Machine Learning Methods for Neural Data Analysis

Lecture 12: EM and Hidden Markov Models

Scott Linderman

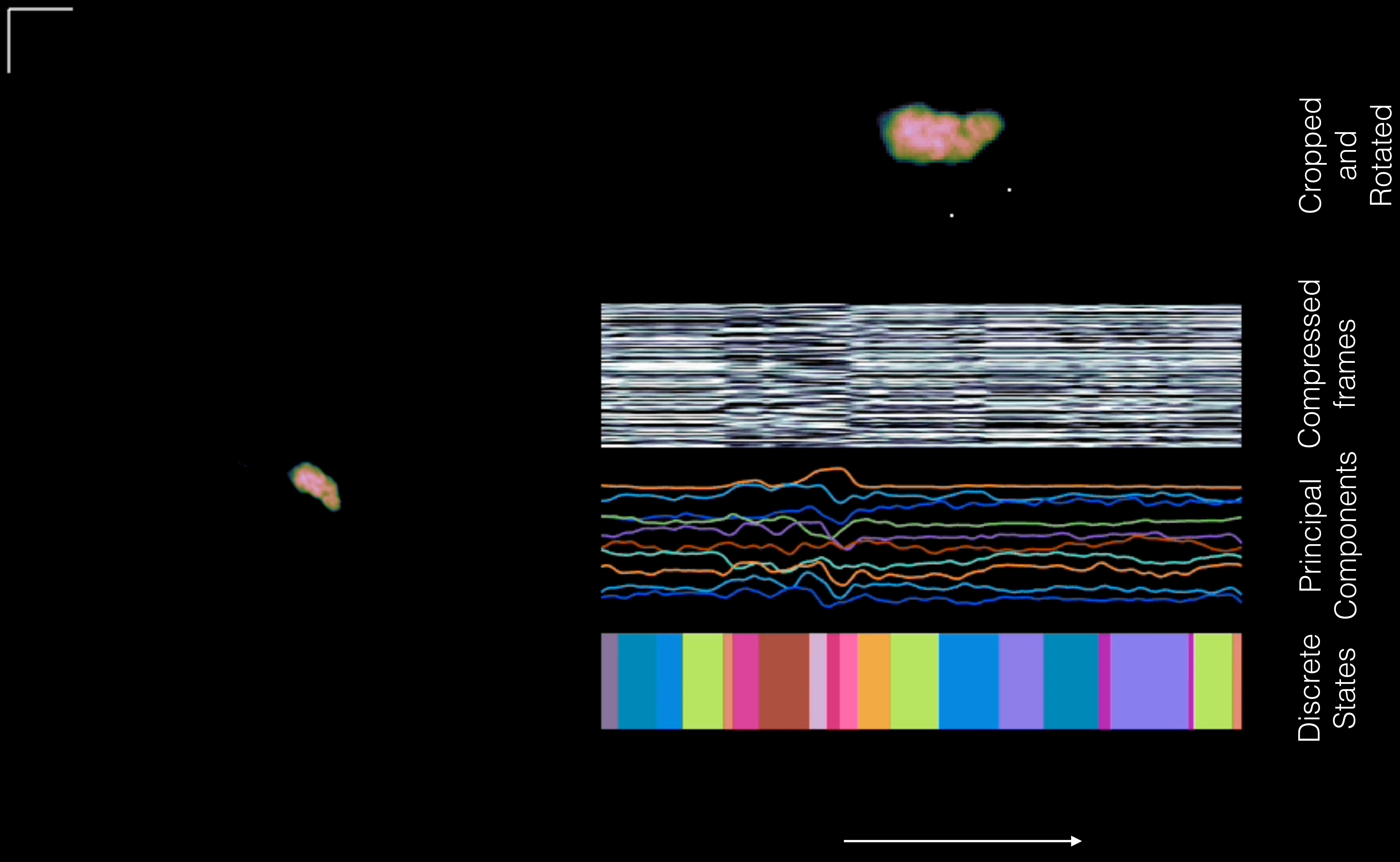
STATS 220/320 (NBIO220, CS339N). Winter 2021.

Announcements

- Apologies for the delay in proposal feedback. Finishing this afternoon.
- **COSYNE** is happening this week! $5 registration and lots of great talks and posters.

Agenda

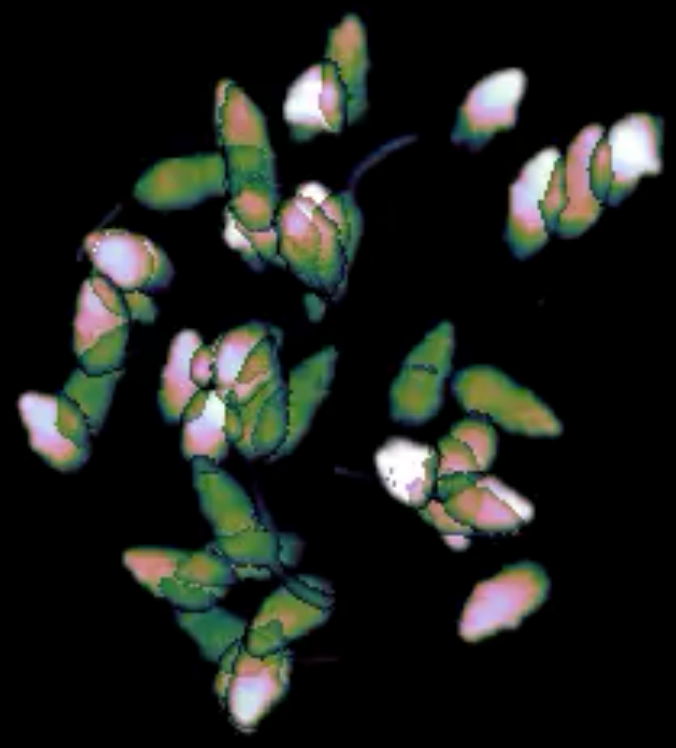
- Expectation-maximization for Gaussian mixture models using expected sufficient statistics
- Hidden Markov models and the forward-backward algorithm



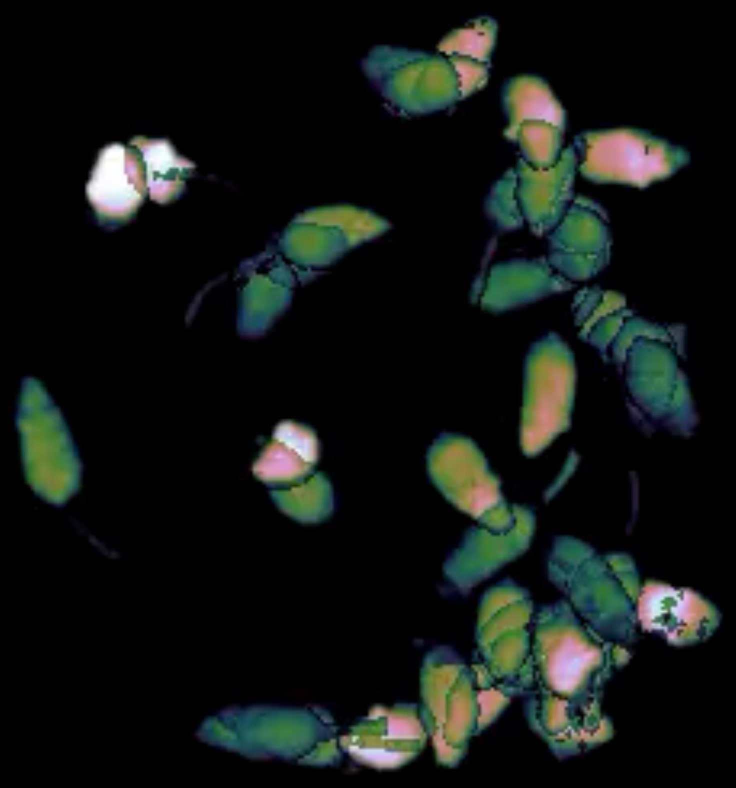
Wiltschko et al. (2015)
Markowitz et al. (2018)

Motivating Example: summarizing videos with behavioral states

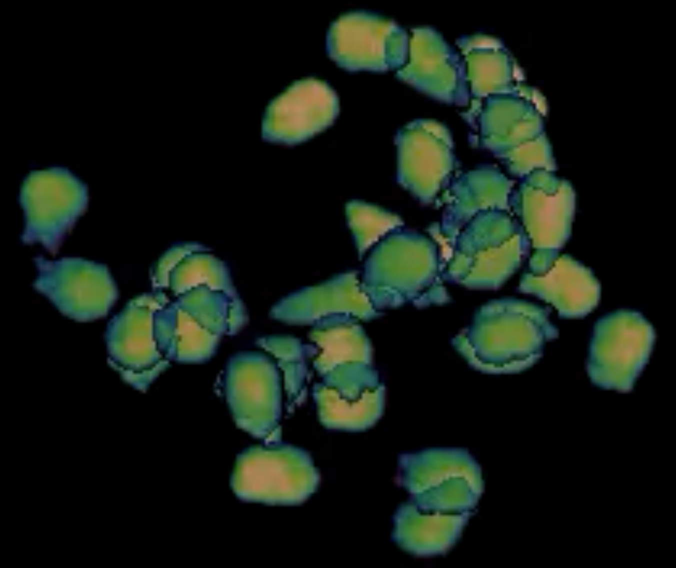
Rear down



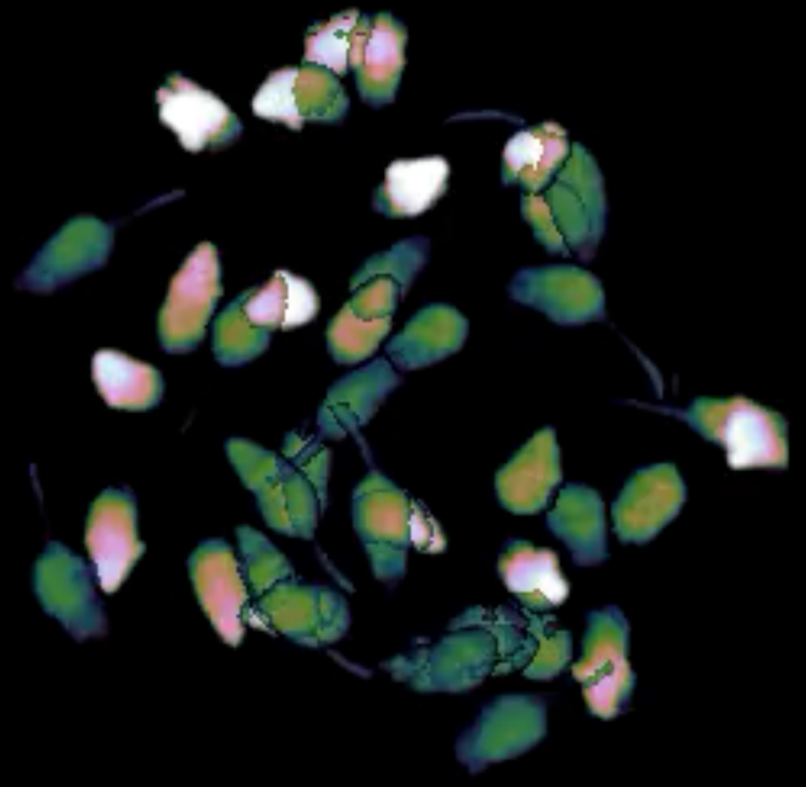
Walk forward



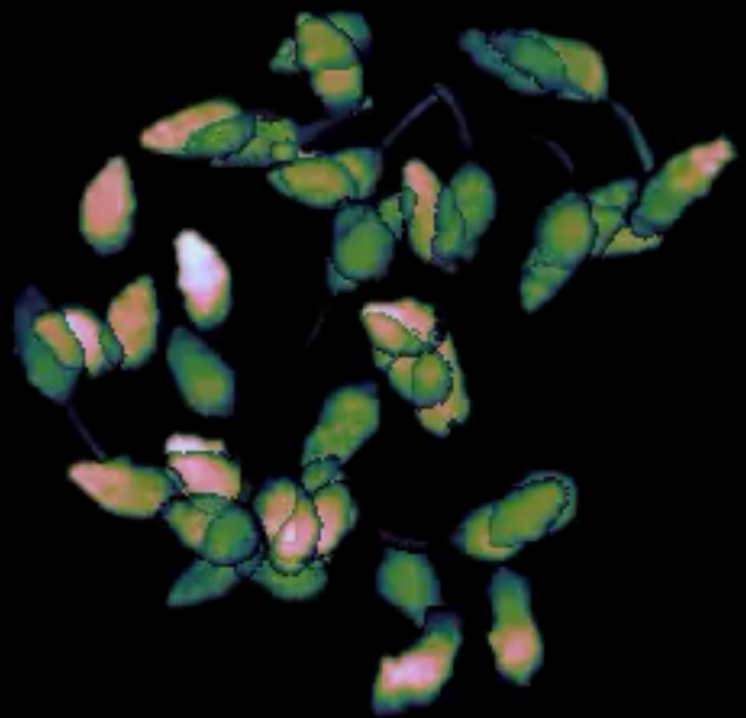
Grooming



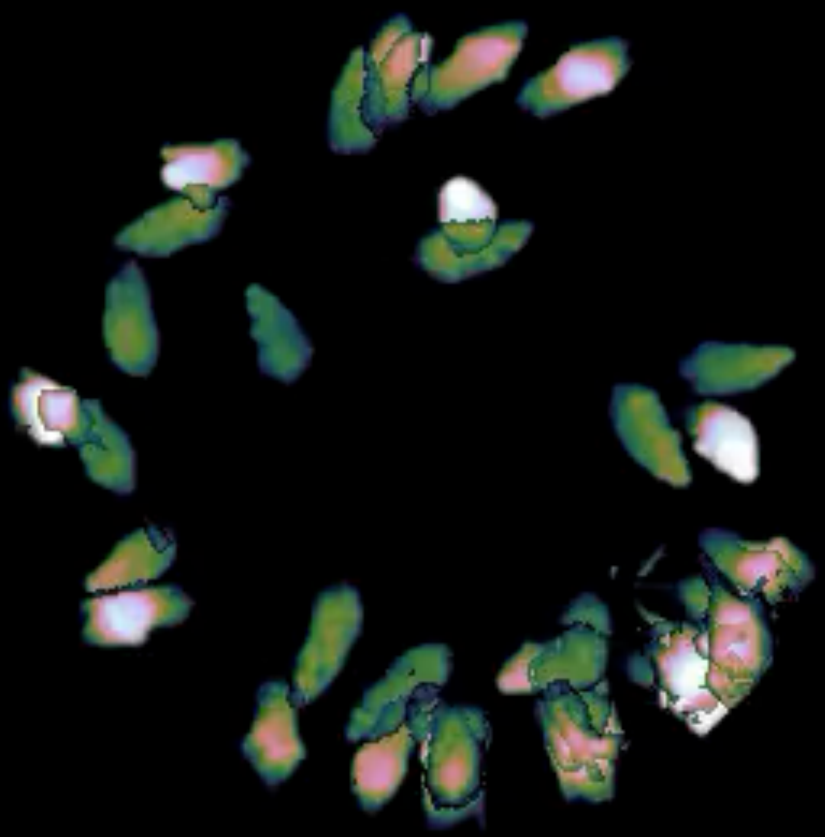
Scrunch



Rear up



Jump



Wiltschko et al, 2015

The Expectation-Maximization (EM) algorithm

Coordinate ascent on parameters and latent variable posteriors

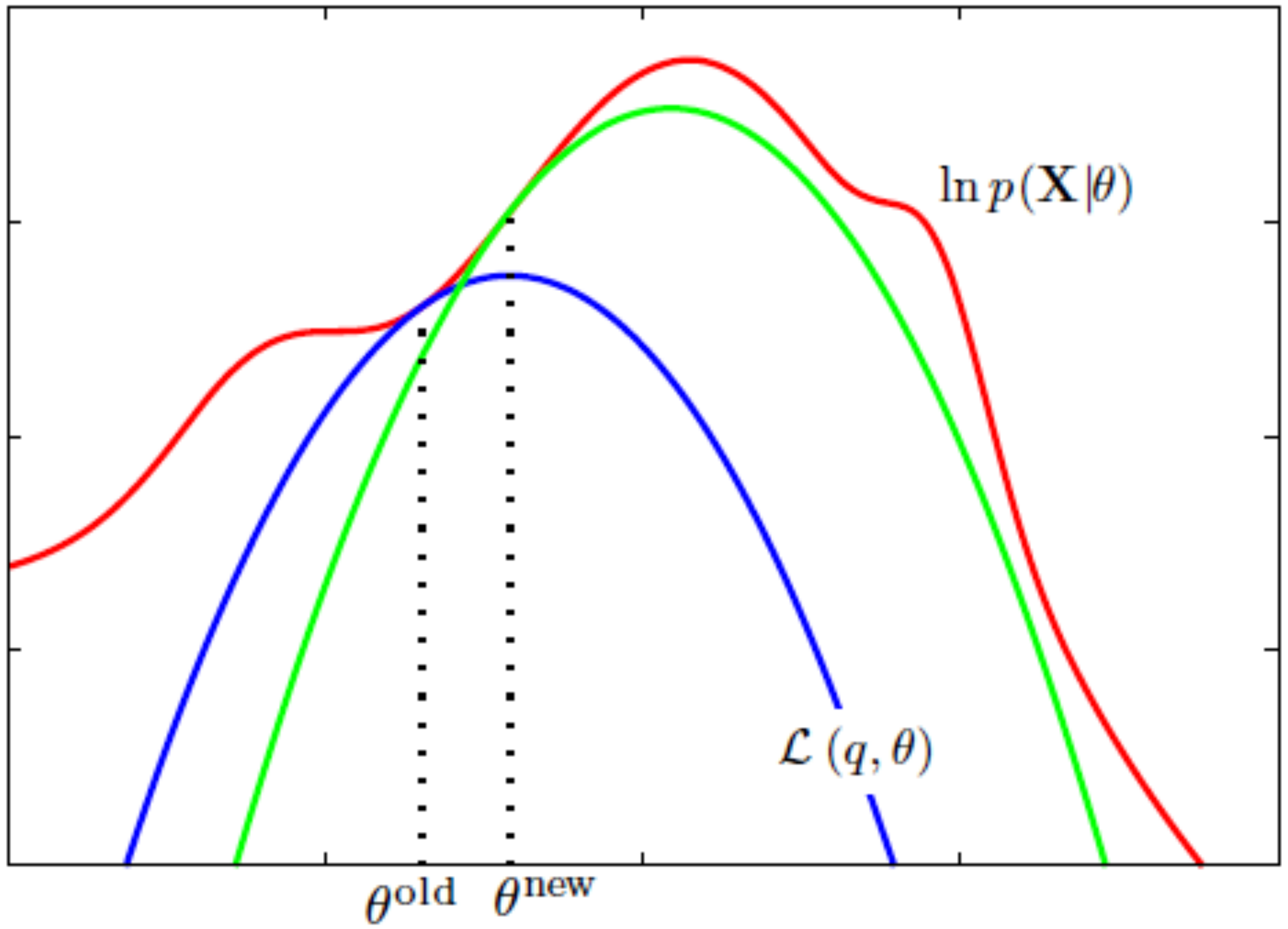
- **M-step**: Maximize the expected log probability

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

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

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

- EM converges to **local optima** of the marginal distribution.



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

The Gaussian Mixture Model

Consider a Gaussian mixture model with discrete states $z_t \in \{1,\ldots,K\}$ and data $x_t \in \mathbb{R}$:

$$\begin{aligned} z_t &\sim \mathrm{Cat}(\pi), \\ x_t \mid z_t &\sim \mathcal{N}(b_{z_t}, Q_{z_t}) \end{aligned}.$$

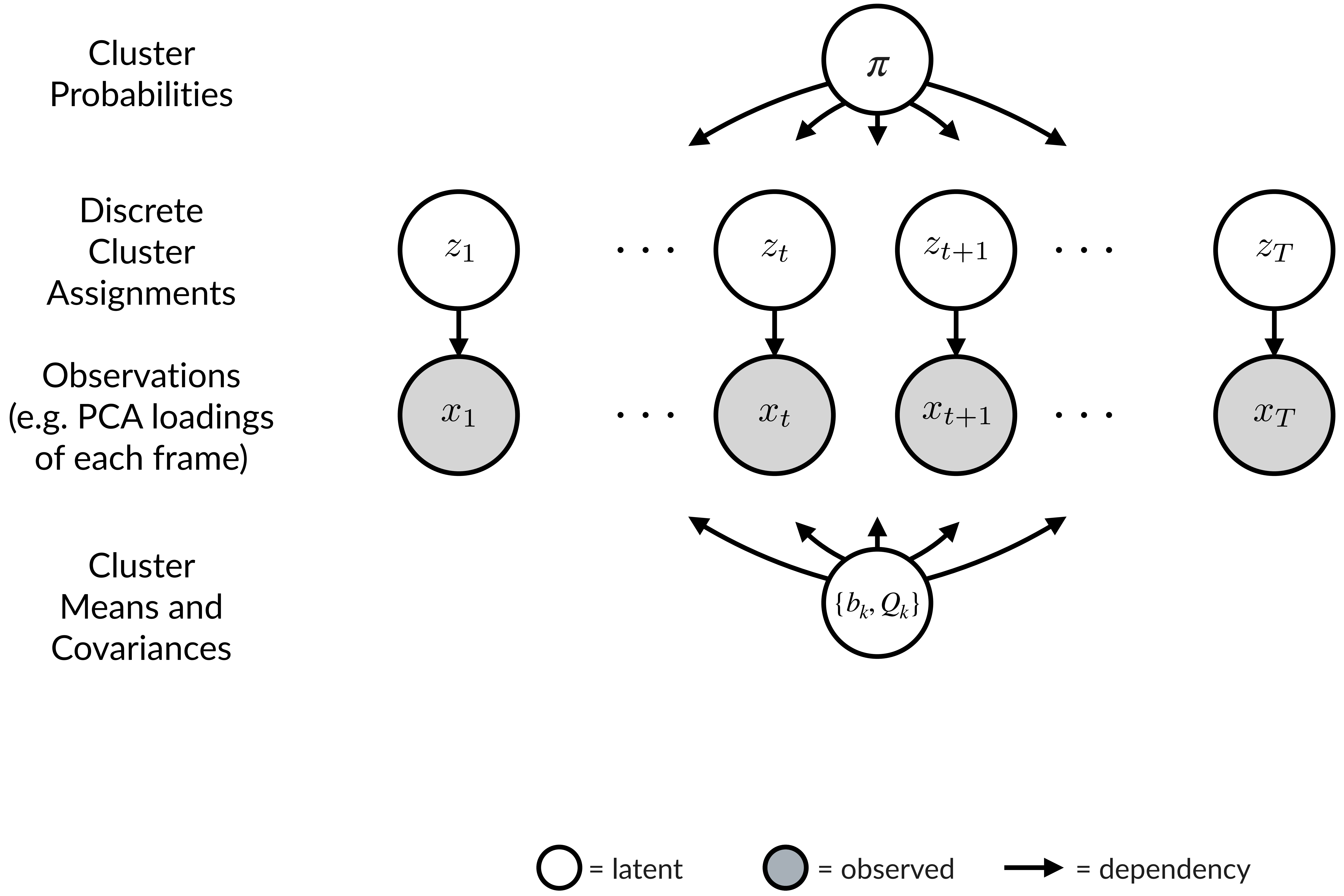
Its parameters are $\Theta = \pi, \{b_k, Q_k\}_{k=1}^K$.

The **joint probability** factors into a product over time bins,

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

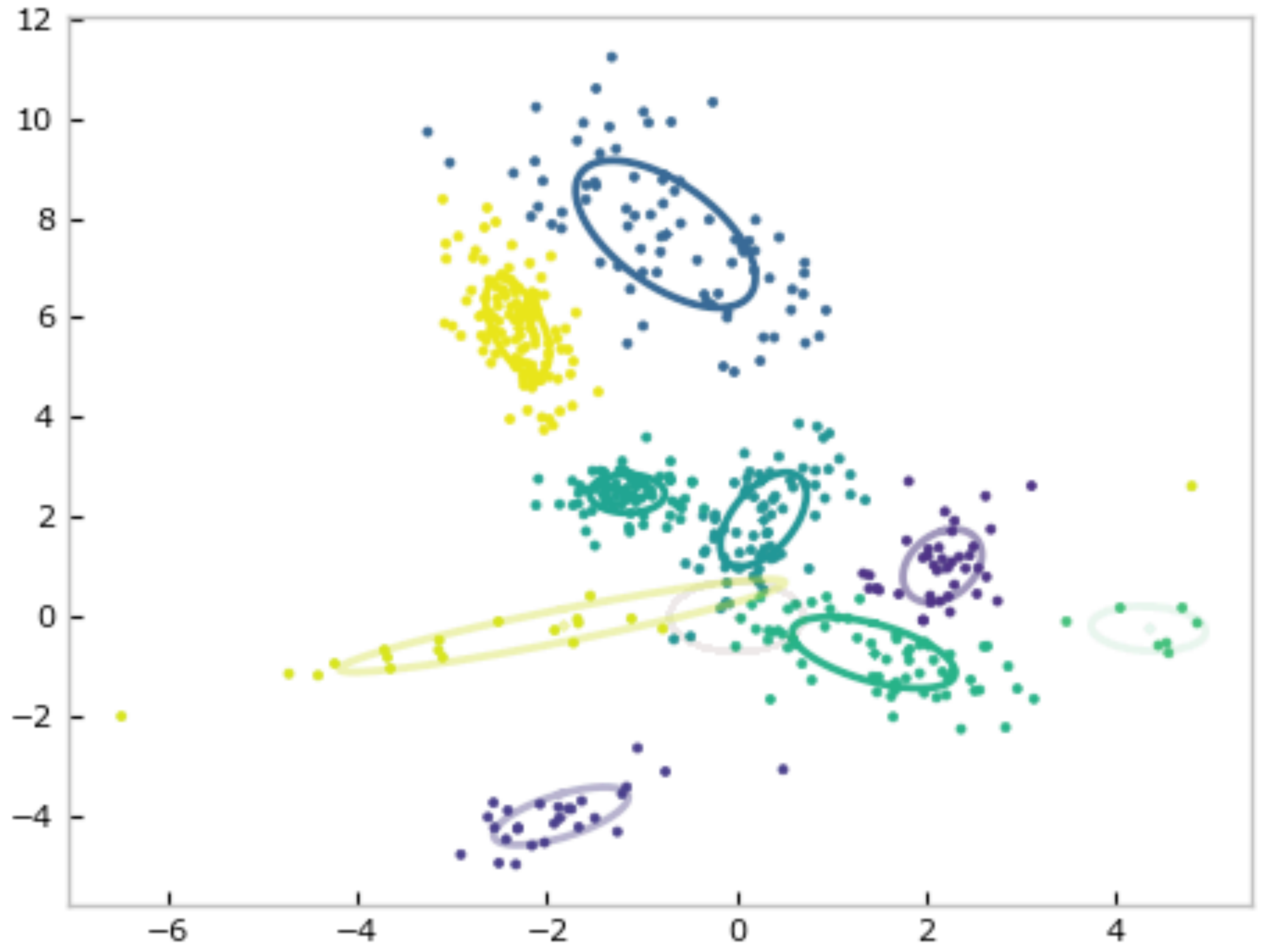
The Gaussian Mixture Model

Graphical Model



The Gaussian Mixture Model

Example draw from a 2D GMM with 10 clusters



EM for the Gaussian mixture model

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

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

- **M-step**: Update the parameters. Let $N_k = \sum_{t=1}^{T} q(z_t = k)$, then

$$\pi_k \leftarrow \frac{N_k}{T}, \qquad b_k \leftarrow \frac{1}{N_k} \sum_{t=1}^{T} q(z_t = k)\, x_t, \qquad Q_k \leftarrow \frac{1}{N_k} \sum_{t=1}^{T} q(z_t = k)\, (x_t - b_k)(x_t - b_k)^\top .$$

i.e. set the parameters to their weighted averages.

EM for GMMs using ESS, ASAP

EM for the Gaussian mixture model

The M-step

Recall the M-step:

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

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 mixture model

The M-step

Recall the M-step:

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

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 mixture model

The M-step

Recall the M-step:

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

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 mixture model

The M-step

Recall the M-step:

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

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 mixture model

Expected sufficient statistics

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

EM for the Gaussian mixture model

Expected sufficient statistics

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

EM for the Gaussian mixture model

Expected sufficient statistics

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

EM for the Gaussian mixture model

Expected sufficient statistics

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

where

$$
N_k = \sum_{t=1}^{T} q(z_t = k) \qquad \bar{\psi}_{k,1} = \sum_{t=1}^{T} q(z_t = k)\, x_t x_t^\top \qquad \bar{\psi}_{k,2} = \sum_{t=1}^{T} q(z_t = k)\, x_t \qquad \bar{\psi}_{k,3} = \sum_{t=1}^{T} q(z_t = k)
$$

are the **expected sufficient statistics.**

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

The objective we're trying to maximize is,

$$\mathcal{J}(b_k, Q_k) = \left\langle -\frac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\frac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle b_k^\top Q_k^{-1}, \bar{\psi}_{k,2} \right\rangle + \left\langle -\frac{1}{2}b_k^\top Q_k^{-1} b_k, \bar{\psi}_{k,3} \right\rangle + \mathrm{c}$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

The objective we're trying to maximize is,

$$\mathcal{J}(b_k, Q_k) = \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle b_k^\top Q_k^{-1}, \bar{\psi}_{k,2} \right\rangle + \left\langle -\tfrac{1}{2}b_k^\top Q_k^{-1} b_k, \bar{\psi}_{k,3} \right\rangle + \mathrm{c}$$

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

$$\frac{\partial}{\partial b_k}\mathcal{J}(b_k, Q_k) = Q_k^{-1}\bar{\psi}_k^{(2)} - Q_k^{-1} b_k \bar{\psi}_k^{(3)} = 0$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

The objective we're trying to maximize is,

$$\mathcal{J}(b_k, Q_k) = \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle b_k^\top Q_k^{-1}, \bar{\psi}_{k,2} \right\rangle + \left\langle -\tfrac{1}{2}b_k^\top Q_k^{-1} b_k, \bar{\psi}_{k,3} \right\rangle + \mathrm{c}$$

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

$$\frac{\partial}{\partial b_k}\mathcal{J}(b_k, Q_k) = Q_k^{-1}\bar{\psi}_k^{(2)} - Q_k^{-1} b_k \bar{\psi}_k^{(3)} = 0$$

$$\implies b_k^\star = \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}} = \frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)\, x_t$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

Plug in the optimum

$$\mathcal{J}(b_k^\star, Q_k) = \left\langle -\frac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\frac{1}{2} Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle \frac{\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} Q_k^{-1}, \bar{\psi}_{k,2} \right\rangle + \left\langle -\frac{1}{2} \frac{\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} Q_k^{-1} \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}}, \bar{\psi}_{k,3} \right\rangle + \mathrm{c}$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

Plug in the optimum

$$\begin{aligned}
\mathcal{J}(b_k^\star, Q_k) &= \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle \frac{\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} Q_k^{-1}, \bar{\psi}_{k,2} \right\rangle + \left\langle -\frac{1}{2}\frac{\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} Q_k^{-1} \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}}, \bar{\psi}_{k,3} \right\rangle + \mathrm{c} \\
&= \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle Q_k^{-1}, \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \left\langle -\frac{1}{2}Q_k^{-1}, \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \mathrm{c}
\end{aligned}$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

Plug in the optimum

$$\begin{aligned}
\mathcal{J}(b_k^\star, Q_k) &= \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle \frac{\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} Q_k^{-1}, \bar{\psi}_{k,2} \right\rangle + \left\langle -\frac{1}{2}\frac{\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} Q_k^{-1} \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}}, \bar{\psi}_{k,3} \right\rangle + \mathrm{c} \\
&= \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} \right\rangle + \left\langle Q_k^{-1}, \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \mathrm{c} \\
&= \left\langle -\tfrac{1}{2}\log|Q_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}Q_k^{-1}, \bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \mathrm{c}
\end{aligned}$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

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

$$\mathcal{J}(b_k^\star, \Lambda_k) = \left\langle \frac{1}{2} \log |\Lambda_k|, N_k \right\rangle + \left\langle -\frac{1}{2}\Lambda_k, \bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \mathrm{c}$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

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

$$\mathcal{J}(b_k^\star, \Lambda_k) = \left\langle \tfrac{1}{2} \log |\Lambda_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}\Lambda_k, \bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \mathrm{c}$$

Taking the partial derivative wrt Λ_k and setting equal to zero,

$$\frac{\partial}{\partial \Lambda_k}\mathcal{J}(b_k^\star, \Lambda_k) = \frac{N_k}{2}\Lambda_k^{-1} - \frac{1}{2}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right) = 0$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

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

$$\mathcal{J}(b_k^\star, \Lambda_k) = \left\langle \tfrac{1}{2}\log|\Lambda_k|, N_k \right\rangle + \left\langle -\tfrac{1}{2}\Lambda_k, \bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right\rangle + \mathrm{c}$$

Taking the partial derivative wrt Λ_k and setting equal to zero,

$$\frac{\partial}{\partial \Lambda_k}\mathcal{J}(b_k^\star, \Lambda_k) = \frac{N_k}{2}\Lambda_k^{-1} - \frac{1}{2}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right) = 0$$

$$\implies (\Lambda_k^{-1})^\star = Q_k^\star = \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right)$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

The result makes sense…

$$
\begin{aligned}
Q_k^\star &= \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right) \\
&= \left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)x_t x_t^\top\right) - \left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)x_t\right)\left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)x_t^\top\right)
\end{aligned}
$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

The result makes sense…

$$Q_k^\star = \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right)$$

$$= \left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)x_t x_t^\top\right) - \left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)x_t\right)\left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k)x_t^\top\right)$$

$$\to \mathbb{E}[xx^\top \mid z = k] - \mathbb{E}[x \mid z = k]\mathbb{E}[x^\top \mid z = k]$$

EM for the Gaussian mixture model

Solving for the optimal Gaussian parameters

The result makes sense…

$$
\begin{aligned}
Q_k^\star &= \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right) \\
&= \left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k) x_t x_t^\top\right) - \left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k) x_t\right)\left(\frac{1}{N_k}\sum_{t=1}^{T} q(z_t = k) x_t^\top\right) \\
&\to \mathbb{E}[xx^\top \mid z = k] - \mathbb{E}[x \mid z = k]\mathbb{E}[x^\top \mid z = k] \\
&= \mathrm{Var}[x \mid z = k]
\end{aligned}
$$

EM for the Gaussian mixture model

In summary…

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

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

Compute **expected sufficient statistics**:

$$N_k = \sum_{t=1}^{T} q(z_t = k) \qquad \psi_{k,1} = \sum_{t=1}^{T} q(z_t = k)\, x_t x_t^\top \qquad \psi_{k,2} = \sum_{t=1}^{T} q(z_t = k)\, x_t \qquad \psi_{k,3} = \sum_{t=1}^{T} q(z_t = k)$$

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

$$\pi_k \leftarrow \frac{N_k}{T}, \qquad b_k \leftarrow \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}} \qquad Q_k \leftarrow \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right)$$

Stochastic EM for the Gaussian mixture model

- Grab a **mini-batch** of M randomly chosen data points.

- **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 **expected sufficient statistics** for the mini-batch and **rescale** as if they came from the whole dataset:

$$\tilde{N}_k = \frac{T}{M} \sum_{m=1}^{M} q(z_m = k) \qquad \tilde{\psi}_{k,1} = \frac{T}{M} \sum_{m=1}^{M} q(z_m = k)\, x_m x_m^\top \qquad \tilde{\psi}_{k,2} = \frac{T}{M} \sum_{m=1}^{M} q(z_m = k)\, x_m \qquad \tilde{\psi}_{k,3} = \frac{T}{M} \sum_{m=1}^{M} q(z_m = k)$$

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

$$N_k \leftarrow (1-\alpha) N_k + \alpha \tilde{N}_k \qquad \bar{\psi}_{k,1} \leftarrow (1-\alpha)\bar{\psi}_{k,1} + \alpha \tilde{\psi}_{k,1} \qquad \bar{\psi}_{k,2} \leftarrow (1-\alpha)\bar{\psi}_{k,2} + \alpha \tilde{\psi}_{k,2} \qquad \bar{\psi}_{k,3} \leftarrow (1-\alpha)\bar{\psi}_{k,3} + \alpha \tilde{\psi}_{k,3}$$

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

$$\pi_k \leftarrow \frac{N_k}{T}, \qquad b_k \leftarrow \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}} \qquad Q_k \leftarrow \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}} \right)$$

Hidden Markov Models

The Gaussian HMM

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

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

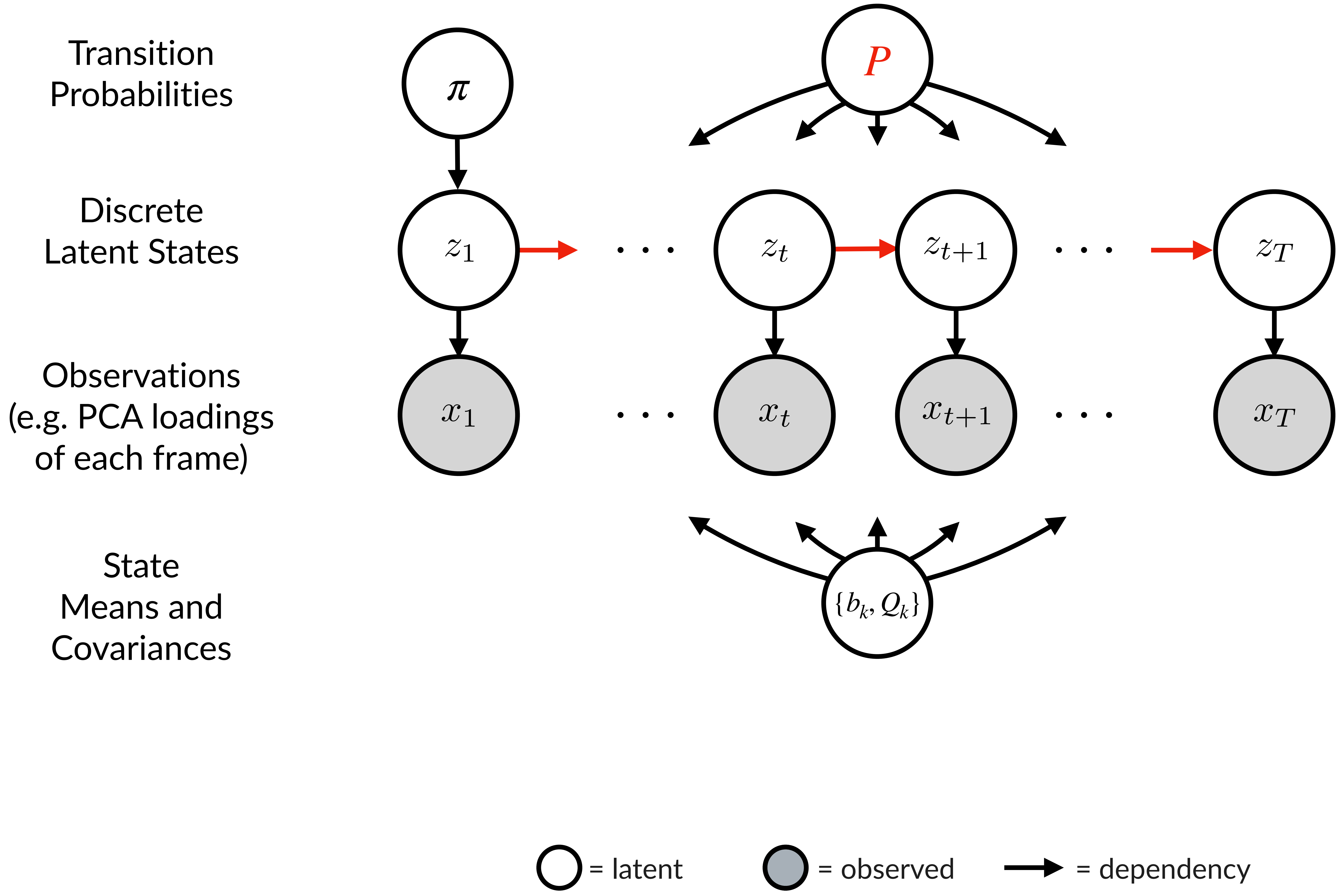
Its parameters are $\Theta = \pi, P, \{b_k, Q_k\}_{k=1}^{K}$ where $P \in [0,1]^{K \times K}$ is a row-stochastic **transition matrix.**

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

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

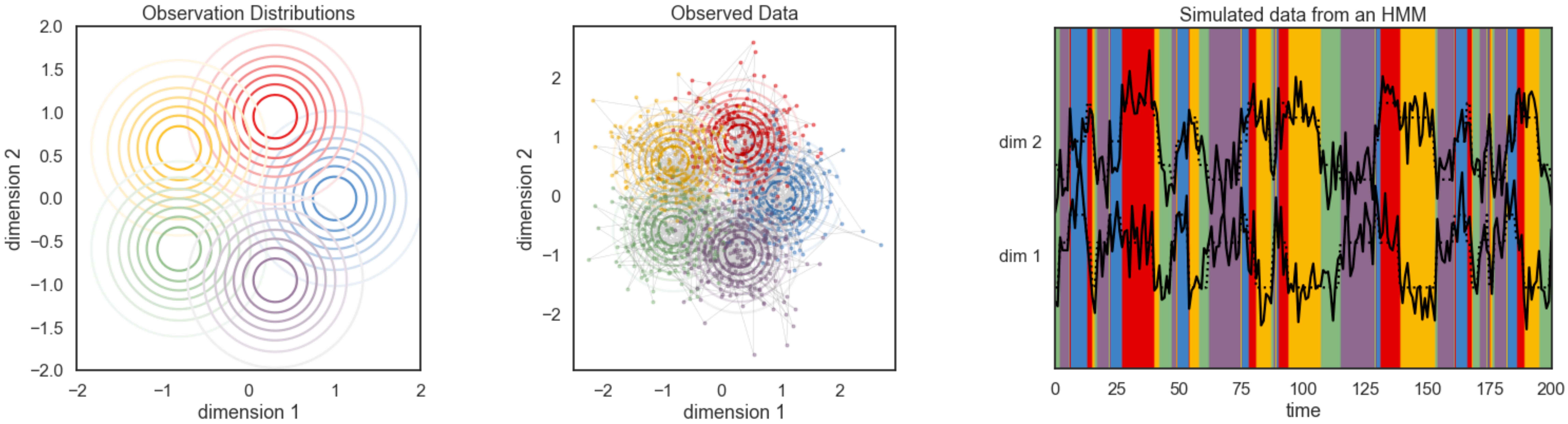
The Gaussian HMM

Graphical Model



The Gaussian HMM

Example draw from a 2D Gaussian HMM with 5 clusters



EM for the Gaussian HMM

The posterior is a little trickier…

- **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 the marginal likelihood

- 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 the marginal likelihood

- 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 the marginal likelihood

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

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

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

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

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

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 α_t and/or β_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

$$
\begin{aligned}
\log p(x \mid \Theta) &= \log \sum_{z_1=1}^{K} \cdots \sum_{z_T=1}^{K} \left[p(z_1) \prod_{t=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=1}^{T} p(x_t \mid z_t) \right] \\
&= \log \sum_{z_T=1}^{K} \alpha_T(z_T)\, p(x_T \mid z_T) \\
&= \log \prod_{t=1}^{T} A_t = \sum_{t=1}^{T} \log A_t
\end{aligned}
$$

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

EM for the Gaussian HMM

Putting it all together

- **E-step**: Run the forward-backward algorithm to compute

$$q(z_t = k) \leftarrow p(z_t = k \mid x_{1:T}, \Theta) = \frac{\alpha_{tk} \ell_{tk} \beta_{tk}}{\sum_{j=1}^{K} \alpha_{tj} \ell_{tj} \beta_{tj}}$$ and the marginal log likelihood $\log p(x_{1:T} \mid \Theta)$.

  Then compute the expected sufficient statistics:

$$N_k = \sum_{t=1}^{T} q(z_t = k) \qquad \bar{\psi}_{k,1} = \sum_{t=1}^{T} q(z_t = k)\, x_t x_t^\top \qquad \bar{\psi}_{k,2} = \sum_{t=1}^{T} q(z_t = k)\, x_t \qquad \bar{\psi}_{k,3} = \sum_{t=1}^{T} q(z_t = k)$$

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

$$b_k \leftarrow \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}} \qquad Q_k \leftarrow \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right)$$

- ***Note:** You can use the forward-backward algorithm to compute $q(z_t = i, z_{t+1} = j)$ too. That's all you need to update the transition matrix P.*

EM for the Exponential Family HMMs

Same thing, different sufficient statistics

- **E-step**: Run the forward-backward algorithm to compute

$$q(z_t = k) \leftarrow p(z_t = k \mid x_{1:T}, \Theta) = \frac{\alpha_{tk}\ell_{tk}\beta_{tk}}{\sum_{j=1}^{K} \alpha_{tj}\ell_{tj}\beta_{tj}}$$ and the marginal log likelihood $\log p(x_{1:T} \mid \Theta)$.

  Then compute the expected sufficient statistics:

$$N_k = \sum_{t=1}^{T} q(z_t = k) \qquad \bar{\psi}_{k,j} = \sum_{t=1}^{T} q(z_t = k) f_j(x_t)$$ for each sufficient statistic $f_1, \ldots, f_J$.

- **M-step**: Update the parameters using the expected sufficient statistics.

$$\Theta \leftarrow \arg\max \mathbb{E}_{q(z)}[\log p(x, z, \Theta)] \quad \text{using} \quad \{N_k, \{\bar{\psi}_{kj}\}_{j=1}^{J}\}_{k=1}^{K}$$

- Works for **autoregressive, Bernoulli, binomial, categorical, multinomial, Poisson, etc**. observation distributions, and with minor adjustment, for **compound distributions** like negative binomial and Student's T also.

Try it out!

SSM: Bayesian learning and inference for state space models

build passing

https://github.com/lindermanlab/ssm/

This package has fast and flexible code for simulating, learning, and performing inference in a variety of state space models. Currently, it supports:

- Hidden Markov Models (HMM)
- Auto-regressive HMMs (ARHMM)
- Input-output HMMs (IOHMM)
- Linear Dynamical Systems (LDS)
- Switching Linear Dynamical Systems (SLDS)
- Recurrent SLDS (rSLDS)
- Hierarchical extensions of the above
- Partial observations and missing data

We support the following observation models:

- Gaussian
- Student's t
- Bernoulli
- Poisson
- Categorical
- Von Mises

```
from ssm.models import HMM
T = 100  # number of time bins
K = 5    # number of discrete states
D = 2    # dimension of the observations

# make an hmm and sample from it
hmm = HMM(K, D, observations="gaussian")
z, y = hmm.sample(T)
```

Fitting an HMM is simple.

```
test_hmm = HMM(K, D, observations="gaussian")
test_hmm.fit(y)
zhat = test_hmm.most_likely_states(y)
```

Conclusion

- EM for mixture models (with exponential family likelihoods) amounts to **computing cluster assignment probabilities** and **expected sufficient statistics**, then updating parameters based on them.
- **Stochastic EM** generalizes this approach to work with mini-batches of data.
- Hidden Markov models (HMMs) are just mixture models with dependencies across time.
- The EM algorithm is nearly the same, but we use the **forward-backward algorithm** to compute latent state probabilities and expected sufficient stats.

Further reading

- For more on the **EM** and **Stochastic EM** algorithms:
  - Bishop (2006). Pattern Recognition and Machine Learning, Ch 9. *[free online]*
  - Cappé, Olivier, and Eric Moulines. 2009. “On-Line Expectation-Maximization Algorithm for Latent Data Models.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 71 (3): 593–613.
- For more on **Hidden Markov Models**:
  - Barber, David. 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press. Ch 23. *[free online]*
  - Rabiner, Lawrence R. 1990. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” In *Readings in Speech Recognition*, edited by Alex Waibel and Kai-Fu Lee, 267–96. San Francisco: Morgan Kaufmann.
- Some code: https://github.com/lindermanlab/ssm

EM for the Gaussian mixture model

Solving for the optimal categorical parameters

As a function of the categorical parameters π, and accounting for the normalization constraint, the Lagrangian is,

$$\begin{aligned}
\mathcal{J}(\pi) &= \mathbb{E}_{q(z)}\left[\log p(x, z, \Theta)\right] - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) \\
&= \mathbb{E}_{q(z)}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{I}[z_t = k] \log \pi_k\right] - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} q(z_t = k) \log \pi_k - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)
\end{aligned}$$

EM for the Gaussian mixture model

Solving for the optimal categorical parameters

Taking the derivative with respect to π_k yields,

$$\frac{\partial}{\partial \pi_k} \mathcal{J}(\pi) = \pi_k^{-1} \sum_{t=1}^{T} q(z_t = k) - \lambda = 0 \implies \pi_k^\star = \lambda^{-1} N_k$$

Imposing the normalization constraint yields $\pi_k^\star = \frac{N_k}{\sum_k N_k} = \frac{N_k}{T}$.

EM using expected sufficient statistics

- **E-step**: First, compute the posterior probabilities:

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

Then compute the expected sufficient statistics:

$$N_k = \sum_{t=1}^{T} q(z_t = k) \qquad \bar{\psi}_{k,1} = \sum_{t=1}^{T} q(z_t = k)\, x_t x_t^\top \qquad \bar{\psi}_{k,2} = \sum_{t=1}^{T} q(z_t = k)\, x_t \qquad \bar{\psi}_{k,3} = \sum_{t=1}^{T} q(z_t = k)$$

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

$$\pi_k \leftarrow \frac{N_k}{T}, \qquad b_k \leftarrow \frac{\bar{\psi}_{k,2}}{\bar{\psi}_{k,3}} \qquad Q_k \leftarrow \frac{1}{N_k}\left(\bar{\psi}_{k,1} - \frac{\bar{\psi}_{k,2}\bar{\psi}_{k,2}^\top}{\bar{\psi}_{k,3}}\right)$$