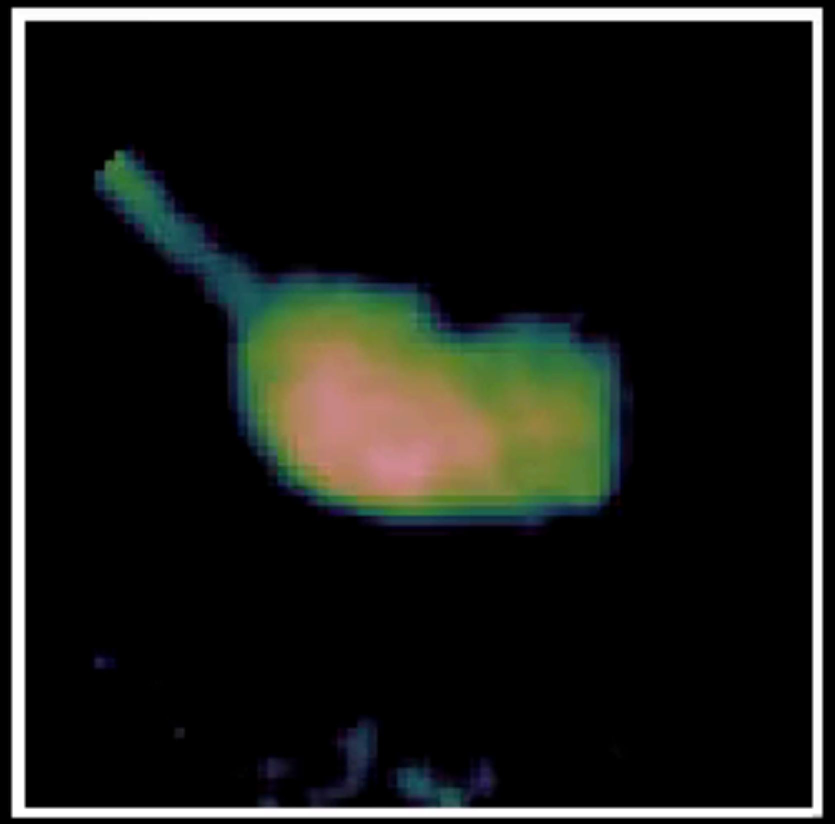
Scott Linderman

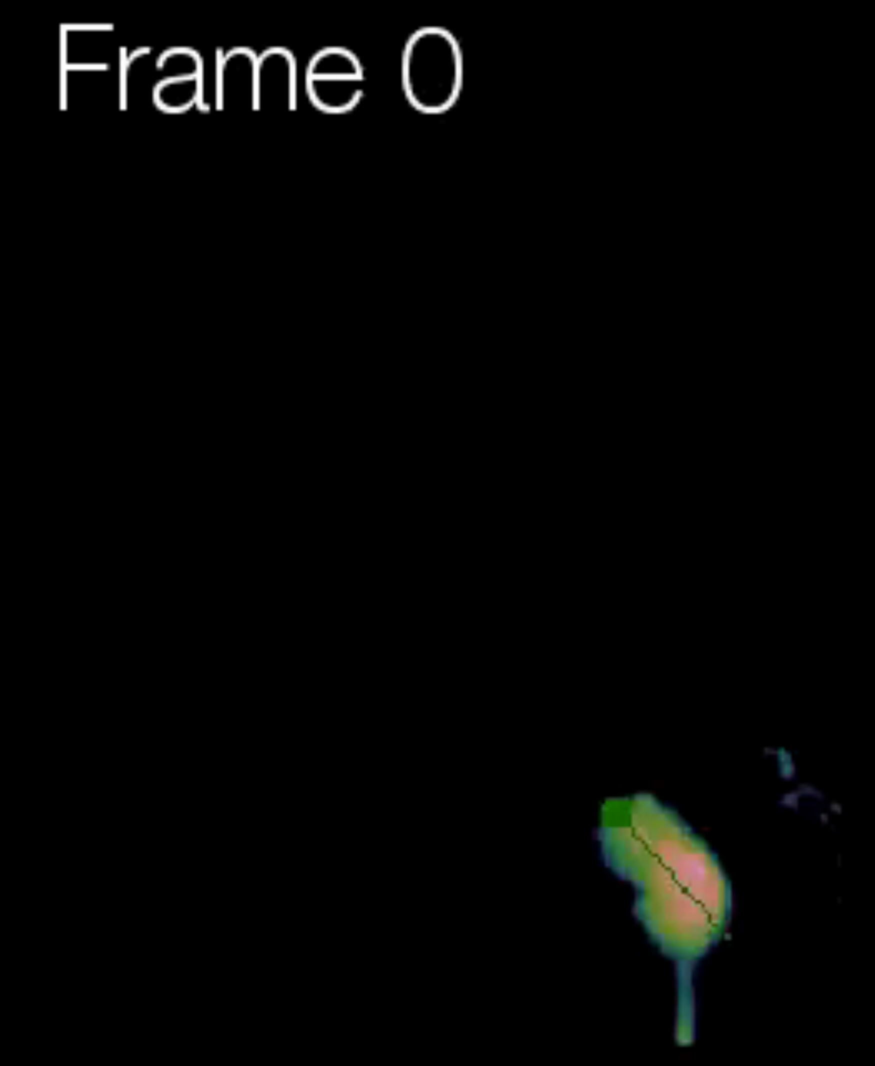
*STATS 220/320 (NBIO220, CS339N). Winter 2023.*

# Announcements

- Lab 6
  - Add `atol=1e-4` to the `allclose` checks.
- Please submit your 1 page **project proposal** tonight.

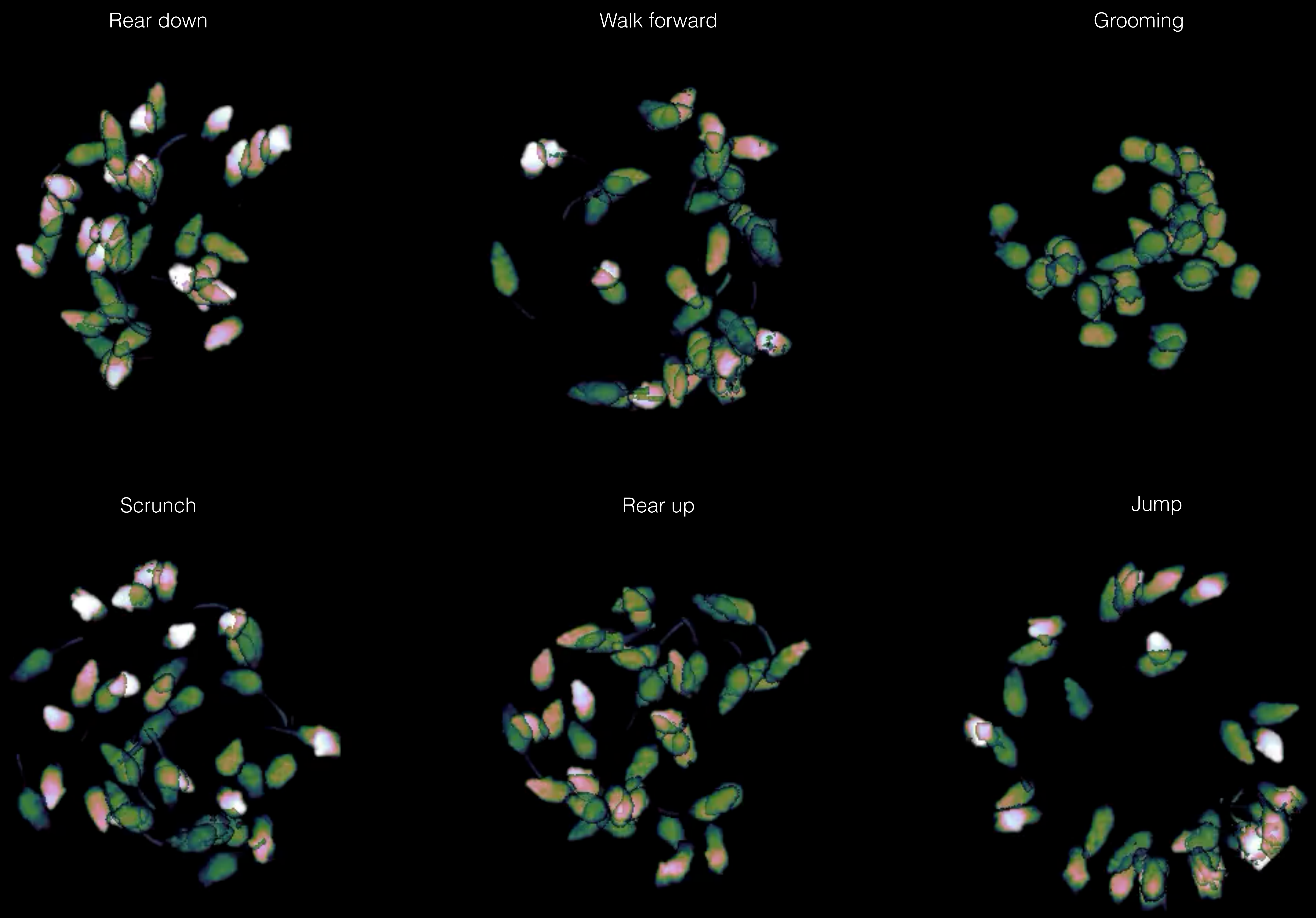
# Motivating Example: summarizing videos with behavioral states






*Wiltschko et al, 2015*

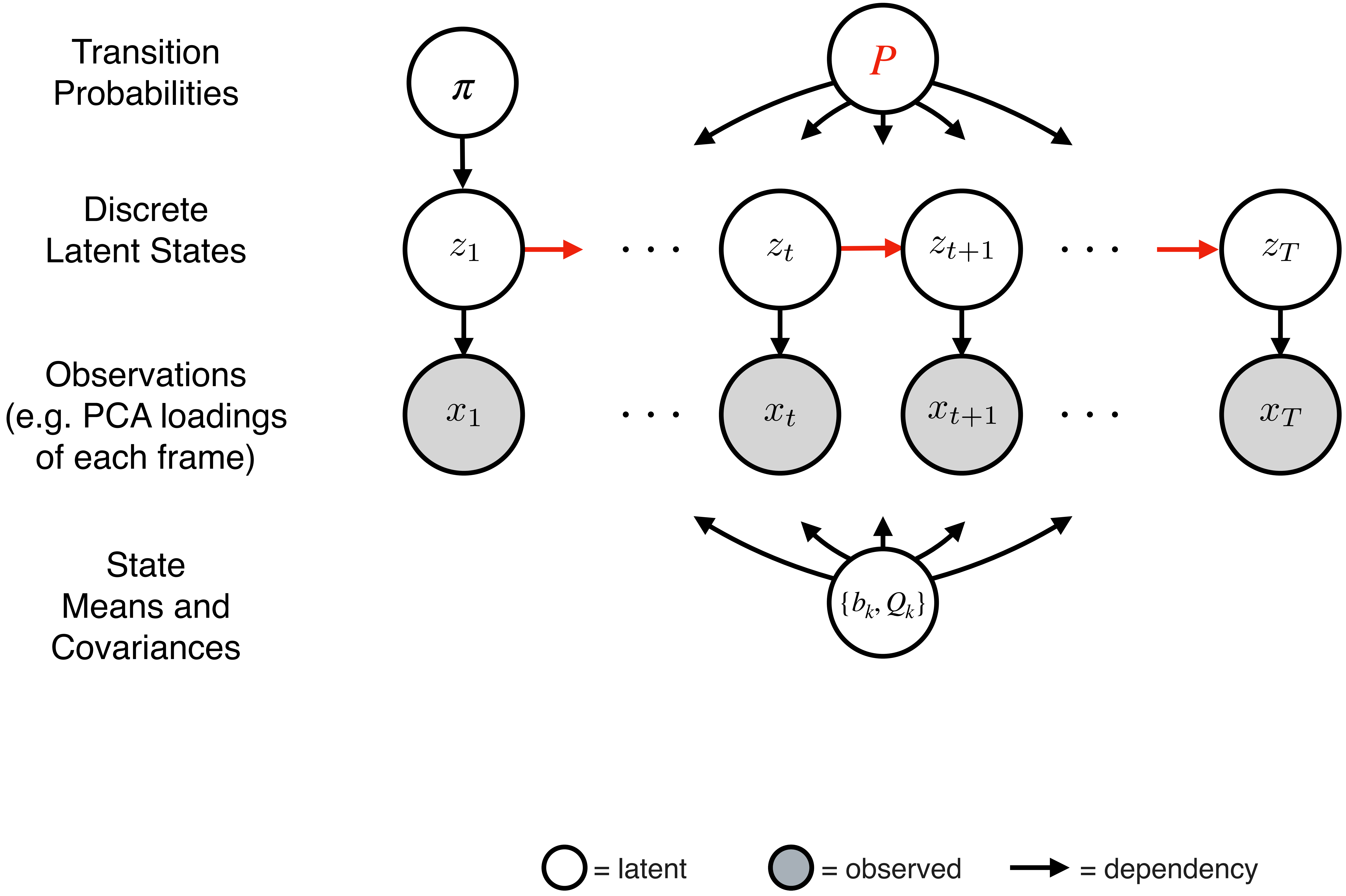
# Motivating Example: summarizing videos with behavioral states




*Wiltschko et al, 2015*

# The Gaussian HMM

## Graphical Model



# The Gaussian HMM

A Gaussian HMM is just a Gaussian mixture model but where cluster assignments are linked across time!

$$
\begin{aligned}
z_1 &\sim \mathrm{Cat}(\pi), \\
z_t \mid z_{t-1} &\sim \mathrm{Cat}(P_{z_{t-1}}), && \text{for } t = 2,\ldots,T. \\
x_t \mid z_t &\sim \mathcal{N}(b_{z_t}, Q_{z_t}) && \text{for } t = 1,\ldots,T
\end{aligned}
$$

Its parameters are $\Theta = \pi, P, \{b_k, Q_k\}_{k=1}^K$ where $P \in [0,1]^{K\times K}$ is a row-stochastic **transition matrix.**

Under this model, the **joint probability** factors as

$$
p(x, z, \Theta) = p(z_1)\prod_{t=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=1}^{T} p(x_t \mid z_t)
$$

# Bayesian inference in latent variable models

## The Expectation-Maximization (EM) algorithm

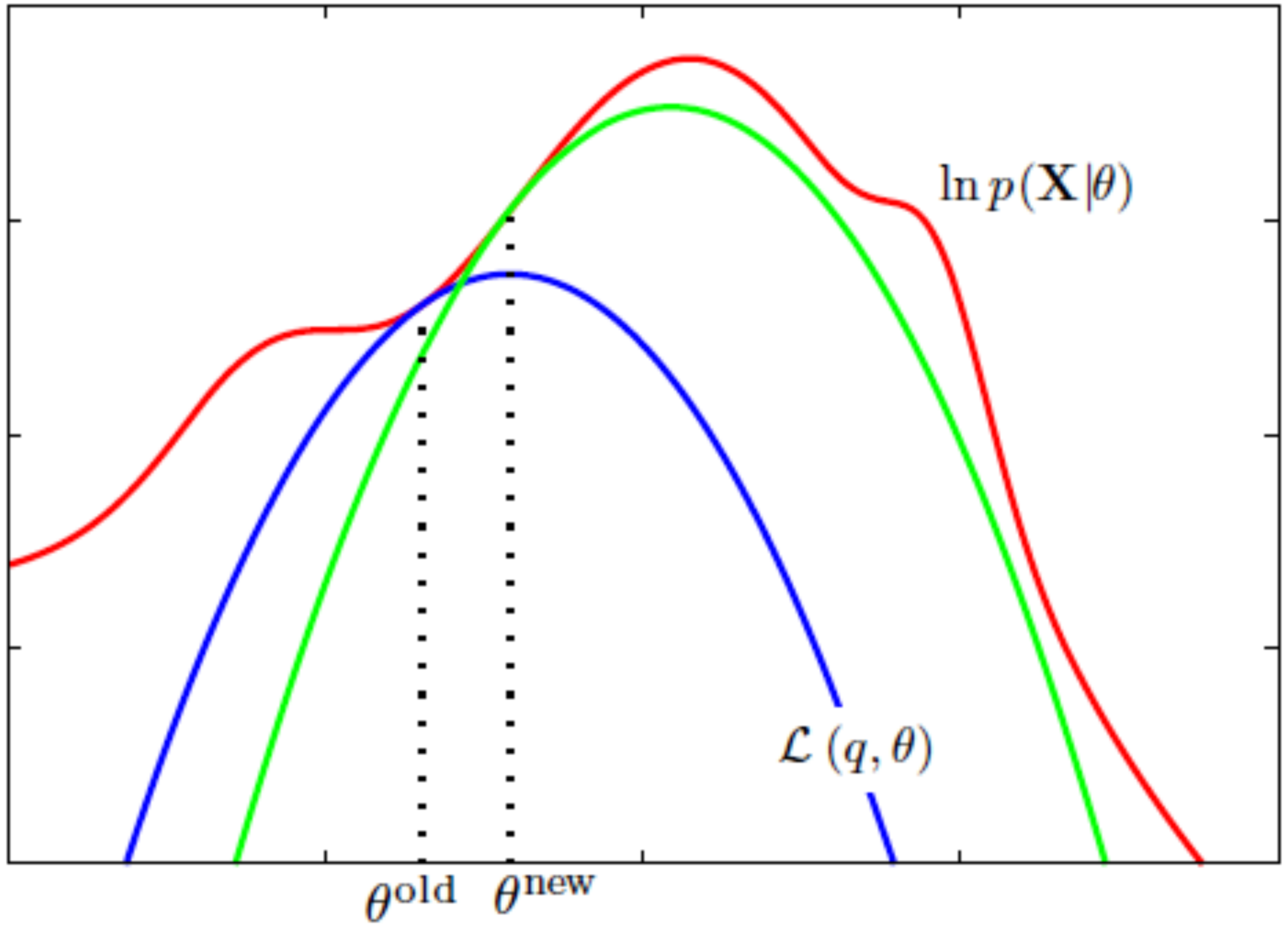
- **M-step**: Maximize the expected log probability

$$\Theta \leftarrow \arg\max_{\Theta} \mathbb{E}_{q(z)}[\log p(x, z, \Theta)]$$

- **E-step**: Update the posterior over latent variables

$$q \leftarrow p(z \mid x, \Theta)$$

- EM converges to **local maxima** of the log marginal likelihood, $\log p(x; \Theta)$




Bishop (2006). Pattern Recognition and Machine Learning, Ch 9.4.

# EM for the Gaussian HMM

## The M-step

In the M-step we set,

$$\Theta = \arg\max_{\Theta} \mathbb{E}_{q(z)}[\log p(x, z, \Theta)]$$

As a function of the mean $b_k$ and variance $Q_k$ for cluster $k$, this objective is,

$$\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] = \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T} \log \mathcal{N}(x_t \mid b_{z_t}, Q_{z_t}) + \log \mathrm{Cat}(z_t \mid \pi)\right]$$

# EM for the Gaussian HMM

## The M-step

In the M-step we set,

$$\Theta = \arg\max_{\Theta} \mathbb{E}_{q(z)}[\log p(x, z, \Theta)]$$

As a function of the mean $b_k$ and variance $Q_k$ for cluster $k$, this objective is,

$$\begin{aligned} \mathbb{E}_{q(z)}[\log p(x, z, \Theta)] &= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T} \log \mathcal{N}(x_t \mid b_{z_t}, Q_{z_t}) + \log \mathrm{Cat}(z_t \mid \pi)\right] \\ &= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T} \sum_{j=1}^{K} \mathbb{I}[z_t = j] \, \log \mathcal{N}(x_t \mid b_j, Q_j)\right] + \text{const} \end{aligned}$$

# EM for the Gaussian HMM

## The M-step

In the M-step we set,

$$\Theta = \arg\max_{\Theta} \mathbb{E}_{q(z)}[\log p(x, z, \Theta)]$$

As a function of the mean $b_k$ and variance $Q_k$ for cluster $k$, this objective is,

$$\begin{aligned}
\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] &= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T} \log \mathcal{N}(x_t \mid b_{z_t}, Q_{z_t}) + \log \mathrm{Cat}(z_t \mid \pi)\right] \\
&= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} \mathbb{I}[z_t = j]\ \log \mathcal{N}(x_t \mid b_j, Q_j)\right] + \text{const} \\
&= \sum_{t=1}^{T} \mathbb{E}_{q(z)}[\mathbb{I}[z_t = k]]\left(-\frac{1}{2}\log|Q_k| - \frac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\right) + \text{const}
\end{aligned}$$

# EM for the Gaussian HMM

## The M-step

In the M-step we set,

$$\Theta = \arg\max_\Theta \mathbb{E}_{q(z)}[\log p(x, z, \Theta)]$$

As a function of the mean $b_k$ and variance $Q_k$ for cluster $k$, this objective is,

$$\begin{aligned}
\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] &= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T} \log \mathcal{N}(x_t \mid b_{z_t}, Q_{z_t}) + \log \mathrm{Cat}(z_t \mid \pi)\right] \\
&= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} \mathbb{I}[z_t = j]\ \log \mathcal{N}(x_t \mid b_j, Q_j)\right] + \text{const} \\
&= \sum_{t=1}^{T} \mathbb{E}_{q(z)}[\mathbb{I}[z_t = k]]\left(-\frac{1}{2}\log|Q_k| - \frac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\right) + \text{const} \\
&= \sum_{t=1}^{T} q(z_t = k)\left(-\frac{1}{2}\log|Q_k| - \frac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\right) + \text{const}
\end{aligned}$$

# EM for the Gaussian HMM

## The M-step

Taking derivatives and setting to zero yields the following updates,

$$T_k = \sum_{t=1}^{T} q(z_t = k)$$

$$b_k = \frac{1}{T_k} \sum_{t=1}^{T} q(z_t = k)\, x_t$$

$$Q_k = \frac{1}{T_k} \sum_{t=1}^{T} q(z_t = k)\, (x_t - b_k)(x_t - b_k)^\top$$

**Note:** we only need the **posterior marginal probabilities** $q(z_t = k)$**!**

# EM for the Gaussian HMM

## The posterior is a little trickier than in the Gaussian mixture model

- **E-step**: Update the posterior over latent variables,

$$q(z) \leftarrow p(z \mid x, \Theta) \propto p(x, z, \Theta) = p(z_1) \prod_{t=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=1}^{T} p(x_t \mid z_t)$$

- The normalized posterior no longer has a simple **closed form!**

- However, we can still **efficiently compute** the **marginal probabilities** for the **M-step**.

# EM for the Gaussian HMM

## Computing posterior marginals

- Consider the marginal probability of state $k$ at time $t$:

$$q(z_t = k) = \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} q(z_, \ldots, z_{t-1}, z_t = k, z_{t+1}, \ldots, z_T)$$

# EM for the Gaussian HMM

## Computing posterior marginals

- Consider the marginal probability of state $k$ at time $t$:

$$q(z_t = k) = \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} q(z_, \ldots, z_{t-1}, z_t = k, z_{t+1}, \ldots, z_T)$$

$$\propto \left[ \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(x_s \mid z_s)\, p(z_{s+1} \mid z_s) \right] \times \left[ p(x_t \mid z_t) \right]$$

$$\times \left[ \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u) \right]$$

# EM for the Gaussian HMM

## Computing posterior marginals

- Consider the marginal probability of state $k$ at time $t$:

$$
\begin{aligned}
q(z_t = k) &= \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} q(z_, \ldots, z_{t-1}, z_t = k, z_{t+1}, \ldots, z_T) \\
&\propto \left[ \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(x_s \mid z_s)\, p(z_{s+1} \mid z_s) \right] \times \left[ p(x_t \mid z_t) \right] \\
&\quad \times \left[ \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u) \right] \\
&\triangleq \alpha_t(z_t) \times p(x_t \mid z_t) \times \beta_t(z_t)
\end{aligned}
$$

# EM for the Gaussian HMM

## Computing the forward messages $\alpha_t(z_t)$

- Consider the **forward messages**:

$$\alpha_t(z_t) \triangleq \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(x_s \mid z_s)\, p(z_{s+1} \mid z_s)$$

# EM for the Gaussian HMM

## Computing the forward messages $\alpha_t(z_t)$

- Consider the forward messages:

$$\alpha_t(z_t) \triangleq \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(x_s \mid z_s)\, p(z_{s+1} \mid z_s)$$

$$= \sum_{z_{t-1}=1}^{K} \left[ \left( \sum_{z_1=1}^{K} \cdots \sum_{z_{t-2}=1}^{K} p(z_1) \prod_{s=1}^{t-2} p(x_s \mid z_s) p(z_{s+1} \mid z_s) \right) p(x_{t-1} \mid z_{t-1})\, p(z_t \mid z_{t-1}) \right]$$

# EM for the Gaussian HMM

## Computing the forward messages $\alpha_t(z_t)$

- Consider the forward messages:

$$\alpha_t(z_t) \triangleq \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(x_s \mid z_s)\, p(z_{s+1} \mid z_s)$$

$$= \sum_{z_{t-1}=1}^{K} \left[ \left( \sum_{z_1=1}^{K} \cdots \sum_{z_{t-2}=1}^{K} p(z_1) \prod_{s=1}^{t-2} p(x_s \mid z_s) p(z_{s+1} \mid z_s) \right) p(x_{t-1} \mid z_{t-1})\, p(z_t \mid z_{t-1}) \right]$$

$$= \sum_{z_{t-1}=1}^{K} \alpha_{t-1}(z_{t-1})\, p(x_{t-1} \mid z_{t-1})\, p(z_t \mid z_{t-1})$$

- We can compute these messages **recursively!**

# EM for the Gaussian HMM

## Computing the forward messages $\alpha_t(z_t)$. Vectorized.

- Let $\alpha_t = [\alpha_t(z_t = 1), \ldots, \alpha_t(z_t = K)]^\top$ denote the column vector of forward messages. Then,

$$\alpha_t = P^\top(\alpha_{t-1} \odot \ell_{t-1})$$

  where

  - $\ell_{t-1} = [p(x_{t-1} \mid z_{t-1} = 1), \ldots, p(x_{t-1} \mid z_{t-1} = K)]^\top$ is the vector of likelihoods,
  - $\odot$ denotes the element-wise product, and
  - $P$ is the transition matrix with $P_{ij} = p(z_t = j \mid z_{t-1} = i)$.

- For the base case, let $\alpha_1(z_1) = p(z_1)$.

# EM for the Gaussian HMM

## Computing the backward messages $\beta_t(z_t)$

- Now take the **backward messages**:

$$\beta_t(z_t) \triangleq \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u)$$

# EM for the Gaussian HMM

## Computing the backward messages $\beta_t(z_t)$

- Now take the **backward messages**:

$$\beta_t(z_t) \triangleq \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u)$$

$$= \sum_{z_{t+1}=1}^{K} p(z_{t+1} \mid z_t)\, p(x_{t+1} \mid z_{t+1}) \sum_{z_{t+2}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+2}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u)$$

# EM for the Gaussian HMM

## Computing the backward messages $\beta_t(z_t)$

- Now take the **backward messages**:

$$\beta_t(z_t) \triangleq \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u)$$

$$= \sum_{z_{t+1}=1}^{K} p(z_{t+1} \mid z_t)\, p(x_{t+1} \mid z_{t+1}) \sum_{z_{t+2}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+2}^{T} p(z_u \mid z_{u-1})\, p(x_u \mid z_u)$$

$$= \sum_{z_{t+1}=1}^{K} p(z_{t+1} \mid z_t)\, p(x_{t+1} \mid z_{t+1})\, \beta_{t+1}(z_{t+1})$$

- Again, we can compute the backward messages recursively!

# EM for the Gaussian HMM

## Computing the backward messages $\beta_t(z_t)$. Vectorized.

- Let $\beta_t = [\beta_t(z_t = 1), \ldots, \beta_t(z_t = K)]^\top$ denote the column vector of backward messages. Then,

$$\beta_t = P(\beta_{t+1} \odot \ell_{t+1})$$

- For the base case, let $\beta_T(z_T) = 1$.

# EM for the Gaussian HMM

## Combining the forward and backward messages

- The posterior marginal probability of state $k$ at time $t$ is,

$$\begin{aligned} q(z_t = k) &\propto \alpha_t(z_t = k) \times p(x_t \mid z_t = k) \times \beta_t(z_t = k) \\ &= \alpha_{tk} \ell_{tk} \beta_{tk} \end{aligned}$$

- The probabilities need to sum to one. Normalizing yields,

$$q(z_t = k) = \frac{\alpha_{tk} \ell_{tk} \beta_{tk}}{\sum_{j=1}^{K} \alpha_{tj} \ell_{tj} \beta_{tj}}$$

- Finally, note the marginal is invariant to multiplying $\alpha_t$ and/or $\beta_t$ by a constant.

# EM for the Gaussian HMM

## Normalizing the messages to prevent underflow

- The messages involve **products of probabilities**, which quickly underflow.
- We can leverage the scale invariance to renormalize the messages. I.e. replace:

$$\alpha_t = P^\top(\alpha_{t-1} \odot \ell_{t-1}) \quad \text{with} \quad \begin{aligned} A_{t-1} &= \textstyle\sum_k \tilde{\alpha}_{t-1,k} \ell_{t-1,k} \\ \tilde{\alpha}_t &= \tfrac{1}{A_{t-1}} P^\top(\tilde{\alpha}_{t-1} \odot \ell_{t-1}) \end{aligned}$$

where $\tilde{\alpha}_t$ are normalized for numerical stability. As before, $\tilde{\alpha}_1 = \pi$.

- This lends a nice **interpretation**: the **forward messages are conditional probabilities** $\tilde{\alpha}_{tk} = p(z_t = k \mid x_{1:t-1})$ and the **normalization constants are the marginal likelihoods** $A_t = p(x_t \mid x_{1:t-1})$.

# EM for the Gaussian HMM

## Computing the marginal likelihood

- Finally, we can compute the marginal likelihood alongside the forward messages

$$\log p(x \mid \Theta) = \log \sum_{z_1=1}^{K} \cdots \sum_{z_T=1}^{K} \left[ p(z_1) \prod_{t=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=1}^{T} p(x_t \mid z_t) \right]$$

$$= \log \sum_{z_T=1}^{K} \alpha_T(z_T)\, p(x_T \mid z_T)$$

$$= \log \prod_{t=1}^{T} A_t = \sum_{t=1}^{T} \log A_t$$

- Again, makes sense since the normalization constants are $A_t = p(x_t \mid x_{1:t-1})$.

# The M-step with sufficient statistics

# EM for the Gaussian HMM

## Sufficient statistics

$$\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] = \sum_{t=1}^{T} q(z_t = k)\left[-\frac{1}{2}\log|Q_k| - \frac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\right] + \mathrm{c}$$

# EM for the Gaussian HMM

## Sufficient statistics

$$\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] = \sum_{t=1}^{T} q(z_t = k)\left[-\frac{1}{2}\log|Q_k| - \frac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\right] + \mathrm{c}$$

$$= \sum_{t=1}^{T} q(z_t = k)\left[-\frac{1}{2}\log|Q_k| - \frac{1}{2}x_t^\top Q_k^{-1}x_t + b_k^\top Q_k^{-1}x_t - \frac{1}{2}b_k^\top Q_k^{-1}b_k\right] + \mathrm{c}$$

# EM for the Gaussian HMM

## Sufficient statistics

$$\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] = \sum_{t=1}^{T} q(z_t = k)\Big[-\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\Big] + \mathrm{c}$$

$$= \sum_{t=1}^{T} q(z_t = k)\Big[-\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}x_t^\top Q_k^{-1}x_t + b_k^\top Q_k^{-1}x_t - \tfrac{1}{2}b_k^\top Q_k^{-1}b_k\Big] + \mathrm{c}$$

$$= \sum_{t=1}^{T} q(z_t = k)\Big[\Big\langle -\tfrac{1}{2}\log|Q_k|, 1\Big\rangle + \Big\langle -\tfrac{1}{2}Q_k^{-1}, x_t x_t^\top\Big\rangle + \Big\langle b_k^\top Q_k^{-1}, x_t\Big\rangle + \Big\langle -\tfrac{1}{2}b_k^\top Q_k^{-1}b_k, 1\Big\rangle\Big] + \mathrm{c}$$

# EM for the Gaussian HMM

## Sufficient statistics

$$
\begin{aligned}
\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] &= \sum_{t=1}^{T} q(z_t = k)\Big[-\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}(x_t - b_k)^\top Q_k^{-1}(x_t - b_k)\Big] + \mathrm{c} \\
&= \sum_{t=1}^{T} q(z_t = k)\Big[-\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}b_k^\top Q_k^{-1} b_k + b_k^\top Q_k^{-1} x_t - \tfrac{1}{2}x_t^\top Q_k^{-1} x_t\Big] + \mathrm{c} \\
&= \sum_{t=1}^{T} q(z_t = k)\Big[\Big\langle -\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}b_k^\top Q_k^{-1} b_k, 1\Big\rangle + \Big\langle b_k^\top Q_k^{-1}, x_t\Big\rangle + \Big\langle -\tfrac{1}{2}Q_k^{-1}, x_t x_t^\top\Big\rangle\Big] + \mathrm{c} \\
&= \Big\langle -\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}b_k^\top Q_k^{-1} b_k, T_k\Big\rangle + \Big\langle b_k^\top Q_k^{-1}, \mathbf{t}_{k,1}\Big\rangle + \Big\langle -\tfrac{1}{2}Q_k^{-1}, \mathbf{t}_{k,2}\Big\rangle + \mathrm{c}
\end{aligned}
$$

where

$$
T_k = \sum_{t=1}^{T} q(z_t = k) \qquad \mathbf{t}_{k,1} = \sum_{t=1}^{T} q(z_t = k)\, x_t \qquad \mathbf{t}_{k,2} = \sum_{t=1}^{T} q(z_t = k)\, x_t x_t^\top
$$

are the **weighted sums of sufficient statistics**.

# EM for the Gaussian HMM

## Solving for the optimal Gaussian parameters

The objective we're trying to maximize is,

$$\mathscr{L}(q, \theta) = \left\langle -\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2} b_k^\top Q_k^{-1} b_k, T_k \right\rangle + \left\langle b_k^\top Q_k^{-1}, \mathbf{t}_{k,1} \right\rangle + \left\langle -\tfrac{1}{2} Q_k^{-1}, \mathbf{t}_{k,2} \right\rangle + \mathrm{c}$$

Taking the partial derivative wrt $b_k$ and setting equal to zero,

$$\frac{\partial}{\partial b_k}\mathscr{L}(q, \theta) = Q_k^{-1}\mathbf{t}_{k,1} - Q_k^{-1} b_k T_k = 0$$

$$\implies b_k^\star = \frac{\mathbf{t}_{k,1}}{T_k} = \frac{1}{T_k}\sum_{t=1}^{T} q(z_t = k)\, x_t$$

# EM for the Gaussian HMM

## Solving for the optimal Gaussian parameters

Plug in the optimum

$$\mathscr{L}(q,\theta) = \left\langle -\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}\frac{\mathbf{t}_{k,1}^\top}{T_k} Q_k^{-1} \frac{\mathbf{t}_{k,1}}{T_k}, T_k \right\rangle + \left\langle -\tfrac{1}{2} Q_k^{-1}, \mathbf{t}_{k,2} \right\rangle + \left\langle \frac{\mathbf{t}_{k,1}^\top}{T_k} Q_k^{-1}, \mathbf{t}_{k,1} \right\rangle + \mathrm{c}$$

$$= \left\langle -\tfrac{1}{2}\log|Q_k|, T_k \right\rangle + \left\langle -\tfrac{1}{2} Q_k^{-1}, \mathbf{t}_{k,2} - \frac{\mathbf{t}_{k,1}\mathbf{t}_{k,1}^\top}{T_k} \right\rangle + \mathrm{c}$$

# EM for the Gaussian HMM

## Solving for the optimal Gaussian parameters

Let $\Lambda_k = Q_k^{-1}$,

$$\mathscr{L}(q, \theta) = \left\langle \tfrac{1}{2} \log |\Lambda_k|, T_k \right\rangle + \left\langle -\tfrac{1}{2}\Lambda_k, \mathbf{t}_{k,2} - \frac{\mathbf{t}_{k,1}\mathbf{t}_{k,1}^\top}{T_k} \right\rangle + \mathrm{c}$$

Taking the partial derivative wrt $\Lambda_k$ and setting equal to zero,

$$\frac{\partial}{\partial \Lambda_k} \mathscr{L}(q, \theta) = \frac{T_k}{2} \Lambda_k^{-1} - \frac{1}{2}\left( \mathbf{t}_{k,2} - \frac{\mathbf{t}_{k,2}\mathbf{t}_{k,1}^\top}{T_k} \right) = 0$$

$$\implies (\Lambda_k^{-1})^\star = Q_k^\star = \frac{1}{T_k}\left( \mathbf{t}_{k,2} - \frac{\mathbf{t}_{k,1}\mathbf{t}_{k,1}^\top}{T_k} \right)$$

# EM for the Gaussian HMM

## In summary…

- **E-step**: Compute the posterior probabilities:

  $q(z_t = k) \leftarrow p(z_t = k \mid x_t, \Theta)$ via the **forward-backward algorithm.**

  Compute **weighted sums of sufficient statistics**:

  $$T_k = \sum_{t=1}^{T} q(z_t = k) \qquad \mathbf{t}_{k,1} = \sum_{t=1}^{T} q(z_t = k)\, x_t \qquad \mathbf{t}_{k,2} = \sum_{t=1}^{T} q(z_t = k)\, x_t x_t^\top$$

- **M-step**: Update the parameters.

  $$b_k \leftarrow \frac{\mathbf{t}_{k,1}}{T_k} \qquad Q_k \leftarrow \frac{1}{T_k}\left(\mathbf{t}_{k,2} - \frac{\mathbf{t}_{k,1}\mathbf{t}_{k,1}^\top}{T_k}\right)$$

- *Note: The updates are equivalent if we use* **normalized sufficient statistics**, each divided by $T$.

# Stochastic EM for the Gaussian mixture model

- On iteration $i$, grab a sub-sequence (aka **mini-batch**) of length $M$.
- **E-step**: Compute the posterior probabilities for each data point in the mini-batch:

$$q(z_m = k) \leftarrow p(z_m = k \mid x_m, \Theta) \propto \frac{\pi_k \mathcal{N}(x_m \mid b_k, Q_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_m \mid b_j, Q_j)}$$

  Compute <u>**normalized**</u> **sufficient statistics** for the mini-batch:

$$\bar{T}_k^{(i)} = \frac{1}{M} \sum_{m=1}^{M} q(z_m = k) \qquad \bar{\mathbf{t}}_{k,1}^{(i)} = \frac{1}{M} \sum_{m=1}^{M} q(z_m = k)\, x_m \qquad \bar{\mathbf{t}}_{k,2}^{(i)} = \frac{1}{M} \sum_{m=1}^{M} q(z_m = k)\, x_m x_m^\top$$

  Fold the normalized stats from this mini-batch into the running average via a **convex combination** with step size $\alpha \in [0,1]$:

$$\bar{T}_k \leftarrow (1-\alpha)\bar{T}_k + \alpha \bar{T}_k^{(i)} \qquad \bar{\mathbf{t}}_{k,1} \leftarrow (1-\alpha)\bar{\mathbf{t}}_{k,1} + \alpha \bar{\mathbf{t}}_{k,1}^{(i)} \qquad \bar{\mathbf{t}}_{k,2} \leftarrow (1-\alpha)\bar{\mathbf{t}}_{k,2} + \alpha \bar{\mathbf{t}}_{k,2}^{(i)}$$

- **M-step**: Update the parameters.

$$b_k \leftarrow \frac{\bar{\mathbf{t}}_{k,1}}{\bar{T}_k} \qquad Q_k \leftarrow \frac{1}{\bar{T}_k} \left( \bar{\mathbf{t}}_{k,2} - \frac{\bar{\mathbf{t}}_{k,1} \bar{\mathbf{t}}_{k,1}^\top}{\bar{T}_k} \right)$$

# Conclusion

- Hidden Markov models (HMMs) are just mixture models with dependencies across time.
- The EM algorithm is nearly the same as for mixture models, but we use the **forward-backward algorithm** to compute posterior marginal probabilities.
- With exponential family likelihoods, the M-step only needs weighted sums of **sufficient statistics**.
- **Stochastic EM** generalizes the EM algorithm to work with **mini-batches** of data and rolling averages of the sufficient statistics. It can be seen as SGD with *natural* gradients.