CS434a/541a: Pattern Recognition
Prof. Olga Veksler

# Lecture 17

Today

- Parametric Unsupervised Learning
  - Expectation Maximization (EM)
    - one of the most useful statistical methods
    - oldest version in 1958 (*Hartley*)
    - seminal paper in 1977 (*Dempster et al.*)
    - can also be used when some samples are missing features

Unsupervised Learning

- In unsupervised learning, where we are only given samples $x_1, \ldots, x_n$ without class labels

- Last 2 lectures: nonparametric approach (clustering)

- Today, parametric approach
    - assume parametric distribution of data
    - estimate parameters of this distribution
    - much "harder" than the supervised learning case
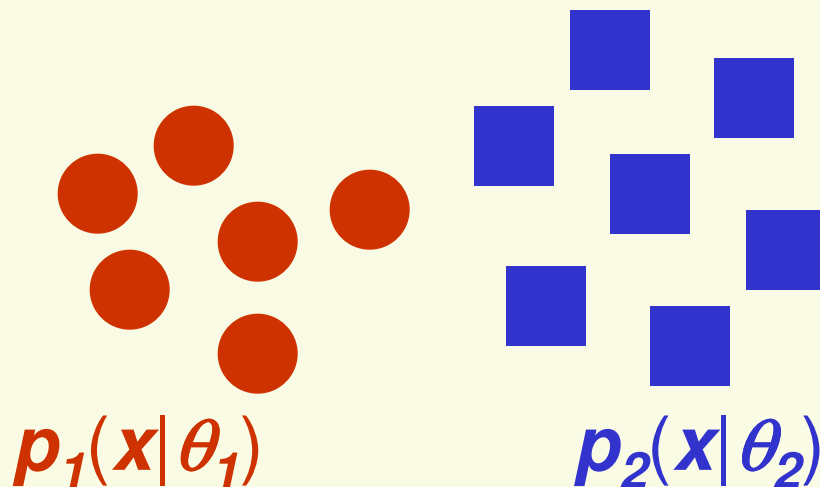
Parametric Unsupervised Learning

- Assume the data was generated by a model with known shape but unknown parameters

$$P(x/\theta)$$

$\longrightarrow$
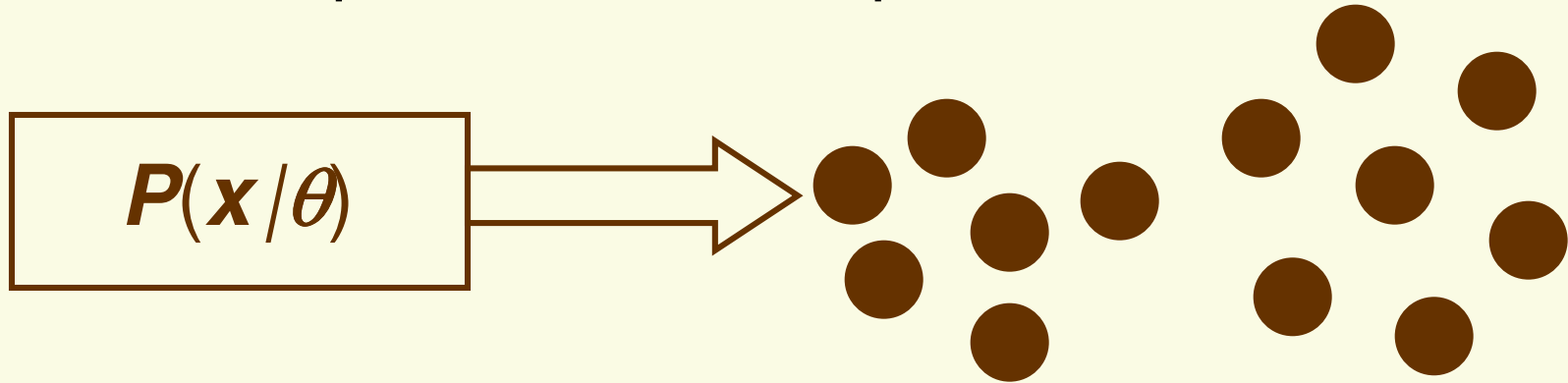
- Advantages of having a model
  - Gives a meaningful way to cluster data
    - adjust the parameters of the model to maximize the probability that the model produced the observed data
  - Can sensibly measure if a clustering is good
    - compute the likelihood of data induced by clustering
  - Can compare 2 clustering algorithms
    - which one gives the higher likelihood of the observed data?

Parametric Supervised Learning

- Let us recall supervised parametric learning
  - have m classes
  - have samples $x_1, \ldots, x_n$ each of class $1, 2, \ldots, m$
  - suppose D_i holds samples from class i
  - probability distribution for class i is $p_i(x|\theta_i)$

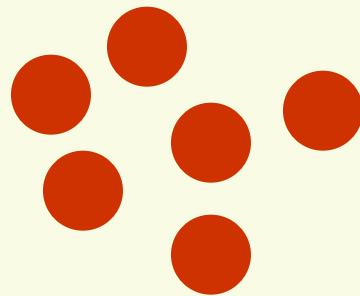$$p_1(x|\theta_1) \qquad p_2(x|\theta_2)$$

Parametric Supervised Learning

- Use the ML method to estimate parameters θ_i
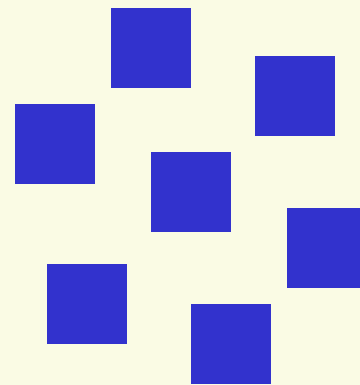 - find θ_i which maximizes the likelihood function $F(\theta_i)$

$$p(D_i \mid \theta_i) = \prod_{x \in D_i} p(x \mid \theta_i) = F(\theta_i)$$

  - or, equivalently, find θ_i which maximizes the log likelihood $l(\theta_i)$

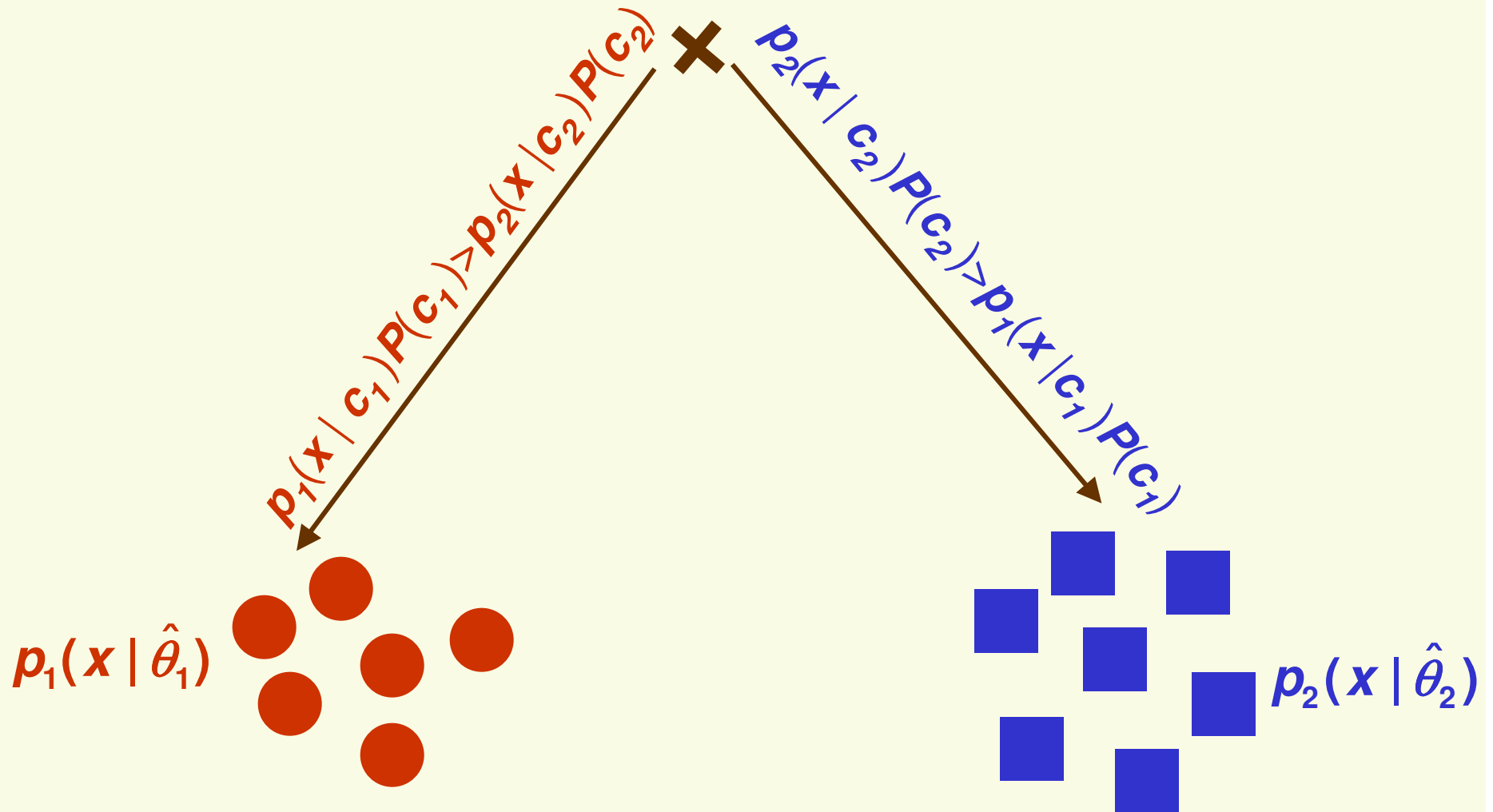$$l(\theta_i) = \ln p(D_i \mid \theta_i) = \sum_{x \in D_i} \ln p(x \mid \theta_i)$$

$$\hat{\theta}_1 = \arg\max_{\theta_1}\left[\ln p(D_1 \mid \theta_1)\right]$$

$$\hat{\theta}_2 = \arg\max_{\theta_2}\left[\ln p(D_2 \mid \theta_2)\right]$$

Parametric Supervised Learning

- now the distributions are fully specified
- can classify unknown sample using MAP classifier

$p_1(x \mid c_1)P(c_1) > p_2(x \mid c_2)P(c_2)$

$p_2(x \mid c_2)P(c_2) > p_1(x \mid c_1)P(c_1)$

$p_1(x \mid \hat{\theta}_1)$
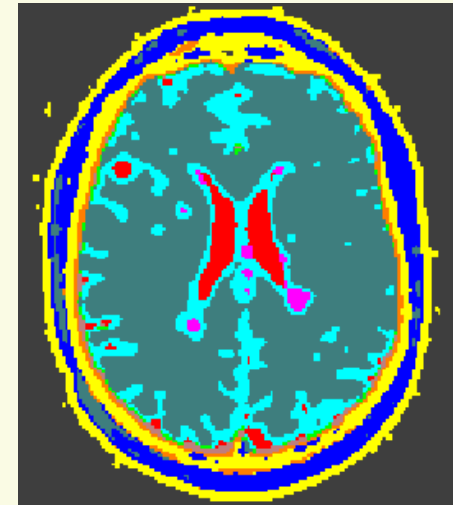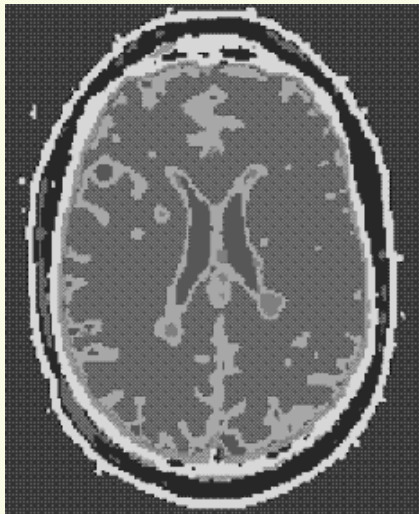
$p_2(x \mid \hat{\theta}_2)$

Parametric Unsupervised Learning

- In unsupervised learning, no one tells us the true classes for samples. We still know
  - have m classes
  - have samples $x_1, \ldots, x_n$ each of **unknown** class
  - probability distribution for class i is $p_i(x|\theta_i)$
- Can we determine the classes and parameters simultaneously?
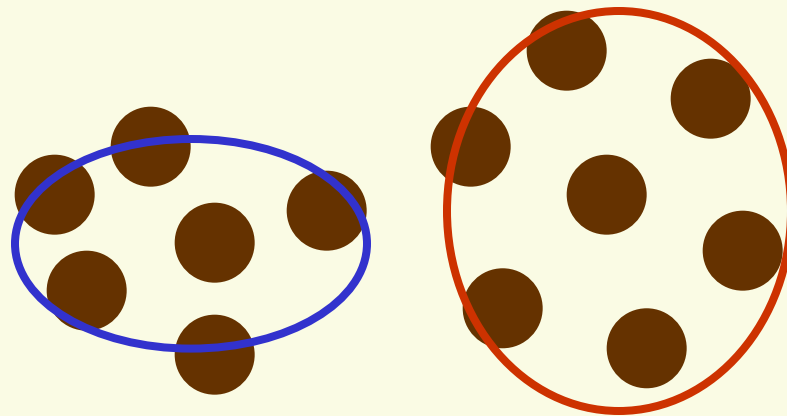
Example: MRI Brain Segmentation

segmentation

Picture from M. Leventon

- In MRI brain image, different brain tissues have different intensities
- Know that brain has 6 major types of tissues
- Each type of tissue can be modeled by a Gaussian $N(\mu_i, \sigma_i^2)$ reasonably well, parameters μ_i, σ_i^2 are unknown
- Segmenting (classifying) the brain image into different tissue classes is very useful
  - don't know which image pixel corresponds to which tissue (class)
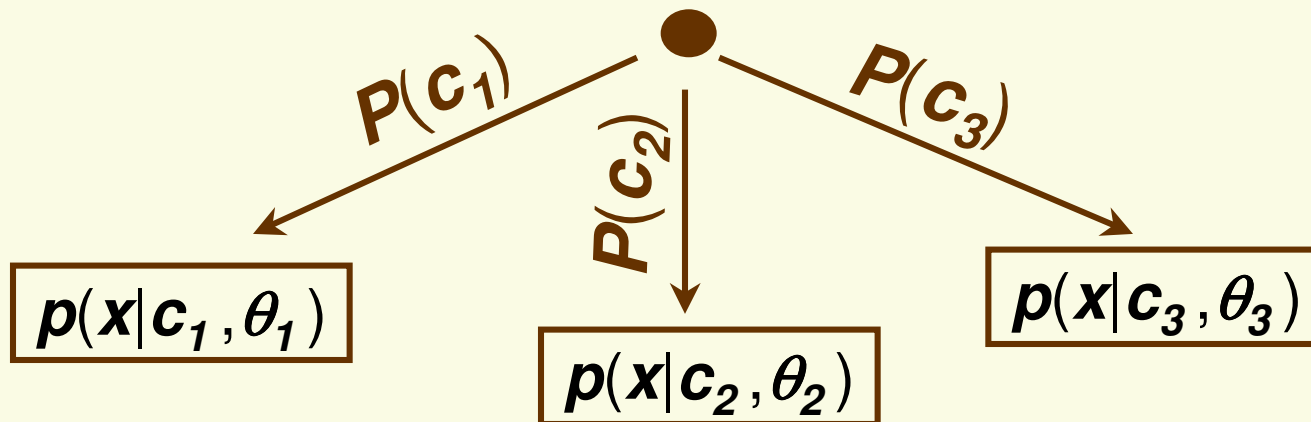 - don't know parameters for each $N(\mu_i, \sigma_i^2)$

Mixture Density Model

- Model data with **mixture density**

component densities

$$p(x \mid \theta) = \sum_{j=1}^{m} p(x \mid c_j, \theta_j) P(c_j)$$

mixing parameters

- where $\theta = \{\theta_1, ..., \theta_m\}$
- $P(c_1) + P(c_2) + ... + P(c_m) = 1$

- To generate a sample from distribution $p(x|\theta)$
  - first select class j with probability $P(c_j)$
  - then generate x according to probability law $p(x|c_j, \theta_j)$

$P(c_1)$ $P(c_2)$ $P(c_3)$

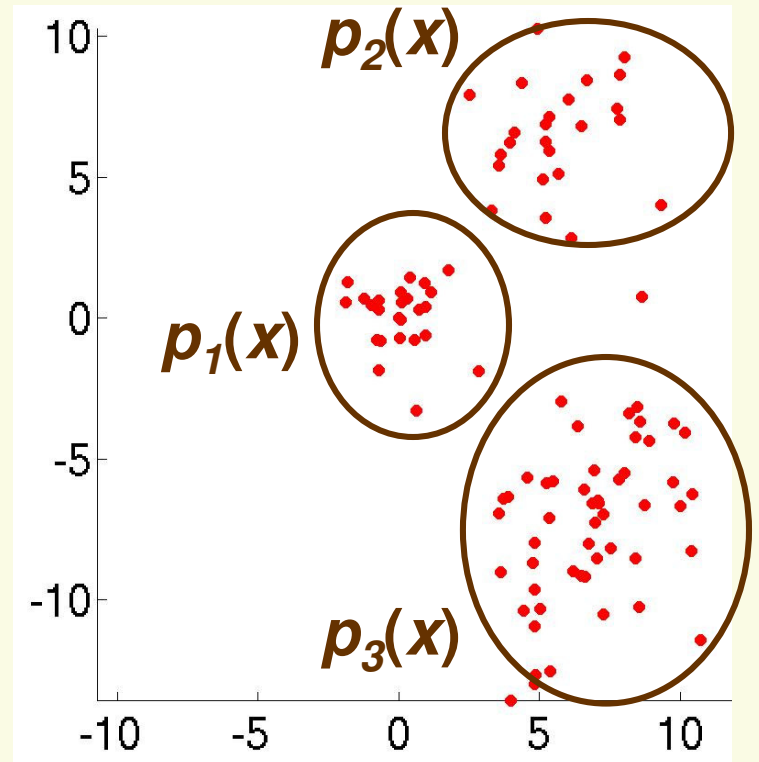$p(x|c_1, \theta_1)$ $p(x|c_2, \theta_2)$ $p(x|c_3, \theta_3)$

Example: Gaussian Mixture Density

- Mixture of 3 Gaussians

$$p_1(x) \cong N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$p_2(x) \cong N\left([6,6], \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}\right)$$

$$p_3(x) \cong N\left([7,-7], \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}\right)$$

$$p(x) = 0.2p_1(x) + 0.3p_2(x) + 0.5p_3(x)$$

Mixture Density

$$p(x \mid \theta) = \sum_{j=1}^{m} p(x \mid c_j, \theta_j) P(c_j)$$

- $P(c_1), \ldots, P(c_m)$ can be known or unknown

- Suppose we know how to estimate $\theta_1, \ldots, \theta_m$ and $P(c_1), \ldots, P(c_m)$

- Can "break apart" mixture $p(x|\theta)$ for classification

- To classify sample x, use MAP estimation, that is choose class i which maximizes

$$P(c_i \mid x, \theta_i) \propto \underbrace{p(x \mid c_i, \theta_i)}_{\substack{\textit{probability of component } i \\ \textit{to generate } x}} \underbrace{P(c_i)}_{\substack{\textit{probability of} \\ \textit{component } i}}$$

ML Estimation for Mixture Density

$$p(x \mid \theta, \rho) = \sum_{j=1}^{m} p(x \mid c_j, \theta_j) P(c_j) = \sum_{j=1}^{m} p(x \mid c_j, \theta_j) \rho_i$$

- Can use Maximum Likelihood estimation for a mixture density; need to estimate
  - $\theta_1, \ldots, \theta_m$
  - $\rho_1 = P(c_1), \ldots, \rho_m = P(c_m)$, and $\rho = \{\rho_1, \ldots, \rho_m\}$
- As in the supervised case, form the logarithm likelihood function

$$I(\theta, \rho) = \ln p(D \mid \theta, \rho) = \sum_{k=1}^{n} \ln p(x_k \mid \theta, \rho) = \sum_{k=1}^{n} \ln \left[\sum_{j=1}^{m} p(x \mid c_j, \theta_j) \rho_i \right]$$

ML Estimation for Mixture Density

$$I(\theta, \rho) = \sum_{k=1}^{n} \ln \left[\sum_{j=1}^{m} p(x \mid c_j, \theta_j) \rho_i \right]$$

- need to maximize $I(\theta, \rho)$ with respect to θ and ρ
- As you may have guessed, $I(\theta, \rho)$ is not the easiest function to maximize
  - If we take partial derivatives with respect to θ, ρ and set them to **0**, typically we have a "coupled" nonlinear system of equation
  - usually closed form solution cannot be found
- We could use the gradient ascent method
  - in general, it is not the greatest method to use, should only be used as last resort
- There is a better algorithm, called **EM**

Mixture Density

- Before EM, let's look at the mixture density again

$$p(x \mid \theta, \rho) = \sum_{j=1}^{m} p\left(x \mid c_j, \theta_j\right) \rho_j$$

- Suppose we know how to estimate $\theta_1, \ldots, \theta_m$ and $\rho_1, \ldots, \rho_m$
  - Estimating the class of x is easy with MAP, maximize

$$p(x \mid c_i, \theta_i) P(c_i) = p(x \mid c_i, \theta_i) \rho_i$$

- Suppose we know the class of samples $x_1, \ldots, x_n$
  - This is just the supervised learning case, so estimating $\theta_1, \ldots, \theta_m$ and $\rho_1, \ldots, \rho_m$ is easy

$$\hat{\theta}_i = \underset{\theta_i}{\operatorname{argmax}} \left[\ln p(D_i \mid \theta_i)\right] \qquad \hat{\rho}_i = \frac{\mid D_i \mid}{n}$$

- This is an example of chicken-and-egg problem
  - *ME* algorithm approaches this problem by adding "hidden" variables

Expectation Maximization Algorithm

- **EM** is an algorithm for ML parameter estimation when the data has missing values. It is used when
  1. data is incomplete (has missing values)
     - some features are missing for some samples due to data corruption, partial survey responces, etc.
     - This scenario is very useful, covered in section 3.9
  2. Suppose data X is complete, but $p(X|\theta)$ is hard to optimize. Suppose further that introducing certain hidden variables U whose values are missing, and suppose it is easier to optimize the "complete" likelihood function $p(X,U|\theta)$. Then EM is useful.
     - This scenario is useful for the mixture density estimation, and is subject of our lecture today
  - Notice that after we introduce artificial (hidden) variables U with missing values, case **2** is completely equivalent to case **1**
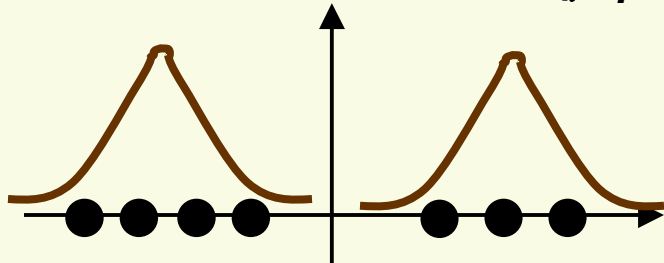
$$p(x \mid \theta) = \sum_{j=1}^{m} p(x \mid c_j, \theta_j) \rho_j$$

- For simplicity, assume component densities are

$$p(x \mid c_j, \theta_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_j)^2}{2\sigma^2} \right)$$

  - assume for now that the variance is known
  - need to estimate $\theta = \{\mu_1, \ldots, \mu_m\}$

- If we knew which sample came from which component (that is the class label), the ML parameter estimation is easy

- Thus to get an easier problem, introduce hidden variables which indicate which component each sample belongs to

EM: Hidden Variables for Mixture Density

- For $1 \leq i \leq n$, $1 \leq k \leq m$, define hidden variables $z_i^{(k)}$

$$z_i^{(k)} = \begin{cases} 1 & \text{if sample } i \text{ was generated by component } k \\ 0 & \text{otherwise} \end{cases}$$

$$\boxed{x_i \rightarrow \left\{ x_i, z_i^{(1)}, ..., z_i^{(m)} \right\}}$$

- $z_i^{(k)}$ are indicator *random* variables, they indicate which Gaussian component generated sample x_i

- Let $z_i = \{ z_i^{(1)}, ..., z_i^{(m)} \}$, indicator r.v. corresponding to sample x_i

- Conditioned on z_i, distribution of x_i is Gaussian

$$p(x_i \mid z_i, \theta) \sim N\left(\mu_k, \sigma^2 \right)$$

  - where k is s.t. $z_i^{(k)} = 1$

EM: Joint Likelihood

- Let $z_i = \{z_i^{(1)},\dots, z_i^{(m)}\}$, and $Z = \{z_1,\dots, z_n\}$

- The complete likelihood is

$$p(X,Z \mid \theta) = p(x_1,\dots, x_n, z_1,\dots, z_n \mid \theta) = \prod_{i=1}^{n} p(x_i, z_i \mid \theta)$$

$$= \prod_{i=1}^{n} \underbrace{p(x_i \mid z_i, \theta)}_{gaussian} \underbrace{p(z_i)}_{part\ of\ \rho_c}$$

- If we actually observed Z, the log likelihood $\ln[p(X,Z|\theta)]$ would be trivial to maximize with respect to θ and ρ_i

- The problem, is, of course, is that the values of Z are missing, since we made it up (that is Z is hidden)

EM Derivation

- Instead of maximizing $ln[p(X,Z|\theta)]$ the idea behind *EM* is to maximize some function of $ln[p(X,Z|\theta)]$, usually its expected value

$$E_z[\ln p(X,Z|\theta)]$$

  - If θ makes $ln[p(X,Z|\theta)]$ large, then θ tends to make $E[ln\,p(X,Z|\theta)]$ large

  - the expectation is with respect to the missing data Z

  - that is with respect to density $p(Z|X,\theta)$

- however θ is our ultimate goal, we don't know θ!

EM Algorithm

- *EM* solution is to iterate

    1. start with initial parameters $\theta^{(0)}$
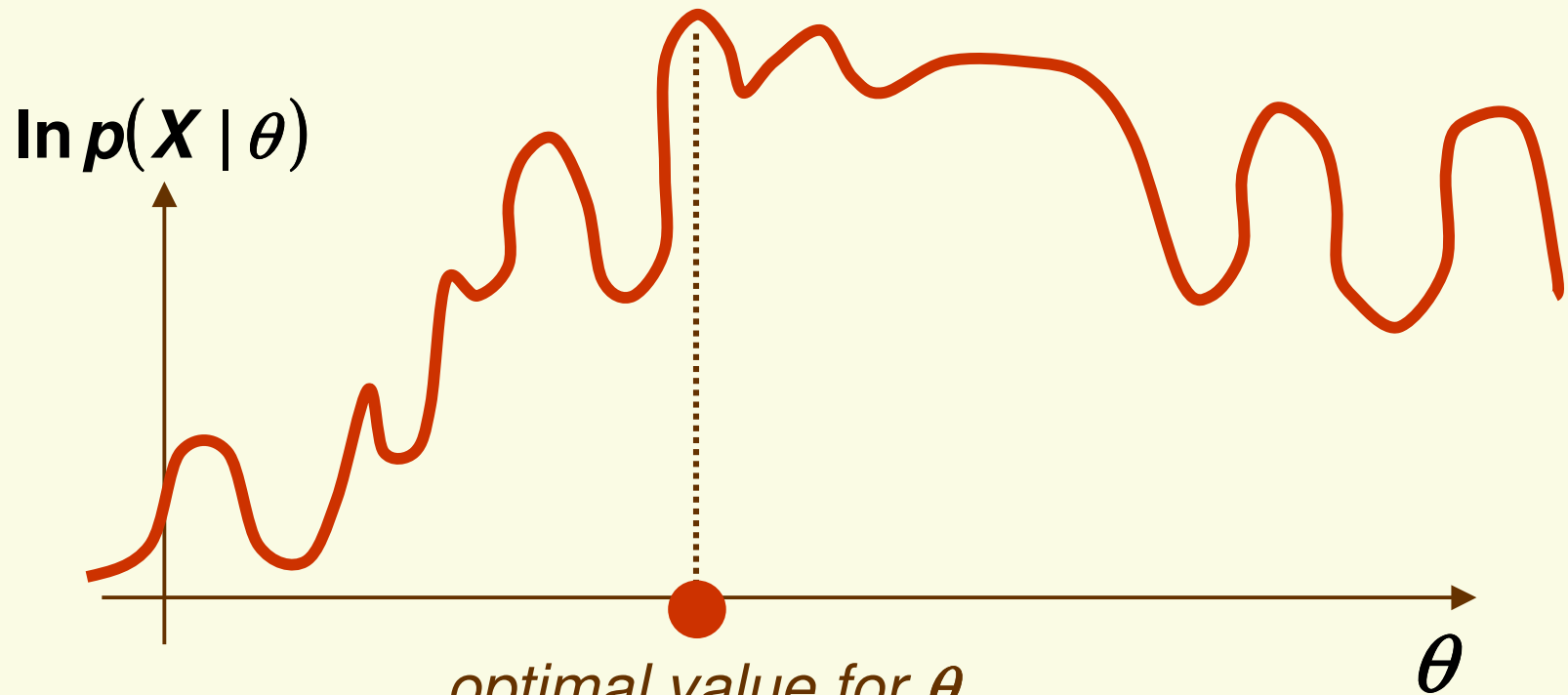
    iterate the following 2 step until convergence

    E. compute the expectation of log likelihood with respect to current estimate $\theta^{(t)}$ and **X**. Let's call it $\boldsymbol{Q}(\theta|\theta^{(t)})$

    $$Q\left(\theta \mid \theta^{(t)}\right) = \boldsymbol{E}_z\left[\ln \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{Z} \mid \theta) \mid \boldsymbol{X}, \theta^{(t)}\right]$$

    M. maximize $\boldsymbol{Q}(\theta|\theta^{(t)})$

    $$\theta^{(t+1)} = \underset{\theta}{\textbf{argmax}}\, \boldsymbol{Q}\left(\theta \mid \theta^{(t)}\right)$$

EM Algorithm: Picture

$\ln p(X \mid \theta)$

θ

optimal value for θ

we'd like to find it but optimizing $p(X \mid \theta)$ is very difficult

EM Algorithm: Picture

$\ln p(X, Z | \theta)$

θ

Z

unobserved Z corresponding to observed data X

This curve corresponds to the correct Z, we should optimize for but Z is not observed

for mixture estimation, there are m^n curves, each curve corresponds to a particular assignment of samples to classes
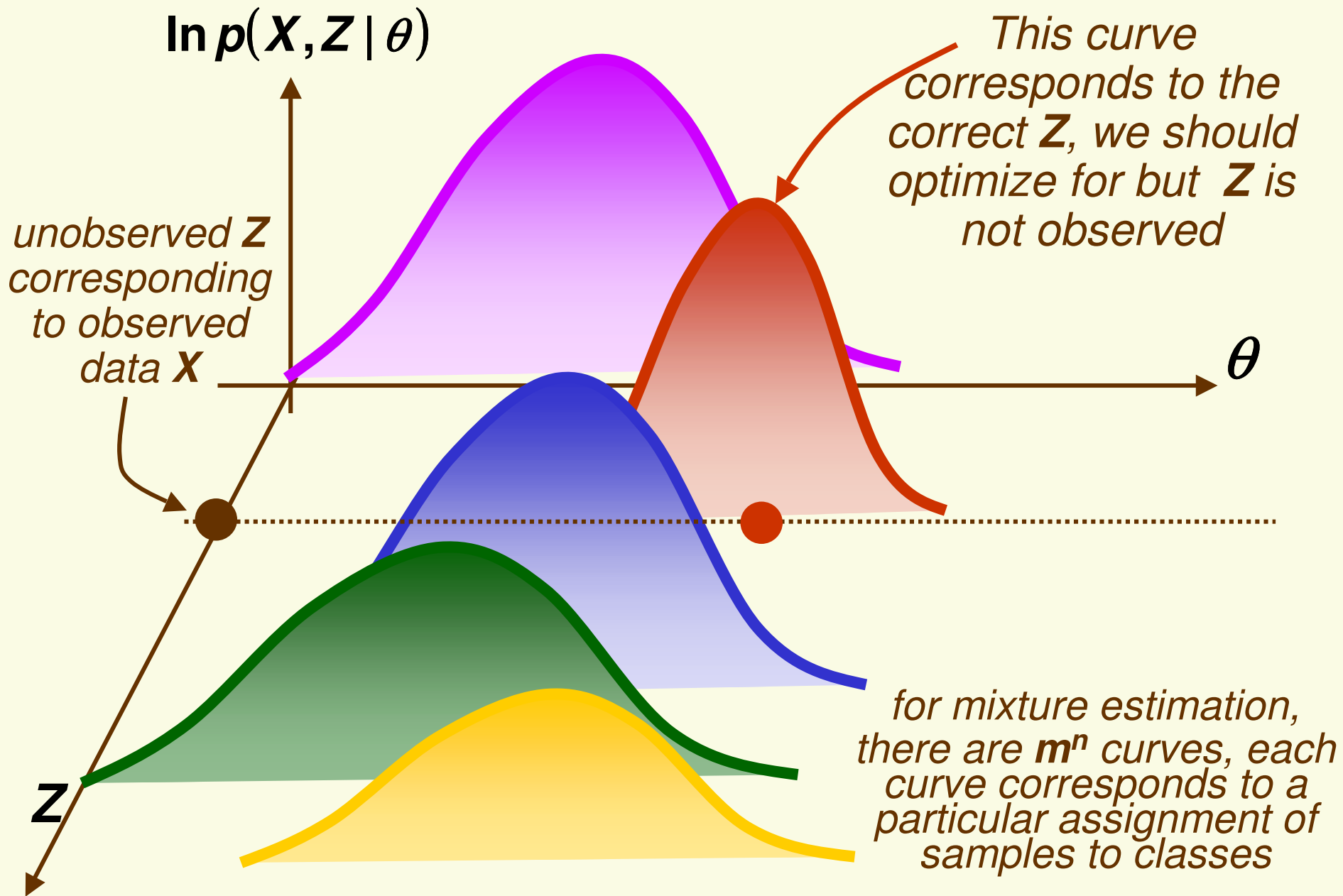
EM Algorithm: Picture
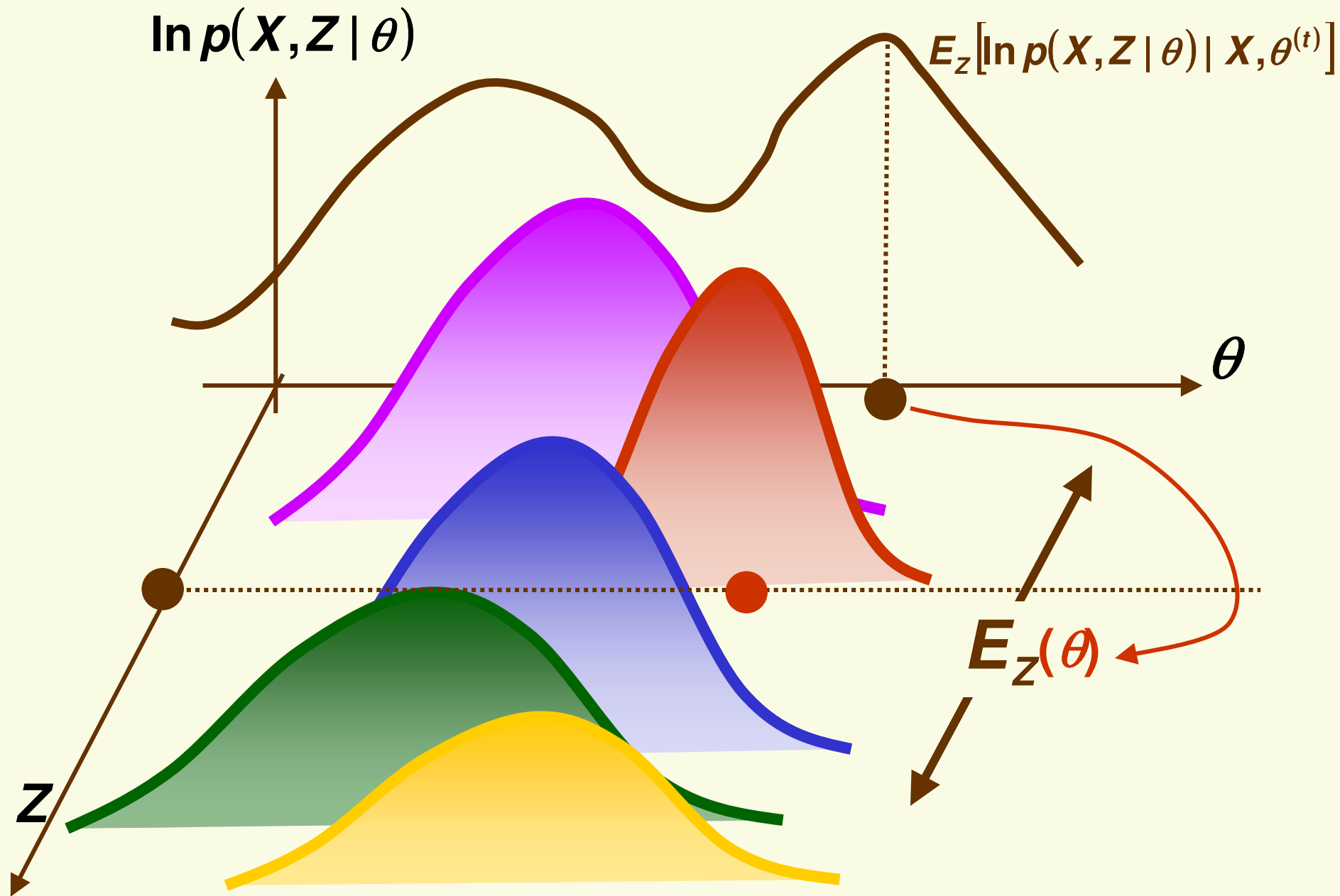
EM Algorithm

- It can be proven that EM algorithm converges to the local maximum of the log-likelihood

$$\ln p(X \mid \theta)$$

- Why is it better than gradient ascent?
    - Convergence of EM is usually **significantly** faster, in the beginning, very large steps are made (that is likelihood function increases rapidly), as opposed to gradient ascent which usually takes tiny steps
    - gradient descent is not guaranteed to converge
    - recall all the difficulties of choosing the appropriate learning rate

- Let's come back to our example $p(x \mid \theta) = \sum_{j=1}^{m} p(x \mid c_j, \theta_j) \rho_j$

$$p(x \mid c_j, \theta_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_j)^2}{2\sigma^2}\right)$$

  - need to estimate $\theta = \{\mu_1, \ldots, \mu_m\}$ and $\rho_1, \ldots, \rho_m$

- for $1 \leq i \leq n$, $1 \leq k \leq m$, define $z_i^{(k)}$

$$z_i^{(k)} = \begin{cases} 1 & \text{if sample } i \text{ was generated by component } k \\ 0 & \text{otherwise} \end{cases}$$

- as before, $z_i = \{z_i^{(1)}, \ldots, z_i^{(m)}\}$, and $Z = \{z_1, \ldots, z_n\}$

- We need log-likelihood of observed X and hidden Z

$$\ln p(X, Z \mid \theta) = \ln \prod_{i=1}^{n} p(x_i, z_i \mid \theta) = \sum_{i=1}^{n} \ln p(x_i \mid z_i, \theta) P(z_i)$$

EM for Mixture of Gaussians: E step

- We need log-likelihood of observed **X** and hidden **Z**

$$\ln p(X,Z \mid \theta) = \ln \prod_{i=1}^{n} p(x_i, z_i \mid \theta) = \sum_{i=1}^{n} \ln p(x_i \mid z_i, \theta) P(z_i)$$

- First let's rewrite $p(x_i \mid z_i, \theta) P(z_i)$

$$p(x_i \mid z_i, \theta) P(z_i) = \begin{cases} p\left(x_i \mid z_i^{(1)} = 1, \theta\right) P\left(z_i^{(1)} = 1\right) & \text{if } z_i^{(1)} = 1 \\ \vdots & \vdots \\ p\left(x_i \mid z_i^{(m)} = 1, \theta\right) P\left(z_i^{(m)} = 1\right) & \text{if } z_i^{(m)} = 1 \end{cases}$$

$$= \prod_{k=1}^{m} \left[p\left(x_i \mid z_i^{(k)} = 1, \theta\right) P\left(z_i^{(k)} = 1\right) \right]^{z_i^{(k)}}$$

$$= \prod_{k=1}^{m} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma^2} \right) P\left(z_i^{(k)} = 1\right) \right]^{z_i^{(k)}}$$

EM for Mixture of Gaussians: E step

- log-likelihood of observed **X** and hidden **Z** is

$$\ln p(X, Z \mid \theta) = \sum_{i=1}^{n} \ln p(x_i \mid z_i, \theta) P(z_i)$$

$$= \sum_{i=1}^{n} \ln \prod_{k=1}^{m} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma^2} \right) P\left(z_i^{(k)} = 1\right) \right]^{z_i^{(k)}}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{m} \ln \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma^2} \right) P\left(z_i^{(k)} = 1\right) \right]^{z_i^{(k)}}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{m} z_i^{(k)} \left[\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \underbrace{P\left(z_i^{(k)} = 1\right)}_{} \right]$$

$$P(\text{sample } x_i \text{ from class } k) = P(c_k) = \rho_k$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{m} z_i^{(k)} \left[\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \rho_k \right]$$

EM for Mixture of Gaussians: E step

- log-likelihood of observed X and hidden Z is

$$\ln p(X,Z \mid \theta) = \sum_{i=1}^{n} \sum_{k=1}^{m} z_i^{(k)} \left[\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \rho_k \right]$$

- For the E step, we must compute

$$Q(\theta \mid \theta^{(t)}) = Q(\theta \mid \mu_1^{(t)}, ..., \mu_m^{(t)}, \rho_1^{(t)}, ..., \rho_m^{(t)}) = E_z \left[\ln p(X,Z \mid \theta) \mid X, \theta^{(t)} \right]$$

$$= E_z \left(\sum_{i=1}^{n} \sum_{k=1}^{m} z_i^{(k)} \left[\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \rho_k^{(t)} \right] \right)$$

$$E_X \left[\sum_i a_i x_i + b \right] = \sum_i a_i E_X[x_i] + b$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{m} E_z \left[z_i^{(k)} \right] \left(\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \rho_k \right)$$

EM for Mixture of Gaussians: E step

$$Q\!\left(\theta\,|\,\theta^{(t)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{m} E_z\!\left[z_i^{(k)}\right]\!\left(\ln\frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i-\mu_k)^2}{2\sigma^2} + \ln\rho_k\right)$$

- need to compute $E_z[z_i^{(k)}]$ in the above expression

$$E_z\!\left[z_i^{(k)}\right] = 0 * P\!\left(z_i^{(k)} = 0\,|\,\theta^{(t)}, x_i\right) + 1 * P\!\left(z_i^{(k)} = 1\,|\,\theta^{(t)}, x_i\right)$$

$$= P\!\left(z_i^{(k)} = 1\,|\,\theta^{(t)}, x_i\right) = \frac{p\!\left(x_i\,|\,\theta^{(t)}, z_i^{(k)} = 1\right) P\!\left(z_i^{(k)} = 1\,|\,\theta^{(t)}\right)}{p\!\left(x_i\,|\,\theta^{(t)}\right)}$$

$$= \frac{\rho_k^{(t)} \exp\!\left(-\frac{1}{2}\left(x_i - \mu_k^{(t)}\right)^2\right)}{\sum_{j=1}^{m} P\!\left(x_i\,|\,\theta^{(t)}, z_i^{(j)} = 1\right) P\!\left(z_i^{(j)} = 1\,|\,\theta^{(t)}\right)} = \frac{\rho_k^{(t)} \exp\!\left(-\frac{1}{2\sigma^2}\left(x_i - \mu_k^{(t)}\right)^2\right)}{\sum_{j=1}^{m} \rho_j^{(t)} \exp\!\left(-\frac{1}{2\sigma^2}\left(x_i - \mu_j^{(t)}\right)^2\right)}$$

- we are finally done with the **E** step
  - for implementation, just need to compute $E_z[z_i^{(k)}]$'s don't need to compute **Q**

EM for Mixture of Gaussians: M step

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{m} E_z[z_i^{(k)}] \left(\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \rho_k \right)$$

- Need to maximize **Q** with respect to all parameters
- First differentiate with respect to μ_k

$$\frac{\partial}{\partial \mu_k} Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} E_z[z_i^{(k)}] \frac{(x_i - \mu_k)}{\sigma^2} = 0$$

$$\Rightarrow new\ \mu_k = \mu_k^{(t+1)} = \boxed{\frac{1}{n} \sum_{i=1}^{n} E_z[z_i^{(k)}] x_i}$$

*the mean for class **k** is weighted average of all samples, and this weight is proportional to the current estimate of probability that the sample belongs to class **k***

EM for Mixture of Gaussians: M step

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{m} E_z[z_i^{(k)}] \left(\ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x_i - \mu_k)^2}{2\sigma^2} + \ln \rho_k \right)$$

- For ρ_k we have to use Lagrange multipliers to preserve constraint $\sum_{j=1}^{m} \rho_j = 1$

- Thus we need to differentiate $F(\lambda, \rho) = Q(\theta \mid \theta^{(t)}) - \lambda \left(\sum_{j=1}^{m} \rho_j - 1 \right)$

$$\frac{\partial}{\partial \rho_k} F(\lambda, \rho) = \sum_{i=1}^{n} \frac{1}{\rho_k} E_z[z_i^{(k)}] - \lambda = 0 \implies \sum_{i=1}^{n} E_z[z_i^{(k)}] - \lambda \rho_k = 0$$

- Summing up over all components: $\sum_{k=1}^{m} \sum_{i=1}^{n} E_z[z_i^{(k)}] = \sum_{k=1}^{m} \lambda \rho_k$

- Since $\sum_{k=1}^{m} \sum_{i=1}^{n} E_z[z_i^{(k)}] = n$ and $\sum_{k=1}^{m} \rho_k = 1$ we get $\lambda = n$

$$\boxed{\rho_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} E_z[z_i^{(k)}]}$$

EM Algorithm

The algorithm on this slide applies ONLY to univariate gaussian case with known variances

1. Randomly initialize $\mu_1,\ldots,\mu_m$, $\rho_1,\ldots,\rho_m$ (with constraint $\Sigma\rho_i = 1$)

iterate until no change in $\mu_1,\ldots,\mu_m$, $\rho_1,\ldots,\rho_m$

E. for all **i**, **k**, compute

$$E_z\left[z_i^{(k)}\right] = \frac{\rho_k \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_k)^2\right)}{\sum_{j=1}^{m} \rho_j \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right)}$$

M. for all **k**, do parameter update

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} E_z\left[z_i^{(k)}\right] x_i \qquad \rho_k = \frac{1}{n}\sum_{i=1}^{n} E_z\left[z_i^{(k)}\right]$$

EM Algorithm

- For the more general case of multivariate Gaussians with unknown means and variances

  - **E** step: $E_z\big[z_i^{(k)}\big] = \dfrac{\rho_k\, p(x\,|\,\mu_k, \Sigma_k)}{\displaystyle\sum_{j=1}^{m} \rho_j p\big(x\,|\,\mu_j, \Sigma_j\big)}$

    where $p(x\,|\,\mu_k, \Sigma_k) = \dfrac{1}{(2\pi)^{d/2}\big|\Sigma_k^{-1}\big|^{1/2}}\, exp\left[-\dfrac{1}{2}(x-\mu_k)^t\, \Sigma_k^{-1}(x-\mu_k)\right]$

  - **M** step:

    $$\rho_k = \frac{1}{n}\sum_{i=1}^{n} E_z\big[z_i^{(k)}\big]$$

    $$\Sigma_k = \frac{\displaystyle\sum_{i=1}^{n} E_z\big[z_i^{(k)}\big](x_i-\mu_k)(x_i-\mu_k)^T}{\displaystyle\sum_{i=1}^{n} E_z\big[z_i^{(k)}\big]}$$

    $$\mu_k = \frac{\displaystyle\sum_{i=1}^{n} E_z\big[z_i^{(k)}\big]x_i}{\displaystyle\sum_{i=1}^{n} E_z\big[z_i^{(k)}\big]}$$

EM Algorithm and K-means

- **k**-means can be derived from EM algorithm

- Setting mixing parameters equal for all classes,

$$E_z\left[z_i^{(k)}\right] = \frac{\rho_k \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_k)^2\right)}{\sum_{j=1}^{m} \rho_j \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right)} = \frac{\exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_k)^2\right)}{\sum_{j=1}^{m} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right)}$$

- If we let $\sigma \to \infty$, then

$$E_z\left[z_i^{(k)}\right] = \begin{cases} 1 & \textit{if } \forall j, \ \| x_i - \mu_k \| > \| x_i - \mu_j \| \\ 0 & \textit{otherwise} \end{cases}$$

  - so at the *E* step, for each current mean, we find all points closest to it and form new clusters

  - at the *M* step, we compute the new means inside current clusters

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} E_z\left[z_i^{(k)}\right] x_i$$

EM Gaussian Mixture Example
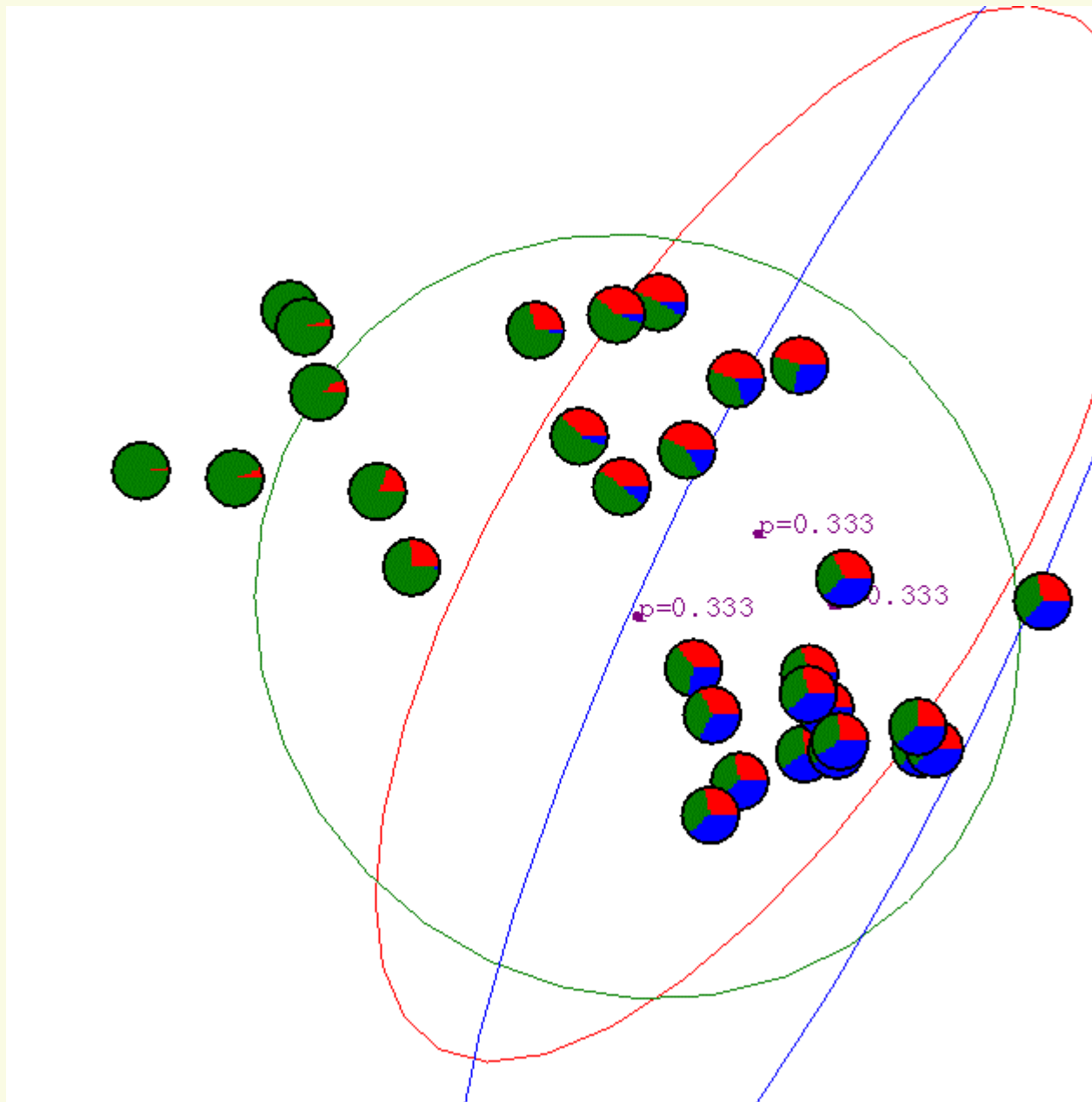
EM Gaussian Mixture Example

After first iteration
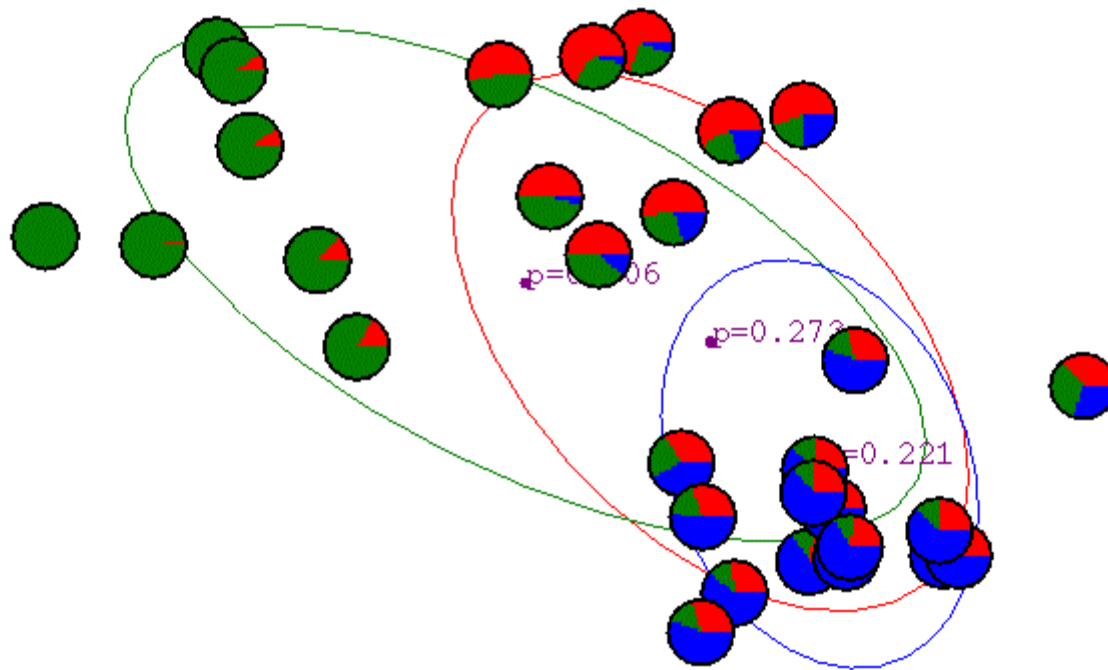
EM Gaussian Mixture Example
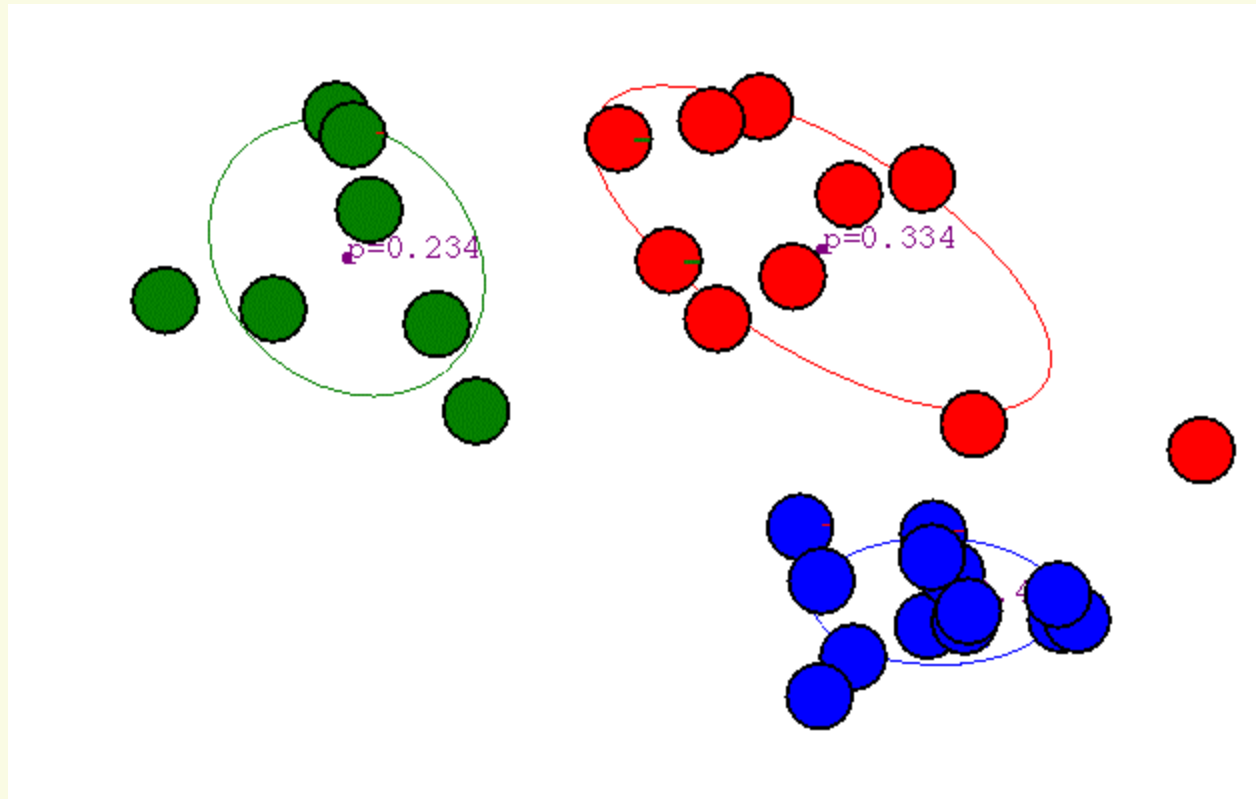
After second iteration
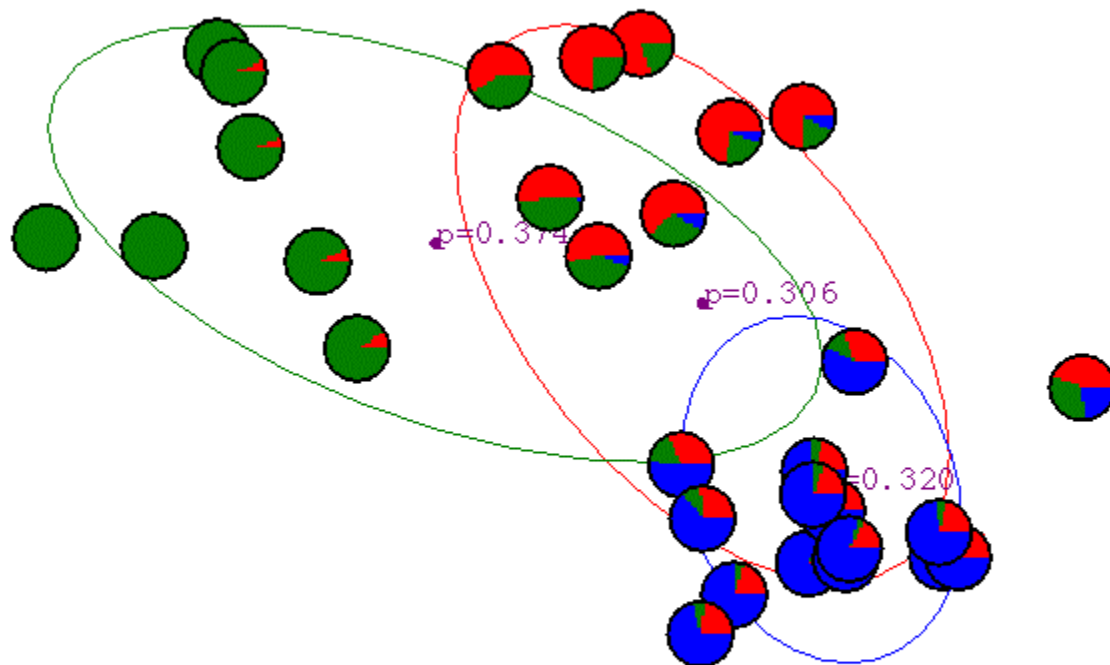
EM Gaussian Mixture Example

After third iteration
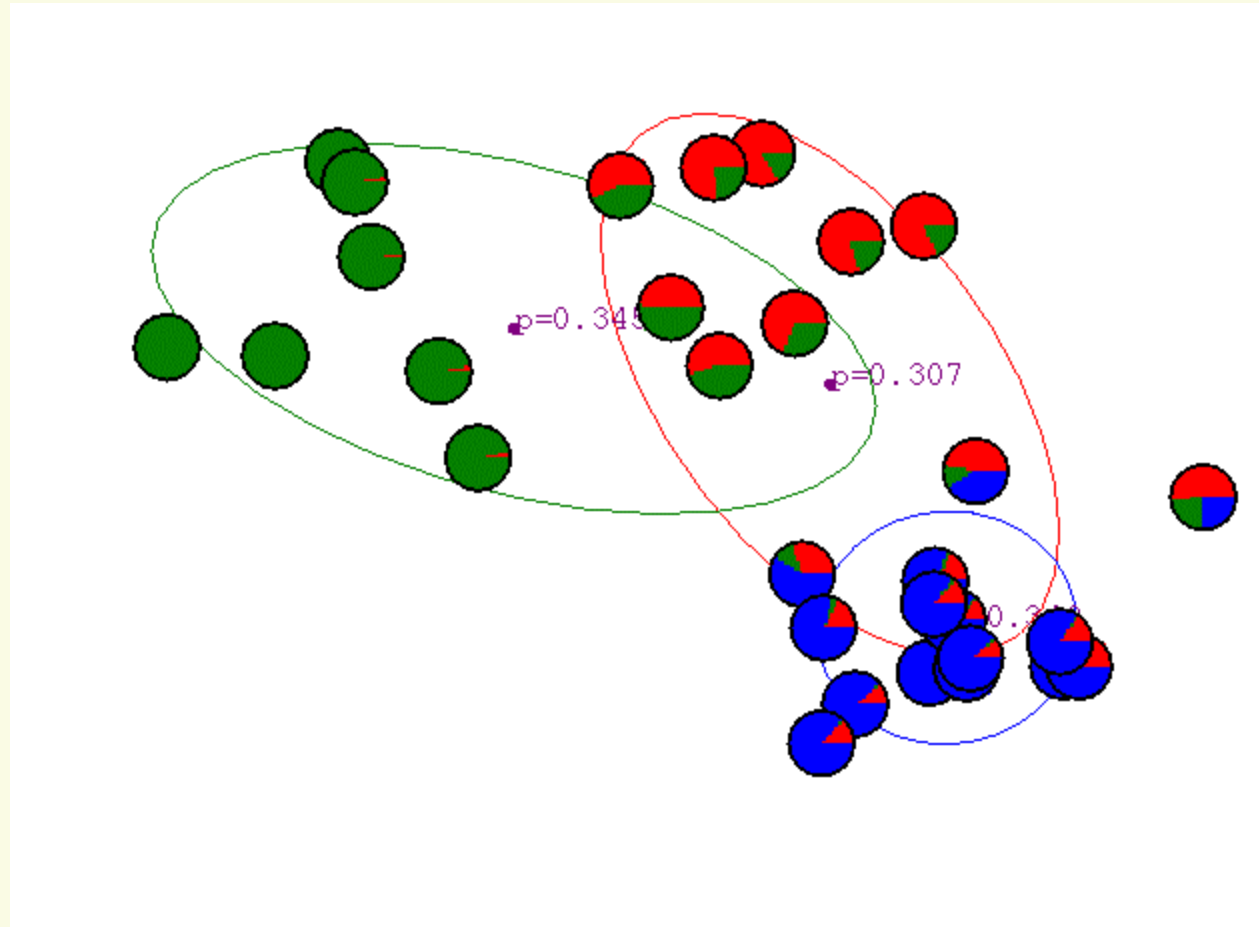
EM Gaussian Mixture Example

After 20th iteration

EM Example

- Example from R. Gutierrez-Osuna
- Training set of 900 examples forming an annulus
- Mixture model with $m = 30$ Gaussian components of unknown mean and variance is used
- Training:
  - Initialization:
    - means to 30 random examples
    - covaraince matrices initialized to be diagonal, with large variances on the diagonal (compared to the training data variance)
  - During EM training, components with small mixing coefficients were trimmed
    - This is a trick to get in a more compact model, with fewer than 30 Gaussian components

EM Example
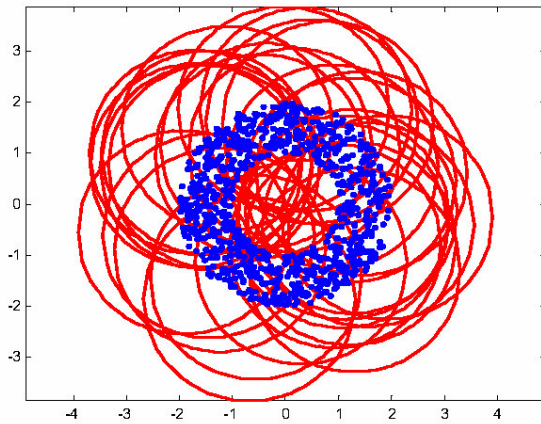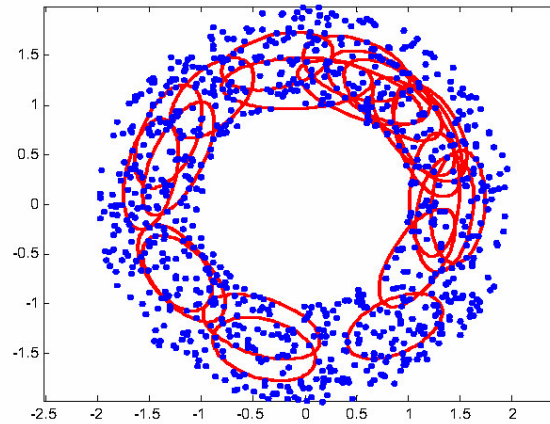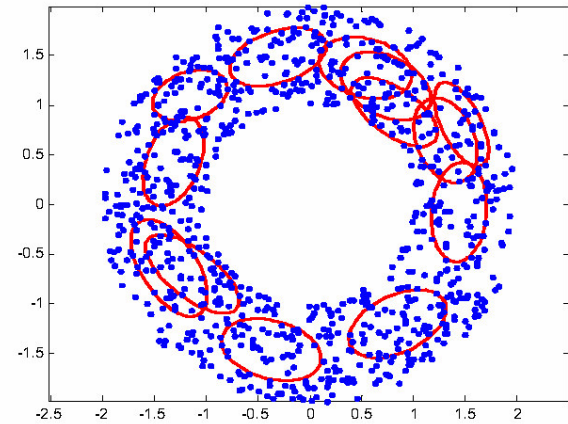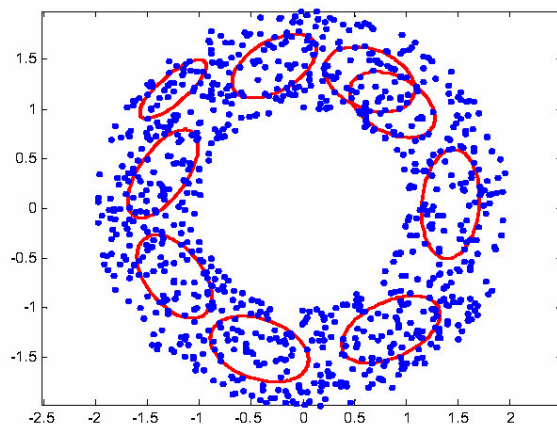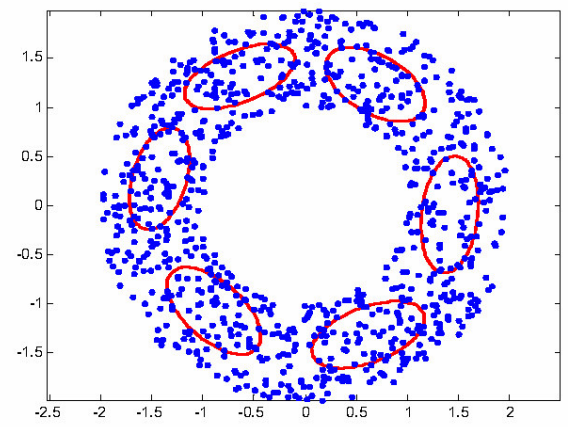
from R. Gutierrez-Osuna
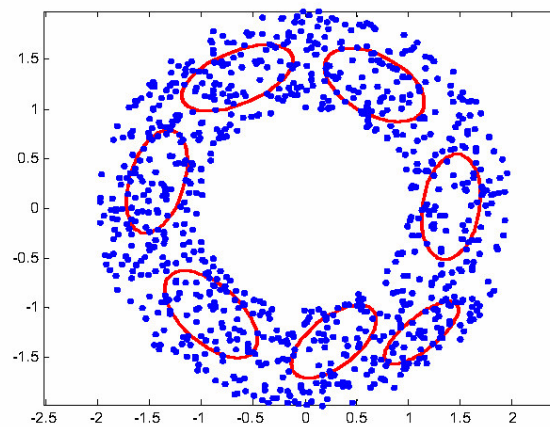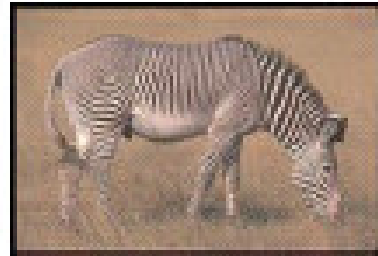
EM Texture Segmentation Example

Figure from "Color and Texture Based Image Segmentation Using EM and Its Application to Content Based Image Retrieval",S.J. Belongie et al., ICCV 1998

EM Motion Segmentation Example

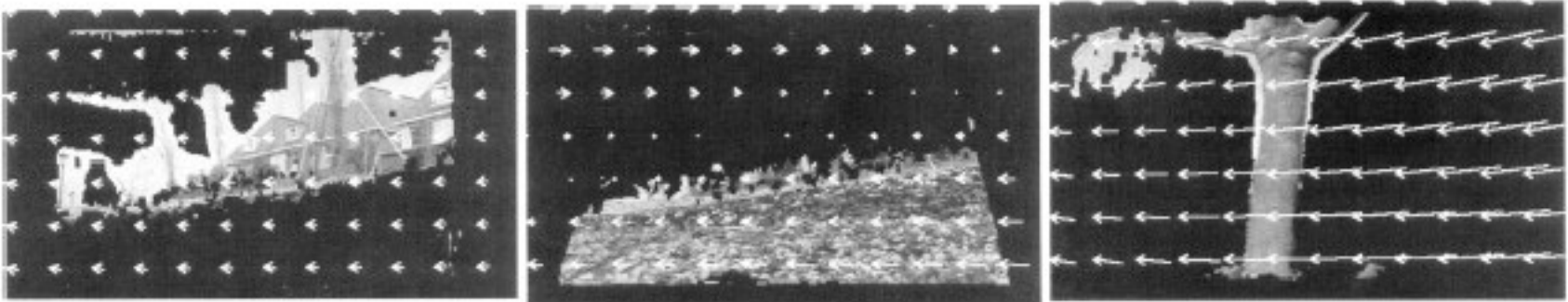**Three frames from the MPEG "flower garden" sequence**

Figure from "Representing Images with layers,", by J. Wang and E.H. Adelson, IEEE Transactions on Image Processing, 1994, c 1994, IEEE

Summary

- Advantages
  - If the assumed data distribution is correct, the algorithm works well
- Disadvantages
  - If assumed data distribution is wrong, results can be quite bad.
    - In particular, bad results if use incorrect number of classes (i.e. the number of mixture components)