Expectation-Maximization Algorithm

Qinliang Su （苏勤亮）

Sun Yat-sen University

suqliang@mail.sysu.edu.cn

Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example 1: Training Gaussian Mixture Models
- Example 2: Training Gaussian Latent-Variable Models

General Form of the Concerned Problem

- Given the joint distribution

$$p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}),$$

where $\boldsymbol{x}$ is the observed variable and $\boldsymbol{z}$ is the latent variable, we need to maximize the log likelihood *w.r.t.* $\boldsymbol{x}$, that is,

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x}; \boldsymbol{\theta}),$$

where

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$$

What we have is the joint pdf $p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$, but what we need to optimize is the marginal pdf $p(\boldsymbol{x}; \boldsymbol{\theta})$

Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example 1: Training Gaussian Mixture Models
- Example 2: Training Gaussian Latent-Variable Models

EM Algorithm

- Algorithm

E-step: Evaluating the expectation

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

M-step: Updating the parameter

$$\boldsymbol{\theta}^{(t+1)}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

- Key integrant in EM

1) The posteriori distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

2) The expectation of joint distribution $\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ *w.r.t.* the posteriori

3) Maximization

Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example 1: Training Gaussian Mixture Models
- Example 2: Training Gaussian Latent-Variable Models

Re-representing the Log-likelihood

- The log-likelihood can be reformulated as

$\forall$ distribution $q(\boldsymbol{z})$

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{x})$$

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}) q(\boldsymbol{z})}$$

$$= \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})}}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})}}_{KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}))}$$

$$\boxed{= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})), \qquad \text{for } \forall\, \boldsymbol{\theta}, q(\boldsymbol{z})}$$

Remark: The KL-divergence is used to *measure the distance* between two distributions q and p, which is defined as

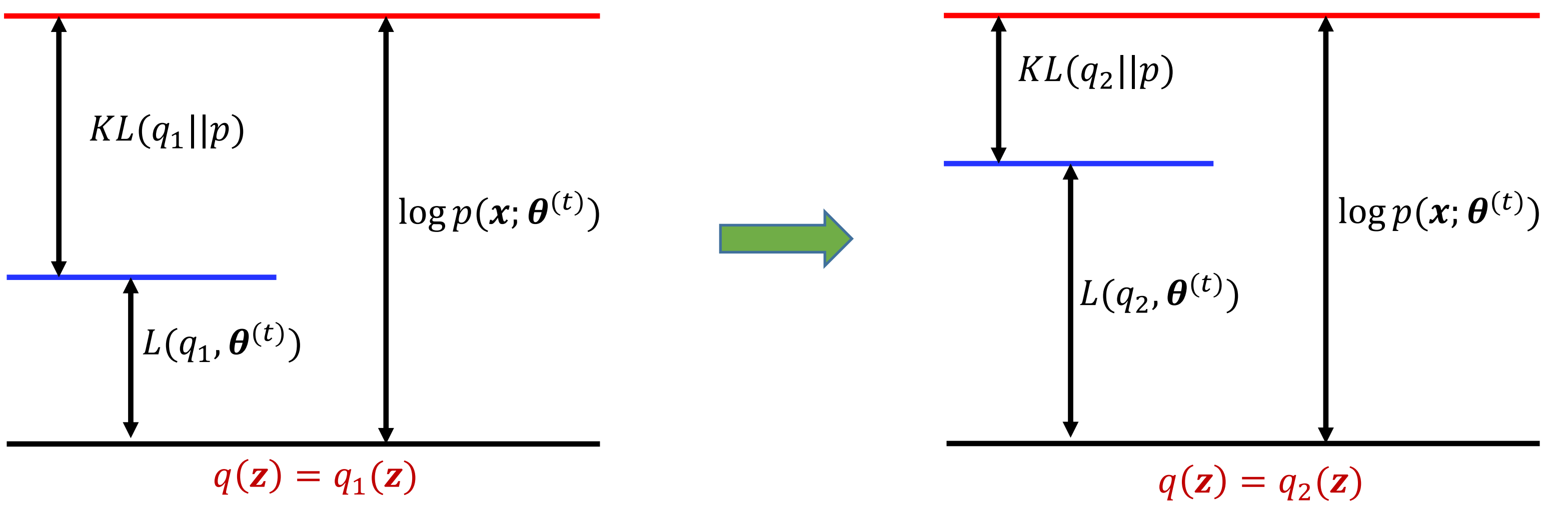
$$KL(q||p) \triangleq \int q(z) \log \frac{q(z)}{p(z)} dz \geq 0$$

- Thus, with the parameter at the t-th iteration denoted as $\boldsymbol{\theta}^{(t)}$, we have

$$\log p\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) = \mathcal{L}\left(q,\boldsymbol{\theta}^{(t)}\right) + KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}))$$

This equality holds for any distribution $q(\boldsymbol{z})$

- Different $q(\boldsymbol{z})$ will lead to different decomposition of $\log p\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)$



Theoretical Justification for EM

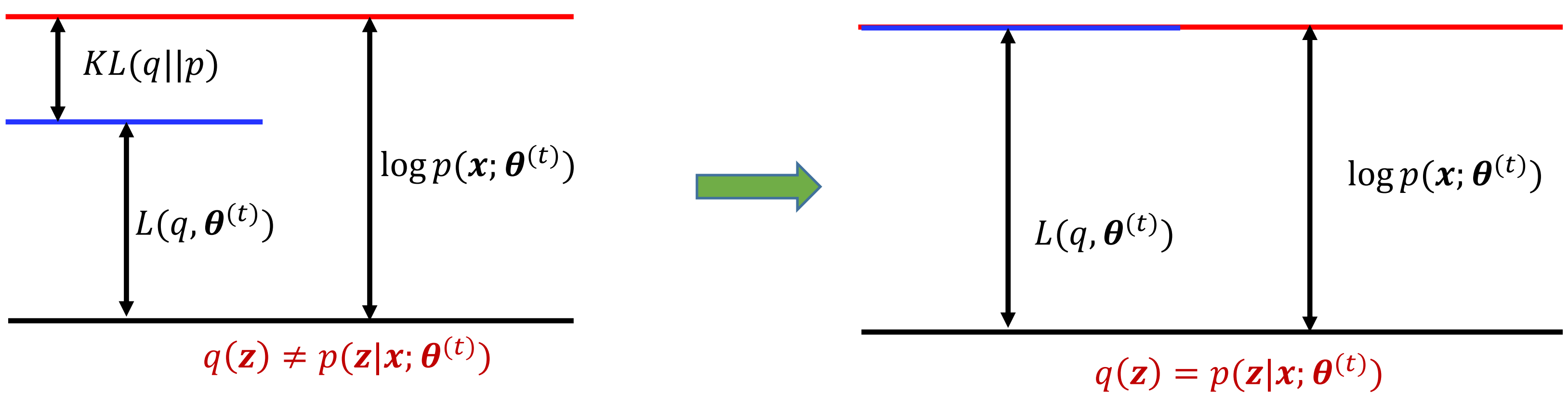
$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)}\right)}{q(\boldsymbol{z})} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

- If we set $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$, then we have

$$KL\left(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right) = 0$$

Thus, we have

$$\begin{aligned}\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) &= \mathcal{L}\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t)}\right) \\ &= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) \log \frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)}\right)}{p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)}\end{aligned}$$



$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}^{(t)}\right)$$

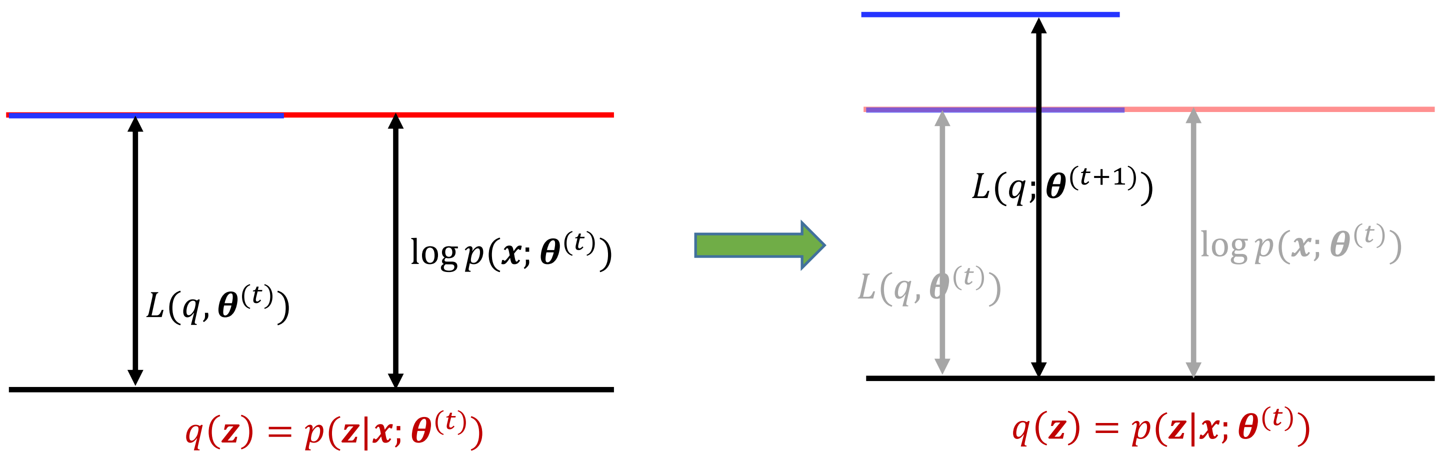
$$= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)\log\frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)}\right)}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

- If we update $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right),$$

then we must have the relation

$$\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}^{(t+1)}\right) \geq \underbrace{\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}^{(t)}\right)}_{=\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

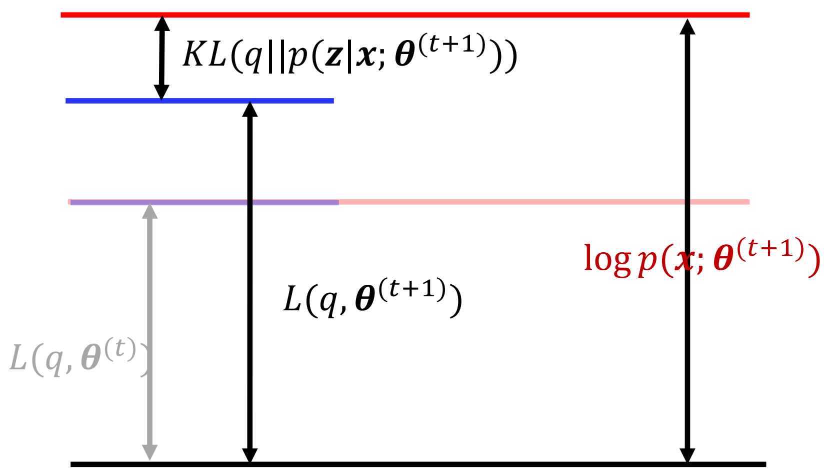


$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t+1)}\right)}{q(\boldsymbol{z})} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)})}$$

- By setting $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$, we obtain

$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}) = \underbrace{\mathcal{L}\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t+1)}\right)}_{\geq \log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})} + \underbrace{KL(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}))}_{\geq 0}$$

The KL-divergence is always non-negative



$q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

- Thus, we can see that

$$\boxed{\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})}$$

$\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z} \middle| \boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right)$ *can guarantee the increase of likelihood at each step*

- Equivalence between EM updating

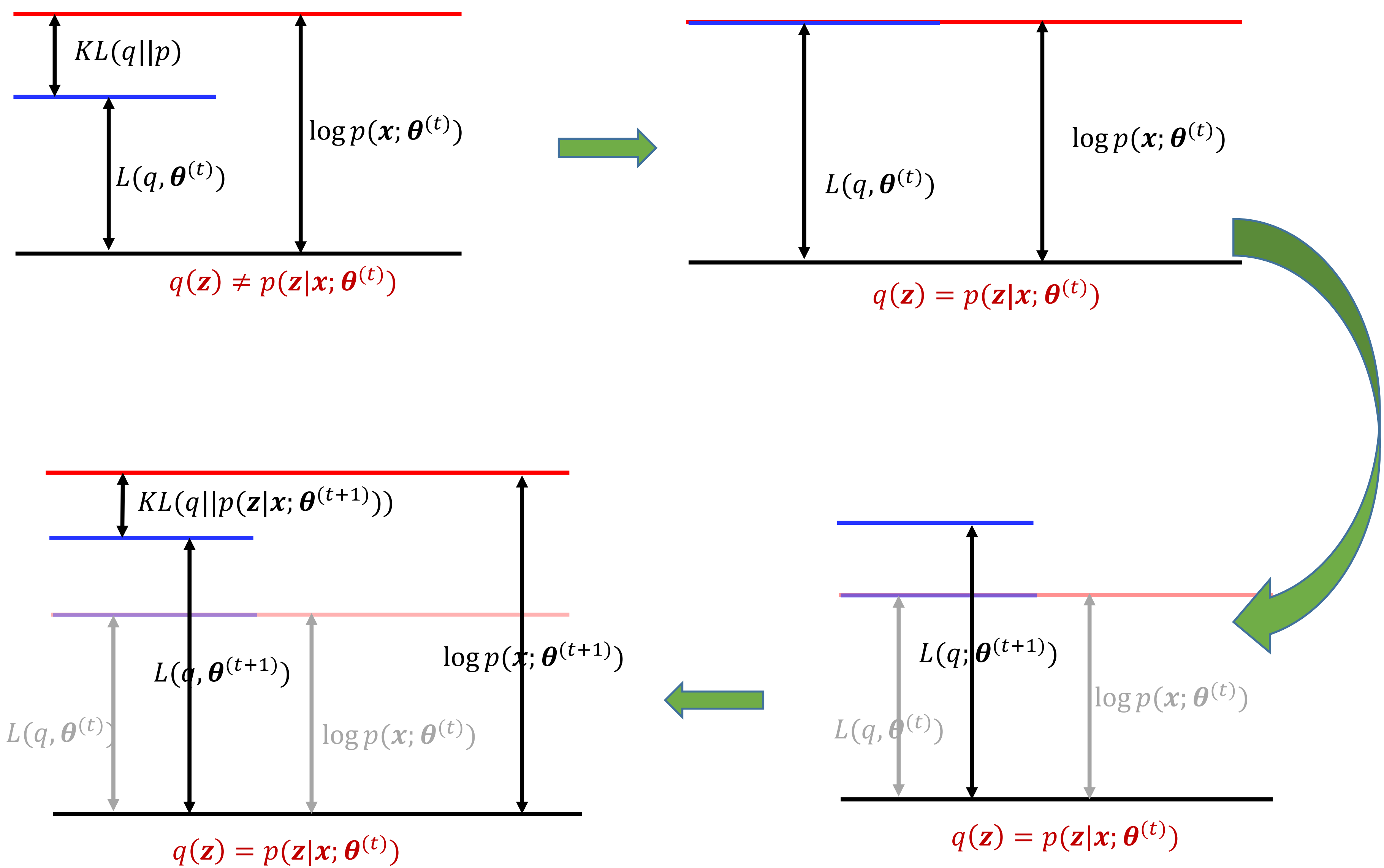
$$\arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) \quad \text{with} \quad Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) \triangleq \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})}[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})]$$

and the updating rule $\arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\mathbf{z}\middle|\mathbf{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right)$

$$\mathcal{L}\left(p\left(\mathbf{z}\middle|\mathbf{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right) = \underbrace{\sum_{\mathbf{z}} p\left(\mathbf{z}\middle|\mathbf{x}; \boldsymbol{\theta}^{(t)}\right) \log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}_{\mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(t)})}[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})]} - \underbrace{\sum_{\mathbf{z}} p\left(\mathbf{z}\middle|\mathbf{x}; \boldsymbol{\theta}^{(t)}\right) \log p\left(\mathbf{z}\middle|\mathbf{x}; \boldsymbol{\theta}^{(t)}\right)}_{constant}$$

Therefore, $\arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\mathbf{z}\middle|\mathbf{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right) \Leftrightarrow \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$

EM algorithm can guarantee the increase of likelihood at each step



$KL(q||p)$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) \neq p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q;\boldsymbol{\theta}^{(t+1)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$KL(q||p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}))$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)})$
$L(q,\boldsymbol{\theta}^{(t+1)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

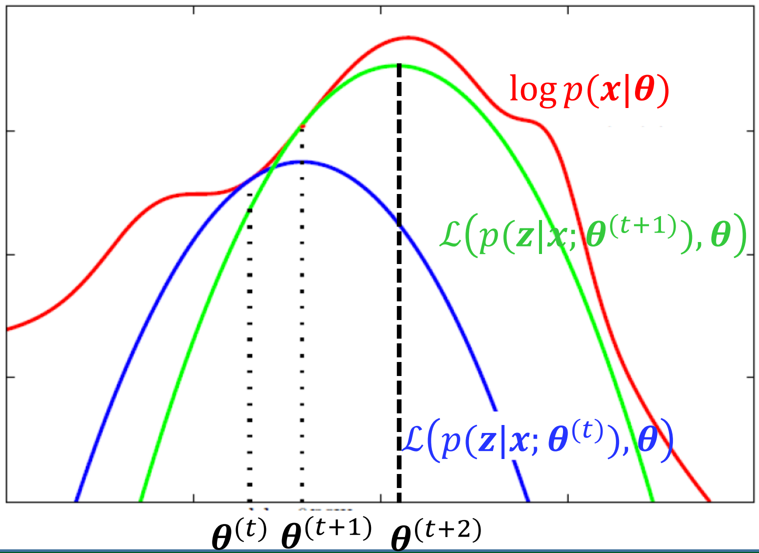
A View in the Parameter Space

1) E-step (*t*): deriving the expression $\mathcal{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$ given the model parameter $\boldsymbol{\theta}^{(t)}$

2) M-step (*t*): computing the optimal value $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$

3) E-step (*t+1*): deriving the expression for $\mathcal{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}),\boldsymbol{\theta})$ given the model parameter $\boldsymbol{\theta}^{(t+1)}$

4) Repeating the above process until convergence



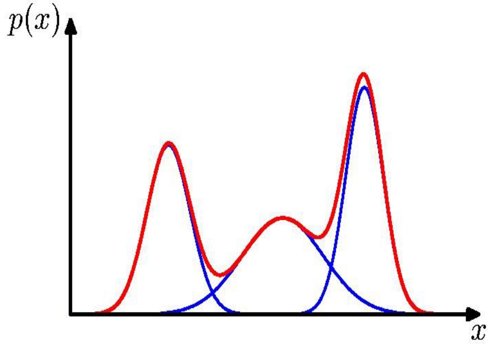
Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example 1: Training Gaussian Mixture Models
- Example 2: Training Gaussian Latent-Variable Models

Gaussian Mixture Model Review

- For a Gaussian mixture distribution, *i.e.*,

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$



it can be represented as the marginal distribution of the joint distribution

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

$$= \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_k}$$

  - $\boldsymbol{z} = [z_1, z_2, \cdots, z_K]$ follows the categorical distribution with parameter $\boldsymbol{\pi}$

EM Two Steps

- It is a latent-variable model, thus we can use EM to optimize it

Remark: maximizing $\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right)$ is equivalent to $\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$

- *Reminder:* Key integrant in EM
  - E-step: Expectation *w.r.t.* the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{p\left(\boldsymbol{z}^{(n)}|\boldsymbol{x}^{(n)};\boldsymbol{\theta}^{(t)}\right)}\left[\log p\left(\boldsymbol{x}^{(n)},\boldsymbol{z}^{(n)};\boldsymbol{\theta}\right)\right]$$

  - M-step: Maximization

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

EM: E-step

- The posteriori distribution

$$p\left(\boldsymbol{z} = \mathbf{1}_k \middle| \boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) = \frac{p\left(\boldsymbol{x}, \boldsymbol{z}=\mathbf{1}_k; \boldsymbol{\theta}^{(t)}\right)}{\sum_{i=1}^{K} p\left(\boldsymbol{x}, \boldsymbol{z}=\mathbf{1}_i; \boldsymbol{\theta}^{(t)}\right)}$$

$$\boxed{= \frac{\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\pi_k^{(t)}}{\sum_{i=1}^{K} \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}\right)\pi_i^{(t)}}}$$

  - $\mathbf{1}_k$ denotes the one-hot vector with the k-th element being 1

- The log of the joint distribution $\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$

$$\boxed{\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \sum_{k=1}^{K} z_k \cdot [\log \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \log \pi_k]}$$

Note that $\boldsymbol{z}$ can only be a one-hot vector

- The expectation

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})]$$

$$= \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[z_k][\log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \log \pi_k]$$

➢ Due to $p\left(\boldsymbol{z} = \mathbf{1}_k \middle| \boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) = \frac{\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}\right)\pi_k^{(t)}}{\sum_{i=1}^{K} \mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_i^{(t)},\boldsymbol{\Sigma}_i^{(t)}\right)\pi_i^{(t)}}$, we have

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[z_k] = \frac{\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}\right)\pi_k^{(t)}}{\sum_{i=1}^{K} \mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_i^{(t)},\boldsymbol{\Sigma}_i^{(t)}\right)\boldsymbol{\pi}_i^{(t)}} \triangleq \gamma_k^{(t)}$$

- Therefore, we have

$$\boxed{\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) = \sum_{k=1}^{K} \gamma_k^{(t)}[\log \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \log \pi_k]}$$

- Substituting $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\frac{1}{(2\mathrm{Pi})^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$ into $\mathcal{Q}(\cdot)$ gives

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\sum_{k=1}^{K}\gamma_k^{(t)}\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)-\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\log\pi_k\right]+C$$

  - C is the constant

- So far, only one data example $\boldsymbol{x}$ is considered

- If data $\boldsymbol{x}^{(n)}$ for $n=1,2,\cdots N$ are considered, the $\mathcal{Q}(\cdot)$ becomes

$$\boxed{\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}^{(t)}\left[-\frac{1}{2}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)-\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\log\pi_k\right]+C}$$

EM: M-step

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}^{(t)}\left[-\frac{1}{2}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)-\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\log\pi_k\right]+C$$

- By taking derivatives *w.r.t.* $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and setting them to zero, we obtain the optimal $\boldsymbol{\theta}$ as

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}^{(t)}\boldsymbol{x}^{(n)}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}^{(t)}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k^{(t+1)}\right)\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k^{(t+1)}\right)^T$$

  - For π_k, we need to consider the optimization under constraint $\sum_{k=1}^{K}\pi_k = 1$, leading to the solution

$$\pi_k^{(t+1)} = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N}\gamma_{nk}^{(t)}$ is the effective number of examples assigned to the k-th class

Summary of EM Algorithm

- Given the current estimate $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^K$, update γ_{nk} as

$$\gamma_{nk} \leftarrow \frac{\mathcal{N}(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{i=1}^K \mathcal{N}(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\pi_i}$$

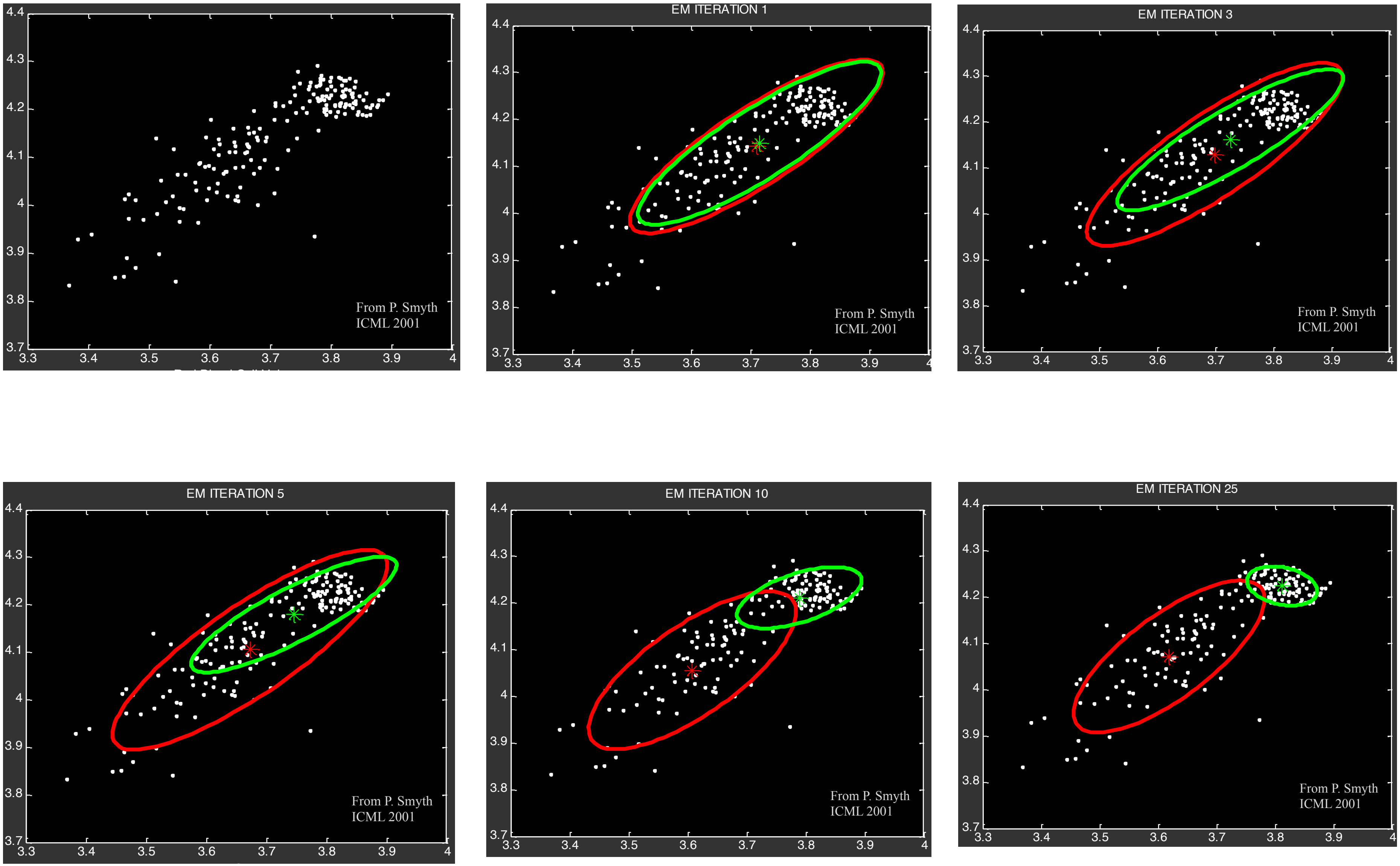
- Given the γ_{nk}, update $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ and π_k as

$$N_k \leftarrow \sum_{n=1}^N \gamma_{nk}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} \boldsymbol{x}^{(n)}$$

$$\boldsymbol{\Sigma}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k)^T$$

$$\pi_k \leftarrow \frac{N_k}{N}$$


From P. Smyth
ICML 2001
EM ITERATION 1
From P. Smyth
ICML 2001
EM ITERATION 3
From P. Smyth
ICML 2001
EM ITERATION 5
From P. Smyth
ICML 2001
EM ITERATION 10
From P. Smyth
ICML 2001
EM ITERATION 25
From P. Smyth
ICML 2001

Relation to Soft K-Means

- When restricting $\boldsymbol{\Sigma}_k$ to the form $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$, the EM updating rules for GMM are

$$\pi_k \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$$

$$\gamma_{nk} \leftarrow \frac{\pi_k e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right\|^2}}{\sum_{i=1}^{K} \pi_i e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_i\right\|^2}}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}^{(n)}}{\sum_{n=1}^{N} \gamma_{nk}}$$

- Updates in soft K-means

Setting π_k and β as
$\pi_k = \frac{1}{K}$, $\beta = \frac{1}{2\sigma^2}$

$$r_{nk} = \frac{e^{-\beta\left\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right\|^2}}{\sum_{i=1}^{K} e^{-\beta\left\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_i\right\|^2}}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{n=1}^{N} r_{nk}\, \boldsymbol{x}^{(n)}}{\sum_{n=1}^{N} r_{nk}}$$

Outline

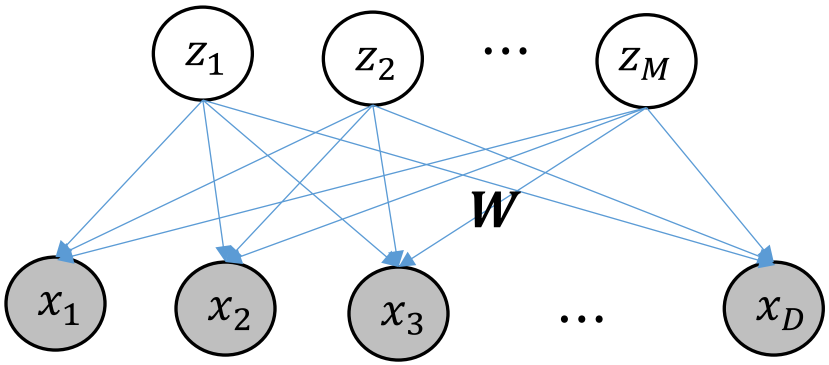
- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example 1: Training Gaussian Mixture Models
- Example 2: Training Gaussian Latent-Variable Models

Probabilistic PCA Review

- Probabilistic PCA model

Prior distribution: $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$

Likelihood function: $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$



- The objective is to maximize the $\log p(\boldsymbol{x})$ *w.r.t.* all training data points $\boldsymbol{x}_n$

EM Two Steps

- It is a latent-variable model, thus we can use EM to optimize it

Remark: maximizing $\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right)$ is equivalent to $\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$

- *Reminder:* Key integrant in EM
  - E-step: Expectation *w.r.t.* the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\sum_{n=1}^{N}\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^{(t)})}\left[\log p(\boldsymbol{x}_n,\boldsymbol{z}_n;\boldsymbol{\theta})\right]$$

  - M-step: Maximization

$$\boldsymbol{\theta}^{(t+1)}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

E-Step: Evaluating $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$

- From

$$p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{D/2}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{\mu}\|^2}{2\sigma^2}} \cdot \frac{1}{(2\pi)^{M/2}} e^{-\frac{\|\boldsymbol{z}\|^2}{2}}$$

  we obtain

$$\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = -\frac{D}{2}\log 2\pi\sigma^2 - \frac{M}{2}\log 2\pi - \frac{\|\boldsymbol{x} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{\mu}\|^2}{2\sigma^2} - \frac{\|\boldsymbol{z}\|^2}{2}$$

- Thus, we have

$$\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) = \sum_{n=1}^{N} \left(-\frac{1}{2\sigma^2}\|\boldsymbol{\mu}\|^2 + \frac{1}{\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{W} \mathbb{E}_{\boldsymbol{z}_n}[\boldsymbol{z}_n] - \frac{1}{2\sigma^2} Tr\left(\boldsymbol{W}^T \boldsymbol{W} \mathbb{E}_{\boldsymbol{z}_n}[\boldsymbol{z}_n \boldsymbol{z}_n^T]\right) + C \right)$$

  - $\mathbb{E}_{\boldsymbol{z}_n}[\cdot]$ denotes the expectation *w.r.t.* the distribution $p(\boldsymbol{z}_n | \boldsymbol{x}_n; \boldsymbol{\theta}^{(t)})$
  - $Tr(\cdot)$ means the trace operation, and C is irrelevant to $\boldsymbol{W}$ *and* $\boldsymbol{\mu}$

M-Step: Maximization

- The global optimal $\boldsymbol{\mu}$ is already known to be $\overline{\boldsymbol{x}} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n}{N}$, so we fix

$$\boldsymbol{\mu} = \overline{\boldsymbol{x}}$$

- By deriving

$$\frac{\partial \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)}{\partial \boldsymbol{W}} = -\frac{1}{\sigma^2} \sum_{n=1}^{N} \left(\boldsymbol{W} \mathbb{E}_{\boldsymbol{z}_n}[\boldsymbol{z}_n \boldsymbol{z}_n^T] - (\boldsymbol{x} - \overline{\boldsymbol{x}}) \mathbb{E}_{\boldsymbol{z}_n}[\boldsymbol{z}_n^T] \right)$$

and setting $\frac{\partial \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)}{\partial \boldsymbol{W}} = 0$, we obtain

$$\boldsymbol{W}^{(t+1)} \leftarrow \left(\sum_{n=1}^{N} (\boldsymbol{x}_n - \overline{\boldsymbol{x}}) \mathbb{E}_{\boldsymbol{z}_n}[\boldsymbol{z}_n^T] \right) \left(\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{z}_n}[\boldsymbol{z}_n \boldsymbol{z}_n^T] \right)^{-1}$$

How to get the expectations $\mathbb{E}_{z_n}[\boldsymbol{z}_n]$ and $\mathbb{E}_{z_n}[\boldsymbol{z}_n\boldsymbol{z}_n^T]$

- Given the data $\boldsymbol{x}_n$, and fixing $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$, it can be derived that the posterior is

$$p(\boldsymbol{z}_n|\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{z}_n; \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_n - \overline{\boldsymbol{x}}), \sigma^2\boldsymbol{M}^{-1})$$

where $\boldsymbol{M} \triangleq \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}$

- From the distribution, we can easily obtain

$$\mathbb{E}_{z_n}[\boldsymbol{z}_n] = \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_n - \overline{\boldsymbol{x}})$$

$$\mathbb{E}_{z_n}[\boldsymbol{z}_n\boldsymbol{z}_n^T] = \sigma^2\boldsymbol{M}^{-1} + \mathbb{E}_{z_n}[\boldsymbol{z}_n]\mathbb{E}_{z_n}[\boldsymbol{z}_n^T]$$

Using 'completing the square' trick to derive the posteriori

$$\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) = \underbrace{-\frac{D}{2}\log 2\pi\sigma^2 - \frac{M}{2}\log 2\pi}_{C_1} - \frac{\|\boldsymbol{x}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{\mu}\|^2}{2\sigma^2} - \frac{\|\boldsymbol{z}\|^2}{2}$$

$$= \underbrace{C_1 - \frac{1}{2\sigma^2}\left(\|\boldsymbol{x}\|^2 - 2\boldsymbol{\mu}^T\boldsymbol{x} + \|\boldsymbol{\mu}\|^2\right)}_{\phi(\boldsymbol{x})} - \frac{1}{2\sigma^2}\left(-2\boldsymbol{x}^T\boldsymbol{W}\boldsymbol{z} + 2\boldsymbol{\mu}^T\boldsymbol{W}\boldsymbol{z} + \|\boldsymbol{W}\boldsymbol{z}\|^2\right) - \frac{1}{2}\|\boldsymbol{z}\|^2$$

$$= \phi(\boldsymbol{x}) + \frac{1}{\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{W}\boldsymbol{z} - \frac{1}{2\sigma^2}\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}$$

$$\boldsymbol{M} \triangleq \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}$$

$$= -\frac{1}{2\sigma^2}\left(\boldsymbol{z} - \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{\mu})\right)^T\boldsymbol{M}\left(\boldsymbol{z} - \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{\mu})\right) + \phi(\boldsymbol{x}) + \|\boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{\mu})\|^2$$

$$\Longrightarrow p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})} = C(\boldsymbol{x}) \cdot e^{-\frac{1}{2}\left(\boldsymbol{z}-\boldsymbol{M}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^T\left(\frac{\boldsymbol{M}}{\sigma^2}\right)\left(\boldsymbol{z}-\boldsymbol{M}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}$$

A term only depending on $\boldsymbol{x}$

$$\Longrightarrow p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1}\right)$$