Gaussian Mixture Models & EM

CE-717: Machine Learning

Sharif University of Technology

M. Soleymani

Fall 2016

Mixture Models: definition

- Mixture models: Linear supper-position of mixtures or components

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{K} P(M_j)\, p\left(\boldsymbol{x}\middle|M_j; \boldsymbol{\theta}_j\right)$$

  - $\sum_{j=1}^{K} P(M_j) = 1$
  - $P(M_j)$: the prior probability of j-th mixture
  - $\boldsymbol{\theta}_j$: the parameters of j-th mixture
  - $p\left(\boldsymbol{x}\middle|M_j; \boldsymbol{\theta}_j\right)$: the probability of $\boldsymbol{x}$ according to j-th mixture

- Framework for finding more complex probability distributions
  - Goal: estimate $p(\boldsymbol{x}|\theta)$ E.g., Multi-modal density estimation

Gaussian Mixture Models (GMMs)

- Gaussian Mixture Models: $p(\boldsymbol{x}|M_j;\boldsymbol{\theta}_j)\sim N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$

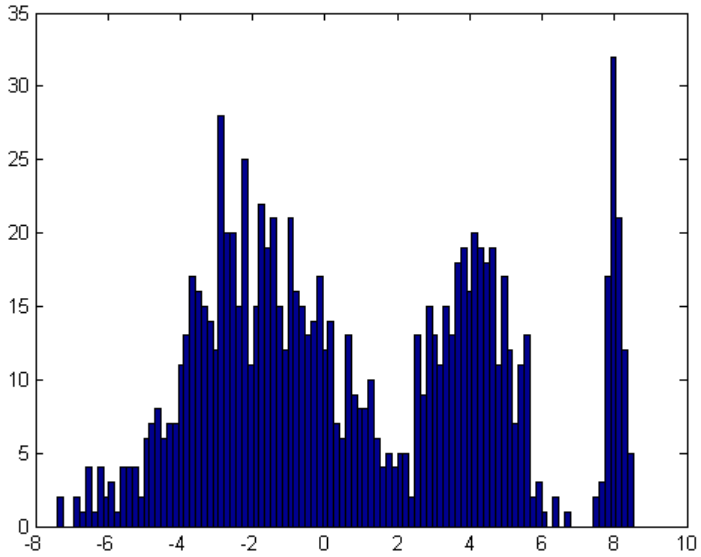
$$p(\boldsymbol{x})=\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

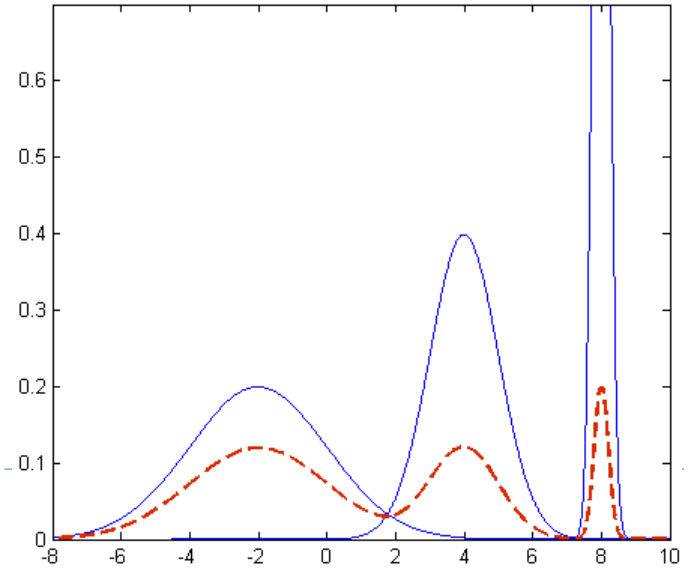
$$0\le\pi_j\le 1$$

$$\sum_{j=1}^{K}\pi_j=1$$

- Fitting the Gaussian mixture model
  - Input: data points $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$
  - Goal: find the parameters of GMM $(\pi_j,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,j=1,\dots,K)$

GMM: 1-D Example





$\mu_1 = -2$

$\sigma_1 = 2$

$\pi_1 = 0.6$

$\mu_2 = 4$

$\sigma_2 = 1$

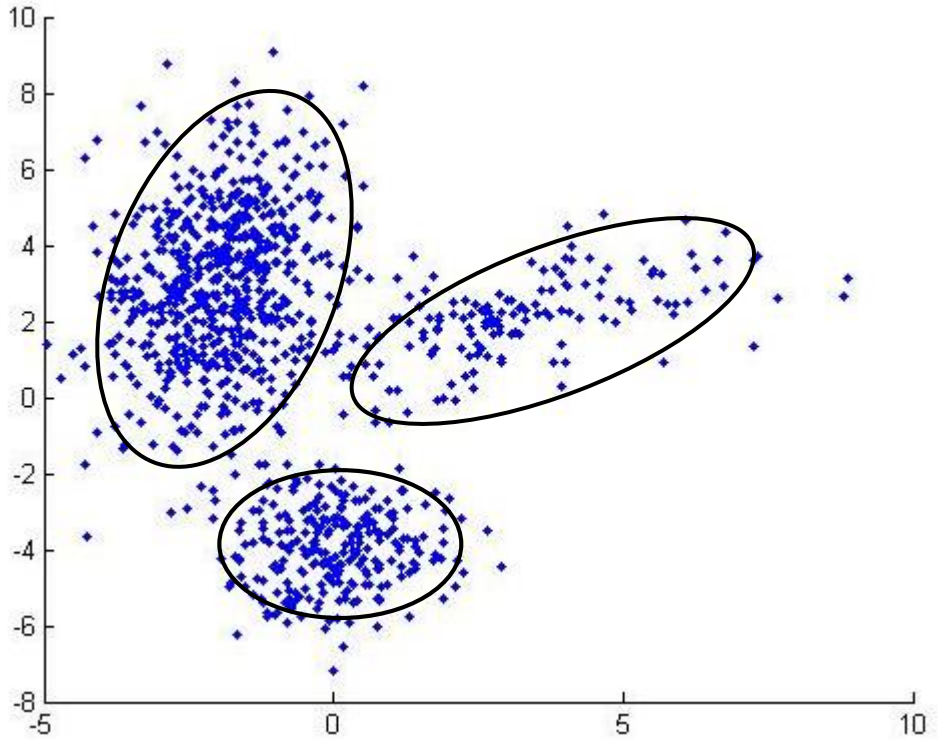
$\pi_2 = 0.3$

$\mu_3 = 8$

$\sigma_3 = 0.2$

$\pi_3 = 0.1$

GMM: 2-D Example



k = 3

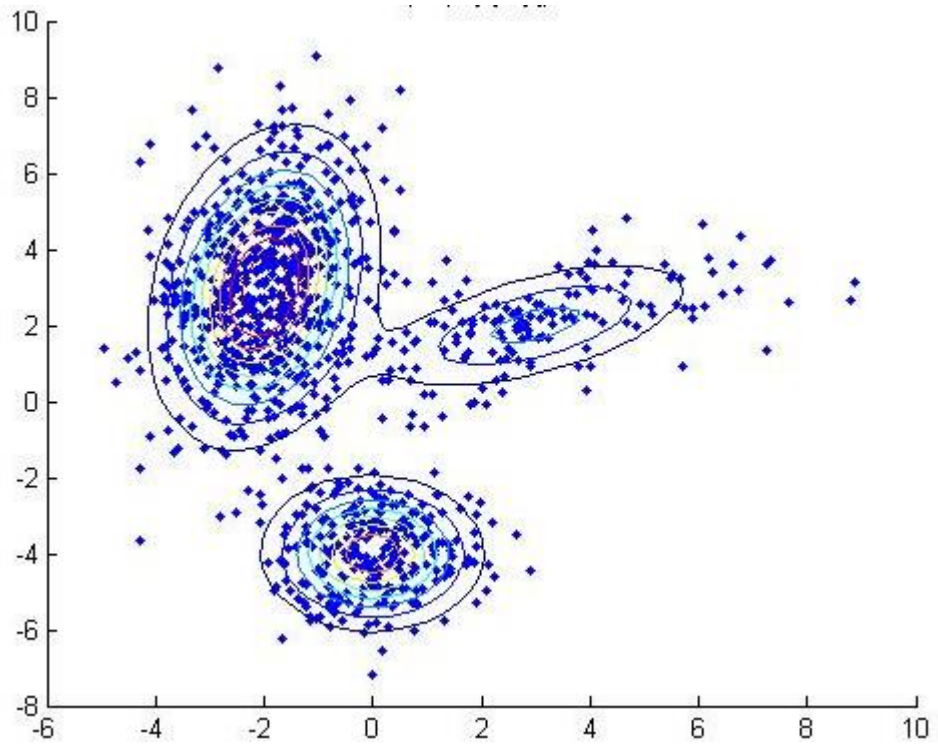
$$\boldsymbol{\mu}_1 = [-2 \quad 3]$$
$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 4 \end{bmatrix}$$
$$\pi_1 = 0.6$$

$$\boldsymbol{\mu}_2 = [0 \quad -4]$$
$$\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\pi_2 = 0.25$$

$$\boldsymbol{\mu}_3 = [3 \quad 2]$$
$$\Sigma_3 = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$
$$\pi_3 = 0.15$$

GMM: 2-D Example

- GMM distribution



k = 3

$$\boldsymbol{\mu}_1 = [-2 \quad 3]$$
$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 4 \end{bmatrix}$$
$$\pi_1 = 0.6$$

$$\boldsymbol{\mu}_2 = [0 \quad -4]$$
$$\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\pi_2 = 0.25$$

$$\boldsymbol{\mu}_3 = [3 \quad 2]$$
$$\Sigma_3 = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$
$$\pi_3 = 0.15$$

How to Fit GMM?

- In order to maximize log likelihood: $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln \left\{ \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\}$$

  - The sum over components appears inside the log and there is no closed form solution for maximum likelihood.

$$\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}$$

$$\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} \qquad k = 1, \dots, K$$

$$\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\left(\sum_{j=1}^{K} \pi_j - 1\right)}{\partial \pi_k} = 0$$

ML for GMM

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N} \boxed{\frac{\pi_k \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}} \boldsymbol{x}^{(i)}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{i=1}^{N} \boxed{\frac{\pi_k \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k^{\text{new}})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

$$N_k = \sum_{i=1}^{N} \boxed{\frac{\pi_k \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}$$

$$\frac{\partial \log|\boldsymbol{A}^{-1}|}{\partial \boldsymbol{A}^{-1}} = \boldsymbol{A}^T \qquad \frac{\partial \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{x}^T$$

EM algorithm

- An iterative algorithm in which each iteration is guaranteed to improve the log-likelihood function

- General algorithm for finding ML estimation when the data is incomplete (missing or unobserved data).
  - EM find the maximum likelihood parameters in cases where the models involve unobserved variables Z in addition to unknown parameters $\boldsymbol{\theta}$ and known data observations X.

Mixture models: discrete latent variables

$$p(\boldsymbol{x}) = \sum P(z_j = 1) p(\boldsymbol{x} | z_j = 1) = \sum_{j=1}^{K} \pi_j \, p(\boldsymbol{x} | z_j = 1)$$

- z: latent or hidden variable
  - specifies the mixture component

- $P(z_j = 1) = \pi_j$
  - $0 \leq \pi_j \leq 1$
  - $\sum_{j=1}^{K} \pi_j = 1$

EM for GMM

$\boldsymbol{\theta} = [\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$

$z^{(i)} \in \{1,2, \dots, K\}$ shows the mixture from which $x^{(i)}$ is generated

- Initialize $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k \quad k = 1, \dots, K$
- **E step**: $i = 1, \dots, N, j = 1, \dots, K$

$$\gamma_j^i = P\left(z_j^{(i)} = 1 | \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{old}\right) = \frac{\pi_j^{old} \mathcal{N}(\boldsymbol{x}^{(i)} | \boldsymbol{\mu}_j^{old}, \boldsymbol{\Sigma}_j^{old})}{\sum_{k=1}^{K} \pi_k^{old} \mathcal{N}(\boldsymbol{x}^{(i)} | \boldsymbol{\mu}_k^{old}, \boldsymbol{\Sigma}_k^{old})}$$

- **M Step**: $j = 1, \dots, K$

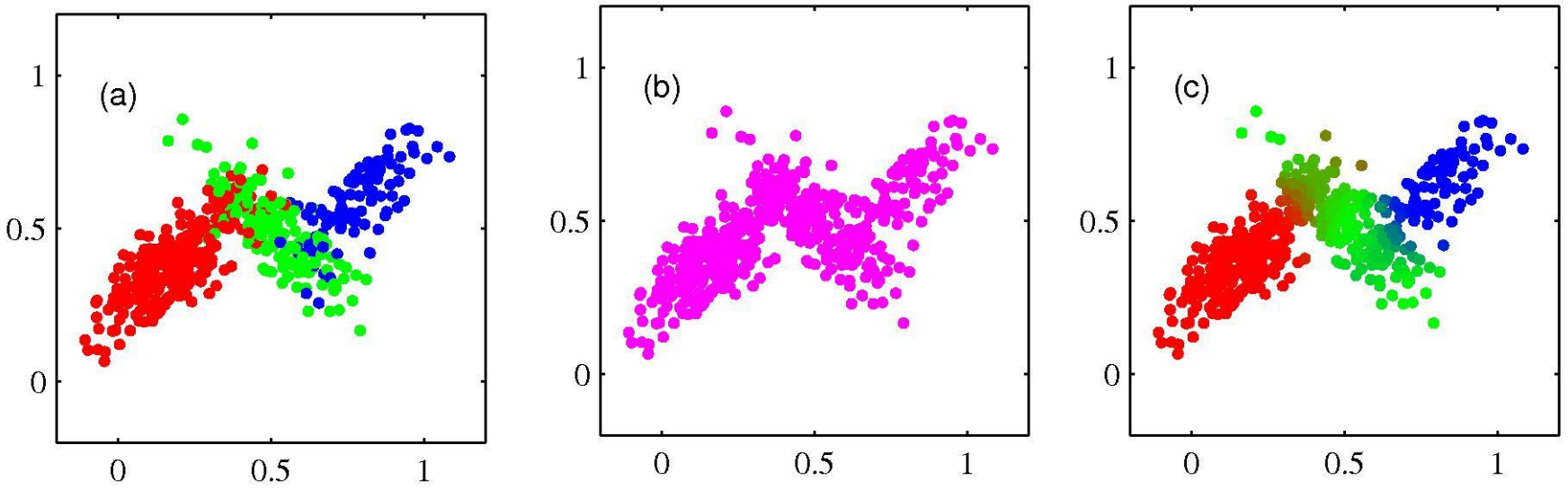
$$\boldsymbol{\mu}_j^{new} = \frac{\sum_{i=1}^{N} \gamma_j^i \boldsymbol{x}^{(i)}}{\sum_{i=1}^{N} \gamma_j^i}$$

$$\boldsymbol{\Sigma}_j^{new} = \frac{1}{\sum_{i=1}^{N} \gamma_j^i} \sum_{i=1}^{N} \gamma_j^i (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j^{\text{new}})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j^{\text{new}})^T$$

$$\pi_j^{\text{new}} = \frac{\sum_{i=1}^{N} \gamma_j^i}{N}$$

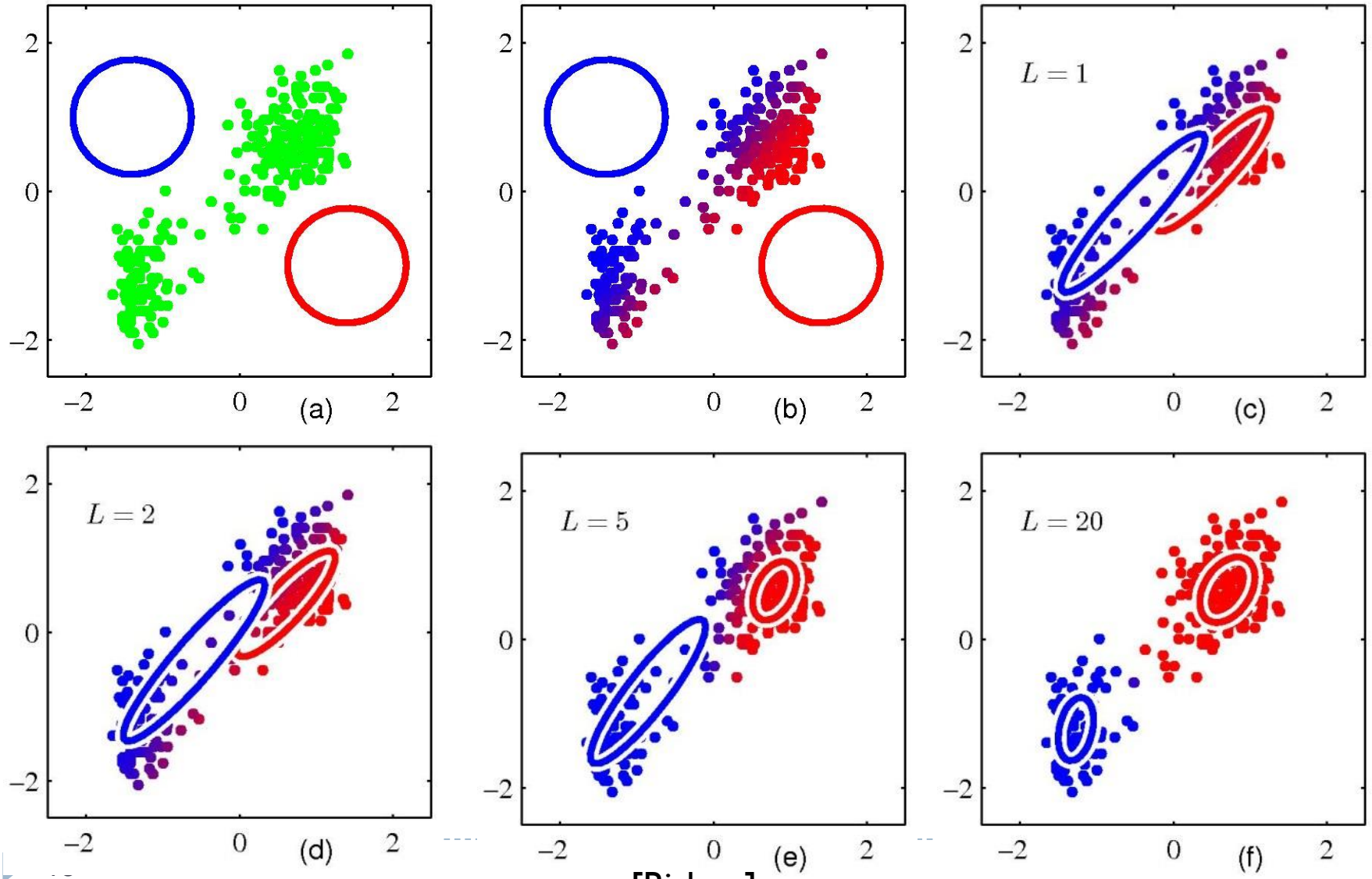
- Repeat E and M steps until convergence

EM & GMM: example



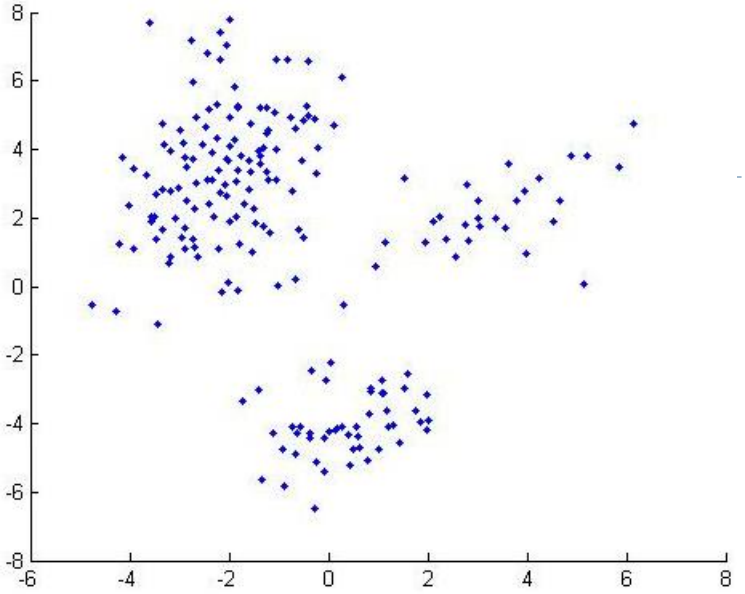
[Bishop]

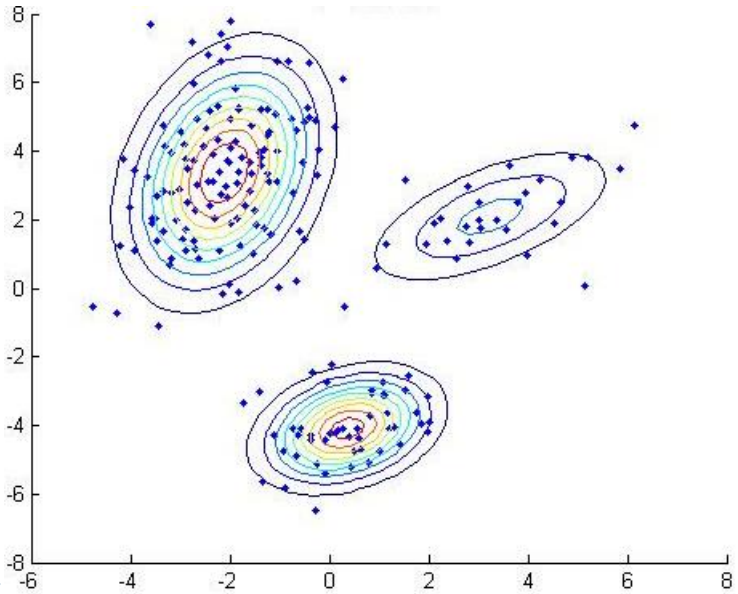
EM & GMM: Example

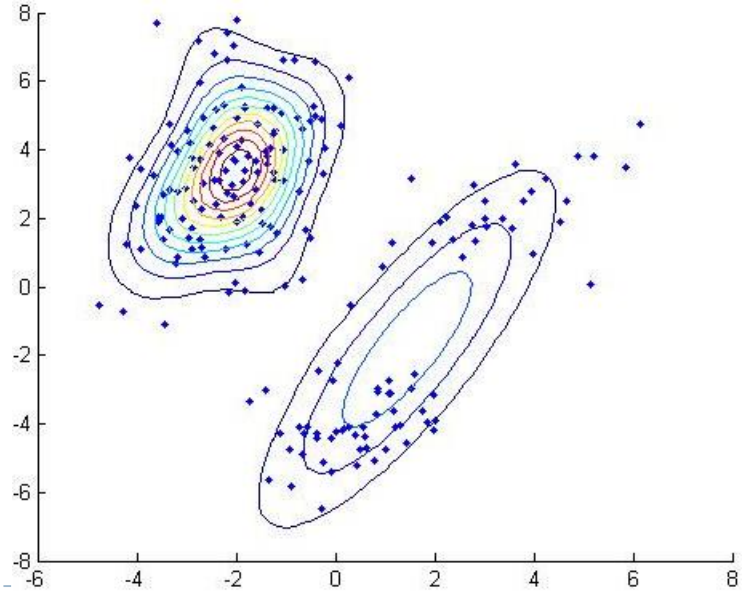


[Bishop]

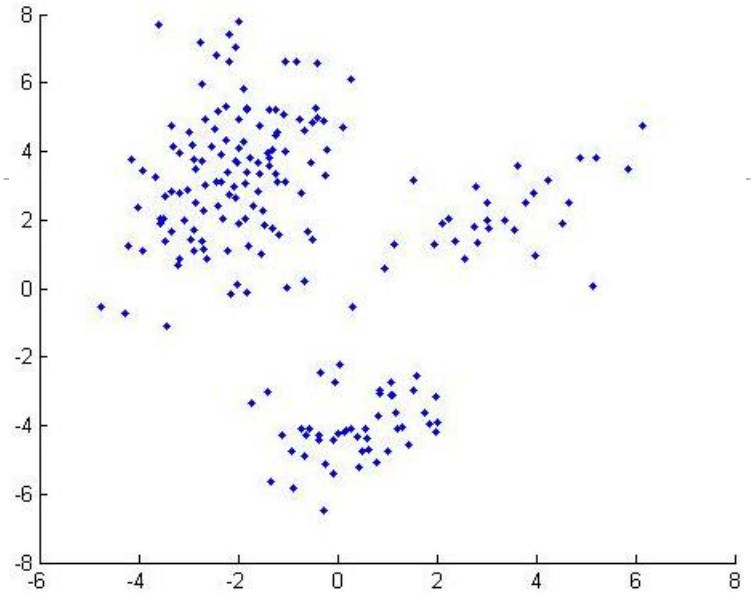
Local Minima







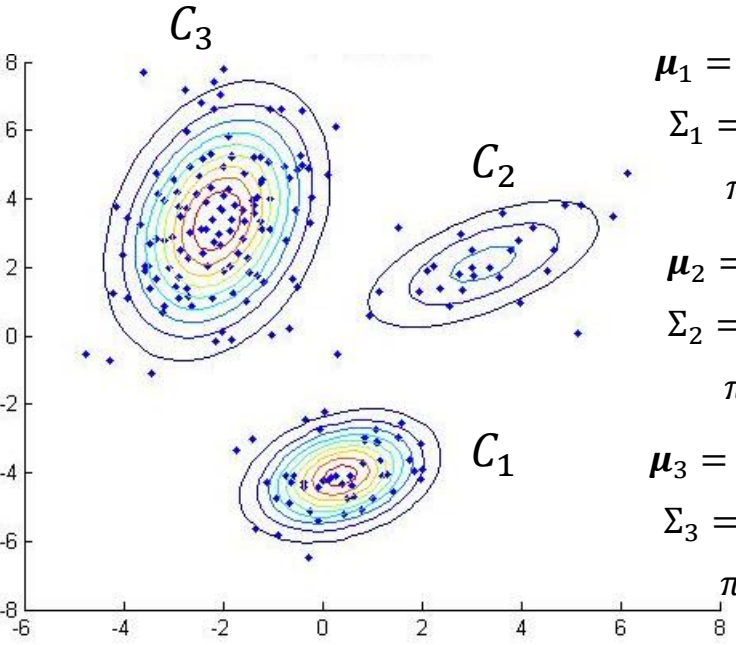
Local Minima



$$\boldsymbol{\mu}_1 = [-2 \quad 3]$$
$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 4 \end{bmatrix}$$
$$\pi_1 = 0.6$$

$$\boldsymbol{\mu}_2 = [0 \quad -4]$$
$$\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\pi_2 = 0.25$$

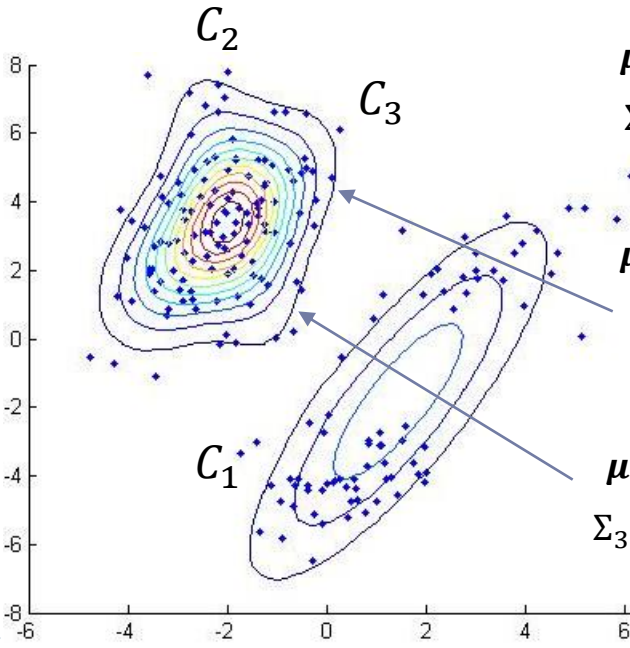
$$\boldsymbol{\mu}_3 = [3 \quad 2]$$
$$\Sigma_3 = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$
$$\pi_3 = 0.15$$



$$\boldsymbol{\mu}_1 = [0.36 \quad -4.09]$$
$$\Sigma_1 = \begin{bmatrix} 0.89 & 0.26 \\ 0.26 & 0.83 \end{bmatrix}$$
$$\pi_1 = 0.249$$

$$\boldsymbol{\mu}_2 = [3.25 \quad 2.09]$$
$$\Sigma_2 = \begin{bmatrix} 2.23 & 1.08 \\ 1.09 & 1.41 \end{bmatrix}$$
$$\pi_2 = 0.146$$

$$\boldsymbol{\mu}_3 = [-2.11 \quad 3.36]$$
$$\Sigma_3 = \begin{bmatrix} 1.12 & 0.61 \\ 0.61 & 3.61 \end{bmatrix}$$
$$\pi_3 = 0.604$$



$$\boldsymbol{\mu}_1 = [1.45 \quad -1.81]$$
$$\Sigma_1 = \begin{bmatrix} 3.30 & 4.76 \\ 4.76 & 10.01 \end{bmatrix}$$
$$\pi_1 = 0.392$$

$$\boldsymbol{\mu}_2 = [-2.20 \quad 3.16]$$
$$\Sigma_2 = \begin{bmatrix} 1.30 & 1.10 \\ 1.10 & 2.80 \end{bmatrix}$$
$$\pi_2 = 0.429$$

$$\boldsymbol{\mu}_3 = [-1.88 \quad 3.74]$$
$$\Sigma_3 = \begin{bmatrix} 5.83 & -0.82 \\ -0.82 & 5.83 \end{bmatrix}$$
$$\pi_3 = 0.178$$

EM+GMM vs. k-means

- k-means:
  - It is not probabilistic
  - Has fewer parameters (and faster)
  - Limited by the underlying assumption of spherical clusters
    - can be extended to use covariance – get "hard EM" (ellipsoidal k-means).
- Both EM and k-means depend on initialization
  - getting stuck in local optima
    - EM+GMM has more local minima
    - Useful trick: first run k-means and then use its result to initialize EM.

EM algorithm: general

General algorithm for finding ML estimation when the data is incomplete (missing or unobserved data).

Incomplete log likelihood

- Complete log likelihood
  - Maximizing likelihood (i.e., $\log P(X, Y|\boldsymbol{\theta})$) for labeled data is straightforward

- Incomplete log likelihood
  - With Z unobserved, our objective becomes the log of a marginal probability $\log P(X|\boldsymbol{\theta}) = \log \sum_Z P(X, Z|\boldsymbol{\theta})$
    - This objective will not decouple and we use EM algorithm to solve it

$$X = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$$
$$Z = \{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(N)}\}$$

EM Algorithm

- Assumptions: X (observed or known variables), Z (unobserved or latent variables), X come from a specific model with unknown parameters $\boldsymbol{\theta}$
  - If Z is relevant to X (in any way), we can hope to extract information about it from X assuming a specific parametric model on the data.

- Steps:
  - Initialization: Initialize the unknown parameters $\boldsymbol{\theta}$
  - Iterate the following steps, until convergence:
    - Expectation step: Find the probability of unobserved variables given the current parameters estimates and the observed data.
    - Maximization step: from the observed data and the probability of the unobserved data find the most likely parameters (a better estimate for the parameters).

EM algorithm intuition

- When learning with hidden variables, we are trying to solve two problems at once:
  - hypothesizing values for the unobserved variables in each data sample
  - learning the parameters

- Each of these tasks is fairly easy when we have the solution to the other.
  - Given complete data, we have the statistics, and we can estimate parameters using the MLE formulas.
  - Conversely, computing probability of missing data given the parameters is a probabilistic inference problem

EM algorithm

X: observed variables

Z: Unobserved variables

$\boldsymbol{\theta}$: parameters

Expectation step (E-step): Given the current parameters, find soft completion of the data, using probabilistic inference.

Maximization step (M-step): We then treat the soft completed data as if it were observed and learn a new set of parameters.

Choose an initial parameters $\boldsymbol{\theta}^1$

$t \leftarrow 1$

Iterate until convergence:

E Step: Calculate $P(Z|X,\boldsymbol{\theta}^t)$

M Step: $\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} E_{P(Z|X,\boldsymbol{\theta}^t)}[\log P(Z,X|\boldsymbol{\theta})]$

$t \leftarrow t+1$

expectation of the log-likelihood evaluated using the current estimate for the parameters $\boldsymbol{\theta}^t$

$$E_{Z\sim P(Z|X,\boldsymbol{\theta}^{\text{old}})}[\log p(X,Z|\boldsymbol{\theta})]$$

$$= \sum_Z P(Z|X,\boldsymbol{\theta}^{\text{old}}) \times \log p(X,Z|\boldsymbol{\theta})$$

EM theoretical analysis

- What is the underlying theory for the use of the expected complete log likelihood in the M-step?

$$E_{P\left(Z|X,\boldsymbol{\theta}^{old}\right)}\left[\log P(X,Z|\boldsymbol{\theta})\right]$$

- Now, we show that maximizing this function also maximizes the likelihood

EM theoretical foundation: Objective function

$$X = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$$
$$Z = \{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}\}$$

$$\ell(\boldsymbol{\theta}; X) = \log P(X|\boldsymbol{\theta}) = \log \sum_{Z} P(X, Z|\boldsymbol{\theta})$$

$$= \log \sum_{Z} Q(Z) \frac{P(X, Z|\boldsymbol{\theta})}{Q(Z)} \overset{\text{Jensen inequality}}{\geq} \sum_{Z} Q(Z) \log \frac{P(X, Z|\boldsymbol{\theta})}{Q(Z)}$$

$$= F[\boldsymbol{\theta}, Q]$$

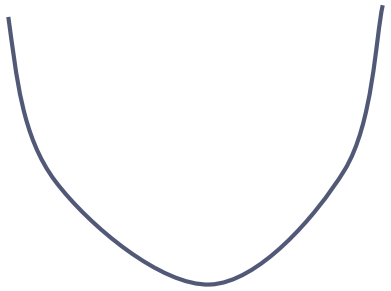
$F[\boldsymbol{\theta}, Q]$ is a lower bound on $\ell(\boldsymbol{\theta}; X)$

EM maximizes $F[\boldsymbol{\theta}, Q]$

Jensen's inequality

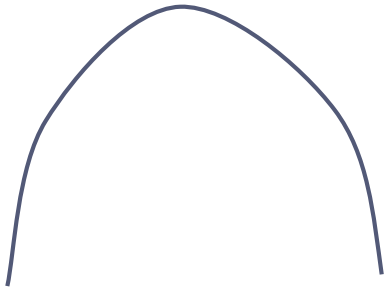
- If f is a convex function

$$E[f(x)] \geq f(E[x])$$



- If f is a concave function

$$E[f(x)] \leq f(E[x])$$



EM theoretical foundation: Algorithm in general form

- EM is a coordinate ascent algorithm on $F[\boldsymbol{\theta}, Q]$. In the t-th iteration,
  - E-step: maximize $F[\boldsymbol{\theta}, Q]$ w.r.t. Q

$$Q^t = \underset{Q}{\operatorname{argmax}}\, F[\boldsymbol{\theta}^t, Q]$$

  - M-step:

$$\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, F[\boldsymbol{\theta}, Q^t]$$

We will show that each iteration improves the log-likelihood

EM theoretical foundation: E-step

$$Q^t = P(Z|X, \boldsymbol{\theta}^t) \Rightarrow Q^t = \underset{Q}{\operatorname{argmax}} F[\boldsymbol{\theta}^t, Q]$$

Proof:

$$F[\boldsymbol{\theta}^t, P(Z|X, \boldsymbol{\theta}^t)] = \sum_Z P(Z|X, \boldsymbol{\theta}^t) \log \frac{P(X, Z|\boldsymbol{\theta}^t)}{P(Z|X, \boldsymbol{\theta}^t)}$$

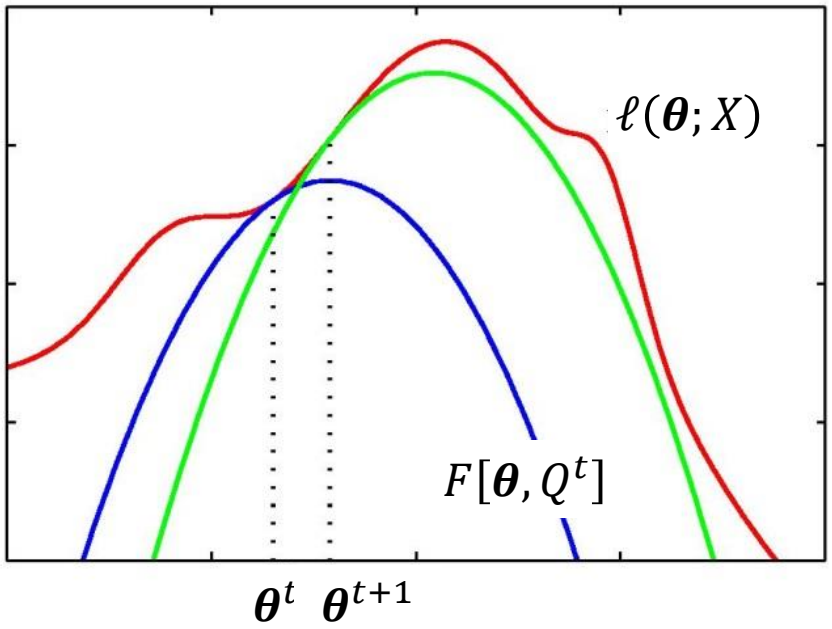
$$= \sum_Z P(Z|X, \boldsymbol{\theta}^t) \log P(X|\boldsymbol{\theta}^t) = \log P(X|\boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}^t; X)$$

- $F[\boldsymbol{\theta}, Q]$ is a lower bound on $\ell(\boldsymbol{\theta}; X)$. Thus, $F[\boldsymbol{\theta}^t, Q]$ has been maximized by setting Q to $P(Z|X, \boldsymbol{\theta}^t)$:

$$F[\boldsymbol{\theta}^t, P(Z|X, \boldsymbol{\theta}^t)] = \ell(\boldsymbol{\theta}^t; X)$$

$$\Rightarrow P(Z|X, \boldsymbol{\theta}^t) = \underset{Q}{\operatorname{argmax}} F[\boldsymbol{\theta}^t, Q]$$

EM algorithm: illustration



EM theoretical foundation: M-step

M-step can be equivalently viewed as maximizing the expected complete log-likelihood:

$$\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, F[\boldsymbol{\theta}, Q^t] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, E_{Q^t}[\log P(X, Z|\boldsymbol{\theta})]$$

Proof:

$$F[\boldsymbol{\theta}, Q^t] = \sum_Z Q^t(Z) \log \frac{P(X, Z|\boldsymbol{\theta})}{Q^t(Z)}$$

$$= \sum_Z Q^t(Z) \log P(X, Z|\boldsymbol{\theta}) - \sum_Z Q^t(Z) \log Q^t(Z)$$

$$\Rightarrow F[\boldsymbol{\theta}, Q^t] = E_{Q^t}[\log P(X, Z|\boldsymbol{\theta})] + H(Q^t(Z))$$

Independent of $\boldsymbol{\theta}$

EM iteration increases $\ell(\boldsymbol{\theta}; X)$

$$\ell(\boldsymbol{\theta}^t; X) = E_{Q^t}[\log P(X, Z|\boldsymbol{\theta}^t)] + H(Q^t(Z))$$

$$\ell(\boldsymbol{\theta}^{t+1}; X) \geq E_{Q^t}[\log P(X, Z|\boldsymbol{\theta}^{t+1})] + H(Q^t(Z))$$

$$\ell(\boldsymbol{\theta}^{t+1}; X) - \ell(\boldsymbol{\theta}^t; X) \geq E_{Q^t}[\log P(X, Z|\boldsymbol{\theta}^{t+1})] - E_{Q^t}[\log P(X, Z|\boldsymbol{\theta}^t)]$$

Moreover, we have:

$$\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} E_{Q^t}[\log P(X, Z|\boldsymbol{\theta})]$$

$$\Rightarrow E_{Q^t}[\log P(X, Z|\boldsymbol{\theta}^{t+1})] \geq E_{Q^t}[\log P(X, Z|\boldsymbol{\theta}^t)]$$

$$\Rightarrow \ell(\boldsymbol{\theta}^{t+1}; X) - \ell(\boldsymbol{\theta}^t; X) \geq 0$$

EM is guaranteed to find a local maxima of the log likelihood

EM for GMM: E step details

$$P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{P(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})}{\sum_{\boldsymbol{z}} P(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})} = \frac{P(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta})P(\boldsymbol{z}|\boldsymbol{\theta})}{\sum_{\boldsymbol{z}} P(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta})P(\boldsymbol{z}|\boldsymbol{\theta})} \qquad \boldsymbol{\theta} = [\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}]$$

$$Q_j(z_j = 1) = P(z_j = 1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{\pi_j \ \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{k=1}^{K} \pi_k \ \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

EM for GMM
M step: details

$$\boldsymbol{\theta} = [\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

$$\boldsymbol{\theta}^{old} = [\boldsymbol{\pi}^{old}, \boldsymbol{\mu}^{old}, \boldsymbol{\Sigma}^{\mathbf{old}}]$$

$$p(X, Z|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i)}, \boldsymbol{\theta}) p(\boldsymbol{z}^{(i)}|\boldsymbol{\pi})$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{K} \mathcal{N}\left(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)^{z_j^{(i)}} {\pi_j}^{z_j^{(i)}}$$

$$\log p(X, Z|\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{K} z_j^{(i)} \left\{\log \mathcal{N}\left(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \log \pi_j\right\}$$

$$E_{Z \sim P\left(Z|X, \boldsymbol{\theta}^{\text{old}}\right)}[\log p(X, Z|\boldsymbol{\theta})] =$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} \underbrace{E_{P\left(z_j^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{\text{old}}\right)}\left[z_j^{(i)}\right]}_{\gamma_j^i} \left\{\log \mathcal{N}\left(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) + \log \pi_j\right\}$$

EM for GMM
M step: details

$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta^{old}})}{\partial \boldsymbol{\mu}_j} = 0 \Rightarrow \boldsymbol{\mu}_j = \frac{\sum_{i=1}^{N} \gamma_j^i \boldsymbol{x}^{(i)}}{\sum_{i=1}^{N} \gamma_j^i}$$

$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta^{old}})}{\partial \boldsymbol{\Sigma}_j} = 0 \Rightarrow \boldsymbol{\Sigma}_j = \frac{1}{\sum_{i=1}^{N} \gamma_j^i} \sum_{i=1}^{N} \gamma_j^i (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)^T$$

$$\frac{\partial \left(Q(\boldsymbol{\theta};\boldsymbol{\theta^{old}}) + \lambda\left(\sum_{l=1}^{k} \pi_l - 1\right)\right)}{\partial \pi_j} = 0 \Rightarrow \pi_j = \frac{\sum_{i=1}^{N} \gamma_j^i}{N}$$

Lagrange multiplier due to the constraint $\sum_{j=1}^{k} \pi_j = 1$

EM algorithm: general

- EM: general procedure for learning from partly observed data
- Define: $\mathrm{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}}) = E_{Z \sim P(Z|X,\boldsymbol{\theta}^{\mathrm{old}})}[\log p(X, Z|\boldsymbol{\theta})]$

$$= \sum_Z P(Z|X, \boldsymbol{\theta}^{\mathrm{old}}) \times \log p(X, Z|\boldsymbol{\theta})$$

expectation of the log-likelihood evaluated using the current estimate for the parameters $\boldsymbol{\theta}^{\mathrm{old}}$

Choose an initial setting $\boldsymbol{\theta}^{\mathrm{old}} = \boldsymbol{\theta}^0$

Iterate until convergence:

E Step: Use X and current $\boldsymbol{\theta}^{\mathrm{old}}$ to calculate $P(Z|X, \boldsymbol{\theta}^{\mathrm{old}})$

M Step: $\boldsymbol{\theta}^{\mathrm{new}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\, \mathrm{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}})$

$\boldsymbol{\theta}^{\mathrm{old}} \leftarrow \boldsymbol{\theta}^{\mathrm{new}}$

EM advantages and disadvantages

- Some good things about EM:
  - no learning rate (step-size) parameter
  - automatically enforces parameter constraints
  - very fast for low dimensions
  - each iteration guaranteed to improve likelihood

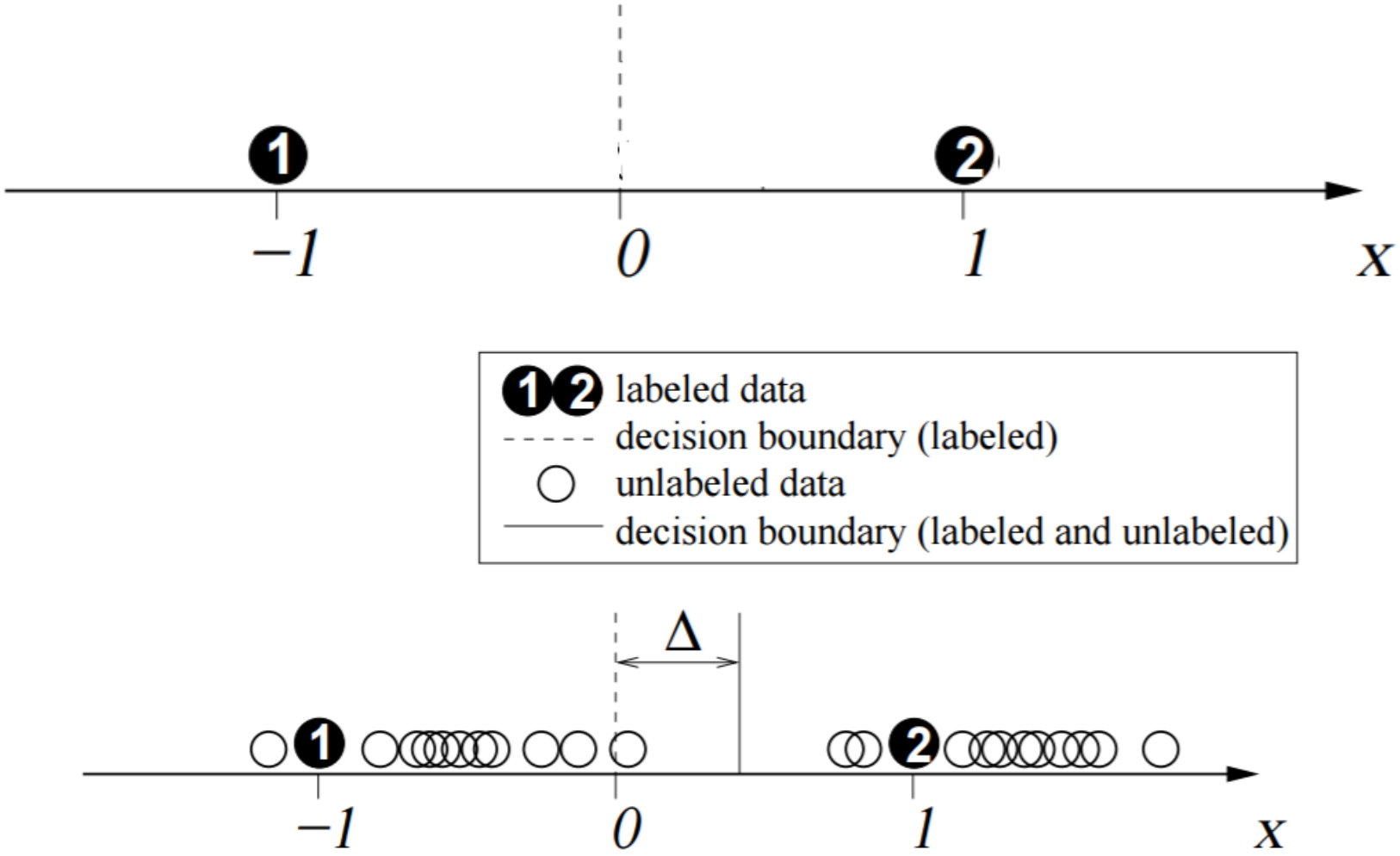
- Some bad things about EM:
  - can be slower than some other iterative gradient-based methods

Semi-supervised learning

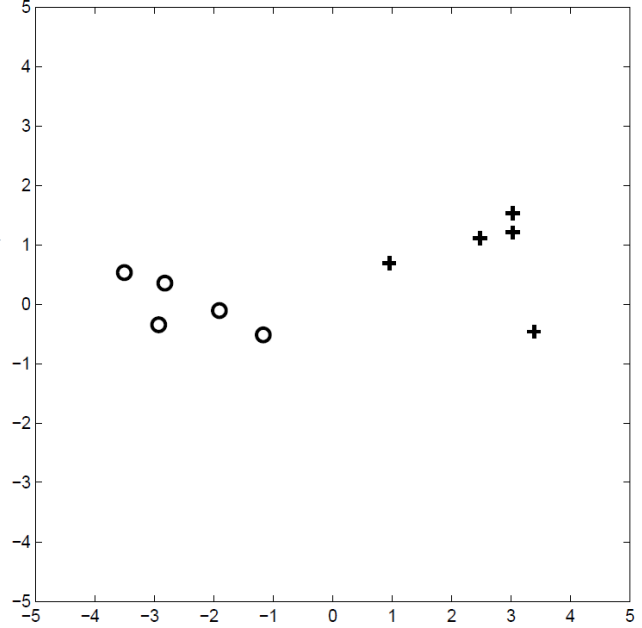
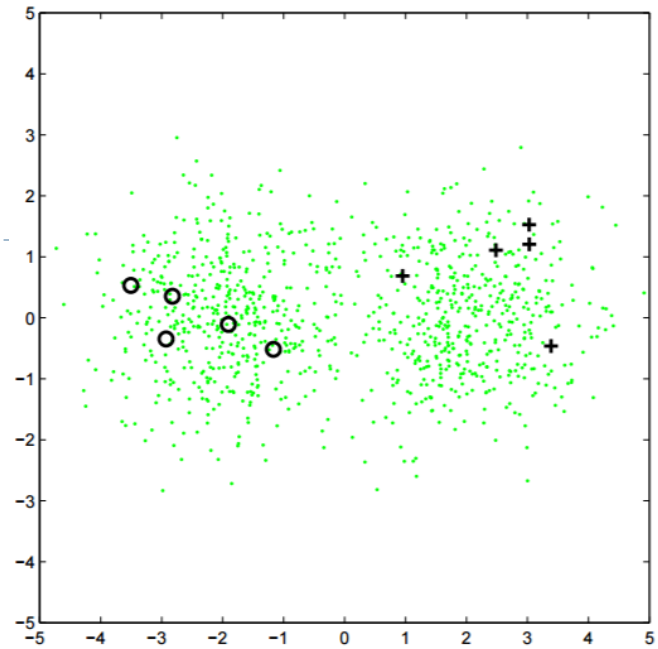
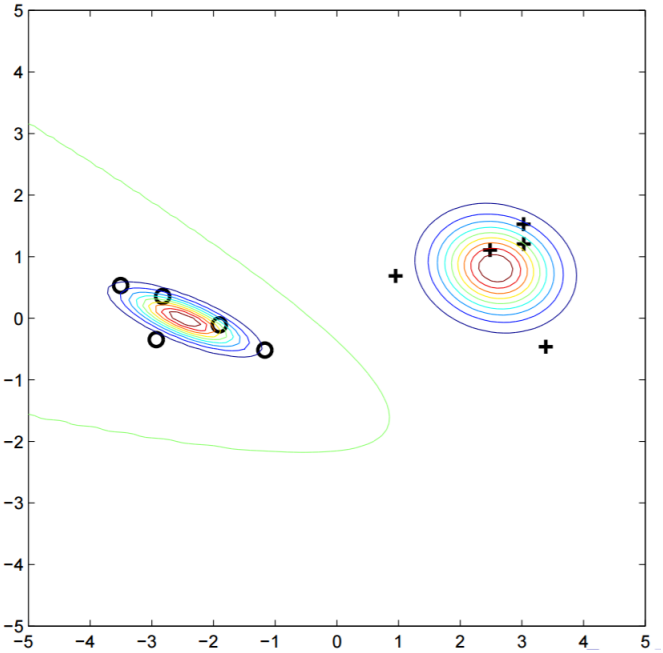
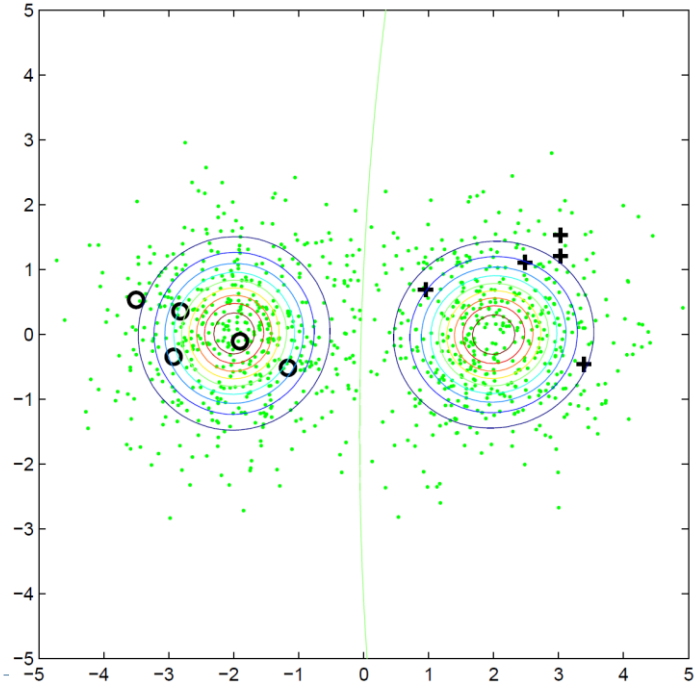
- Supervised Learning models require labeled data
  - Supervised learning usually requires plenty of labeled data
    - It is usually expensive to have a large set of labeled data
    - Unlabeled data is often abundant with no or low cost

- Learning from both labeled and unlabeled data
  - Labeled training data: $\mathcal{L} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{l=1}^{L}$
  - Unlabeled data available during training: $\mathcal{U} = \left\{\boldsymbol{x}^{(n)}\right\}_{n=L+1}^{L+U}$

Semi-supervised learning: example



Zhu, Semi-Supervised Learning Tutorial, ICML, 2007.

Zhu, Semi-Supervised Learning Tutorial, ICML, 2007.

Semi-supervised generative model

- Start from MLE $\boldsymbol{\theta} = [\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$ on $\mathcal{L} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{l=1}^{L}$
- Repeat:
  - E-step: compute $p(y^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{\theta})$ for $n = L + 1$ to $n = L + U$
  - M-step: compute the parameters $\boldsymbol{\theta} = [\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$ considering both labeled data and unlabeled data using the distribution found on their labels in the E-step

Resource

- C. Bishop, “Pattern Recognition and Machine Learning”, Chapter 9.