A Tutorial of Gaussian Mixture Model (GMM)

Nov. 2015

Qiuqiang Kong
q.kong@qmul.ac.uk

- Gaussian Distribution
- Gaussian Mixture Model (GMM)
- A General PGM View of EM algorithm
- GMM Revise
- Tricks

Gaussian Distribution

The probability density function of Gaussian distribution is

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \qquad (1)$$

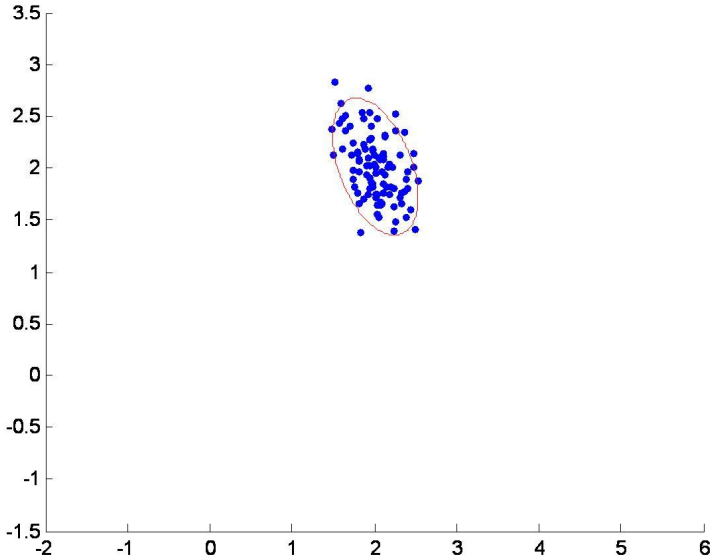
Gaussian distribution is denominated by parameters: $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

where

$\boldsymbol{\mu}_{p\times 1}$ is mean $\boldsymbol{\Sigma}_{p\times p}$ is covariance

Q: How to estimate the parameters?

A: Using maximum likelihood estimation.



Demonstration of 2-D Gaussian distribution

Applying maximum likelihood estimation method

$$\{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\} = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\mathbf{x}_1, \ldots, \mathbf{x}_N \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2}$$

where $L(\mathbf{x}_1, \ldots, \mathbf{x}_N \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ Is likelihood function

$$L(\mathbf{x}_1, \ldots, \mathbf{x}_N \,|\, \theta) = \ln \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \theta)$$

$$= \sum_{n=1}^{N} \left(-\frac{D}{2} \ln(2\pi) - \frac{1}{2} |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \right) \tag{3}$$

Optimize (3) with respect to **μ**, **Σ** simultaneously is difficult. However, sample mean is independent of sample variance. So the optimization can be done with respect to **μ**, **Σ** separately.

Let the derivative of likelihood L with respect to **μ, Σ** be zero. We get estimation of **μ, Σ.**

$$\frac{\partial L}{\partial \boldsymbol{\mu}} = \sum_{n=1}^{N} -\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}} = \sum_{n=1}^{N}\left(-\frac{1}{2}\boldsymbol{\Sigma}^{-\mathrm{T}} + \frac{1}{2}\boldsymbol{\Sigma}^{-\mathrm{T}}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-\mathrm{T}}\right) = 0$$

$$\Rightarrow \quad \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad (4)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \hat{\boldsymbol{\mu}})(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$$

Because $L(\mathbf{x}_1, \ldots, \mathbf{x}_N \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is convex function,

The extreme point is global maximum point.

$\hat{\boldsymbol{\mu}}$ is unbiased estimation

$\hat{\boldsymbol{\Sigma}}$ is biased estimation, when n-> ∞, it is asymptotic unbiased

Hint[1]:

$$\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

$$\frac{\partial |\mathbf{X}|}{\mathbf{X}} = |\mathbf{X}|\mathbf{X}^{-\mathrm{T}}$$

$$\frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{X}^{-1}\mathbf{b}}{\mathbf{X}} = -\mathbf{X}^{-\mathrm{T}}\mathbf{a}\mathbf{b}^{\mathrm{T}}\mathbf{X}^{-\mathrm{T}}$$

[1] Petersen, Kaare Brandt, and Michael Syskind Pedersen. "The matrix cookbook." *Technical University of Denmark* 7 (2008): 15.

Gaussian Mixture Model (GMM)

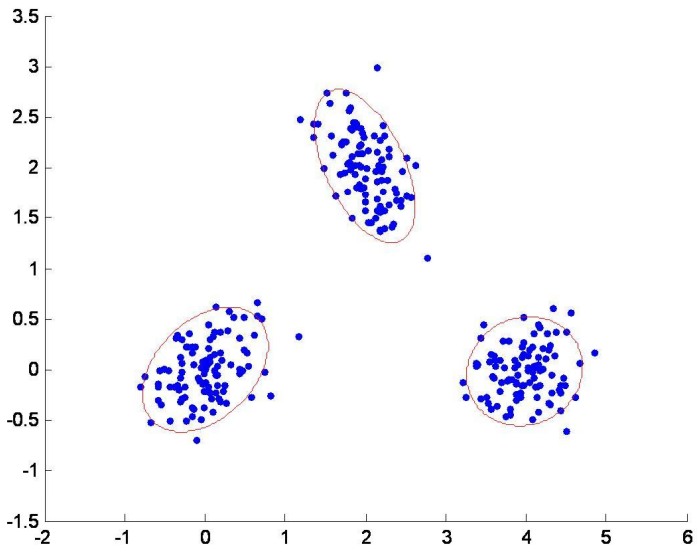
Intuitively, GMM is the sum of several weighted Gaussian distributions.

Probability density function of GMM:

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (5)$$

where $\sum_{k=1}^{K} \pi_k = 1$

Parameters: $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\} \quad k = 1, \ldots, K$



Demostration of GMM, mix number: 3

Illustration

It is unknown which mixture produce $\mathbf{x}_n$.

Define event A_k: "the k-th mixture occurs", k=1, …, K

with probability on A_k: $p(A_k) = \pi_k$

According to Total Probability Theorem

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|A_k)p(A_k) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (6)$$

which has same form as (5).

GMM parameter estimation using maximum likelihood

$$L(\mathbf{x}_1,...,\mathbf{x}_N|\theta)=\ln\prod_{n=1}^{N}p(\mathbf{x}_n|\theta)=\sum_{n=1}^{N}\ln\sum_{k=1}^{K}\pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \qquad (7)$$

Take constraint $\sum_{k=1}^{K}\pi_k=1$ Into account. Introduce Lagrange function

$$R(\theta,\lambda)=\sum_{n=1}^{N}\ln\sum_{k=1}^{K}\pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right) \qquad (8)$$

To optimize $R(\theta,\lambda)$ Let the derivative of $R(\theta,\lambda)$ with respect to

$\theta=\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}\quad k=1,...,K$ be zero.

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \left(-\boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) = 0 \quad k = 1, \ldots, K$$

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \frac{\pi_k N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-\mathrm{T}} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-\mathrm{T}} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-\mathrm{T}} \right) = 0 \quad k = 1, \ldots, K \tag{9}$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda = 0 \quad k = 1, \ldots, K$$

This is a transcendental equation set, can not be solved analytically.

Solve: introduce iterative method (EM algorithm)

while (not converge)

E step: use old parameters to calculate

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \left(-\boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) = 0 \quad k = 1, \ldots, K \tag{10}$$

M step: update parameters

notice $$\frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j N(\mathbf{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}=\frac{p(\mathbf{x}|A_k)p(A_k)}{p(\mathbf{x})}=p(A_k|\mathbf{x}) \tag{11}$$

EM algorithm

1. Choose initial θold

2. E step: (Expectation)

$$\eta_{nk}=p(A_k|\mathbf{x}_n\theta^{old}) \tag{12}$$

where $\theta=\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}\quad k=1,\ldots,K$

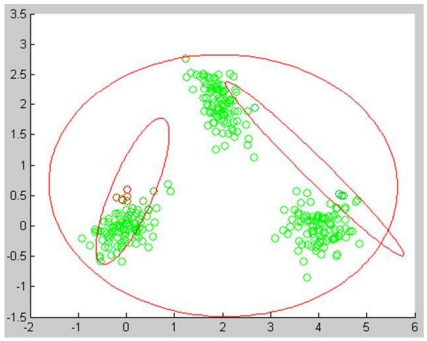
3. M step: (Maximization)

$$\pi_k=\frac{1}{N}\sum_{n=1}^{N}\gamma_{nk}\qquad \boldsymbol{\mu}_k=\frac{\sum_{n=1}^{N}\gamma_{nk}\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_{nk}}\qquad \boldsymbol{\Sigma}_k=\frac{\sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N}\gamma_{nk}} \tag{13}$$

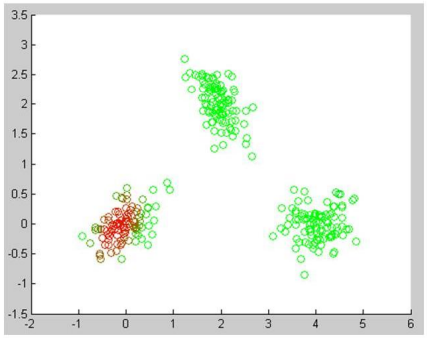
4. If converge then stop. Otherwise goto 2.

Q: Will likelihood decrease using EM algorithm?
Q: Will likelihood converge using EM algorithm?

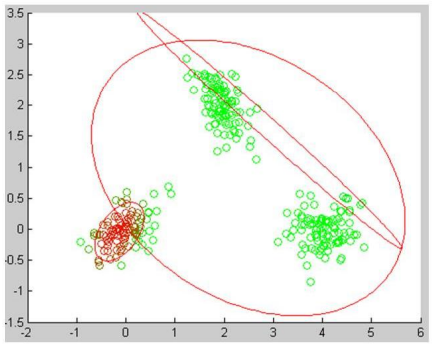
Example of EM algorithm



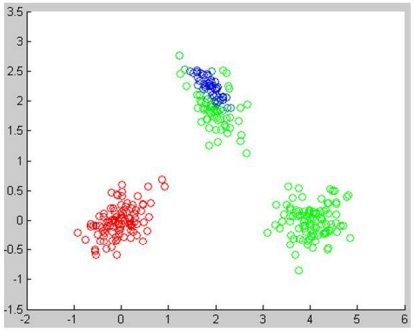
Initialize parameters



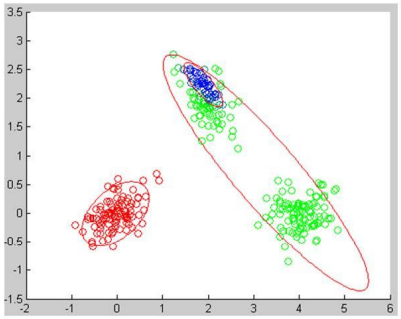
E step



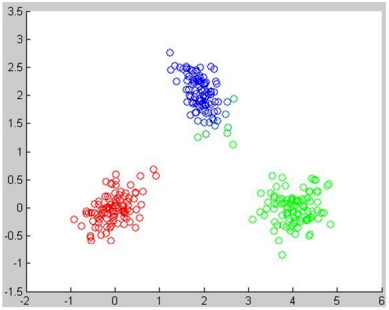
M step



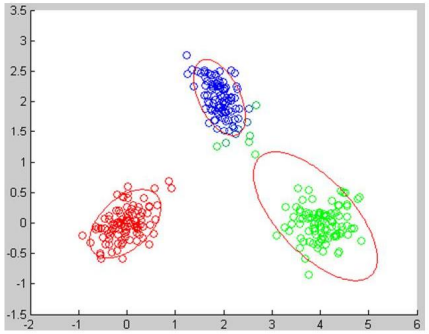
E step



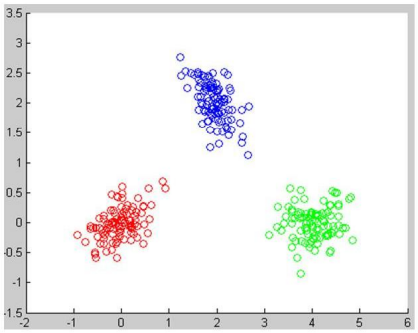
M step



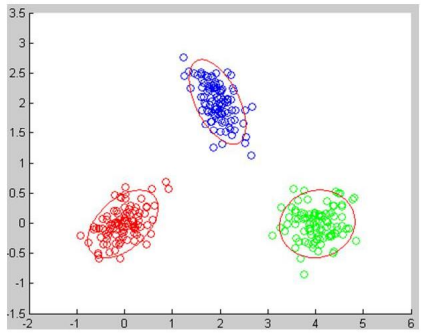
E step



M step



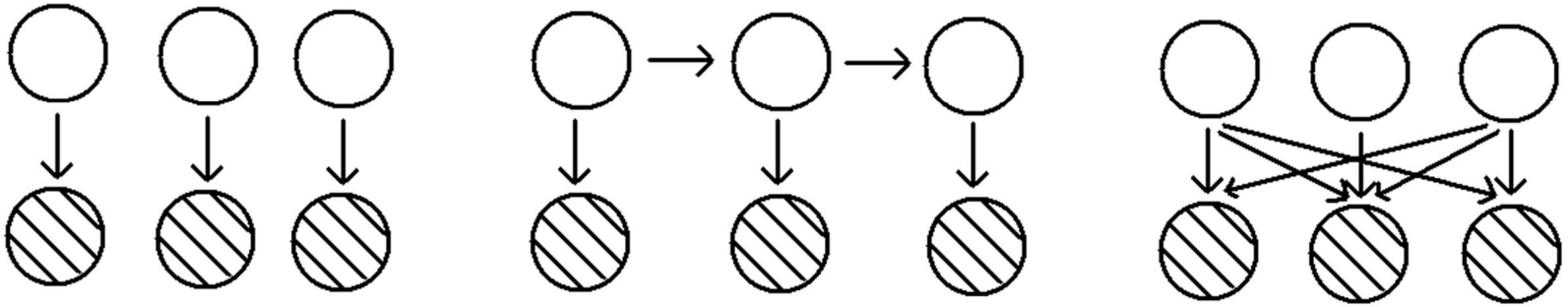
E step



M step

A General View of EM Algorithm

In general, EM algorithm is a method to estimate parameters in a probabilistic graphic model (PGM) with latent variables.



GMM **HMM** **Others**

Black circle is the observed variable. White circle is latent variable.

The estimation of θ will be $\hat{\theta} = \arg\max_{\theta} \ln p(\mathbf{X}|\theta)$

where $\ln p(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} p(\mathbf{XZ}|\theta)$

Obviously, it is needed to sum all **Z** before applying log, which is difficult.
EM algorithm is aimed at solving this problem.

Introduce auxiliary function q(**Z**) over **Z**

$$\ln p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}|\theta)$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{XZ}|\theta)}{p(\mathbf{Z}|\mathbf{X}\theta)}$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{XZ}|\theta) q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}\theta) q(\mathbf{Z})}$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{XZ}|\theta)}{q(\mathbf{Z})} + \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}\theta)}$$

$$= \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{XZ}|\theta)}{q(\mathbf{Z})}}_{①} + \underbrace{\mathrm{KL}\left(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}\theta)\right)}_{②} \qquad (14)$$

② is KL divergence between q(**Z**) and p(**Z**|**X**θ), which is non negative.
It only equals 0 iff q(**Z**) and p(**Z**|**X**θ).

E step: Let q(**Z**) = p(**Z**|**X**θ^{old})

M step: optimize

$$\ln p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{XZ}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) = Q(\theta, \theta^{\text{old}}) + \text{const} \qquad (15)$$

with respect to θ

where $Q(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{XZ}|\theta)$ (16)

EM algorithm for PGM with latent variables

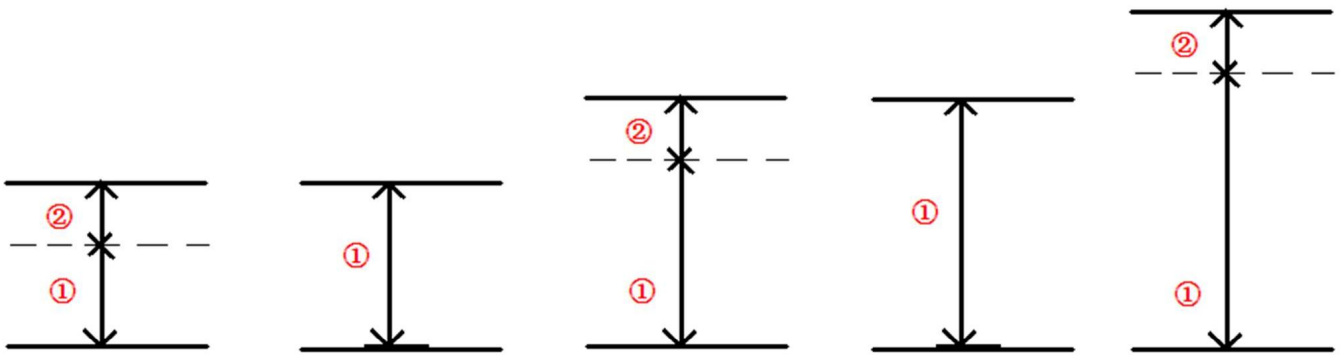
1. Init parameters

2. E step: $q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X}\theta^{old})$ (17)

3. M step: $\hat{\theta} = \arg\max_{\theta} Q(\theta, \theta^{\text{old}})$ (18)

where $Q(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{XZ}|\theta)$

4. If converge then stop, otherwise goto 2



Initial paras **E step** **M step** **E step** **M step**

Illustration[1]: For a function which is difficult to find extreme point (red curve), E step is find the complete data function (blue curve), M step is update θ.

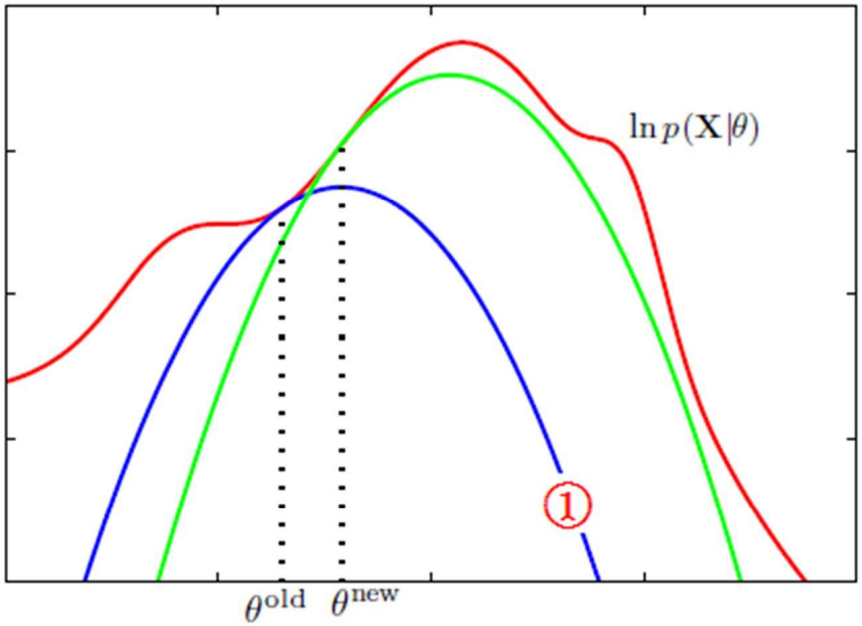


Illustration of EM algorithm

Q: Will EM algorithm converge?

A: Under some conditions, EM algorithm will converge to stationary point[2].

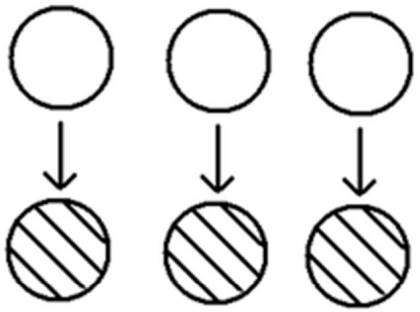
[1] Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006. P453

[2] Wu, CF Jeff. "On the convergence properties of the EM algorithm." *The Annals of statistics* (1983): 95-103.

GMM Revise

Define event Ak: “the k-th mixture occurs”, k=1, …, K

with probability on A_k: $p(A_k) = \pi_k$



GMM

Let z be K-dimensional binary random variable having 1-of-K representation

$$\mathbf{z} : A_k \rightarrow (0,...,0,1,0,...,0) \quad k = 1,...,K \qquad (19)$$

(the k-th position is 1)

The probability over **z** is

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad (20)$$

Thinking:

Q: Why choose random variable in this form?

Can we choose others, eg.

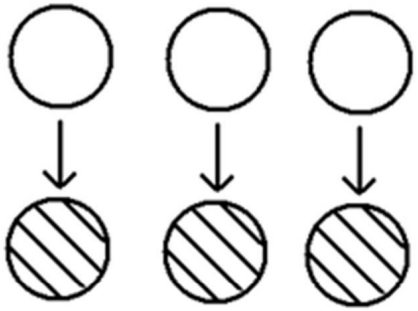
z: A_k -> k, k=1,…,K?

A: 1-of-K representation belongs to exponential family, which will make computation feasible

For GMM model, $Q(\theta,\theta^{old})$ is

$$
\begin{aligned}
Q(\theta,\theta^{\text{old}}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{XZ}|\theta) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln \prod_{n=1}^{N} p(\mathbf{x}_n\mathbf{z}_n|\theta) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \sum_{n=1}^{N} \ln p(\mathbf{x}_n\mathbf{z}_n|\theta) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{x}_1\mathbf{z}_1|\theta) + \ldots + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}\theta^{\text{old}}) \ln p(\mathbf{x}_N\mathbf{z}_N|\theta) \\
&= \sum_{\mathbf{z}_1} p(\mathbf{z}_1|\mathbf{x}_1\theta^{\text{old}}) \ln p(\mathbf{x}_1\mathbf{z}_1|\theta) + \ldots + \sum_{\mathbf{z}_n} p(\mathbf{z}_N|\mathbf{x}_N\theta^{\text{old}}) \ln p(\mathbf{x}_N\mathbf{z}_N|\theta) \\
&= \sum_{n=1}^{N} \sum_{\mathbf{z}_n} p(\mathbf{z}_n|\mathbf{x}_n\theta^{\text{old}}) \ln p(\mathbf{x}_n\mathbf{z}_n|\theta)
\end{aligned}
\qquad (21)
$$

Independence of PGM



GMM

substitute

$$p(\mathbf{x}_n z_{nk}|\theta) = p(z_{nk}|\theta)p(\mathbf{x}_n|z_{nk}\theta) = \pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (22)$$

into (21), and denote

$$\gamma_{nk} = p(z_{nk}|\mathbf{x}_n\theta^{\text{old}}) \qquad (23)$$

We obtain

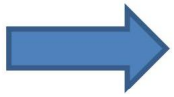
$$Q(\theta,\theta^{\text{old}})=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\left[\ln\pi_k-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_k\right|-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\right] \tag{24}$$

Take constraint $\sum_{k=1}^{K}\pi_k=1$ into consideration, introduce Lagrange function

$$R(\theta,\lambda)=Q(\theta,\theta^{\text{old}})+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right) \tag{25}$$

Let the derivative of R(θ,λ) be zero,
we get the estimation of $\theta=\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}\quad k=1,\ldots,K$

$$\frac{\partial R(\theta,\lambda)}{\partial\pi_k}=0\quad(k=1,\ldots,K)$$
$$\frac{\partial R(\theta,\lambda)}{\partial\boldsymbol{\mu}_k}=0\quad(k=1,\ldots,K)$$
$$\frac{\partial R(\theta,\lambda)}{\partial\boldsymbol{\Sigma}_k}=0\quad(k=1,\ldots,K)$$



$$\pi_k=\frac{1}{N}\sum_{n=1}^{N}\gamma_{nk}\quad(k=1,\ldots,K)$$
$$\boldsymbol{\mu}_k=\frac{\sum_{n=1}^{N}\gamma_{nk}\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_{nk}}\quad(k=1,\ldots,K) \tag{26}$$
$$\boldsymbol{\Sigma}_k=\frac{\sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N}\gamma_{nk}}\quad(k=1,\ldots,K)$$

EM algorithm for GMM

1. Choose initial θold

2. E step: (Expectation)

$$\eta_{nk} = p(A_k \left| \mathbf{x}_n \theta^{old} \right.)$$

where $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\} \quad k = 1, \ldots, K$

3. M step: (Maximization)

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk} \qquad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}} \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma_{nk}}$$

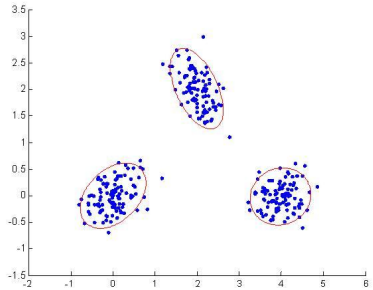
4. If converge then stop. Otherwise goto 2.

This is the same as previous result.

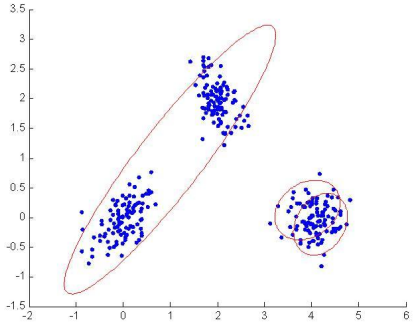
Tricks

1. GMM is a non convex problem. There are many local minimum.

   Solve: Use kmeans to initialize parameter.
   Initialize parameters and run EM algorithm for several times and choose the best model.



Local minimum



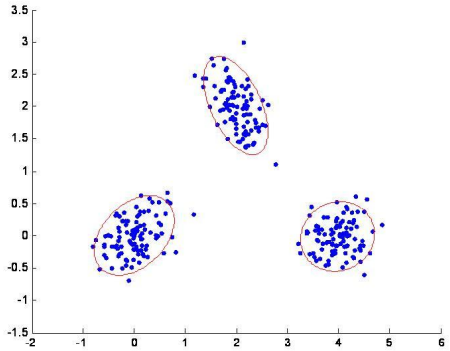
Local minimum

2. if eig($\mathbf{\Sigma}$) is too small. Then $\mathbf{\Sigma}^{-1}$ will be unstable.

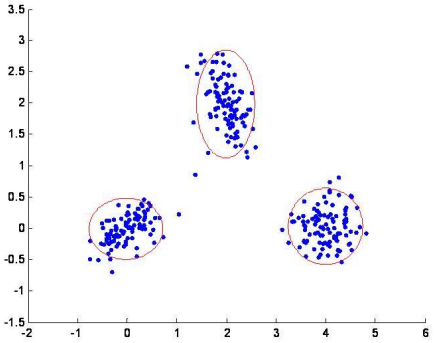
   Solve: if eig($\mathbf{\Sigma}$) < ε then $\mathbf{\Sigma} = \mathbf{\Sigma} + \sigma\mathbf{I}$

3. The number of parameters of full GMM is K+K*D+K*D*D. If dataset is small then model will overfit.

Solve: Use diagonal **Σ** instead. The number of parameters will decrease to K+K*D+K*D.



Full covariance



Diagnoal covariance

4. Gaussian probability density function can be underflow exponentially.

Solve: use log and normalization to solve this problem.

Matlab Code

https://github.com/qiuqiangkong/matlab-gmm

THANK YOU!