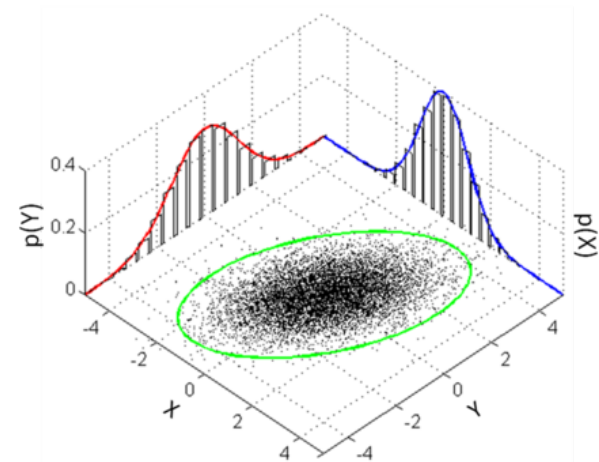
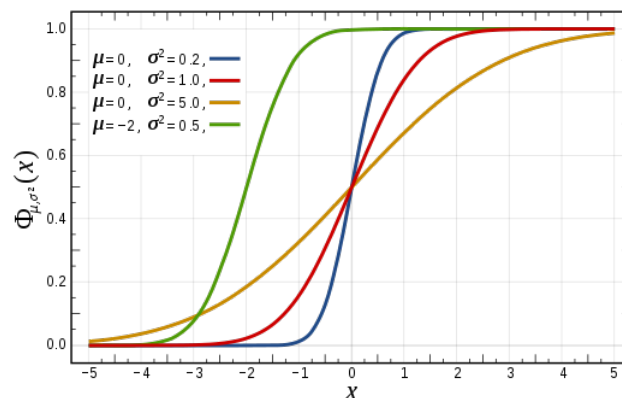
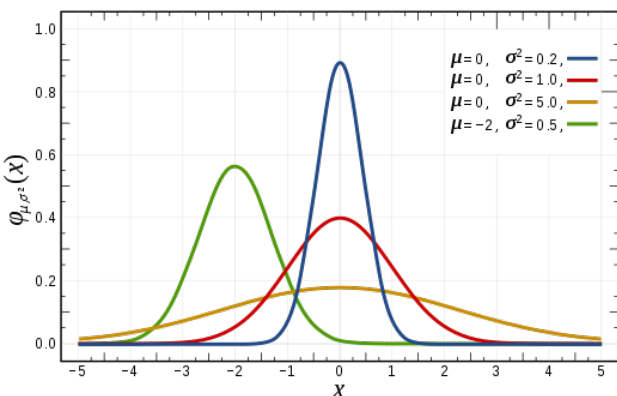


CS489/698: Intro to ML

Lecture 10: Mixture of Gaussians

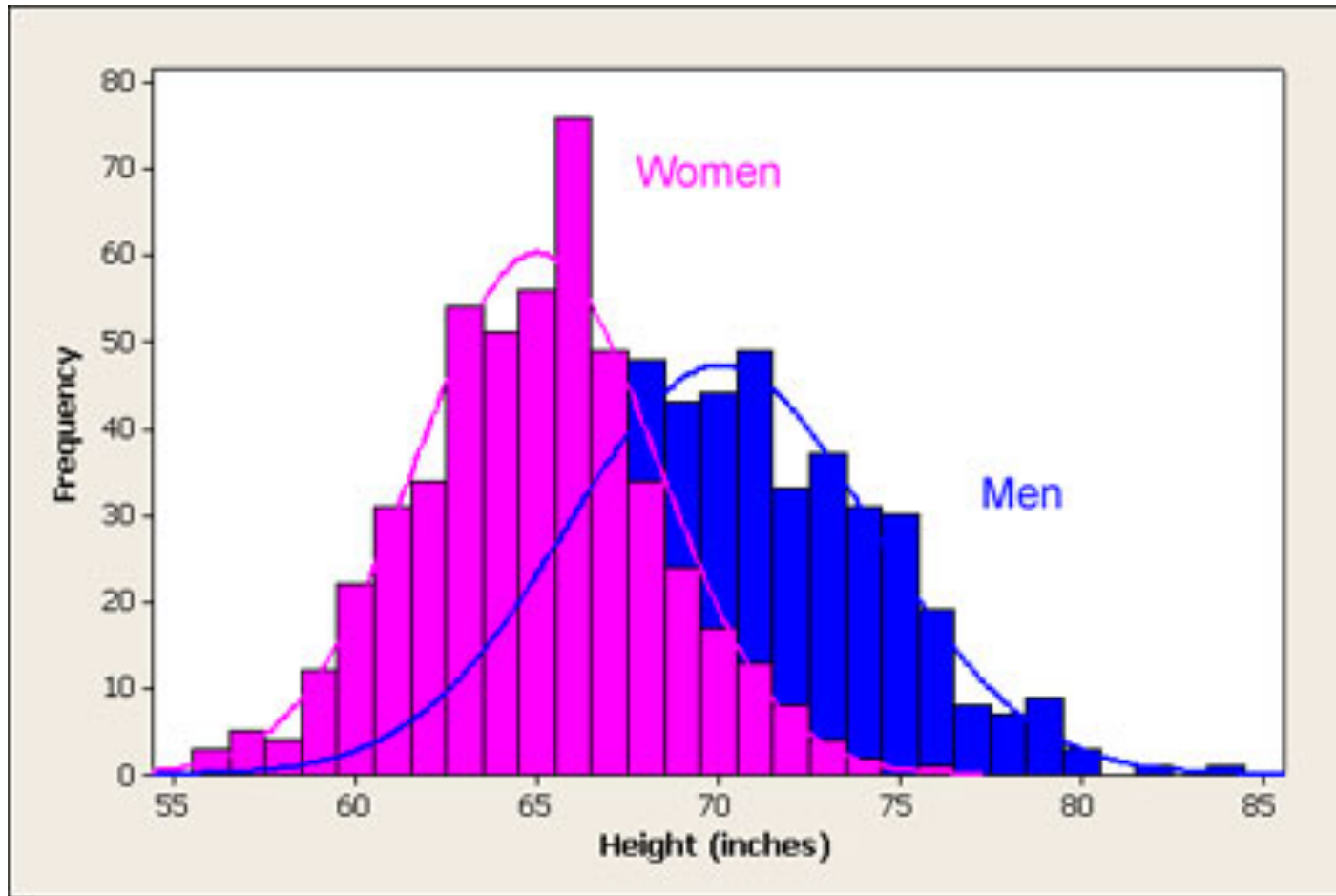


Recap: Gaussian distribution



$$p(\mathbf{x}) = (2\pi)^{-d/2} |\mathbf{S}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

Multi-modality



Mixture models

$$p(x|\theta) = \sum_{k=1}^K \underbrace{p(z=k)}_{\pi_k} \underbrace{p(x|z=k, \theta)}_{p_k(x|\theta)}$$

θ parameters K # of components π_k mixing distr. $p_k(x|\theta)$ k-th component distr.

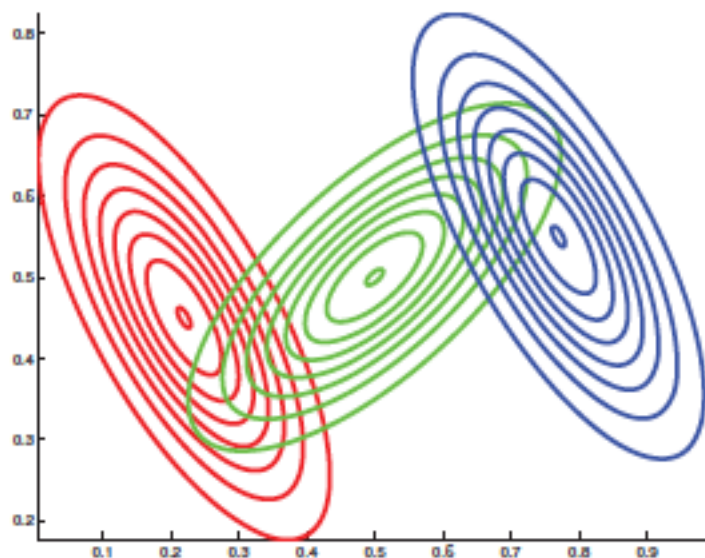
$$\pi_k \geq 0, \sum_k \pi_k = 1$$

- Where did we see a similar idea?

Example: Gaussian Mixture Models

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k, S_k)$$

$(\pi_1, \mu_1, S_1, \mu_2, S_2, \dots, \mu_K, S_K)$

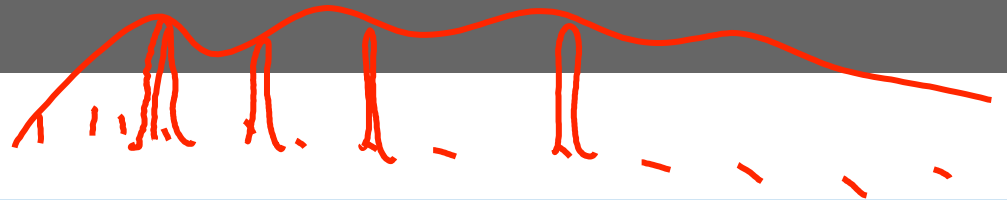


(a)



(b)

Universality $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Theorem. GMM with **sufficiently many** components can **approximate** any probability density function on \mathbb{R}^d .

- How many is many?
- Nothing special about Gaussian here, except computationally (later).

Example: Mixture of Experts

Where did we see a similar idea?

$$\frac{\exp(\mathbf{x}^\top \mathbf{v}_k)}{\sum_{c=1}^K \exp(\mathbf{x}^\top \mathbf{v}_c)}$$

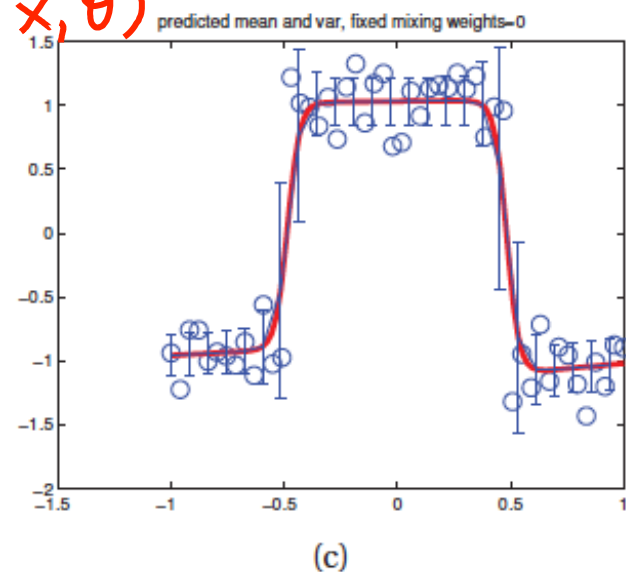
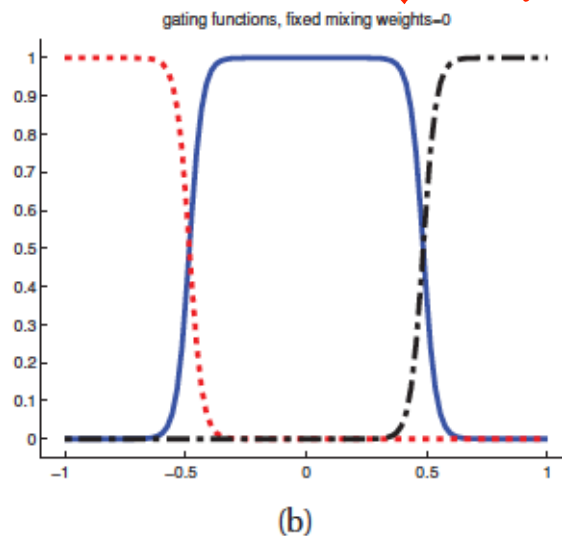
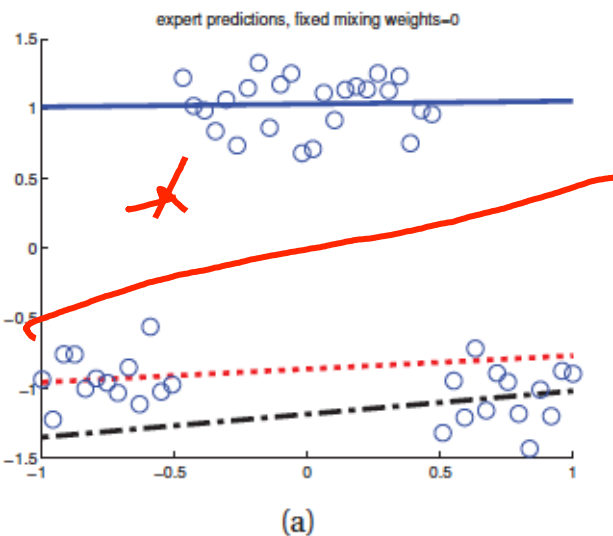
$\mathcal{N}(y | \mathbf{w}_k^\top \mathbf{x}, \sigma_k^2)$

K mixing distr.

k -th component distr.

$$p(y | \mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^K p(z = k | \mathbf{x}, \boldsymbol{\theta}) p(y | \mathbf{x}, z = k, \boldsymbol{\theta})$$

$P(y, z=k | \mathbf{x}, \boldsymbol{\theta})$



Inference problem

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^K p(z = k)p(x|z = k, \boldsymbol{\theta})$$

latent (unobserved)

- Given iid sample X_1, X_2, \dots, X_n from $p(x|\theta)$
- Need to estimate θ
- Maximum likelihood is NP-hard...



Soft clustering

$$p(z = k | \mathbf{x}, \boldsymbol{\theta}) = \frac{p(z = k | \boldsymbol{\theta}) p(\mathbf{x} | z = k, \boldsymbol{\theta})}{\sum_{c=1}^K p(z = c | \boldsymbol{\theta}) p(\mathbf{x} | z = c, \boldsymbol{\theta})}$$



(Stauffer & Grimson, CVPR'98)

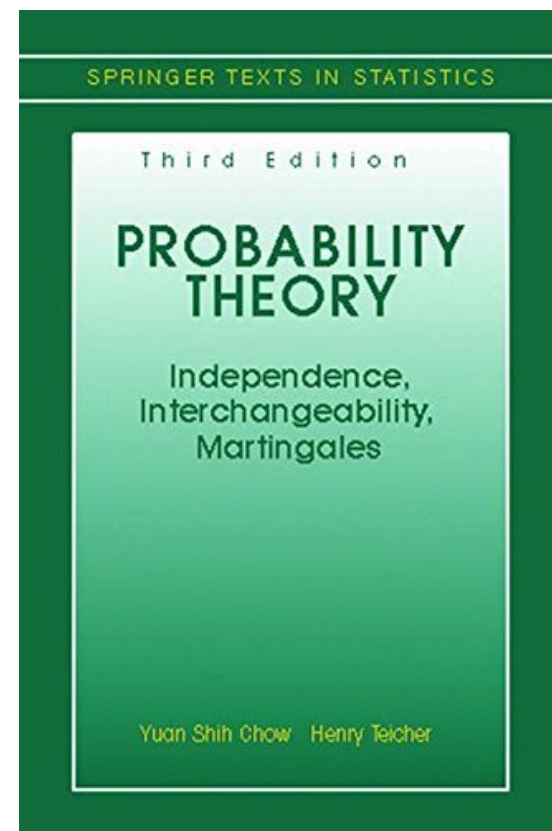
Bigger issue: identifiability?

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^K p(z = k)p(x|z = k, \boldsymbol{\theta})$$

- Is this factorization even unique?
- Yes, for GMMs!



Yao-Liang Yu



Variational form of Max Likelihood

$$\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) := \sum_{i=1}^n -\log p(\mathbf{x}_i | \boldsymbol{\theta}) = \sum_{i=1}^n -\log \underbrace{\sum_{z_i} p(\mathbf{x}_i, z_i | \boldsymbol{\theta})}_{=1}$$

$$\min_{q_i(z_i) \geq 0, \sum_{z_i} q_i(z_i) = 1} - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) + \underbrace{\sum_{z_i} q_i(z_i) \log q_i(z_i)}_{\text{neg. entropy}}$$

Handwritten red notes:
 $\sum_{z_i} q_i(z_i) \log \frac{q_i(z_i)}{p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) / \sum_{z_i} p(\mathbf{x}_i, z_i | \boldsymbol{\theta})} = \sum_{z_i} q_i(z_i) \log \frac{q_i(z_i)}{p(\mathbf{x}_i, z_i | \boldsymbol{\theta})}$

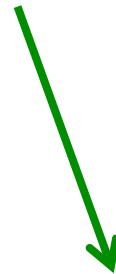
$$\min_{q_i(z_i)} \min_{\boldsymbol{\theta}} \sum_{i=1}^n \left[\sum_{z_i} q_i(z_i) \log q_i(z_i) - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) \right]$$

Handwritten red notes:
 $\sum_{z_i} q_i(z_i) \log \frac{1}{\sum_{z_i} p(\mathbf{x}_i, z_i | \boldsymbol{\theta})}$

KL divergence

$$\text{KL}(\mathbf{p} \parallel \mathbf{q}) := \sum_{i=1}^n p_i \log \left(\frac{p_i}{q_i} \right) \geq 0$$

- Both \mathbf{p} and \mathbf{q} are nonnegative and sum to 1



Jensen's inequality
 $E(\log(X)) \leq \log(E(X))$

- Equality holds iff $\mathbf{p} = \mathbf{q}$
- Measures difference between distributions; **asymmetric**

The EM algorithm

$$\min_{q_i(z_i)} \min_{\theta} \sum_{i=1}^n \left[\sum_{z_i} q_i(z_i) \log q_i(z_i) - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \theta) \right]$$

- Fix q , solve θ

$$\min_{\theta} - \sum_{i=1}^n \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \theta)$$

often closed-form

- Fix θ , solve q

$$\min_{q_i(z_i) \geq 0, \sum_{z_i} q_i(z_i) = 1} \underbrace{- \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \theta) + \sum_{z_i} q_i(z_i) \log q_i(z_i)}_{\text{KL}(q_i || p(z_i | \mathbf{x}_i, \theta))}$$

$p(z_i | \mathbf{x}_i; \theta) \cdot p(\mathbf{x}_i | \theta)$
 \parallel
 $q_i(z_i)$

$$\boxed{q_i(z_i) = p(z_i | \mathbf{x}_i, \theta)}$$

$\text{KL}(q_i || p(z_i | \mathbf{x}_i, \theta))$
 $= -\log(p(\mathbf{x}_i | \theta))$

EM for GMM: step 1 $q_i(z_i=k) = \gamma_{ik} \geq 0$

$$\min_{r_{ik} \geq 0, \sum_k r_{ik} = 1} \min_{\theta} \sum_{i=1}^n \left[\sum_{k=1}^K r_{ik} \log r_{ik} - \sum_{k=1}^K r_{ik} \log p(\mathbf{x}_i, z_i | \theta) \right]$$

$\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, S_k)$

$$\min_{\theta} \sum_{i=1}^n \sum_{k=1}^K r_{ik} \left[-\log \pi_k + \frac{1}{2} \log |S_k| + \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top S_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right]$$

$$\pi_k = \frac{\sum_i r_{ik}}{n} = \frac{\sum_k \sum_i r_{ik}}{n}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^n r_{ik} \mathbf{x}_i}{\sum_{i=1}^n r_{ik}}$$

$$S_k = \frac{\sum_{i=1}^n r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top}{\sum_{i=1}^n r_{ik}} = \frac{\sum_{i=1}^n r_{ik} \mathbf{x}_i \mathbf{x}_i^\top}{\sum_{i=1}^n r_{ik}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$$

Just in case

OK if you don't understand

$$\frac{\partial}{\partial \mu_k} = \sum_{i=1}^n r_{ik} S_k^{-1} (x_i - \mu_k) = 0$$

\Downarrow multiplying S_k

$$\sum_{i=1}^n r_{ik} (x_i - \mu_k) = 0$$

$$\Downarrow \sum_{i=1}^n r_{ik} x_i$$

$$\mu_k = \frac{\sum_{i=1}^n r_{ik} x_i}{\sum_{i=1}^n r_{ik}}$$

$$\frac{\partial}{\partial S_k} = \sum_{i=1}^n r_{ik} \left[\frac{1}{2} S_k^{-1} - \frac{1}{2} S_k^{-1} (x_i - \mu_k) (x_i - \mu_k)^T S_k^{-1} \right] = 0$$

\Downarrow multiplying S_k twice

$$\left(\sum_{i=1}^n r_{ik} \right) \cdot S_k = \sum_{i=1}^n r_{ik} (x_i - \mu_k) (x_i - \mu_k)^T$$

$$\Downarrow \frac{\sum_{i=1}^n r_{ik} (x_i - \mu_k) (x_i - \mu_k)^T}{\sum_{i=1}^n r_{ik}}$$

$$S_k =$$

Multiplying constant does not change min

adding constant does not change min

$$\min_{\pi_k \geq 0, \sum_k \pi_k = 1} \sum_{i=1}^n \sum_{k=1}^K r_{ik} \log \frac{1}{\pi_k} = \sum_{k=1}^K \underbrace{\left(\sum_{i=1}^n r_{ik} \right)}_{\text{def } r_k} \log \frac{1}{\pi_k} \equiv \min_{\pi_k \geq 0, \sum_k \pi_k = 1} \frac{1}{n} \sum_{k=1}^K r_k \log \frac{1}{\pi_k} + \sum_{k=1}^K \frac{r_k}{n} \log \frac{r_k}{n}$$

$$\sum_{k=1}^K \pi_k = 1 \Rightarrow \min_{\pi_k \geq 0, \sum_k \pi_k = 1} \text{KL} \left(\frac{r}{n} \parallel \pi \right) \Rightarrow \pi_k = \frac{r_k}{n}$$



EM for GMM: step 2

$$\min_{r_{ik} \geq 0, \sum_k r_{ik} = 1} \min_{\theta} \sum_{i=1}^n \left[\sum_{k=1}^K r_{ik} \log r_{ik} - \sum_{k=1}^K r_{ik} \log p(\mathbf{x}_i, z_i | \theta) \right]$$

$\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, S_k)$

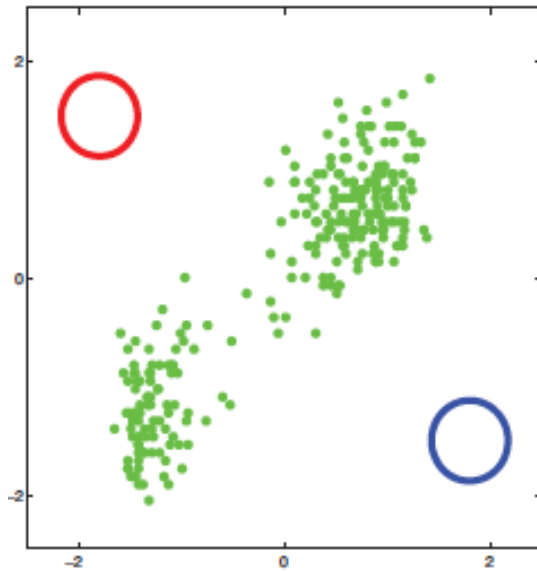
$$r_{ik} = p(z_i = k | \mathbf{x}_i, \theta) \rightarrow \text{posterior}$$

$$\propto p(z_i = k) \cdot p(\mathbf{x}_i | z_i = k, \theta)$$

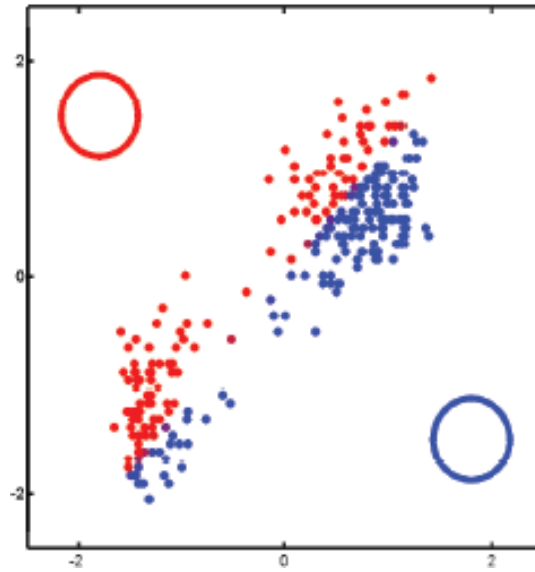
$$p(z|x) = \frac{p(z,x)}{p(x)} \quad \text{prior} \leftarrow = \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, S_k) \rightarrow \text{likelihood}$$

$$r_{ik} = \frac{\pi_k |S_k|^{-1/2} \exp(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top S_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k))}{\sum_{c=1}^K \pi_c |S_c|^{-1/2} \exp(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_c)^\top S_c^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_c))}$$

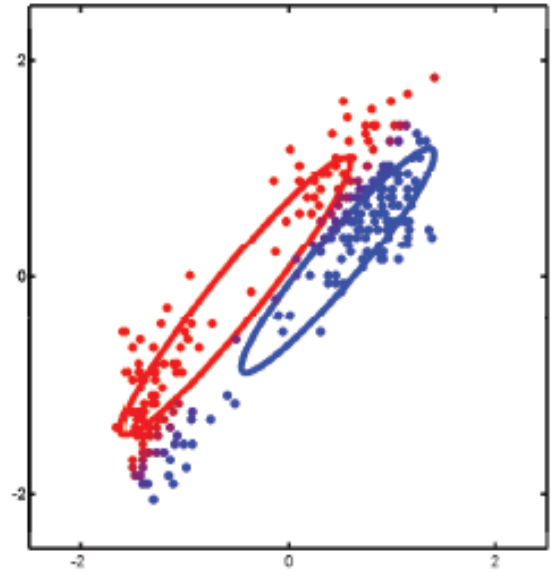
Example



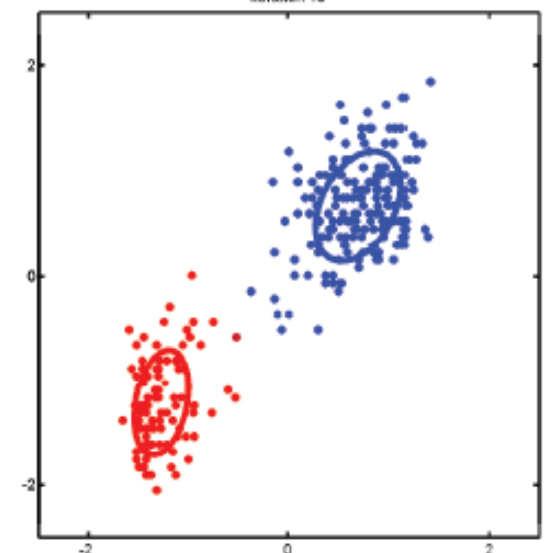
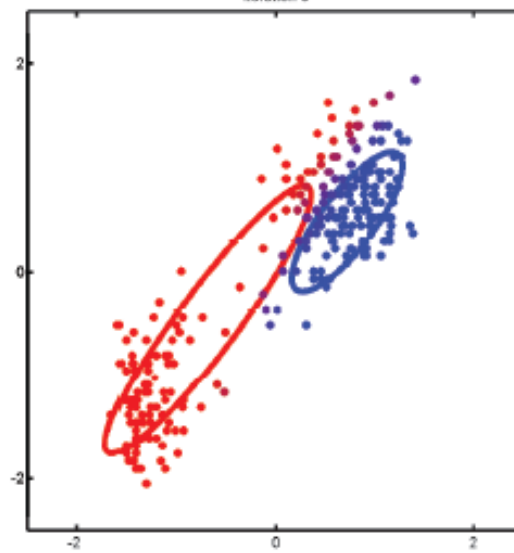
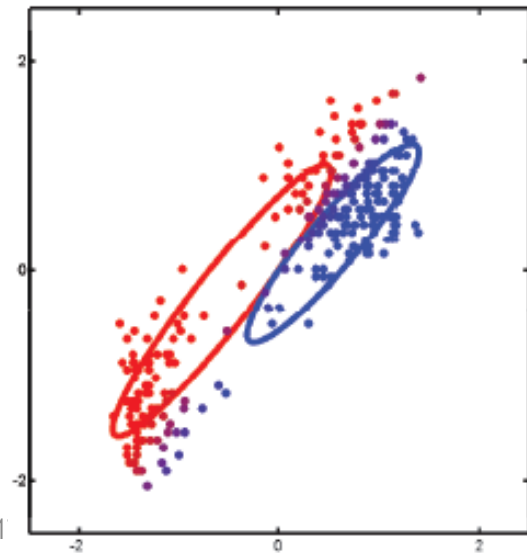
iteration 3



iteration 5



iteration 16



Other uses of EM

- Simplify computation
 - t-distribution as a Gaussian scale-mixture
- Missing data



Questions?

