Joint Institute of Engineering

Design and Implementation of Speech Recognition Systems

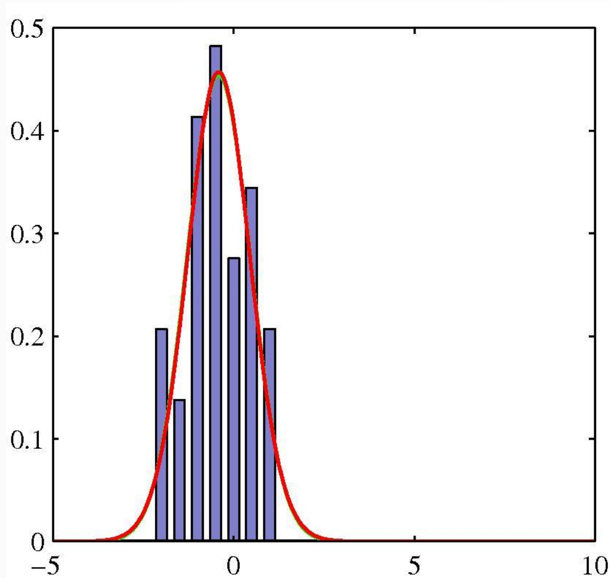
Fall 2014
Ming Li

Special topic: the Expectation-Maximization algorithm and GMM

Sep 30 2014

Some equations are from the slides edited by Muneem S.

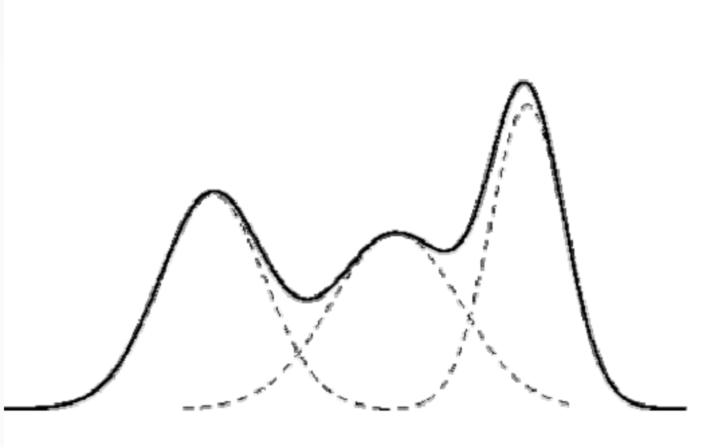
Parametric density estimation

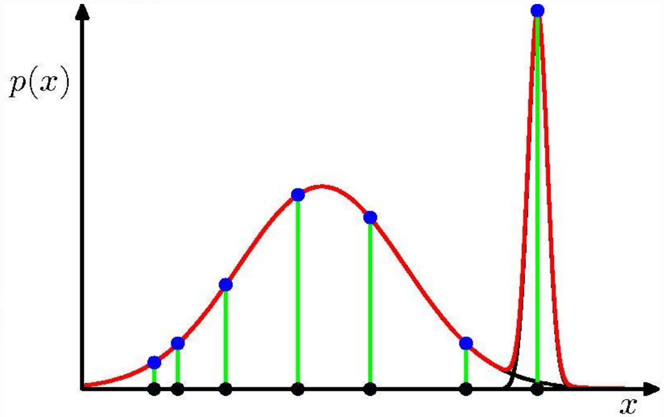


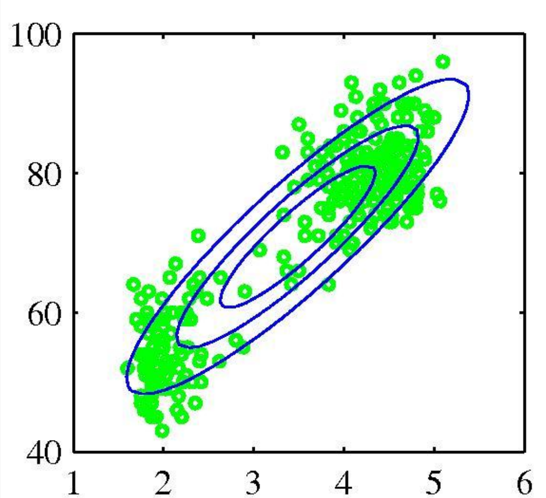
(a)

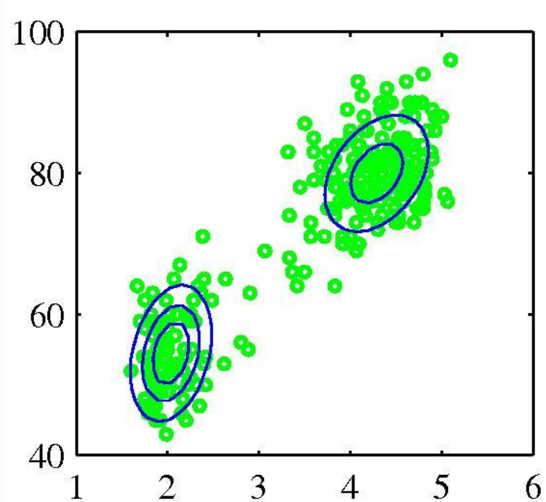
- **How to estimate those parameters?**
  - **ML**
  - **MAP**

How about Multimodal cases?









Mixture Models

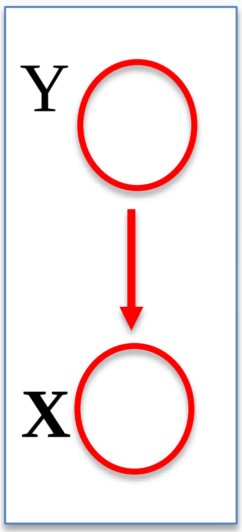
- If you believe that the data set is comprised of several distinct populations
- It has the following form:

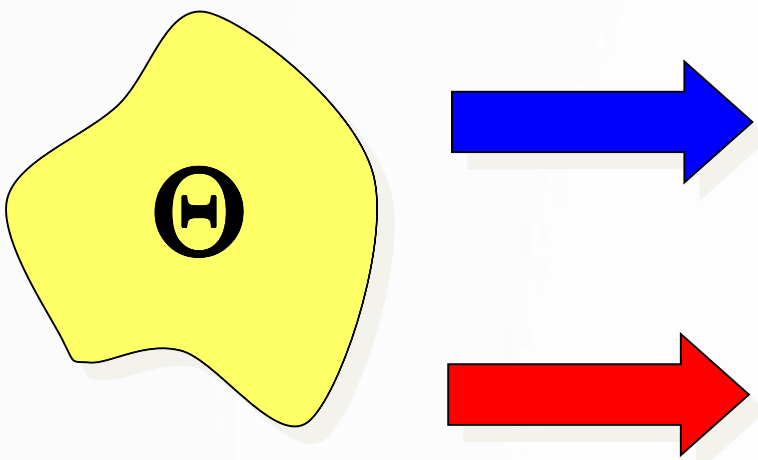
$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{M} \alpha_j p_j(\mathbf{x}|\theta_j) \quad \text{with} \quad \sum_{j=1}^{M} \alpha_j = 1$$

$$\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M)$$

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(\mathbf{x}|\theta_j)$$

Mixture Models





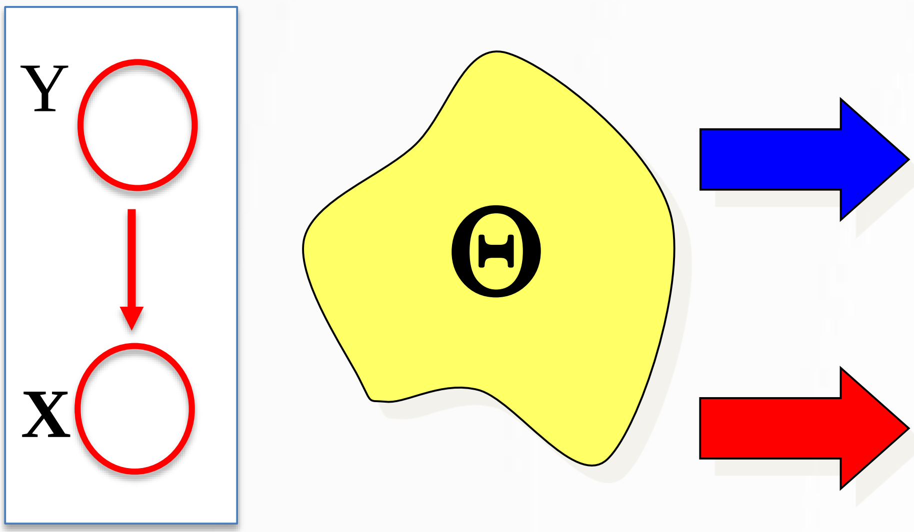
$$\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$$

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

Let $y_i \in \{1, \ldots, M\}$ represents the source that generates the data.

$$p(\mathbf{x}_i \mid \mathbf{\Theta}) = \sum_{j=1}^{M} \alpha_j p_j(\mathbf{x}_i \mid \theta_j)$$

Mixture Models



$$y_i = j$$

$$p(y_i = j \mid \mathbf{\Theta}) = \alpha_j$$

$$\mathbf{x}_i$$

$$p(\mathbf{x}_i \mid y = j, \mathbf{\Theta}) = p_j(\mathbf{x}_i \mid \theta_j)$$

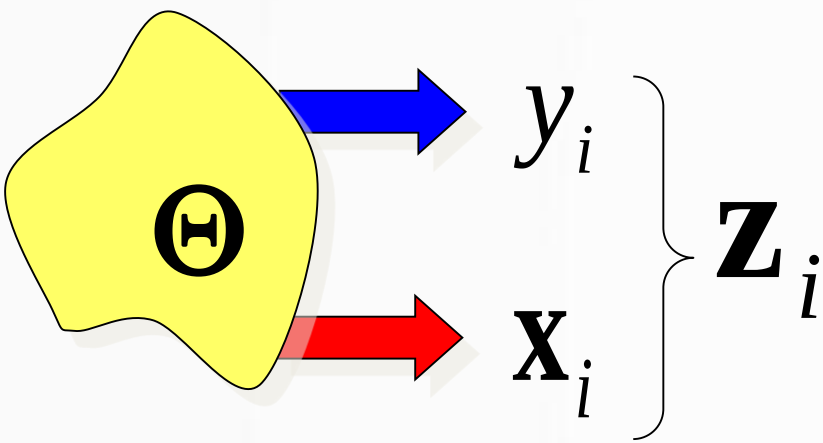
Let $y_i \in \{1, \ldots, M\}$ represents the source that generates the data.

Mixture Models

$$p(\mathbf{x}_i \mid y = j, \boldsymbol{\Theta}) = p_j(\mathbf{x}_i \mid \theta_j)$$

$$p(y_i = j \mid \boldsymbol{\Theta}) = \alpha_j$$

$$p(\mathbf{x}_i \mid \boldsymbol{\Theta}) = \sum_{y_i=1}^{M} p(\mathbf{x}_i, y_i \mid \boldsymbol{\Theta}) = \sum_{y_i=1}^{M} p(y_i = j \mid \boldsymbol{\Theta})\, p(\mathbf{x}_i \mid y = j, \boldsymbol{\Theta}) = \sum_{j=1}^{M} \alpha_j p_j(\mathbf{x}_i \mid \theta_j)$$



$$p(\mathbf{z}_i \mid \boldsymbol{\Theta}) = p(\mathbf{x}_i, y_i \mid \boldsymbol{\Theta}) = p(y_i \mid \mathbf{x}_i, \boldsymbol{\Theta})\, p(\mathbf{x}_i \mid \boldsymbol{\Theta})$$

Mixture Models

$$\boxed{p(\mathbf{x}_i \mid y = j, \mathbf{\Theta}) = p_j(\mathbf{x}_i \mid \theta_j)} \quad \boxed{p(y_i = j \mid \mathbf{\Theta}) = \alpha_j}$$

$$\boxed{p(\mathbf{z}_i \mid \mathbf{\Theta}) = p(\mathbf{x}_i, y_i \mid \mathbf{\Theta}) = p(y_i \mid \mathbf{x}_i, \mathbf{\Theta})\, p(\mathbf{x}_i \mid \mathbf{\Theta})}$$

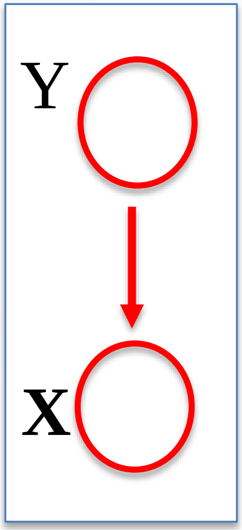
$$p(y_i \mid \mathbf{x}_i, \mathbf{\Theta}) = \frac{p(\mathbf{x}_i, y_i, \mathbf{\Theta})}{p(\mathbf{x}_i, \mathbf{\Theta})} = \frac{p(\mathbf{x}_i \mid y_i, \mathbf{\Theta})\, p(y_i, \mathbf{\Theta})}{p(\mathbf{x}_i, \mathbf{\Theta})}$$

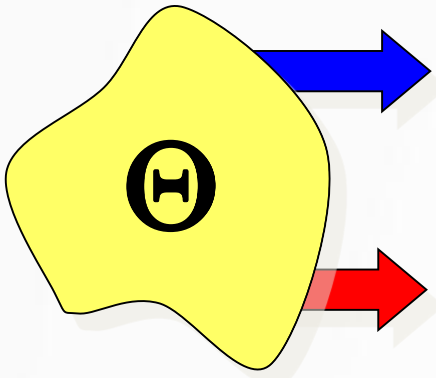
$$= \frac{p(\mathbf{x}_i \mid y_i, \mathbf{\Theta})\, p(y_i \mid \mathbf{\Theta})\, p(\mathbf{\Theta})}{p(\mathbf{x}_i \mid \mathbf{\Theta})\, p(\mathbf{\Theta})} = \frac{p(\mathbf{x}_i \mid y_i, \mathbf{\Theta})\, p(y_i \mid \mathbf{\Theta})}{p(\mathbf{x}_i \mid \mathbf{\Theta})}$$

$$= \frac{p_{y_i}(\mathbf{x}_i \mid \theta_{y_i})\, \alpha_{y_i}}{\sum_{j=1}^{M} \alpha_j p_j(\mathbf{x} \mid \theta_j)}$$

Given $\mathbf{x}$ and $\mathbf{\Theta}$, the conditional density of y can be computed.

Complete-Data Likelihood Function





$$\mathbf{y} = \{y_1, \ldots, y_N\}$$

$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

$$\mathbf{z}_i = (\mathbf{x}_i, y_i)$$

$$\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N)$$

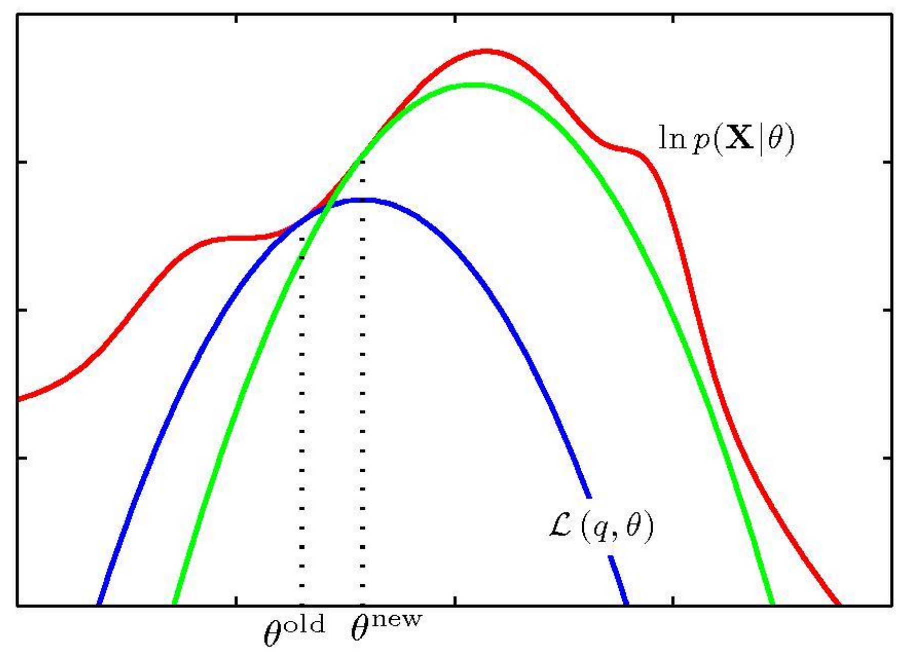
But we don’t know the latent variable y

$$p(\mathbf{Z} \mid \Theta) = p(\mathbf{X}, \mathbf{y} | \Theta$$

$$= \prod_{i=1}^{N} \underbrace{p(\mathbf{x}_i \mid y_i, \Theta)}_{\theta_{y_i}} \underbrace{p(y_i \ldots)}_{\alpha_{y_i}} \ldots \prod_{i=1} \ldots p_{y_i}(\mathbf{x}_i \mid \theta_{y_i})$$

Expectation-Maximization (1)

- If we want to find the local maxima of $f(x)$
  - Define an auxiliary function $A(x, x^t)$
  - Satisfy $f(x) \ge A(x, x^t), \forall x$ and $f(x^t) = A(x^t, x^t)$



$$x^{t+1} = \arg\max A(x, x^t)$$

$$\Downarrow$$

$$A(x^{t+1}, x^t) \ge A(x^t, x^t) = f(x^t)$$

$$f(x^{t+1}) \ge A(x^{t+1}, x^t)$$

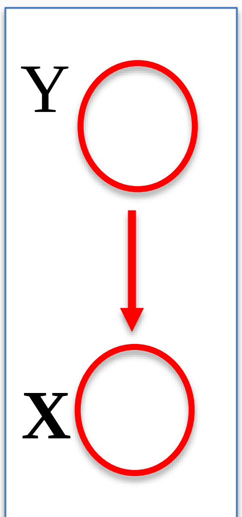
$$\Downarrow$$

$$f(x^{t+1}) \ge f(x^t)$$

$$f(x^{t+2}) \ge f(x^{t+1})$$

Expectation-Maximization (2)

- The goal: finding ML solutions for probabilistic models with latent variables



$$\log p(\mathbf{x} \mid \boldsymbol{\Theta}) = \log \sum_y p(\mathbf{x}, y \mid \boldsymbol{\Theta})$$

Difficult, auxiliary function

$$\log \sum_y q(y) \frac{p(\mathbf{x}, y \mid \boldsymbol{\Theta})}{q(y)} \geq \sum_y q(y) \log\left(\frac{p(\mathbf{x}, y \mid \boldsymbol{\Theta})}{q(y)}\right)$$

$$\sum_y q(y) =$$

Auxiliary function

$$\log p(\mathbf{x} \mid \boldsymbol{\Theta}) = \log \sum_y p(\mathbf{x}, y \mid \boldsymbol{\Theta}) \geq \sum_y q(y) \log\left(\frac{p(\mathbf{x}, y \mid \boldsymbol{\Theta})}{q(y)}\right)$$

Expectation-Maximization (2)

- What is the gap?

$$\log p(\mathbf{x}\mid\Theta) - \sum_y q(y)\log\left(\frac{p(\mathbf{x},y\mid\Theta)}{q(y)}\right) =$$

$$\sum_y q(y)\left[\log p(\mathbf{x}\mid\Theta) - \log\left(\frac{p(\mathbf{x},y\mid\Theta)}{q(y)}\right)\right] =$$

$$= \sum_y q(y)\log\left(\frac{q(y)\,p(\mathbf{x}\mid\Theta)}{p(\mathbf{x},y\mid\Theta)}\right) = \sum_y q(y)\log\left(\frac{q(y)}{p(y\mid x,\Theta)}\right)$$

$$= KL(q(y)\,\|\,p(y\mid x,\Theta)) \geq 0$$

$$D_{KL}(P\|Q) = \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

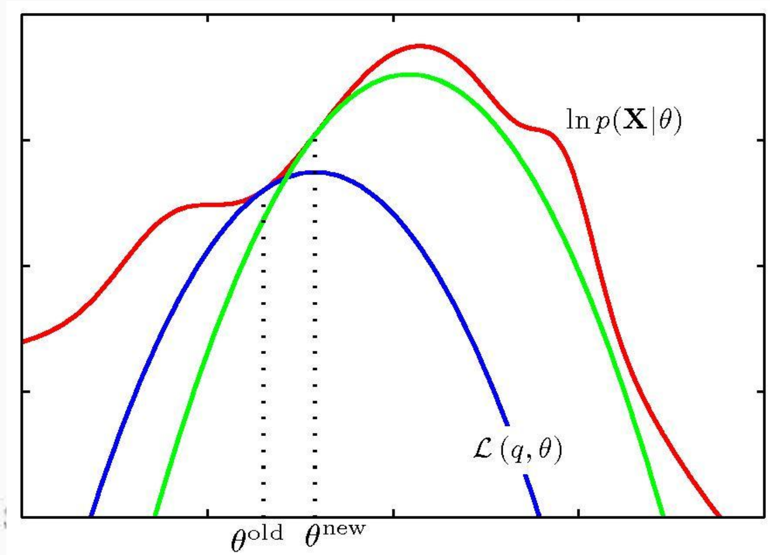
$$KL(q(y)\,\|\,p(y\mid x,\Theta)) = 0 \quad iff\ q(y) = p(y\mid x,\Theta)$$

Expectation-Maximization (2)

- We have shown that

$$\log p(\mathbf{x}\,|\,\mathbf{\Theta}) - \underbrace{\sum_y q(y)\log(\frac{p(\mathbf{x},y\,|\,\mathbf{\Theta})}{q(y)})}_{A(\mathbf{\Theta},\mathbf{\Theta}^{old})} = KL(q(y)\,\|\,p(y\,|\,x,\mathbf{\Theta})) \ge 0$$

$$\boxed{q(y) = p(y\,|\,x,\mathbf{\Theta}^{old})}$$



$$\sum_y q(y)\log(\frac{p(\mathbf{x},y\,|\,\mathbf{\Theta}^{old})}{q(y)}) =$$

$$\sum_y p(y\,|\,x,\mathbf{\Theta}^{old})\log(\frac{p(\mathbf{x},y\,|\,\mathbf{\Theta}^{old})}{p(y\,|\,x,\mathbf{\Theta}^{old})}) =$$

$$\sum_y p(y\,|\,x,\mathbf{\Theta}^{old})\log(p(\mathbf{x}\,|\,\mathbf{\Theta}^{old})) = \log(p(\mathbf{x}\,|\,\mathbf{\Theta}^{old}))$$

University

Expectation-Maximization (3)

- Any physical meaning of the auxiliary function?

$$A(\boldsymbol{\Theta},\boldsymbol{\Theta}^{old}) = \sum_y p(y \mid x,\boldsymbol{\Theta}^{old}) \log\left(\frac{p(\mathbf{x}, y \mid \boldsymbol{\Theta})}{p(y \mid x,\boldsymbol{\Theta}^{old})}\right) =$$

$$\sum_y p(y \mid x,\boldsymbol{\Theta}^{old}) \log(p(\mathbf{x}, y \mid \boldsymbol{\Theta})) - \sum_y \underbrace{p(y \mid x,\boldsymbol{\Theta}^{old}) p(y \mid x,\boldsymbol{\Theta}^{old})}_{\text{Independent of } \boldsymbol{\Theta}}$$

$$= \sum_y p(y \mid x,\boldsymbol{\Theta}^{old}) \log(p(\mathbf{x}, y \mid \boldsymbol{\Theta})) - const$$

Conditional expectation of the complete data log likelihood

Expectation-Maximization (4)

1. Choose an initial setting for the parameters $\boldsymbol{\Theta}^{old}$
2. E step: evaluate $p(y \mid x, \boldsymbol{\Theta}^{old})$

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{old}) = \sum_{y} p(y \mid x, \boldsymbol{\Theta}^{old}) \log(p(\mathbf{x}, y \mid \boldsymbol{\Theta}))$$

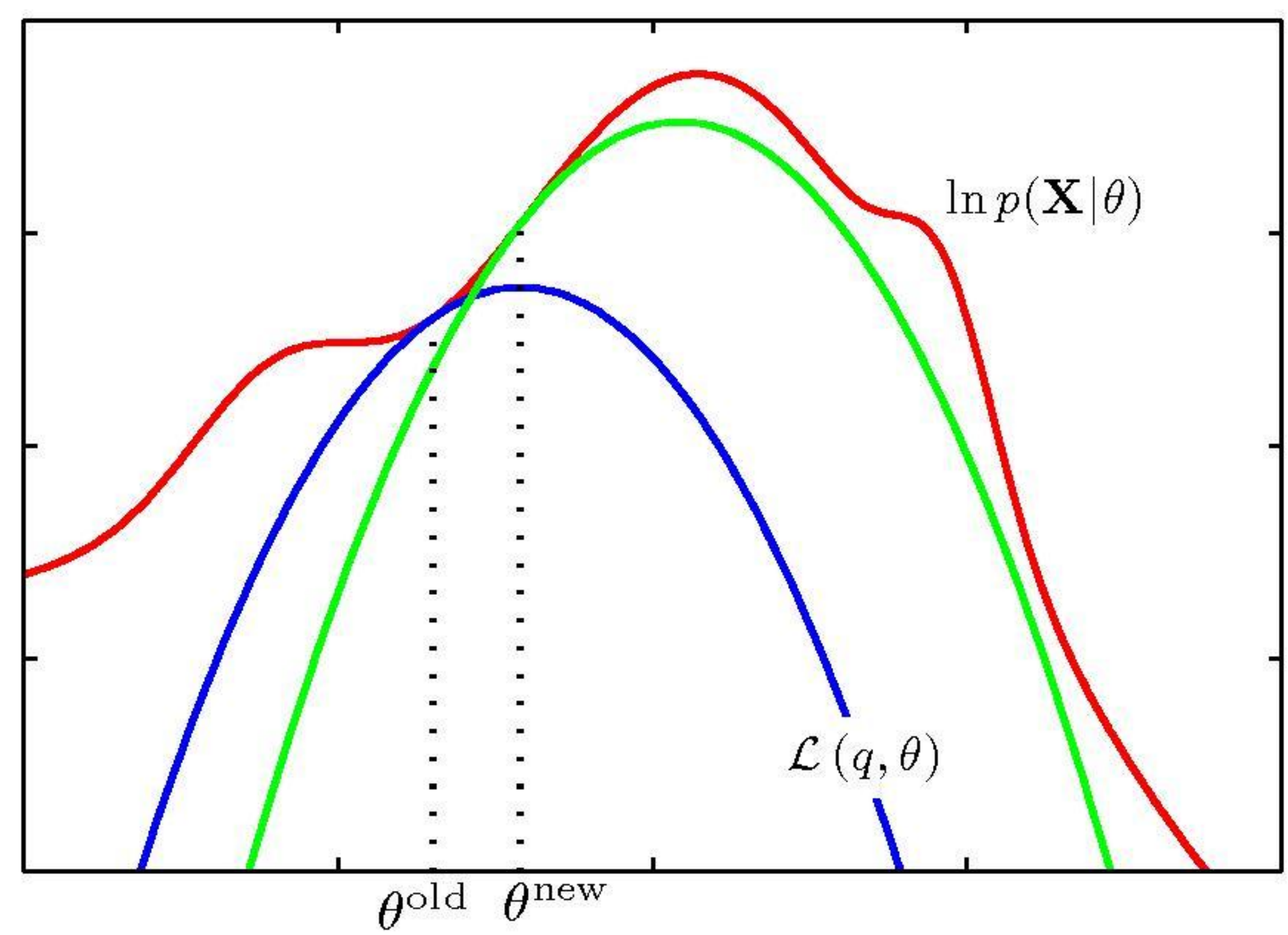
3. M step: evaluate $\boldsymbol{\Theta}^{new}$ given by

$$\boldsymbol{\Theta}^{new} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{old})$$

4. Check for convergence

   if not, $\boldsymbol{\Theta}^{old} \leftarrow \boldsymbol{\Theta}^{new}$ and return to step 2

Expectation-Maximization (4)



Guassian Mixture Model (GMM)

Guassian model of a d-dimensional source, say j :

$$p_j(\mathbf{x}\,|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)=\frac{1}{(2\pi)^{d/2}\,|\,\boldsymbol{\Sigma}_j\,|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]$$

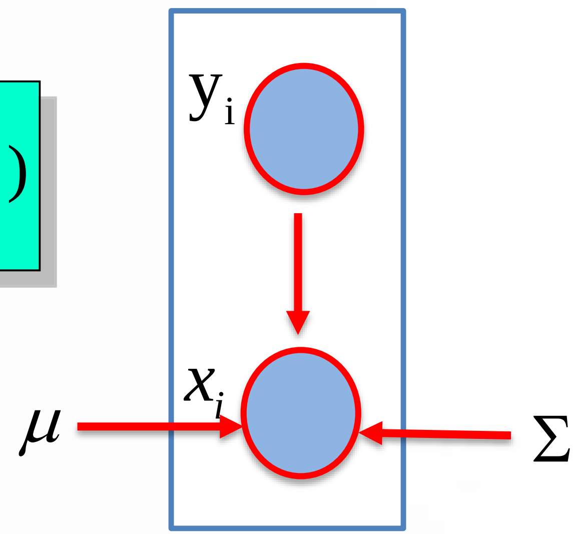
$$\theta_j=(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

GMM with M sources:

$$p_j(\mathbf{x}\,|\,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\cdots,\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M)=\sum_{j=1}^{M}\alpha_j p_j(\mathbf{x}\,|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

$$\alpha_j\geq 0$$

$$\sum\alpha_j=1$$



Mixture Model

$$p(\mathbf{x}|\Theta)=\sum_{l=1}^{M}\alpha_l p_l(\mathbf{x}|\theta_l)$$

$\Theta=(\alpha \ldots \theta_M)$ subj

Correlated with α_l only.

Correlated with θ_l only.

$$Q(\ldots \sum_{y} p(y|x,\Theta^{old})\log(p(\mathbf{x},y|\Theta))$$

$$Q(\Theta,\Theta^g)=\sum_{l=1}^{M}p(l|\mathbf{x}_i,\Theta^g)\sum_{i=1}^{N}\log(\alpha_l p_l(\mathbf{x}_i|\theta_l))$$

$$p(\mathbf{X},\mathbf{y}|\Theta)=\prod_{i=1}^{N}\alpha_{y_i}p_{y_i}(\mathbf{x}_i|\theta_{y_i})$$

$$Q(\Theta,\Theta^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l|\mathbf{x}_i,\Theta^g)+\sum_{l=1}^{M}\sum_{i=1}^{N}\log[p_l(\mathbf{x}_i|\theta_l)]p(l|\mathbf{x}_i,\Theta^g)$$

M step: $\Theta^{new}=\arg\max_{\Theta} Q(\Theta,\Theta^{old})$

$$Q(\mathbf{\Theta},\mathbf{\Theta}^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\sum_{l=1}^{M}\sum_{i=1}^{N}\log[\,p_l(\mathbf{x}_i\,|\,\theta_l)]p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)$$

Finding α_l

Due to the constraint on α_l's, we introduce *Lagrange Multiplier* λ, and solve the following equation.

$$\frac{\partial}{\partial\alpha_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\lambda\left(\sum_{l=1}^{M}\alpha_l-1\right)\right]=0,\qquad l=1,\ldots,M$$

$$\sum_{i=1}^{N}\frac{1}{\alpha_l}p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\lambda=0,\qquad l=1,\ldots,M$$

$$\sum_{i=1}^{N}p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\alpha_l\lambda=0,\qquad l=1,\ldots,M$$

$$Q(\mathbf{\Theta},\mathbf{\Theta}^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\sum_{l=1}^{M}\sum_{i=1}^{N}\log[\,p_l(\mathbf{x}_i\,|\,\theta_l)]p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)$$

Finding α_l

$$\lambda=-N$$

$$\sum_{l=1}^{M}\sum_{i=1}^{N}p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\lambda\sum_{l=1}^{M}\alpha_l=0$$

$$\underbrace{\sum_{i=1}^{N}\underbrace{\sum_{l=1}^{M}p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)}_{1}}_{N}+\lambda\underbrace{\sum_{l=1}^{M}\alpha_l}_{1}=0$$

$$\sum_{i=1}^{N}p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\alpha_l\lambda=0,\qquad l=1,\ldots,M$$

$$Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)+\sum_{l=1}^{M}\sum_{i=1}^{N}\log[\,p_l(\mathbf{x}_i\,|\,\theta_l)]p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)$$

Finding α_l

$$\lambda=-N$$

$$\alpha_l=\frac{1}{N}\sum_{i=1}^{N}p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)$$

$$p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)=\frac{\alpha_l^g\,p_l(\mathbf{x}_i\,|\,\theta_l^g)}{\sum_{j=1}^{M}\alpha_j^g\,p_j(\mathbf{x}\,|\,\theta_j^g)}$$

$$\sum_{i=1}^{N}p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)+\alpha_l\lambda=0,\qquad l=1,\ldots,M$$

$$Q(\mathbf{\Theta},\mathbf{\Theta}^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)+\underbrace{\sum_{l=1}^{M}\sum_{i=1}^{N}\log[\,p_l(\mathbf{x}_i\,|\,\theta_l)]p(l\,|\,\mathbf{x}_i,\mathbf{\Theta}^g)}_{\text{Only need to maximize this term}}$$

Finding θ_l

Consider GMM

$$p_l(\mathbf{x}\,|\,\boldsymbol{\mu}_l,\mathbf{\Sigma}_l)=\frac{1}{(2\pi)^{d/2}\,|\,\mathbf{\Sigma}_l\,|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\mathbf{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)\right]$$

$$\theta_l=(\boldsymbol{\mu}_l,\mathbf{\Sigma}_l)$$

$$\log[\,p_l(\mathbf{x}\,|\,\boldsymbol{\mu}_l,\mathbf{\Sigma}_l)]=\underbrace{-\tfrac{d}{2}\log 2\pi}_{\text{unrelated}}-\tfrac{1}{2}\log|\,\mathbf{\Sigma}_l\,|^{1/2}-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\mathbf{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)$$

$$Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)+\underbrace{\sum_{l=1}^{M}\sum_{i=1}^{N}\log[\,p_l(\mathbf{x}_i\,|\,\theta_l)]p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)}_{\text{Only need to maximize this term}}$$

Finding θ_l

Therefore, we want to maximize:

$$Q'(\boldsymbol{\Theta},\boldsymbol{\Theta}^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\left(-\tfrac{1}{2}\log|\boldsymbol{\Sigma}_l|^{1/2}-\tfrac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_l)\right)p(l\,|\,\mathbf{x}_i,\boldsymbol{\Theta}^g)$$

How?

$$p(l \mid \mathbf{x}_i, \mathbf{\Theta}^g) = \frac{\alpha_l^g \, p_l(\mathbf{x}_i \mid \theta_l^g)}{\sum_{j=1}^{M} \alpha_j^g \, p_j(\mathbf{x} \mid \theta_j^g)}$$

Finding θ_l

Therefore, we want to maximize:

$$Q'(\mathbf{\Theta}, \mathbf{\Theta}^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \left(-\tfrac{1}{2} \log |\mathbf{\Sigma}_l|^{1/2} - \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_l)^T \mathbf{\Sigma}_l^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_l) \right) p(l \mid \mathbf{x}_i, \mathbf{\Theta}^g)$$

$$\boldsymbol{\mu}_l = \frac{\sum_{i=1}^{N} \mathbf{x}_i \, p(l \mid \mathbf{x}_i, \mathbf{\Theta}^g)}{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \mathbf{\Theta}^g)} \qquad \Sigma_l = \frac{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \mathbf{\Theta}^g)(\mathbf{x}_i - \boldsymbol{\mu}_l)(\mathbf{x}_i - \boldsymbol{\mu}_l)^T}{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \mathbf{\Theta}^g)}$$

Summary

EM algorithm for GMM

Given an initial guess $\boldsymbol{\Theta}^g$, find $\boldsymbol{\Theta}^{new}$ as follows

$$p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g) = \frac{\alpha_l^g p_l(\mathbf{x}_i \mid \theta_l^g)}{\sum_{j=1}^{M} \alpha_j^g p_j(\mathbf{x} \mid \theta_j^g)}$$

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)$$

$$\boldsymbol{\mu}_l^{new} = \frac{\sum_{i=1}^{N} \mathbf{x}_i p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)}{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)(\mathbf{x}_i - \boldsymbol{\mu}_l^{new})(\mathbf{x}_i - \boldsymbol{\mu}_l^{new})^T}{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)}$$

Not converge

$$\boldsymbol{\Theta}^g \leftarrow \boldsymbol{\Theta}^{new}$$