Deep Generative Models: Variational Inference and Expectation Maximization

Fall Semester 2026

René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE



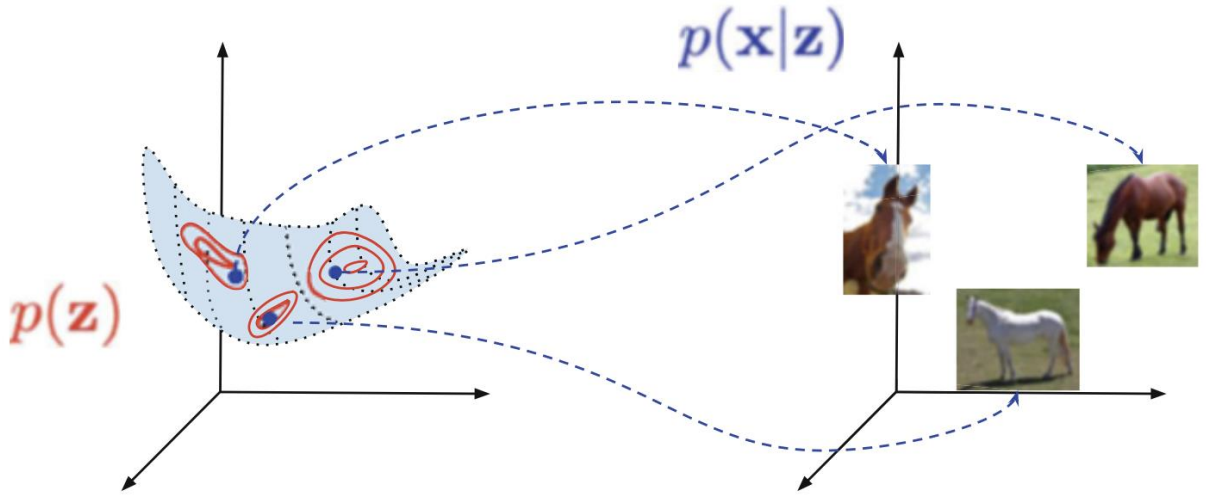
Outline

- **Latent Variable Models**
  - Probabilistic PCA
  - Beyond PPCA?
- Variational Inference
  - Principle
  - Derivation
- Expectation Maximization
  - Derivation
  - EM for a Mixture of Gaussians

Latent Variable Models

- X = observed variable
- Z = latent variable

- $\boldsymbol{z} \sim p(\boldsymbol{z})$
- $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{z})$



A latent variable model and a generative process. Note the low-dimensional manifold (here 2D) embedded in the high-dimensional space (here 3D)

- Factorization of the joint model

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$$

- Marginalization of the model

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Probabilistic Principal Component Analysis (PPCA)

- Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z} \mid \mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma^2 \boldsymbol{I})$ be independent.
- In PPCA, the dependency between $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{z} \in \mathbb{R}^d$, $d \ll D$, is defined by a linear Gaussian additive model

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{\varepsilon}$$

- **Theorem.** Let $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_{\mathrm{N}}$ be, respectively, the ML estimates for the mean and the covariance of the data. Let U_1 be the matrix with the top d eigenvectors of $\boldsymbol{\Sigma}_{\mathrm{N}}$, Λ_1 be the matrix with the corresponding top d eigenvalues, and λ_i be the ith largest eigenvalue of $\boldsymbol{\Sigma}_{\mathrm{N}}$. The ML estimates for the PPCA parameters $(\boldsymbol{b}, \boldsymbol{W}, \sigma)$ are given by

$$\boldsymbol{b} = \boldsymbol{\mu}_N,\ \boldsymbol{W} = \boldsymbol{U}_1(\boldsymbol{\Lambda}_1 - \sigma^2 I)^{1/2}\mathbf{R} \text{ and } \sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D} \lambda_i$$

where $\mathbf{R} \in \mathbb{R}^{d \times d}$ is an arbitrary orthogonal matrix.

Outline

- Latent Variable Models
  - Probabilistic PCA
  - Beyond PPCA?

- **Variational Inference**
  - Principle
  - Derivation

- Expectation Maximization
  - Derivation
  - EM for a Mixture of Gaussians

What about Latent Variable Models other than PPCA

- We would like to learn the parameters of the model via Maximum Likelihood.
- Since $\boldsymbol{z}$ is latent, we need to marginalize $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$, i.e.

$$\max_\theta \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i) = \max_\theta \sum_{i=1}^{N} \log \int p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

- For PPCA, we could compute $p(\boldsymbol{x})$ in closed form and solve for θ analytically.
- In general, we need many samples of $\boldsymbol{z}$ for each $\boldsymbol{x}_i$ to approximate the integral.

$$\max_\theta \sum_{i=1}^{N} \log \sum_{j=1} p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}_j)$$

- We address this challenge using **Variational Inference**, which we describe next.

Variational Inference

- **Old ML learning objective**:

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i) = \max_{\theta} \sum_{i=1}^{N} \log \int p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$$

- **Theorem**: the log likelihood can be written as

$$\log p_\theta(\boldsymbol{x}) = \max_{q(\cdot|\boldsymbol{x}):\, q(\cdot|\boldsymbol{x}) \geq 0,\, \int q(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} = 1} \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}.$$

and the maximizing distribution is given by $q^*(\boldsymbol{z} \mid \boldsymbol{x}) = p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$

- **New ML learning objective**:

$$\max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i), \forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- Before going through the derivation, what is the gain here?

Variational Inference

- **New ML learning objective**:

$$\max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- **Expectation Maximization**:
  - If finding q^* given θ is easy, then we can alternate between finding q given θ and vice versa
  - Here $q^*(\boldsymbol{z} \mid \boldsymbol{x}) = p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$. Thus, if the integral w.r.t. $\boldsymbol{z}$ is easy to evaluate for a fixed θ, we can alternate between computing the integral (E-step) and maximizing w.r.t. θ (M-step).

- **Variational AutoEncoders:** represent $q(\cdot \mid x_i)$ with a function with parameters ψ (e.g., a DNN) that maps x_i to a distribution $q_\psi(\cdot \mid x_i)$, and find (θ, ψ) via SGD

$$\max_{\theta} \max_{\psi} \sum_{i=1}^{N} \int q_\psi(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q_\psi(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

Variational Inference

- **New ML learning objective**:

$$\max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- We will use VI for many latent variable models
  - Mixtures of Gaussians (a.k.a. Gaussian Mixture Models) → EM
  - Probabilistic Principal Component Analysis (PPCA) → EM
  - Mixtures of PPCA → EM
  - Variational Auto-Encoders (VAE) → VI
  - Diffusion models → VI
  - …

Variational Inference: Derivation

- Proof: Let $q(\boldsymbol{z}|\boldsymbol{x})$ be the variational distribution. Observe that

$$\log p_\theta(\boldsymbol{x}) = \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log p_\theta(\boldsymbol{x}) dz = \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

$$= \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

$$= \boxed{\int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}} + \boxed{\int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}}$$

Evidence Lower Bound (ELBO)

$\mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}|\boldsymbol{x})]$

- Thus, $\max_{q(\cdot|\boldsymbol{x})} \int q(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} = \max_{q(\cdot|\boldsymbol{x})} \log p_\theta(\boldsymbol{x}) - \mathrm{KL}(q(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))$

$$= \log p_\theta(\boldsymbol{x}) - \min_{q(\cdot|\boldsymbol{x})} \mathrm{KL}\big(q(\boldsymbol{z} \mid \boldsymbol{x}) || \, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})\big) = \log p_\theta(\boldsymbol{x})$$

- The last step follows because $\mathrm{KL}(q \,||\, p) \geq 0$ and $\mathrm{KL}(q \,||\, p) = 0$ iff $p = q$.

Outline

- Latent Variable Models
  - Probabilistic PCA
  - Beyond PPCA?

- Variational Inference
  - Principle
  - Derivation

- **Expectation Maximization**
  - Derivation
  - EM for a Mixture of Gaussians

Expectation Maximization

- ML objective

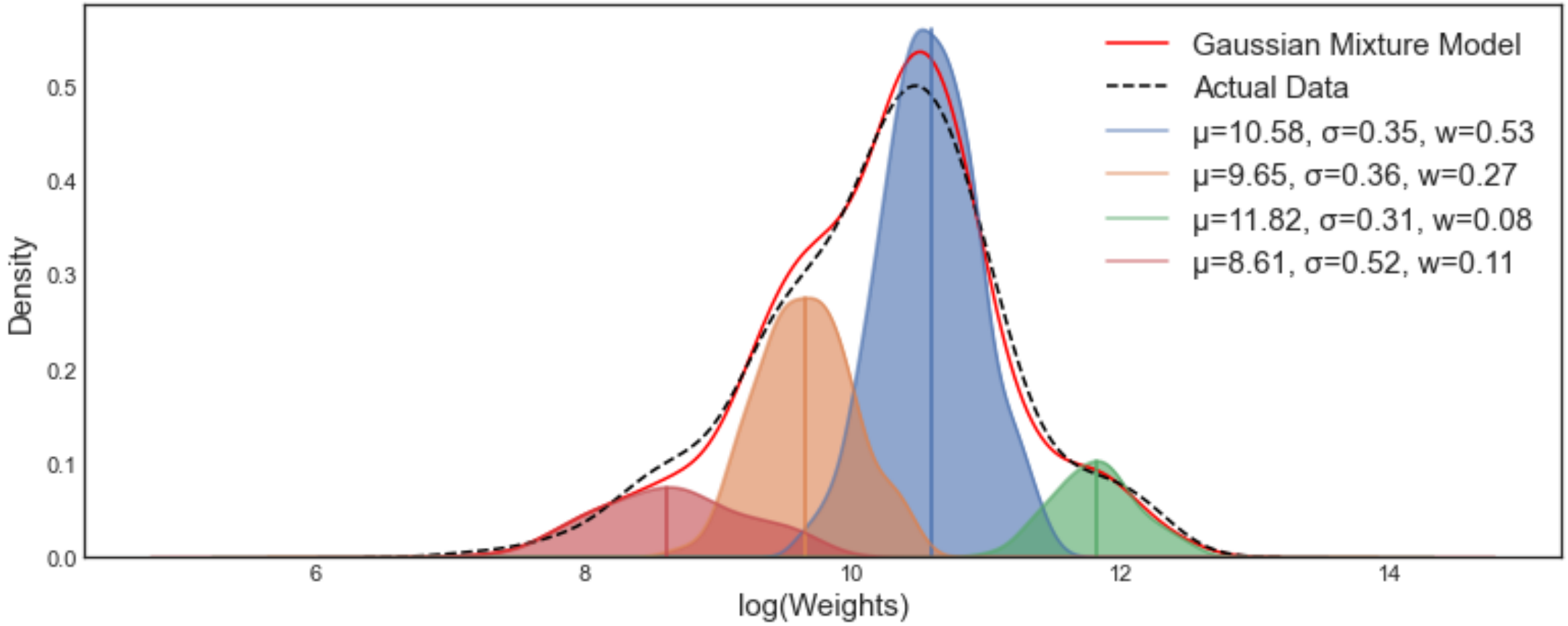
$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i) \equiv \max_{\theta} \max_{q(\cdot|\boldsymbol{x}_i),\forall i} \sum_{i=1}^{N} \int q(\boldsymbol{z} \mid \boldsymbol{x}_i) \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{x}_i)} d\boldsymbol{z}$$

- Expectation Maximization alternates between two steps (k: iteration)
  - **E-step**: $q^k(\boldsymbol{z} \mid \boldsymbol{x}_i) = p_{\theta^k}(\boldsymbol{z} \mid \boldsymbol{x}_i)$ — maximize w.r.t. q with θ fixed
  - **E-step**: $Q(\theta \mid \theta_k) = \sum_{i=1}^{N} \int_z q^k(\boldsymbol{z} \mid \boldsymbol{x}_i) \log p_\theta(\boldsymbol{x}_i, \boldsymbol{z})\, dz$ — integrate w.r.t. q with θ fixed
  - **M-step**: $\theta^{k+1} = \text{argmax}_\theta\, Q(\theta \mid \theta^k)$ — maximize w.r.t. θ with q fixed

- Examples
  - For PPCA and for a mixture of Gaussians, E & M steps are closed-form (HW1, next slide).
  - E-step often done by sampling (MCMC) and M-step often done by optimization (SGD).
  - The E-step is the hardest: it requires an “easy to sample from” distribution.

Example: EM for Gaussian Mixture Model (GMM)

- **GMM**: $p_\theta(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n} p(\boldsymbol{x} \mid z = j)p(z = j) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \cdots + \pi_n p_{\theta_n}(\boldsymbol{x})$
  - $\pi_j \coloneqq p(z = j) > 0$: prior probability of drawing a point from the j-th model; $\sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $p_{\theta_j}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})$, $\theta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})$ are the mean and covariance of the j-th Gaussian
  - $\theta = (\theta_1, \ldots, \theta_n, \pi_1, \ldots, \pi_n)$ are the parameters of the mixture model



Example: EM for Gaussian Mixture Model (GMM)

- **GMM**: $p_\theta(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n} p(\boldsymbol{x} \mid z = j) p(z = j) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \cdots + \pi_n p_{\theta_n}(\boldsymbol{x})$
  - $\pi_j \coloneqq p(z = j) > 0$: prior probability of drawing a point from the j-th model; $\sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $p_{\theta_j}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})$, $\theta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})$ are the mean and covariance of the j-th Gaussian
  - $\theta = (\theta_1, \ldots, \theta_n, \pi_1, \ldots, \pi_n)$ are the parameters of the mixture model
- **Goal**: estimate θ from N i.i.d. samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from p_θ using EM
  - For $i = 1, \ldots, N$, let $\boldsymbol{z}_i = j$ if $\boldsymbol{x}_i$ belongs to class j , and let $q_{ij} = q(\boldsymbol{z}_i = j \mid \boldsymbol{x}_i)$

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i) = \max_{\theta} \max_{q_{ij} \geq 0 : \sum_j q_{ij} = 1, \forall i} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij} \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i = j)}{q_{ij}}$$

- **E-step**: compute $q_{ij}^k = p_{\theta^k}(\boldsymbol{z}_i = j \mid \boldsymbol{x}_i) = \dfrac{p_{\theta^k}(\boldsymbol{x}_i | \boldsymbol{z}_i = j) p_{\theta^k}(\boldsymbol{z}_i = j)}{p_{\theta^k}(\boldsymbol{x}_i)} = \boxed{\dfrac{p_{\theta_j^k}(\boldsymbol{x}_i) \pi_j^k}{\sum_{l=1}^{n} p_{\theta_l^k}(\boldsymbol{x}_i) \pi_l^k}}$
- **E-step**: the expectation w.r.t $\boldsymbol{z}$ is easy to compute because $\boldsymbol{z}$ takes n values.

Example: EM for Gaussian Mixture Model (GMM)

- **GMM**: $p_\theta(\boldsymbol{x}) = \sum_{\mathrm{j}=1}^{n} p(\boldsymbol{x} \mid z = j) p(z = j) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \cdots + \pi_n p_{\theta_n}(\boldsymbol{x})$
  - $\pi_j \coloneqq p(z = j) > 0$: prior probability of drawing a point from the j-th model; $\sum_{\mathrm{j}=1}^{n} \pi_j = 1$
  - $p_{\theta_j}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})$, $\theta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})$ are the mean and covariance of the j-th Gaussian
  - $\theta = (\theta_1, \ldots, \theta_n, \pi_1, \ldots, \pi_n)$ are the parameters of the mixture model
- **M-step**:

$$\theta^{k+1} = \operatorname*{argmax}_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i = j)}{q_{ij}^{k}} = \operatorname*{argmax}_{\{\theta_j\}_{j=1}^{n}, \{\pi_j\}_{j=1}^{n}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log\left(\pi_j \cdot p_{\theta_j}(\boldsymbol{x}_{\mathrm{i}})\right)$$

  - π_j's: $\max_{\{\pi_j\}_{j=1}^{n} : \sum_{\mathrm{j}=1}^{n} \pi_j = 1} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log(\pi_j)$. The solution is given by $\boxed{\pi_j^{k+1} = \frac{\sum_{i=1}^{N} q_{ij}^{k}}{\sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k}}}$
  - θ_j's: $\max_{\mu_j, \Sigma_{\mathrm{j}}} \sum_{\mathrm{i}=1}^{N} q_{ij}^{k} \left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^{\mathsf{T}} \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) - \frac{1}{2} \operatorname{logdet}(\boldsymbol{\Sigma}_j)\right)$. The solution is

$$\boxed{\boldsymbol{\mu}_j^{k+1} = \frac{\sum_{i=1}^{N} q_{ij}^{k} \boldsymbol{x}_i}{\sum_{i=1}^{N} q_{ij}^{k}}} \text{ and } \boxed{\Sigma_j^{k+1} = \frac{\sum_{i=1}^{N} q_{ij}^{k} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)^{\mathsf{T}}}{\sum_{i=1}^{N} q_{ij}^{k}}}$$

E.g.: EM for Gaussian Mixture Model

- **M-step**:

$$\pi_j^{k+1} = \underset{\{\pi_j\}_{j=1}^n : \sum_{j=1}^n \pi_j = 1}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^k \log(\pi_j)$$

- Form the Lagrangian

$$L\left(\{\pi_j\}_{j=1}^n, \lambda\right) = \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^k \log(\pi_j) - \lambda\left(\sum_{j=1}^n \pi_j - 1\right)$$

- First order conditions

$$\frac{\partial L}{\partial \lambda} = 0 \Leftrightarrow \sum_{j=1}^n \pi_j = 1 \,, \forall j: \frac{\partial L}{\partial \pi_j} = 0 \Rightarrow \sum_{i=1}^{N} \frac{q_{ij}^k}{\pi_j} - \lambda = 0 \Rightarrow \pi_j = \frac{\sum_{i=1}^N q_{ij}^k}{\lambda}$$

- Therefore, $\lambda = \sum_{i=1}^N \sum_{j=1}^n q_{ij}^k \Rightarrow \pi_j = \frac{\sum_{i=1}^N q_{ij}^k}{\sum_{i=1}^N \sum_{j=1}^n q_{ij}^k} =: \pi_j^{k+1}$

E.g.: EM for Gaussian Mixture Model

- **M-step**: $\theta^{k+1} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log \frac{p_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i = j)}{q_{ij}^{k}} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log p_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i = j)$
  - If we spell out θ and $p_\theta(\boldsymbol{x}_i, z_i = j)$, the above becomes $\underset{\{\theta_j\}_{j=1}^{n}, \{\pi_j\}_{j=1}^{n}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{n} q_{ij}^{k} \log\left(\pi_j \cdot p_{\theta_j}(\boldsymbol{x}_i)\right)$
  - For θ_j's: $\max_{\mu_j, \Sigma_j} \boxed{\sum_{i=1}^{N} q_{ij}^{k} \left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j) - \frac{1}{2}\operatorname{logdet}(\boldsymbol{\Sigma}_j)\right)} =: \mathcal{L}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$
  - $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_j} = \sum_{i=1}^{N} q_{ij}^{k} \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j) = \Sigma_j^{-1}\left(\sum_{i=1}^{N} q_{ij}^{k} \boldsymbol{x}_i - \left(\sum_{i=1}^{N} q_{ij}^{k}\right)\boldsymbol{\mu}_j\right).$
  - Setting it to $0 \Rightarrow \boldsymbol{\mu}_j = \frac{\sum_{i=1}^{N} q_{ij}^{k} \boldsymbol{x}_i}{\sum_{i=1}^{N} q_{ij}^{k}} =: \boldsymbol{\mu}_j^{k+1}$
  - $\mathcal{L}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_j^{-1} \sum_{i=1}^{N} q_{ij}^{k} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top\right) - \frac{1}{2}\left(\sum_{i=1}^{N} q_{ij}^{k}\right) \operatorname{logdet} \boldsymbol{\Sigma}_j$
  - Reusing our derivation in MLE for Gaussian: $\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{N} q_{ij}^{k} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{k+1}\right)^\top}{\sum_{i=1}^{N} q_{ij}^{k}} =: \boldsymbol{\Sigma}_j^{k+1}$