Expectation-Maximization Algorithm

Shangsong Liang

Sun Yat-sen University

Outline

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

General Form of the Concerned Problem

- Given the joint distribution

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

where is the observed variable and is the latent variable, we need to maximize the log likelihood *w.r.t.* , that is,

$$\boldsymbol{\theta}=\underset{\boldsymbol{\theta}}{arg\,max}\log p\left(\boldsymbol{x};\boldsymbol{\theta}\right),$$

where

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

What we have is the joint pdf , but what we need to optimize is the marginal pdf

Outline

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

EM Algorithm

- Algorithm

  E-step: Evaluating the expectation

  M-step: Updating the parameter

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

- Key integrant in EM

  1) The posteriori distribution

  2) The expectation of joint distribution *w.r.t.* the posteriori

  3) Maximization

Outline

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

Re-representing the Log-likelihood

- The log-likelihood can be reformulated as

$$\text{¿}\,\mathcal{L}(q,\boldsymbol{\theta})+KL\,(q\vee\text{¿}\,p(\boldsymbol{z}\vee\boldsymbol{x};\boldsymbol{\theta})),for\ \forall\,\boldsymbol{\theta},q(\boldsymbol{z})$$

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

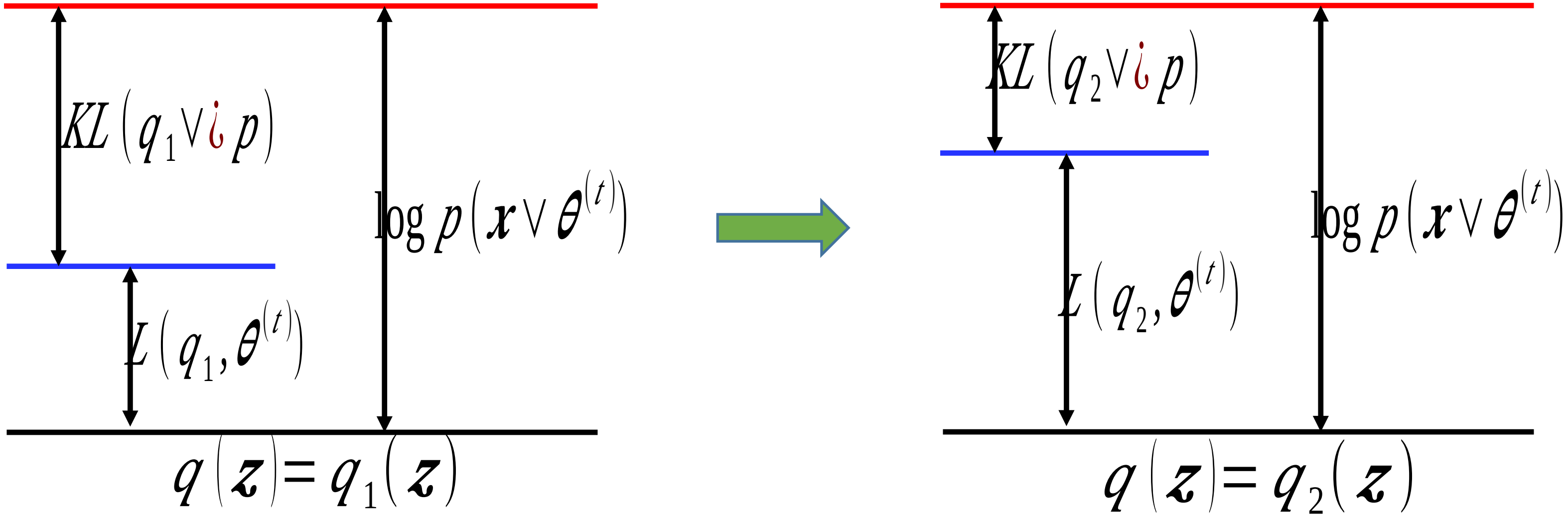
$$KL\quad\text{¿}$$

- Thus, with the parameter at the -th iteration, we have

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

This equality holds for any distribution

- Different will lead to different decomposition of


$KL\left(q_1\vee ¿\, p\right)$
$\log p\left(\boldsymbol{x}\vee \boldsymbol{\theta}^{(t)}\right)$
$L\left(q_1,\boldsymbol{\theta}^{(t)}\right)$
$q\left(\boldsymbol{z}\right)=q_1\left(\boldsymbol{z}\right)$
$KL\left(q_2\vee ¿\, p\right)$
$\log p\left(\boldsymbol{x}\vee \boldsymbol{\theta}^{(t)}\right)$
$L\left(q_2,\boldsymbol{\theta}^{(t)}\right)$
$q\left(\boldsymbol{z}\right)=q_2\left(\boldsymbol{z}\right)$

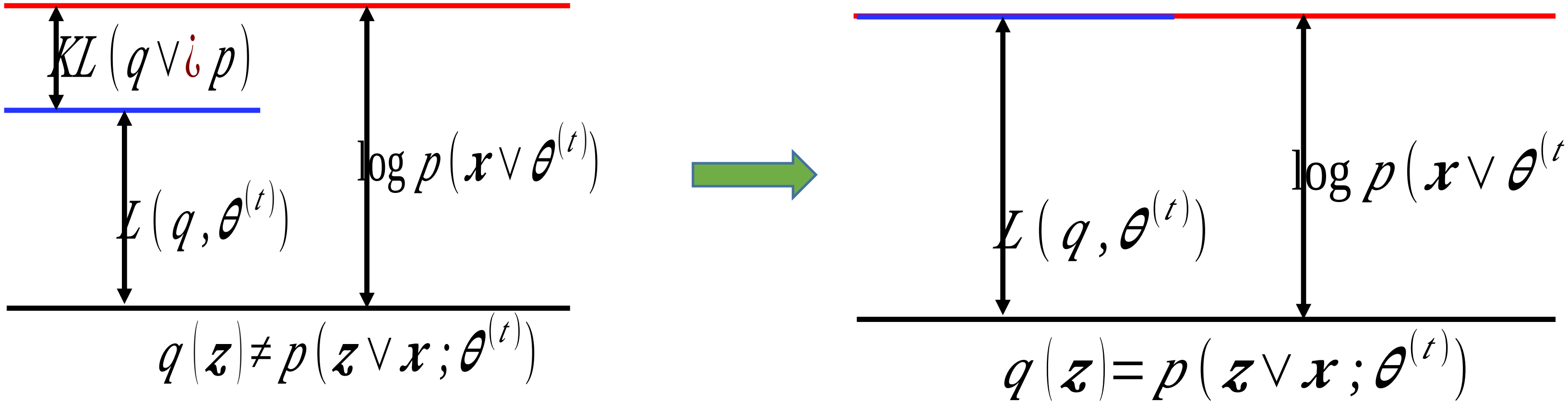
Theoretical Justification for EM

- If we set , then we have

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

Thus, we have

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


$KL\left(q \vee \text{¿}\, p\right)$
$L\left(q, \theta^{(t)}\right)$
$\log p\left(\boldsymbol{x} \vee \theta^{(t)}\right)$
$q\left(\boldsymbol{z}\right) \neq p\left(\boldsymbol{z} \vee \boldsymbol{x} ; \theta^{(t)}\right)$
$L\left(q, \boldsymbol{\theta}^{(t)}\right)$
$\log p\left(\boldsymbol{x} \vee \boldsymbol{\theta}^{(t)}\right)$
$q\left(\boldsymbol{z}\right) = p\left(\boldsymbol{z} \vee \boldsymbol{x} ; \boldsymbol{\theta}^{(t)}\right)$

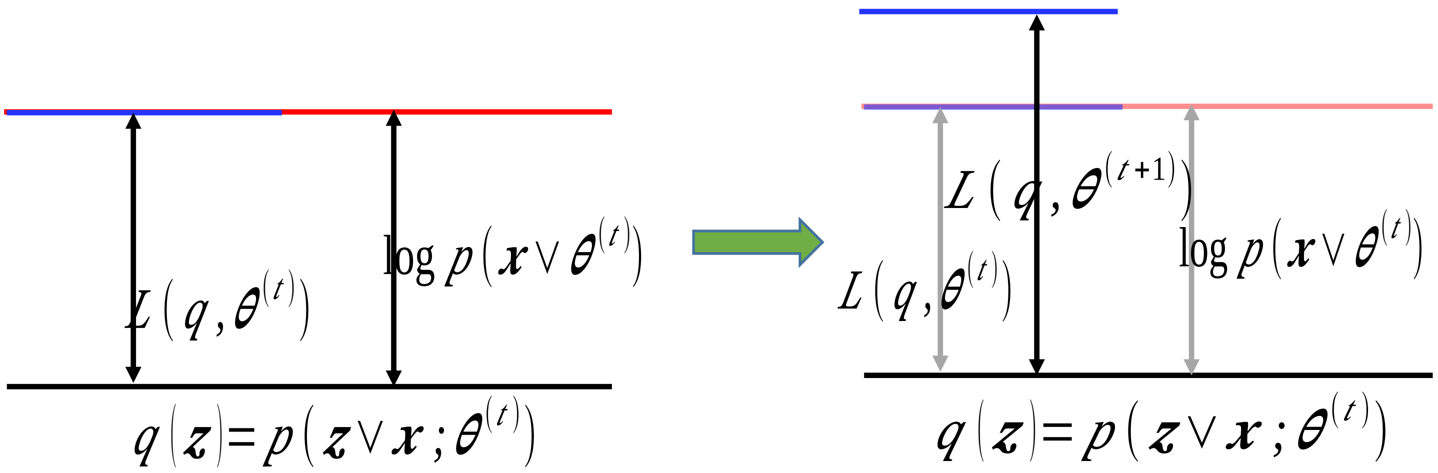
$$\log p\left(\boldsymbol{x} \vee \theta^{(t)}\right) = \mathscr{L}\left(p\left(\boldsymbol{z} \middle| \boldsymbol{x};\theta^{(t)}\right), \theta^{(t)}\right)$$

- If we update as

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

then we must have the relation

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



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

- From the nonnegative property of KL-divergence, we known that

$$KL\left(p\left(\boldsymbol{z} \vee \boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) \vee ¿\, p\left(\boldsymbol{z} \vee \boldsymbol{x};\boldsymbol{\theta}^{(t+1)}\right)\right) \geq 0$$

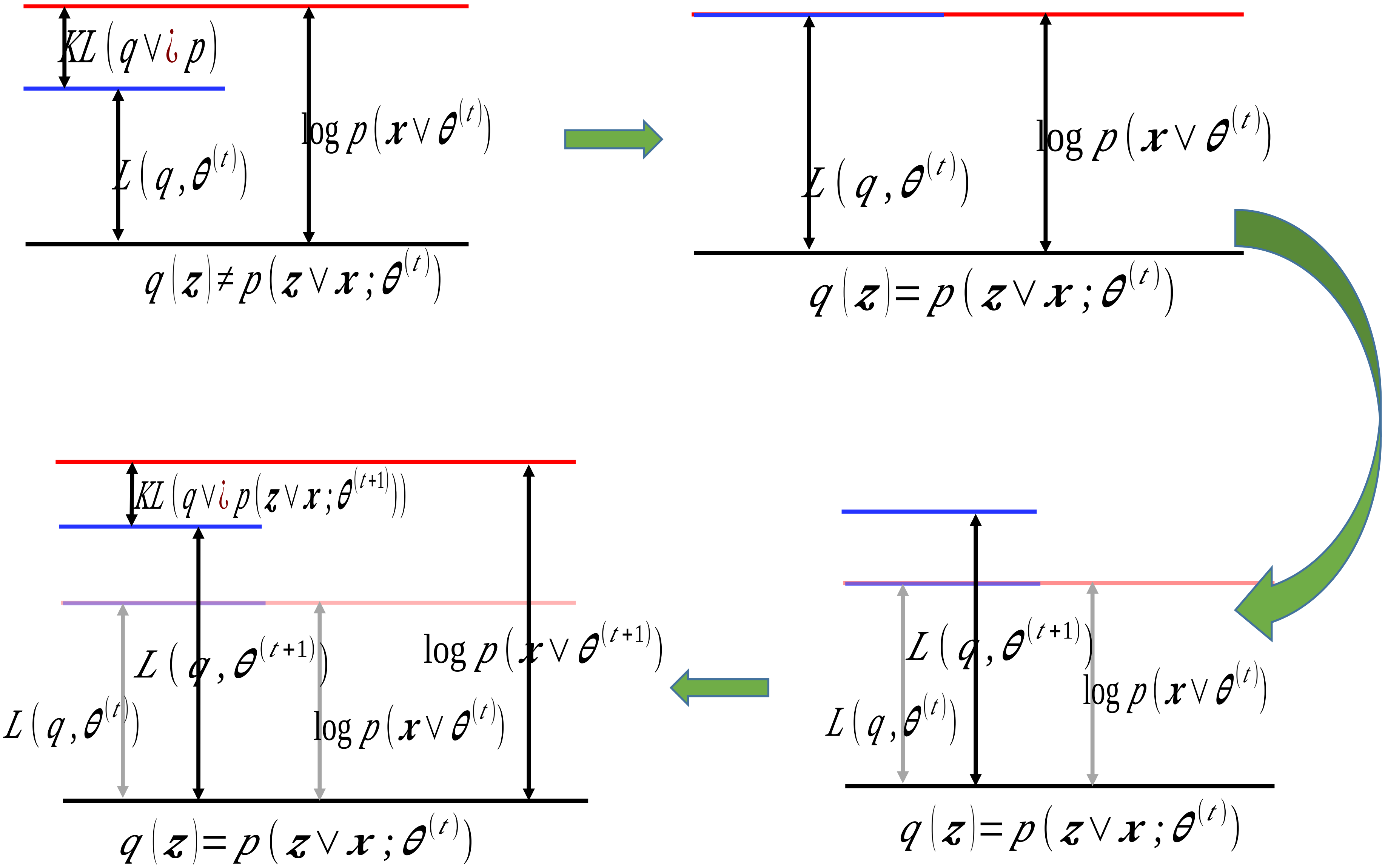
- Because holds for any , thus we have

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

- Thus, we can see that

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

EM algorithm can guarantee the increase of likelihood at each step


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

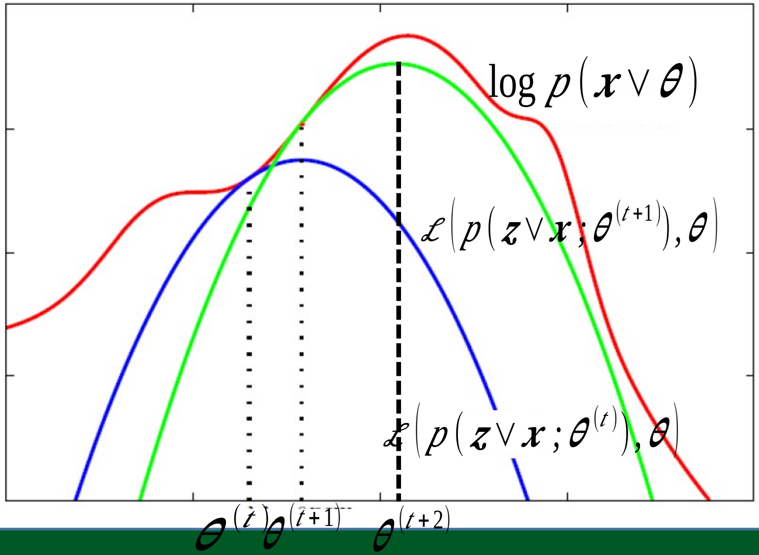
A View in the Parameter Space

1) E-step (*t*): deriving the expression given the model parameter

2) M-step (*t*): computing the optimal value

3) E-step (*t+1*): deriving the expression for given the model parameter

4) Repeating the above process until convergence

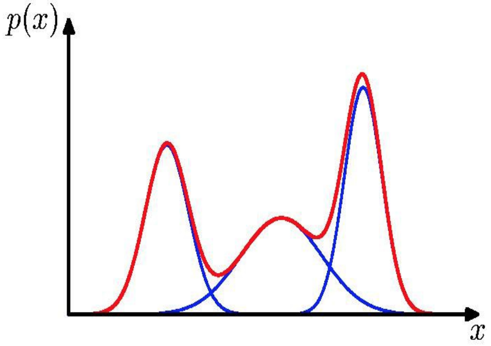


Outline

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

Gaussian Mixture Model Review

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



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

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

- follows the categorical distribution with parameter

EM: E-step

- The posteriori distribution
    - denotes the one-hot vector with the -th element being 1
- The log of the joint distribution

Note that can only be a one-hot vector

- The expectation
  - Due to , we have
- Therefore, we have

- Taking into gives

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

  - is the constant

- So far, only one data example is considered

- If data for are considered, the becomes

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

EM: M-step

- By taking derivatives *w.r.t.* , and and setting them to zero, we obtain the optimal as

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

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

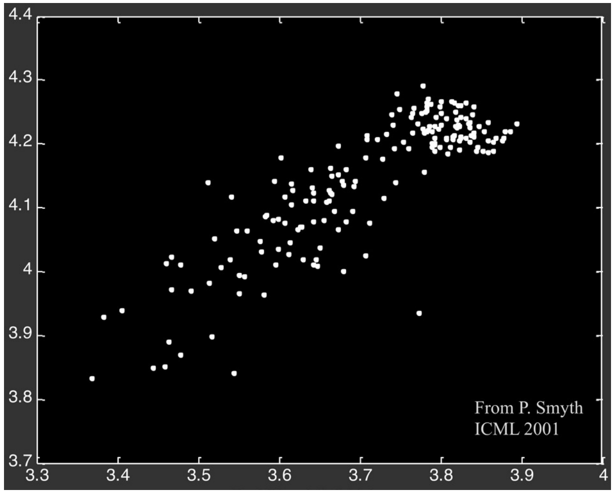
$$\pi_k^{(t+1)} = \frac{N_k}{N}$$

where is the effective number of examples assigned to the *k*-th class

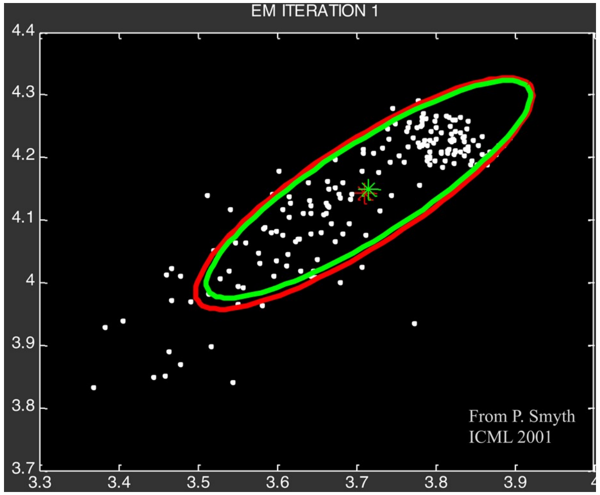
Summary of EM Algorithm

- Given the current estimate , update as

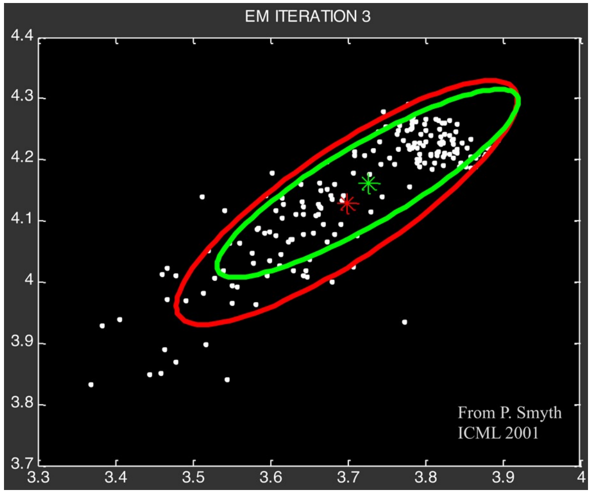
- Given the , update and as



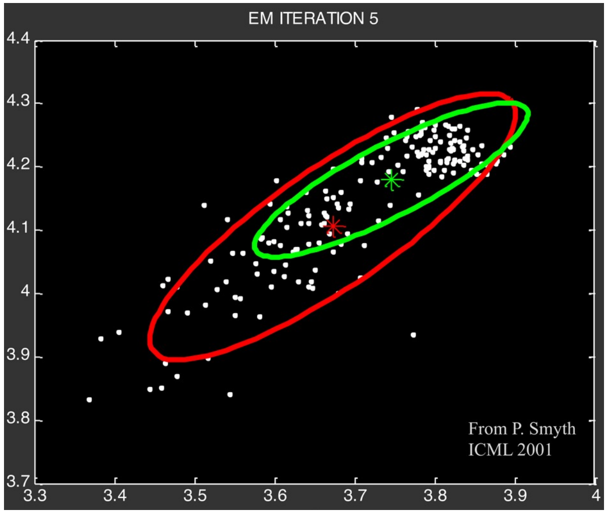
From P. Smyth
ICML 2001


EM ITERATION 1

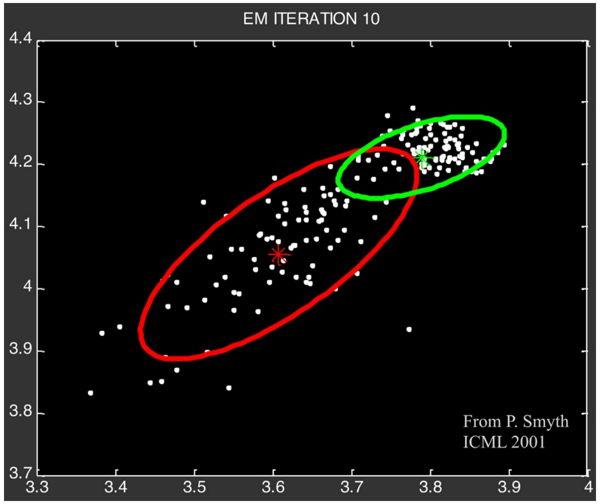
From P. Smyth
ICML 2001


EM ITERATION 3

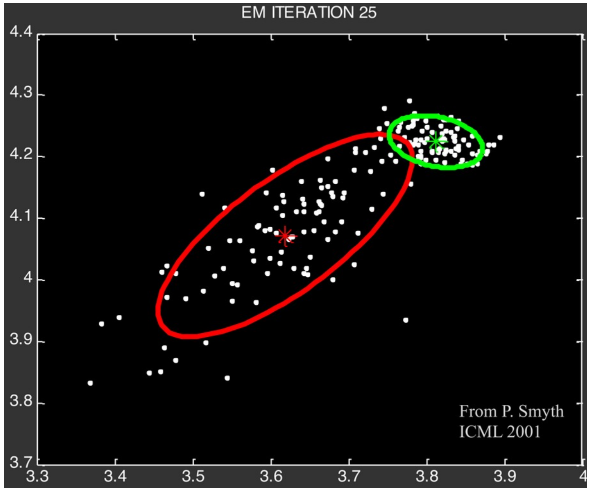
From P. Smyth
ICML 2001


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 , the updating of GMM becomes

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

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

where

- Updates in soft *K*-means

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

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

Outline

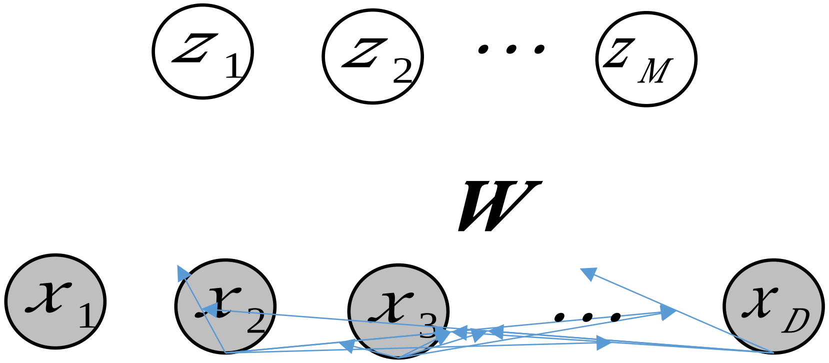
- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example 1: Training Gaussian Latent-Variable Models
- Example 2: Training Probabilistic PCA Models

Probabilistic PCA Review

- Probabilistic PCA model

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

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



- The objective is to maximize the *w.r.t.* all training data points

Two EM Steps

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

Remark: maximizing is equivalent to

- *Reminder:* Key integrant in EM
  - E-step: Expectation *w.r.t.* the posteriori

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

  - M-step: Maximization

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

E-Step: Evaluating

- From

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

we obtain

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

- Thus, we have

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

  - denotes the expectation *w.r.t.* the distribution
  - means the trace operation, and is irrelevant to *and*

M-Step: Maximization

- The global optimal is already known to be , so we fix

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

- By deriving

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

and setting , we obtain

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

How to get the expectations and

- Given the data , and fixing , it can be derived that the posterior is

$$p(\boldsymbol{z}_n|\boldsymbol{x}_n) = \mathcal{N}\text{¿}$$

where

- From the distribution, we can easily obtain

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

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

Using ‘completing the square’ trick to derive the posteriori

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

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

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

$$\text{¿} -\frac{1}{2\sigma^2}\left(\boldsymbol{z} - \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu})\right)^T\boldsymbol{M}\left(\boldsymbol{z} - \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu})\right) + \eta(\boldsymbol{x})$$

$\Longrightarrow$

$\Longrightarrow$

- Thank You!