# Machine Learning Methods for Neural Data Analysis

## Expectation-Maximization (EM)

Scott Linderman

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

# Final project

- **Friday, May 16**: Project proposal due (1pg).
  - Groups (2 people)
  - Project Goal / Key Idea
  - Dataset(s)
- **Weds, May 21**: Lab 7 (last lab) due.
- **Monday, June 9:** Final report due.

# Types of final projects

- The final project is an opportunity to **apply what you’ve learned** to a **problem of interest** to you. If you’re a PhD student, we hope it will advance your research.
- For example, you could:
  - **Propose and implement an extension** to an existing method to address its limitations.
  - **Implement a related method** from a recent research paper and reproduce its results.
  - **Adapt a method** from class to study a scientific question of interest on a new dataset.
- Your project must involve some modeling and/or algorithmic innovations, more than just applying code from lab to a new dataset.

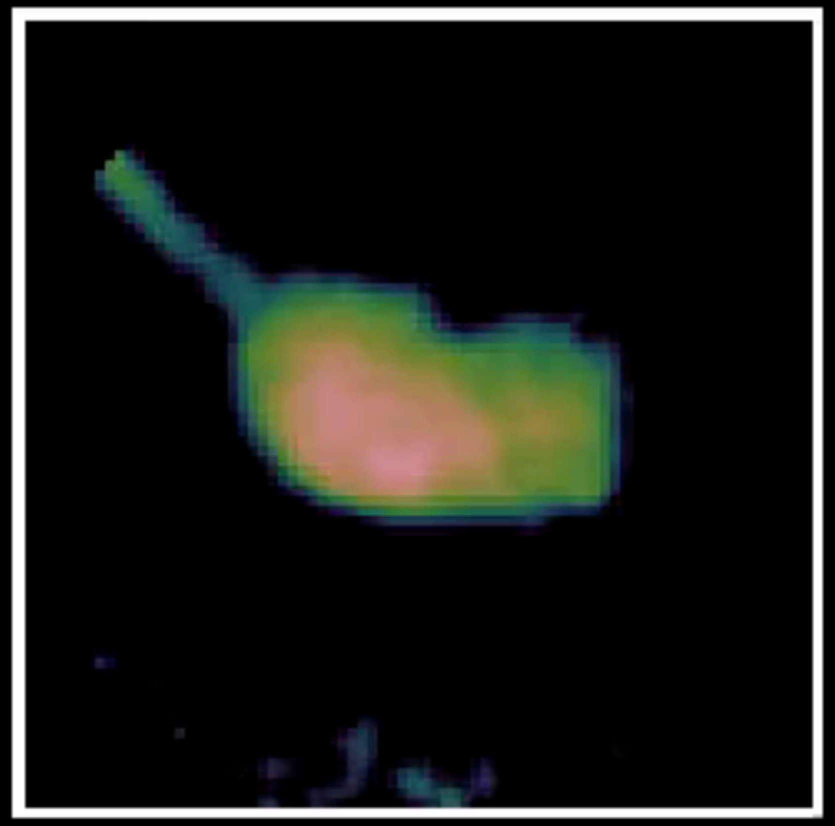
# Datasets

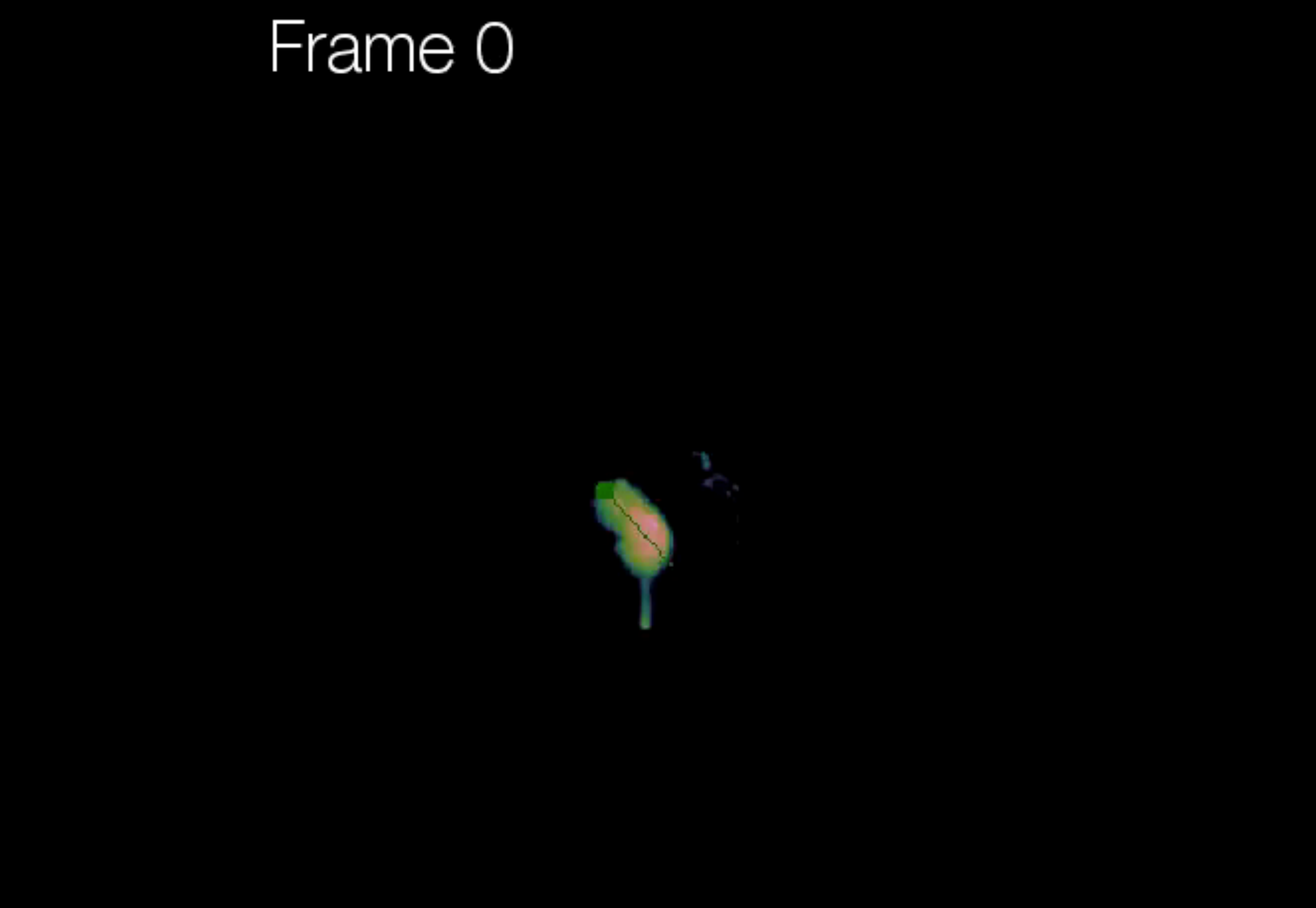
- Most recent papers in Nature, Science, Cell, etc., have publicly available data.
- If you are a graduate student and have access to a relevant dataset from your research, that should be fine too! (But talk to me first about EEG/MEG/fMRI data since this course didn't cover those modalities.)
- NeuroMatch Academy has curated some **large-scale neural and behavioral datasets**:
  - https://compneuro.neuromatch.io/projects/neurons/README.html
  - https://compneuro.neuromatch.io/projects/behavior/README.html
- NeuroMatch also has many **project templates** that could spark ideas:
  - https://compneuro.neuromatch.io/projects/docs/project_templates.html

# Rough Rubric for Final Projects

- A+: Outstanding work. (Anticipating top ~10%.) Ambitious undertaking with challenging implementation of a new idea, a well-motivated question and choice of dataset. Rigorous, quantitative evaluation. Clear and detailed report.
- A: Excellent work. (~60-90th percentile.) Typically characterized by: Strong proposal that required considerable work and/or showed great creativity. Rigorous, quantitative evaluation. Clear and detailed report.
- B: Good work. (~30-60th percentile.) Typically characterized by: An interesting question, but a smaller leap from the labs. Comparisons may be more qualitative in nature. Reports may be less polished.
- C: Fair work. (~10-30th percentile.) Typically characterized by: Relatively small or extension of lab assignments. Comparisons often more qualitative in nature. Reports left much room for improvement.
- D: Bottom (~10%.) This category tends to be more variable, but is typically characterized by: Nearly a direct application of lab work. Qualitative or non-existent comparisons. Very short or sloppy report.

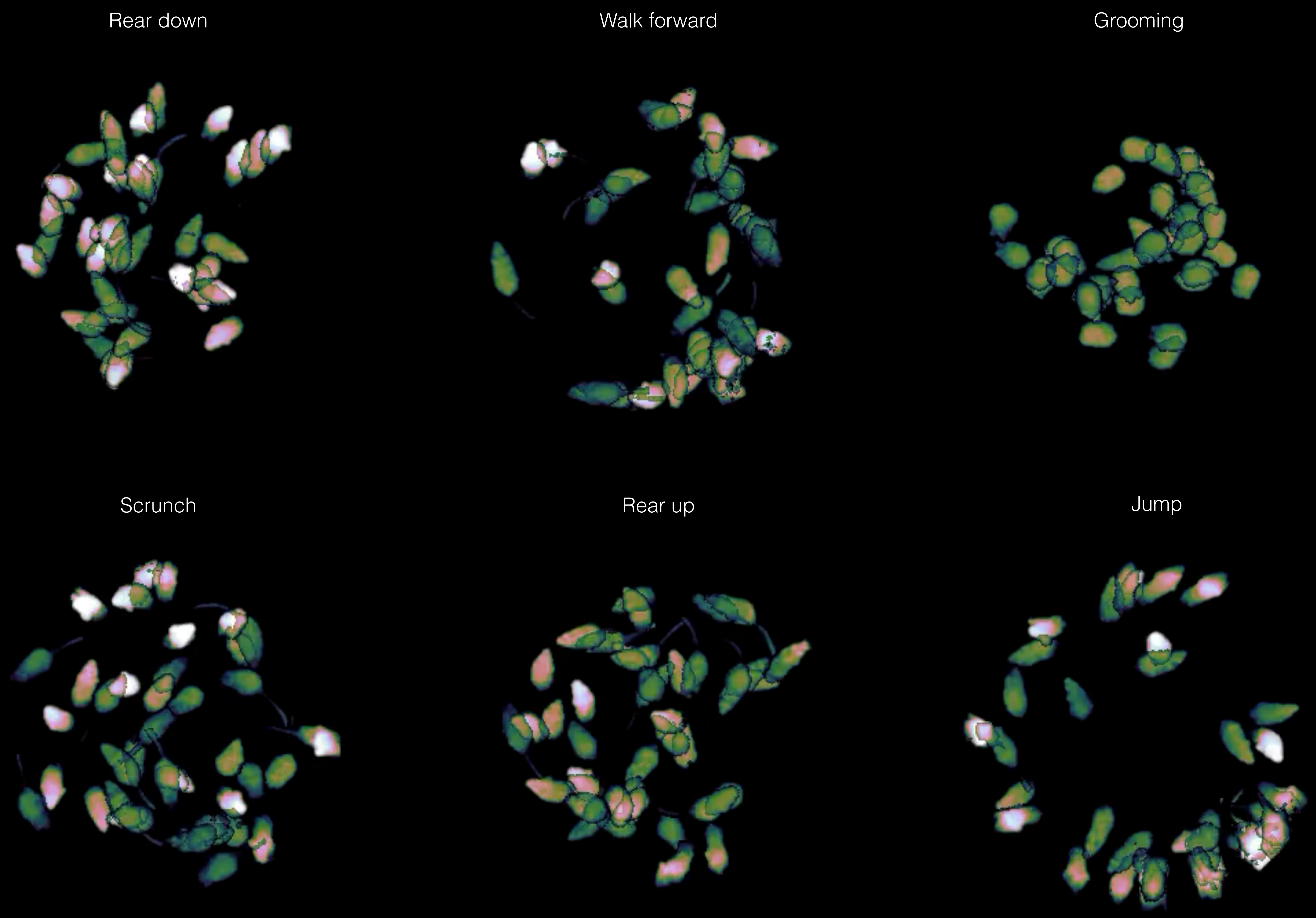
# Motivating Example: summarizing videos with behavioral states






*Wiltschko et al, 2015*

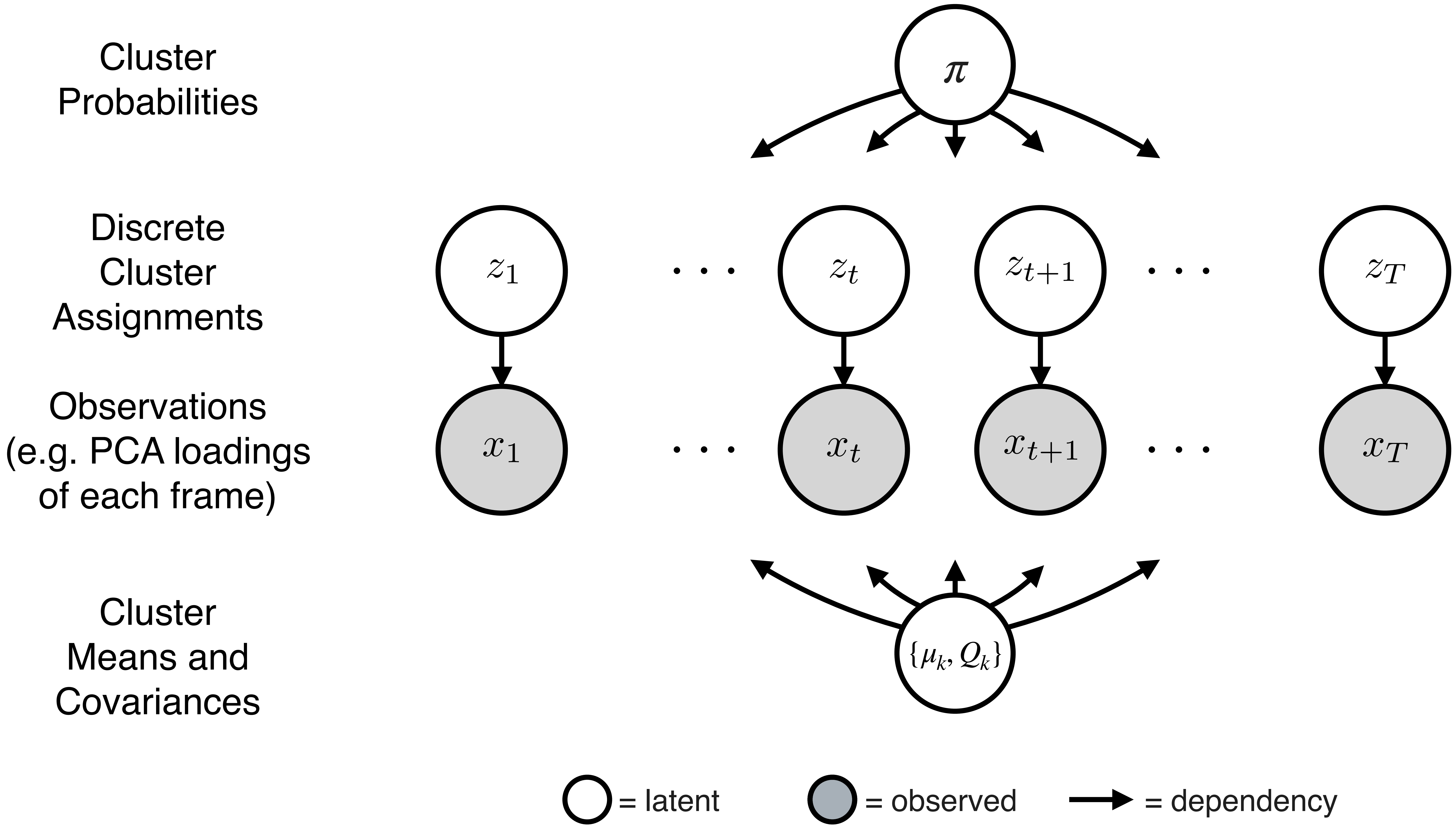
# Motivating Example: summarizing videos with behavioral states




*Wiltschko et al, 2015*

# What We've Seen: The Gaussian Mixture Model

## Graphical Model



# What We've Seen: The Gaussian Mixture Model

## Joint probability

Cluster assignments are independent across time

$$z_t \sim \mathrm{Cat}(\pi), \qquad \text{for } t = 1,\ldots,T$$
$$x_t \mid z_t \sim \mathcal{N}(\mu_{z_t}, Q_{z_t}) \qquad \text{for } t = 1,\ldots,T$$

Its parameters are $\Theta = \pi, \{\mu_k, Q_k\}_{k=1}^K$ where $\pi$ is the prior distribution over clusters and $(\mu_k, Q_k)$ are the cluster means and covariances, respectively.

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

$$\begin{aligned}\log p(x, z, \Theta) &= \sum_{t=1}^{T} \log p(z_t) + \log p(x_t \mid z_t) \\ &= \sum_{t=1}^{T} \log \mathrm{Cat}(z_t \mid \pi) + \log \mathcal{N}(x_t \mid \mu_{z_t}, Q_{z_t})\end{aligned}$$

# Fitting a Gaussian Mixture Model

## Coordinate Ascent

Find the cluster assignments $z_t^\star$ and parameters $\Theta^\star$ that maximize the (log) joint prob:

1. Update the cluster assignments for all time steps,

$$
\begin{aligned}
z_t &\leftarrow \arg\max_k \; \log \mathrm{Cat}(k \mid \pi) + \log \mathcal{N}(x_t \mid \mu_k, Q_k) \\
&= \arg\max_k \; \log \pi_k - \frac{D}{2} \log |Q_k| - \frac{1}{2}(x_t - \mu_k)^\top Q_k^{-1}(x_t - \mu_k)
\end{aligned}
$$

   Note that all clusters can be updated independently!

2. Update the parameters for each cluster based on the assignments. For example,

$$
\begin{aligned}
\mu_k &\leftarrow \arg\max_\mu \sum_{t: z_t = k} \log \mathcal{N}(x_t \mid \mu, Q_k) \\
&= \frac{1}{T_k} \sum_{t: z_t = k} x_t
\end{aligned}
$$

   where $T_k = \sum_t \mathbb{I}[z_t = k]$ is the number of data points assigned to cluster $k$.

# Fitting a Gaussian Mixture Model

## K-Means

Consider the special case where $\pi = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$ and $Q_k = I$. Then coordinate ascent is equivalent to the **K-Means algorithm:**

1. Update the cluster assignments for all time steps,

$$\begin{aligned} z_t &\leftarrow \arg\max_k \; -\frac{1}{2}(x_t - \mu_k)^\top (x_t - \mu_k) \\ &= \arg\min_k \; \|x_t - \mu_k\|_2^2 \end{aligned}$$

   *Assign each data point to the nearest cluster.*

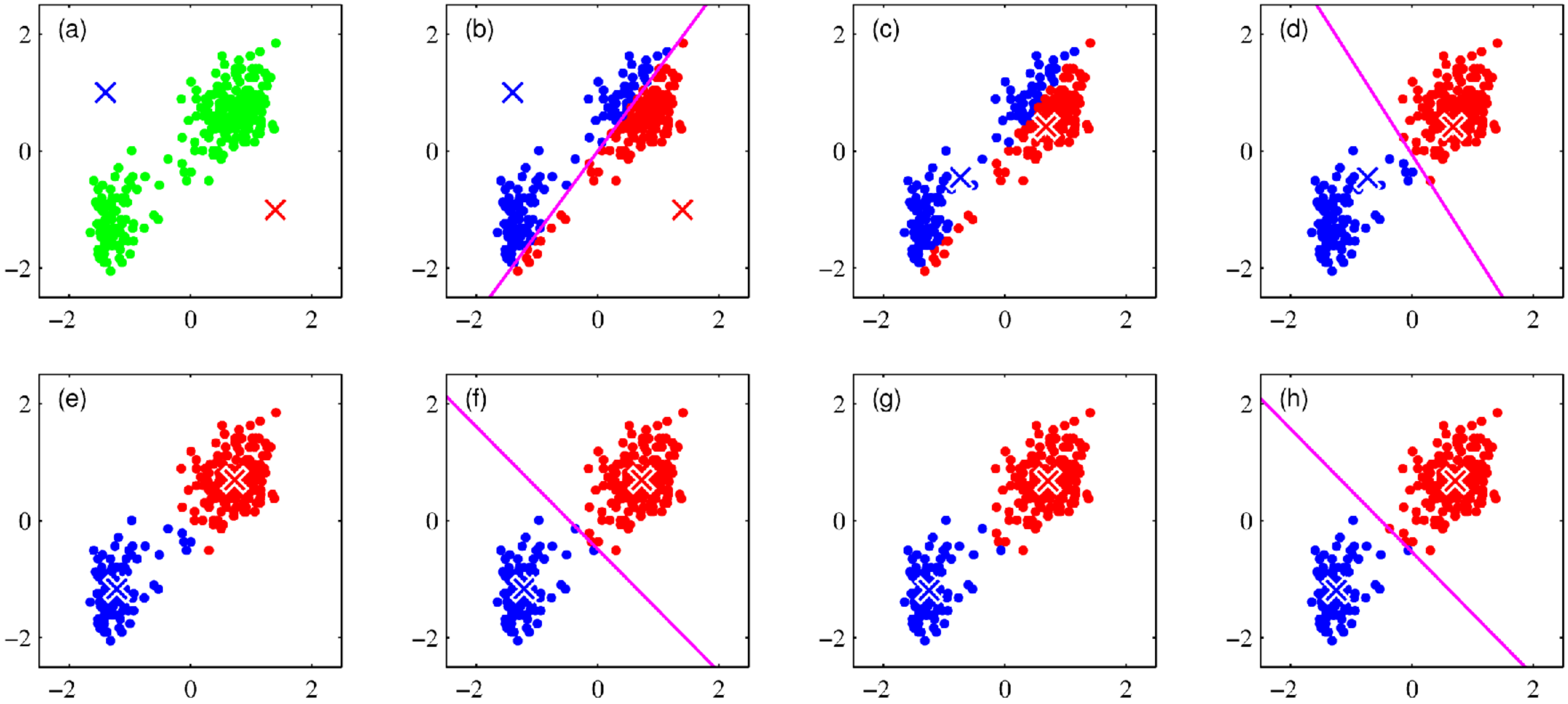
2. Update the parameters for each cluster based on the assignments. For example,

$$\begin{aligned} \mu_k &\leftarrow \arg\max_\mu \sum_{t: z_t = k} \log \mathcal{N}(x_t \mid \mu, Q_k) \\ &= \frac{1}{T_k} \sum_{t: z_t = k} x_t \end{aligned}$$

   *Set each cluster centroid to the mean of the assigned data points.*

# Fitting a Gaussian Mixture Model

## K-Means in Pictures (Bishop, Ch 9.1)



# Fitting a Gaussian Mixture Model

## Expectation-Maximization (EM)

Does it really make sense to use **hard assignments** of data points to the nearest cluster? What about points on the boundary? Consider the following alternative:

1. Compute the **responsibilities**

$$q_{t,k} \triangleq q(z_t = k) \triangleq \Pr(z_t = k \mid x_t, \theta) \propto \pi_k \mathcal{N}(x_t \mid \mu_k, Q_k)$$

2. Update the parameters for each cluster based on the responsibilities. For example,

$$\mu_k \leftarrow \frac{1}{T_k} \sum_t q_{t,k} x_t \qquad Q_k \leftarrow \frac{1}{T_k} \sum_t q_{t,k} (x_t - \mu_k)(x_t - \mu_k)^\top$$

where $T_k = \sum_t q_{t,k}$ is the *expected* number of data points assigned to cluster $k$.

# Fitting a Gaussian Mixture Model

## Expectation-Maximization (EM)

Does it really make sense to use **hard assignments** of data points to the nearest cluster? What about points on the boundary? Consider the following alternative:

1. Compute the **responsibilities**

$$q(z_t = k) \triangleq \Pr(z_t = k \mid x_t, \theta) \propto \pi_k \mathcal{N}(x_t \mid \mu_k, Q_k)$$

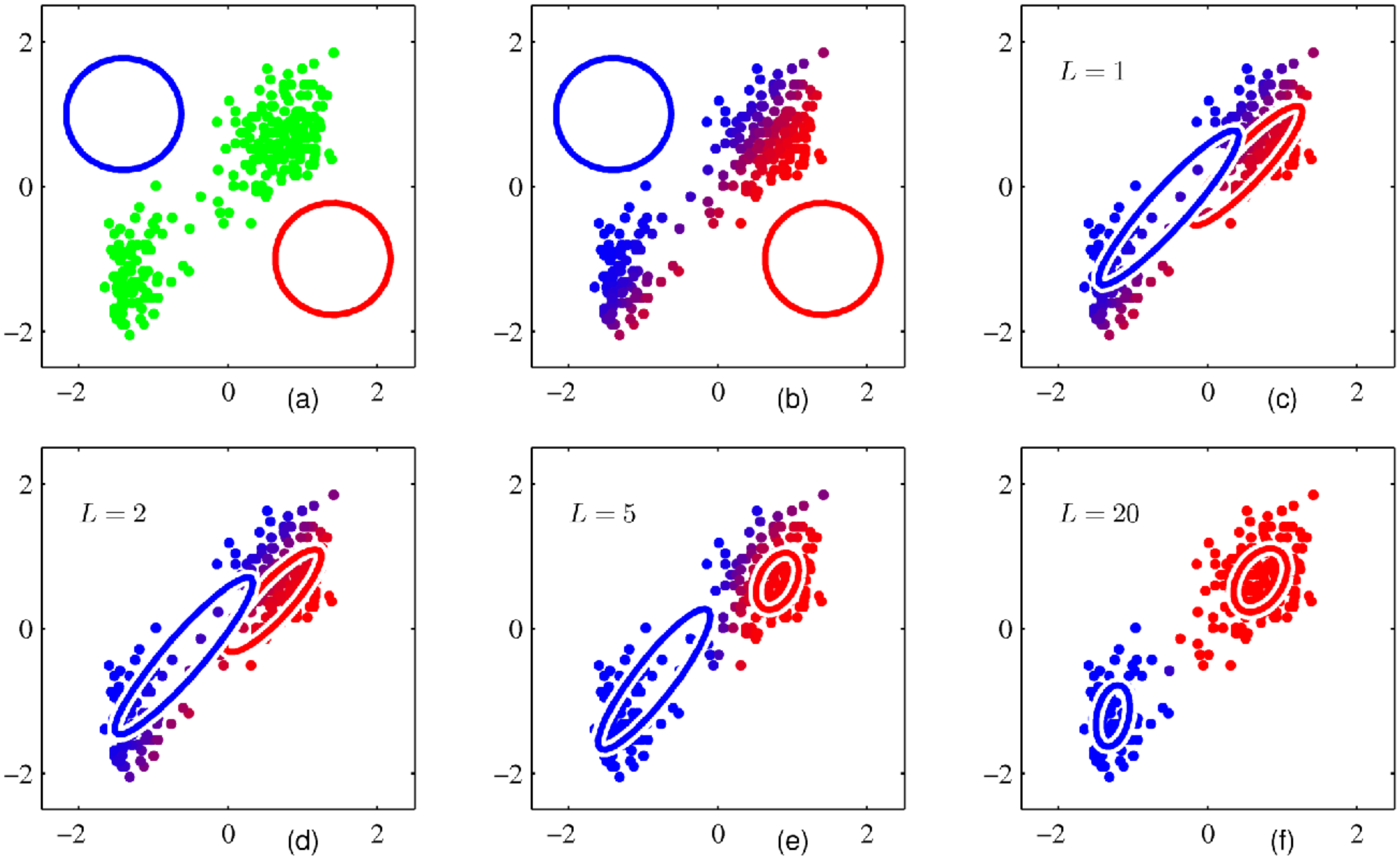
2. Update the parameters for each cluster based on the responsibilities. More generally,

$$\theta \leftarrow \arg\max_{\theta} \mathbb{E}_{q(z)} \left[ \log p(x \mid z; \theta) \right]$$

This is called the **expectation-maximization (EM) algorithm.**

# Fitting a Gaussian Mixture Model

## EM for the GMM in Pictures (Bishop, Ch 9.1)



# Another view of the EM algorithm

## EM finds a local maximum of the log marginal likelihood

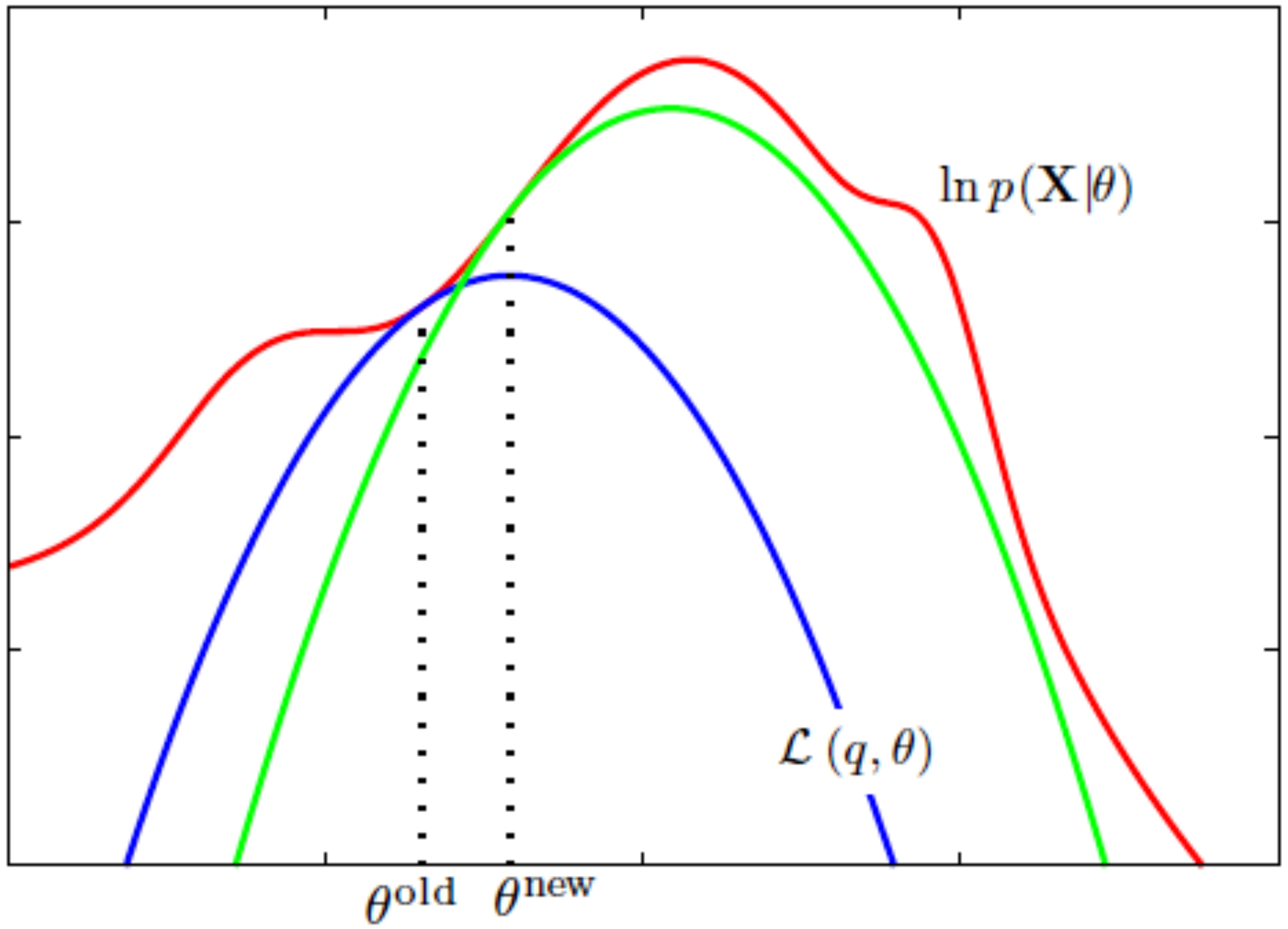
- **M-step**: Maximize the expected log probability

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

- **E-step**: Update the posterior over latent variables *(these are the responsibilities)*

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

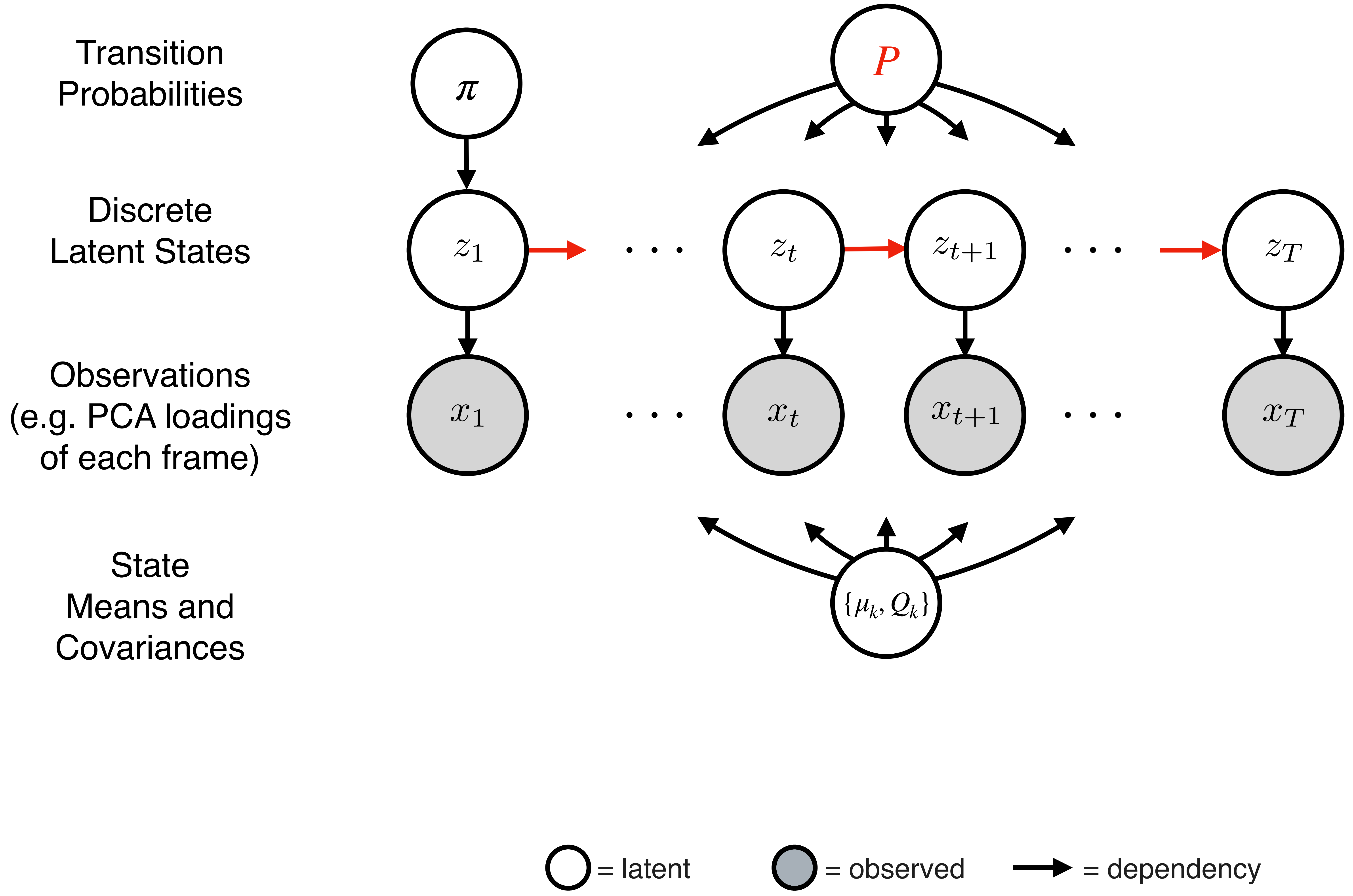
- EM converges to **local maxima** of the log marginal likelihood, $\log p(x; \Theta)$




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

# 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!

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

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

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

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

# 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 $\mu_k$ and variance $Q_k$ for cluster $k$, this objective is,

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

# 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 $\mu_k$ and variance $Q_k$ for cluster $k$, this objective is,

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

# 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 $\mu_k$ and variance $Q_k$ for cluster $k$, this objective is,

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

# 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 $\mu_k$ and variance $Q_k$ for cluster $k$, this objective is,

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

# EM for the Gaussian HMM

## The M-step

Taking derivatives and setting to zero yields the following updates,

$$T_k = \sum_{t=1}^{T} q_{t,k}$$

$$\mu_k = \frac{1}{T_k} \sum_{t=1}^{T} q_{t,k}\, x_t$$

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

**Note:** we only need the **posterior marginal probabilities** $q_{t,k} = \Pr(z_t = k \mid x_{1:T}; \Theta)$**!**

# 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 $t$:

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

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

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

$$\triangleq \alpha_t(z_t) \times p(x_t \mid z_t) \times \beta_t(z_t)$$

# The M-step with sufficient statistics

# EM for the Gaussian HMM

## Sufficient statistics

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

# EM for the Gaussian HMM

## Sufficient statistics

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

# EM for the Gaussian HMM

## Sufficient statistics

$$
\begin{aligned}
\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] &= \sum_{t=1}^{T} q_{t,k} \left[ -\frac{1}{2} \log |Q_k| - \frac{1}{2}(x_t - \mu_k)^\top Q_k^{-1}(x_t - \mu_k) \right] + \mathrm{c} \\
&= \sum_{t=1}^{T} q_{t,k} \left[ -\frac{1}{2} \log |Q_k| - \frac{1}{2} x_t^\top Q_k^{-1} x_t + \mu_k^\top Q_k^{-1} x_t - \frac{1}{2} \mu_k^\top Q_k^{-1} \mu_k \right] + \mathrm{c} \\
&= \sum_{t=1}^{T} q_{t,k} \left[ \left\langle -\frac{1}{2} \log |Q_k|, 1 \right\rangle + \left\langle -\frac{1}{2} Q_k^{-1}, x_t x_t^\top \right\rangle + \left\langle \mu_k^\top Q_k^{-1}, x_t \right\rangle + \left\langle -\frac{1}{2} \mu_k^\top Q_k^{-1} \mu_k, 1 \right\rangle \right] + \mathrm{c}
\end{aligned}
$$

# EM for the Gaussian HMM

## Sufficient statistics

$$
\begin{aligned}
\mathbb{E}_{q(z)}[\log p(x, z, \Theta)] &= \sum_{t=1}^{T} q_{t,k} \left[-\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}(x_t - \mu_k)^\top Q_k^{-1}(x_t - \mu_k)\right] + \mathrm{c} \\
&= \sum_{t=1}^{T} q_{t,k} \left[-\tfrac{1}{2}\log|Q_k| - \tfrac{1}{2}x_t^\top Q_k^{-1} x_t + \mu_k^\top Q_k^{-1} x_t - \tfrac{1}{2}\mu_k^\top Q_k^{-1}\mu_k\right] + \mathrm{c} \\
&= \sum_{t=1}^{T} q_{t,k} \left[\left\langle -\tfrac{1}{2}\log|Q_k|, 1\right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, x_t x_t^\top\right\rangle + \left\langle \mu_k^\top Q_k^{-1}, x_t\right\rangle + \left\langle -\tfrac{1}{2}\mu_k^\top Q_k^{-1}\mu_k, 1\right\rangle\right] + \mathrm{c} \\
&= \left\langle -\tfrac{1}{2}\log|Q_k|, N_k\right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1}\right\rangle + \left\langle \mu_k^\top Q_k^{-1}, \bar{\psi}_{k,2}\right\rangle + \left\langle -\tfrac{1}{2}\mu_k^\top Q_k^{-1}\mu_k, \bar{\psi}_{k,3}\right\rangle + \mathrm{c}
\end{aligned}
$$

where

$$
T_k = \sum_{t=1}^{T} q_{t,k} \qquad \mathbf{t}_{k,1} = \sum_{t=1}^{T} q_{t,k}\, x_t \qquad \mathbf{t}_{k,2} = \sum_{t=1}^{T} q_{t,k}\, x_t x_t^\top
$$

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,

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

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

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

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

# 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:

$$q_{t,k} \leftarrow p(z_t = k \mid x_{1:T}, \Theta)$$

  via the **forward-backward algorithm.**

  Compute **weighted sums of sufficient statistics**:

$$T_k = \sum_{t=1}^{T} q_{t,k} \qquad \mathbf{t}_{k,1} = \sum_{t=1}^{T} q_{t,k}\, x_t \qquad \mathbf{t}_{k,2} = \sum_{t=1}^{T} q_{t,k}\, x_t x_t^\top$$

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

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

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

# 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 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:

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

  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.

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