Plan

Cool stuff

- Expectation-Maximization algorithm
  - Gaussian mixture models for clustering
- Kernels
  - Linear regression
  - Support vector machines
- Duality
  - Support vector machines

Course Update

Out

Current Plan (updated)

Due

- HW 8 (online)
- Mini-project proposal
- HW 9 (online)
- HW 10 (written/prog)
- Midterm 2
- Mini-project

Sun	Mon	Tue	Wed	Thu	Fri	Sat
2	3	4	5 HW8 Proj	6	7	8
9 HW8	10 HW9 HW10	11	12	13	14	15
16	17 HW9	18	19 Prop	20	21	22 HW10
23	24	25	26 MT2	27	28	29
30	1	2	3	4	5 Proj	6

Poll 1

How many people are currently in your mini-project group, including yourself?

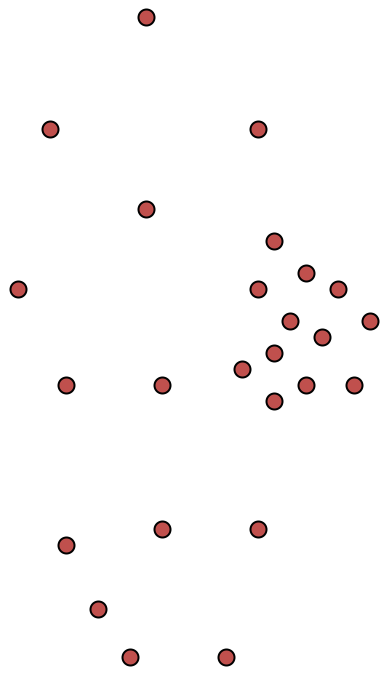
A. 0 (don’t choose this; it doesn’t make sense)
B. 1 (haven’t started looking)
C. 1 (started looking)
D. 2 (haven’t started looking)
E. 2 (started looking)
F. 3
G. 4
H. 5+

10-315
Introduction to ML

Gaussian Mixture Models and Expectation Maximization

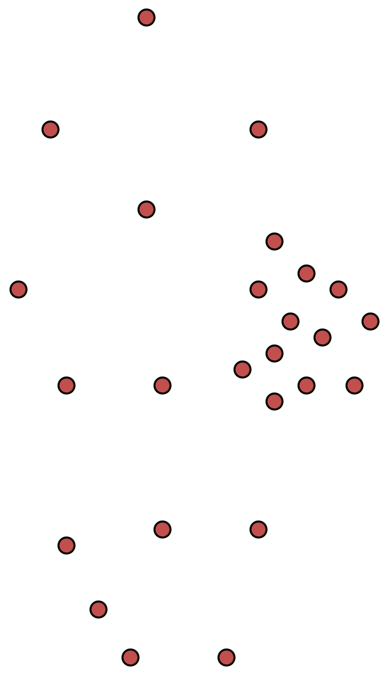
Instructor: Pat Virtue

(One) bad case for K-means



- Clusters may overlap
- Some clusters may be “wider” than others
- Clusters may not be linearly separable

(One) bad case for K-means



- Clusters may overlap
- Some clusters may be "wider" than others
- Clusters may not be linearly separable

Slide credit: CMU MLD Aarti Singh

Partitioning Algorithms

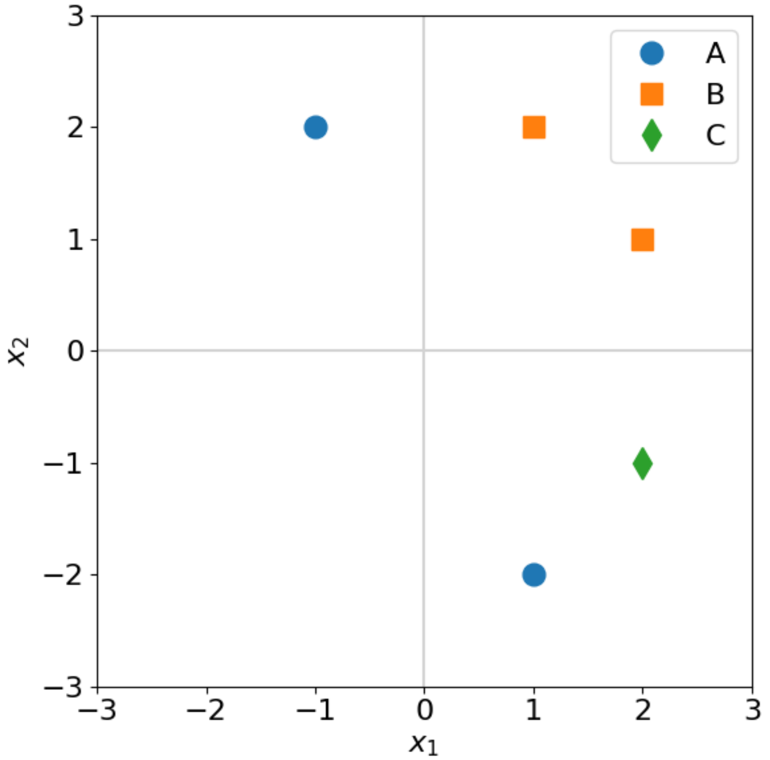
- K-means
  - **hard assignment**: each object belongs to only one cluster

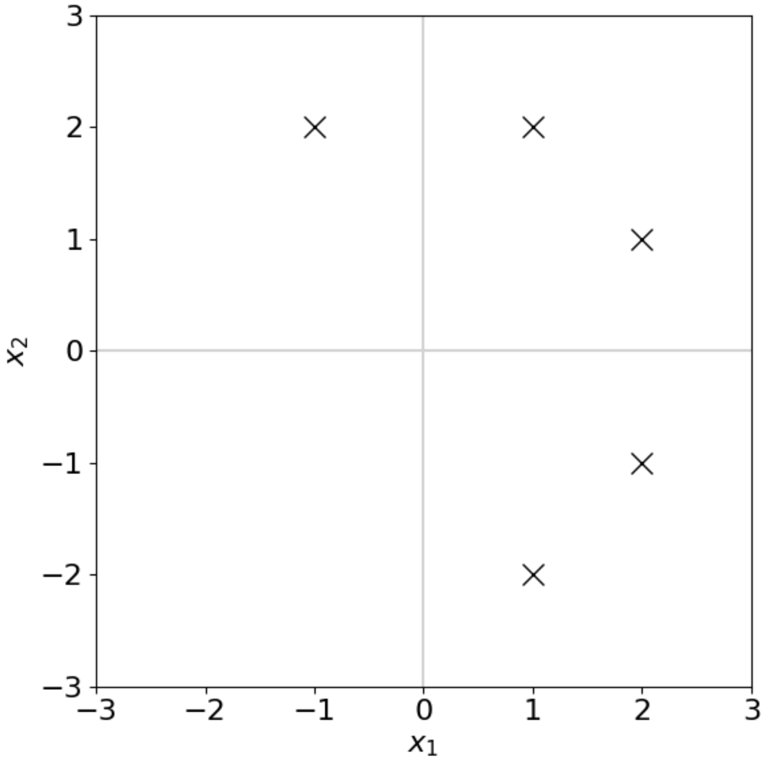
- Mixture modeling
  - **soft assignment**: probability that an object belongs to a cluster

Generative approach

Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models





Poll 2

Which of these terms is the likelihood?

Select all that apply

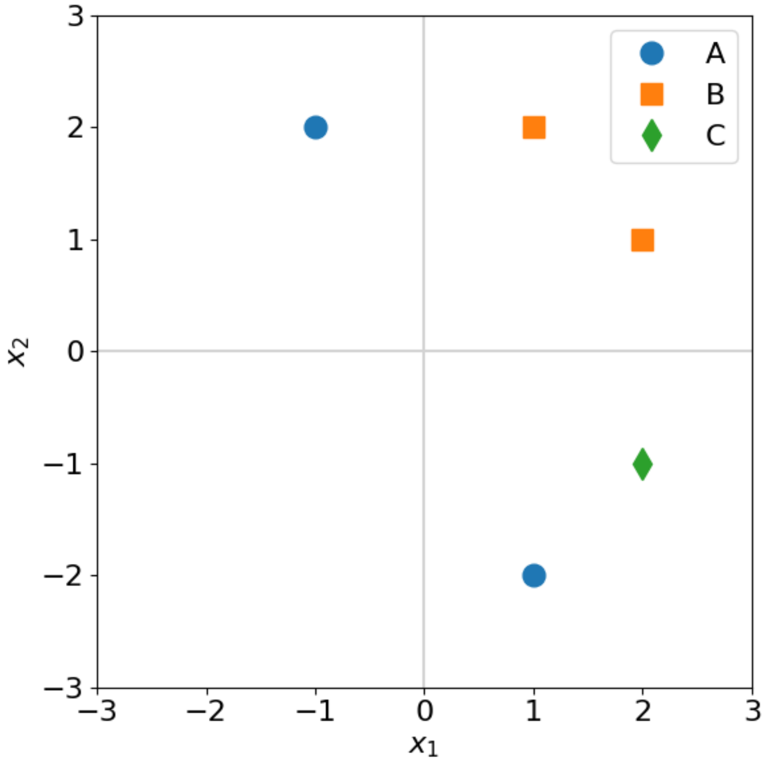
A B C E F G

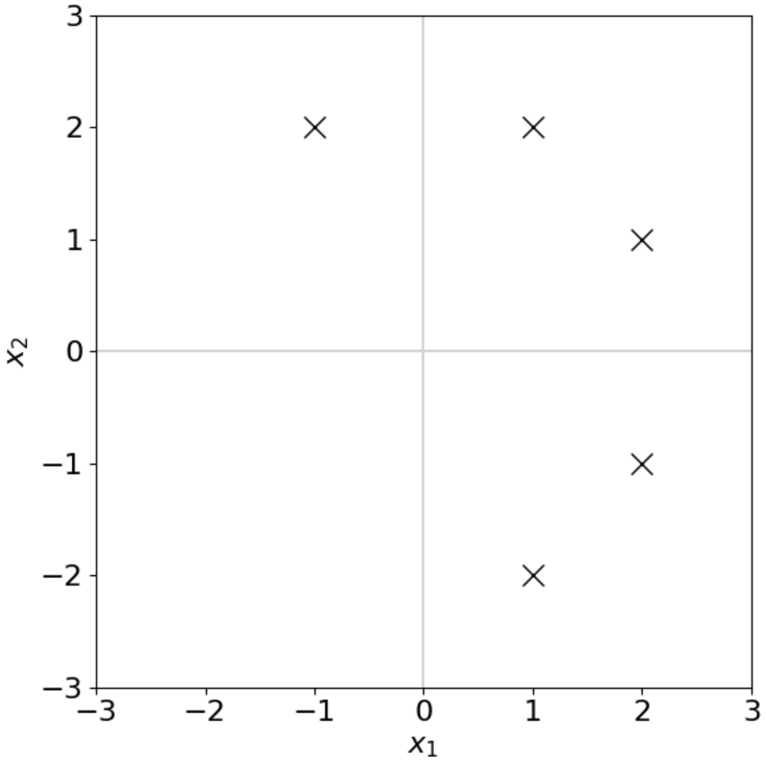
$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})} \qquad p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

D H

Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models





Generative Model: Supervised

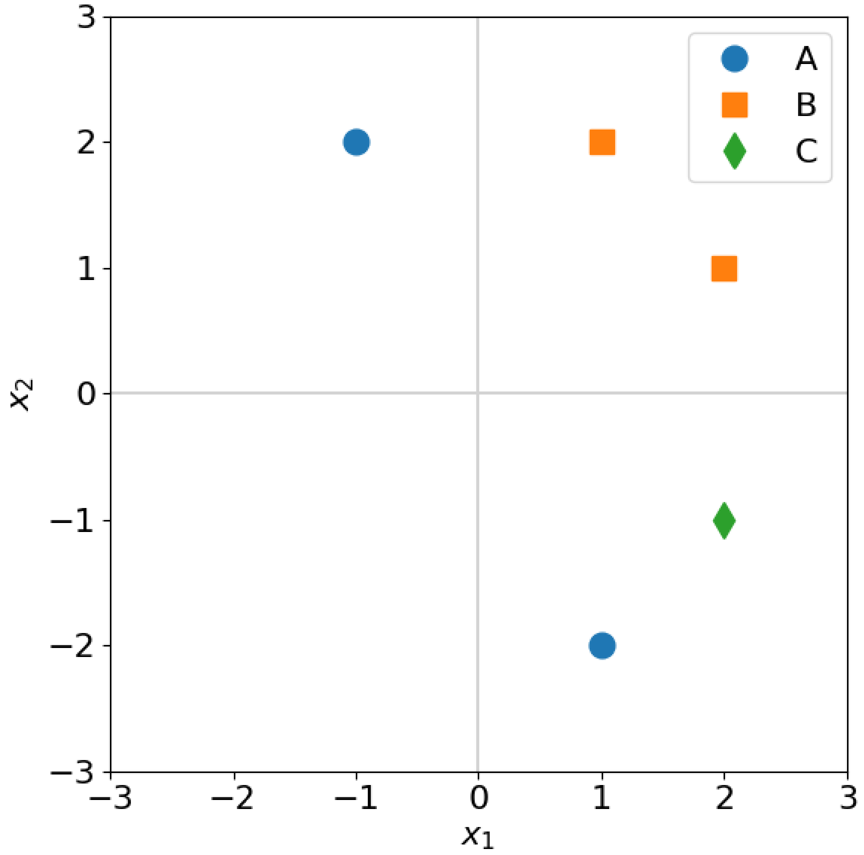
MLE: Discriminant analysis

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \theta\right)$$

$$Y \sim Categorical(\pi_1, \pi_2, \pi_3)$$

$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

$$\mathcal{D} = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{N}$$



Generative Model: Supervised

MLE: Discriminant analysis

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} p\left(\mathbf{x}^{(i)}, y_k^{(i)} = 1 \mid \theta\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \left(p\left(y_k^{(i)} = 1\right)\ p\left(\mathbf{x}^{(i)} \mid y_k^{(i)} = 1\right)\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \left(\pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)}\right)^{y_k^{(i)}}$$

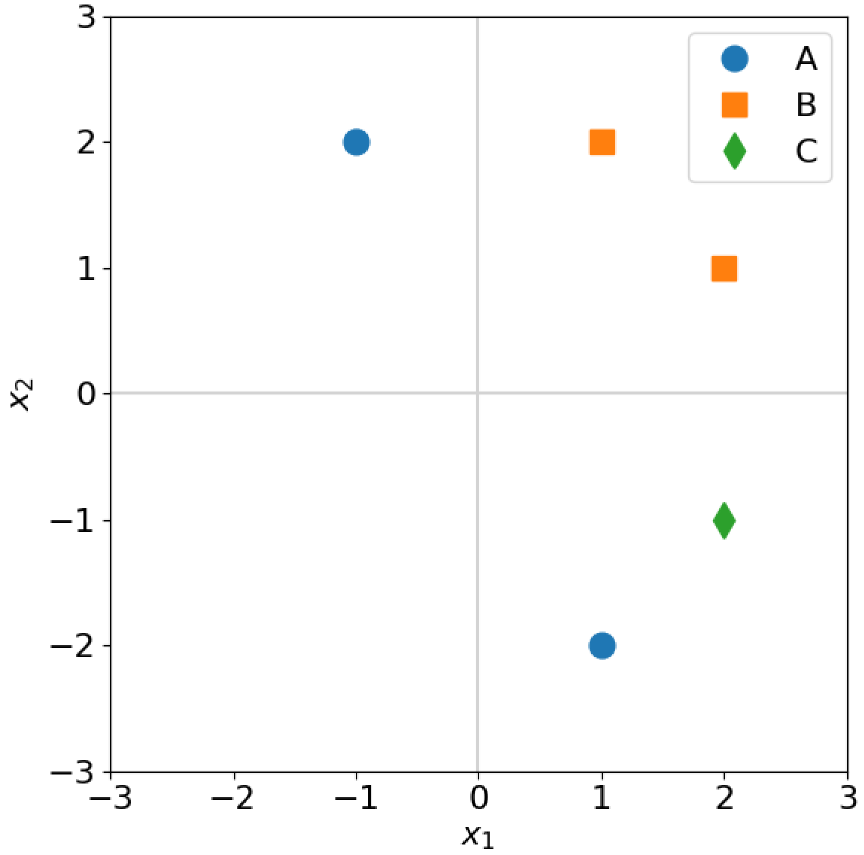
$$= \underset{\theta}{\operatorname{argmax}} \log \prod_{i}^{N} \prod_{k}^{K} \left(\pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)}\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i}^{N} \sum_{k}^{k} y_k^{(i)} \log \left(\pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)}\right)$$

$$Y \sim Categorical(\pi_1, \pi_2, \pi_3)$$

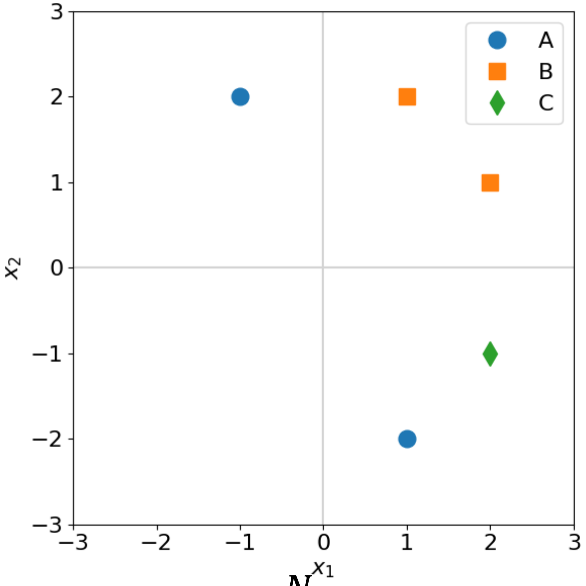
$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

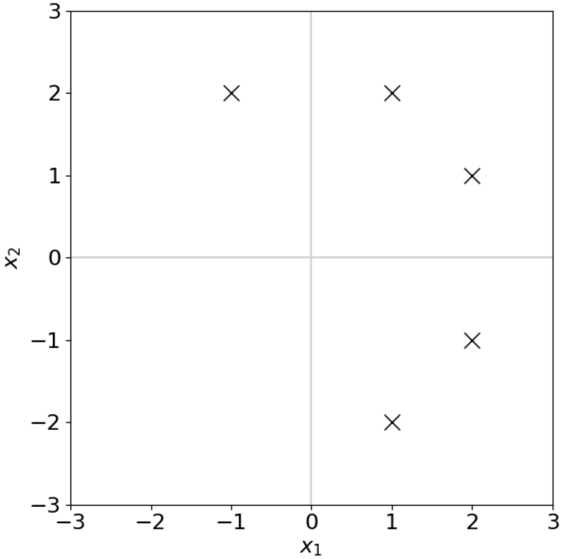
$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$$



Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models





$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} p\left(\mathbf{x}^{(i)}, y_k^{(i)} = 1 \mid \theta\right)^{y_k^{(i)}}$$

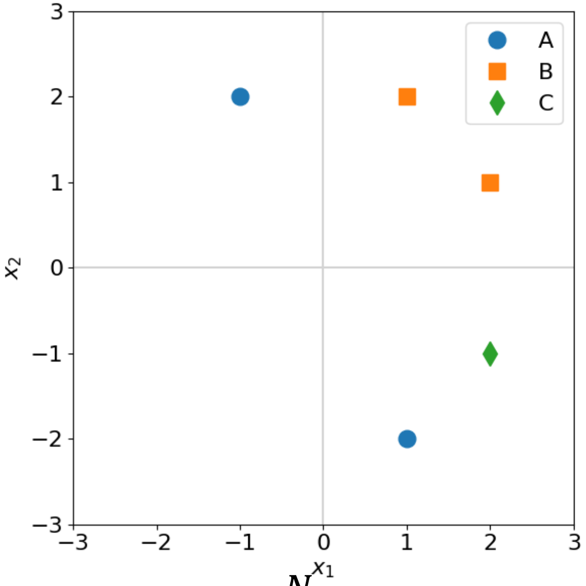
$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \left(p\left(y_k^{(i)} = 1\right) \; p\left(\mathbf{x}^{(i)} \mid y_k^{(i)} = 1\right)\right)^{y_k^{(i)}}$$

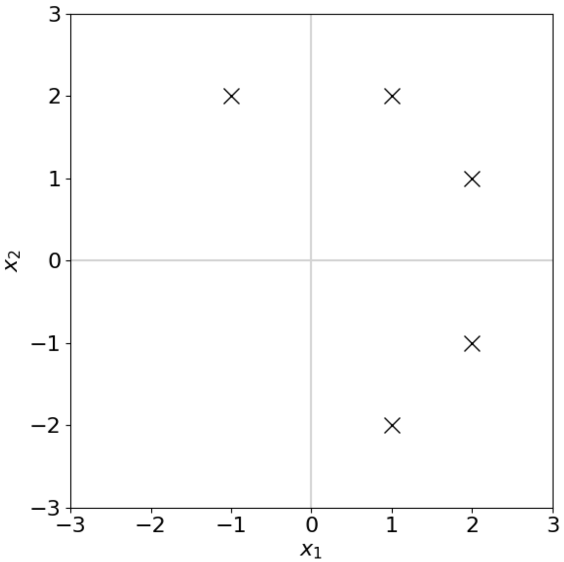
$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N}$$

Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models





$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} p\left(\mathbf{x}^{(i)}, y_k^{(i)} = 1 \mid \theta\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \left(p\left(y_k^{(i)} = 1\right) \; p\left(\mathbf{x}^{(i)} \mid y_k^{(i)} = 1\right)\right)^{y_k^{(i)}}$$

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} p\left(\mathbf{x}^{(i)}, z_k^{(i)} = 1 \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} p\left(z_k^{(i)} = 1\right) \; p\left(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1\right)$$

Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} p\left(\mathbf{x}^{(i)}, y_k^{(i)} = 1 \mid \theta\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \left(p\left(y_k^{(i)} = 1\right)\ p\left(\mathbf{x}^{(i)} \mid y_k^{(i)} = 1\right)\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \left(\pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \log \prod_{i}^{N} \prod_{k}^{K} \left(\pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}\right)^{y_k^{(i)}}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i}^{N} \sum_{k}^{k} y_k^{(i)} \log\left(\pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}\right)$$

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} p\left(\mathbf{x}^{(i)}, z_k^{(i)} = 1 \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} p\left(z_k^{(i)} = 1\right)\ p\left(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} \pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}$$

$$= \underset{\theta}{\operatorname{argmax}} \log \prod_{i}^{N} \sum_{k=1}^{K} \pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i}^{N} \log \sum_{k=1}^{K} \pi_k\ \ |\Sigma_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}$$

Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

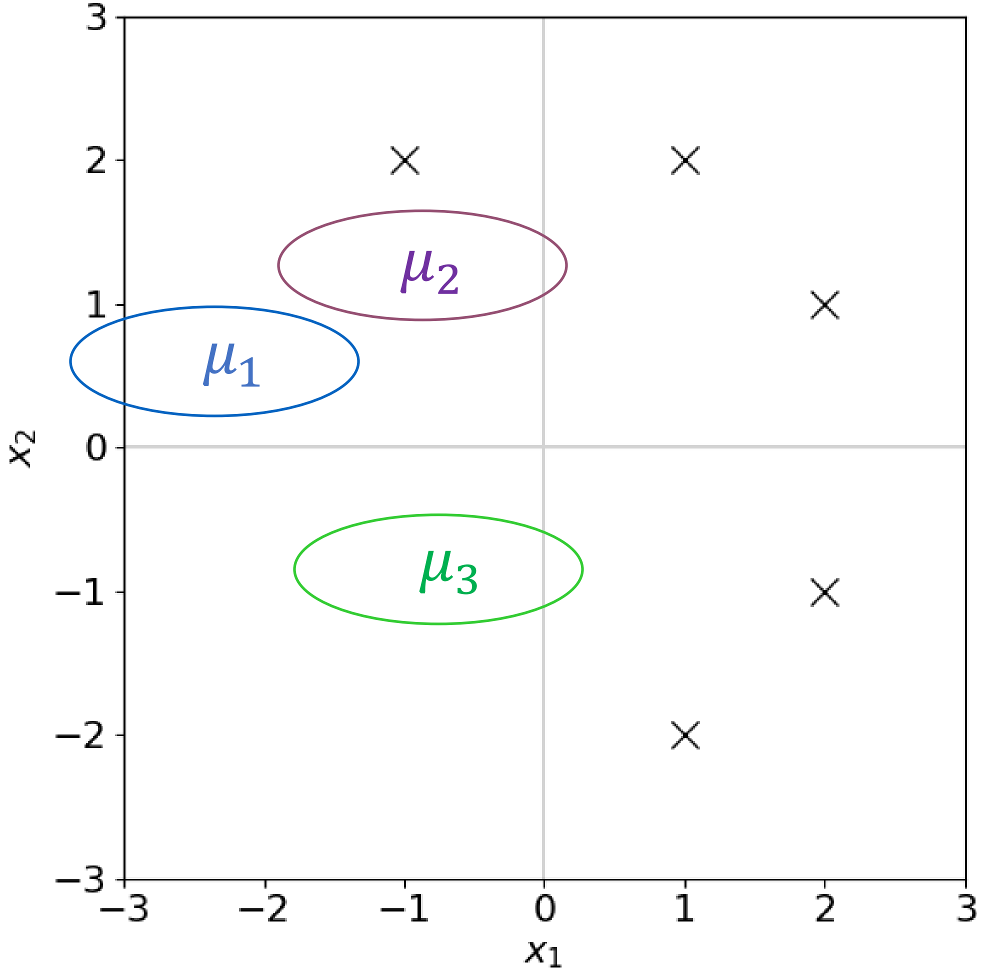
(for simplicity: fixed covariance, Σ, across all three Gaussians)

$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid z_k = 1) p(z_k = 1)$$

Mixture component

Mixture proportion



Gaussian Mixture Model

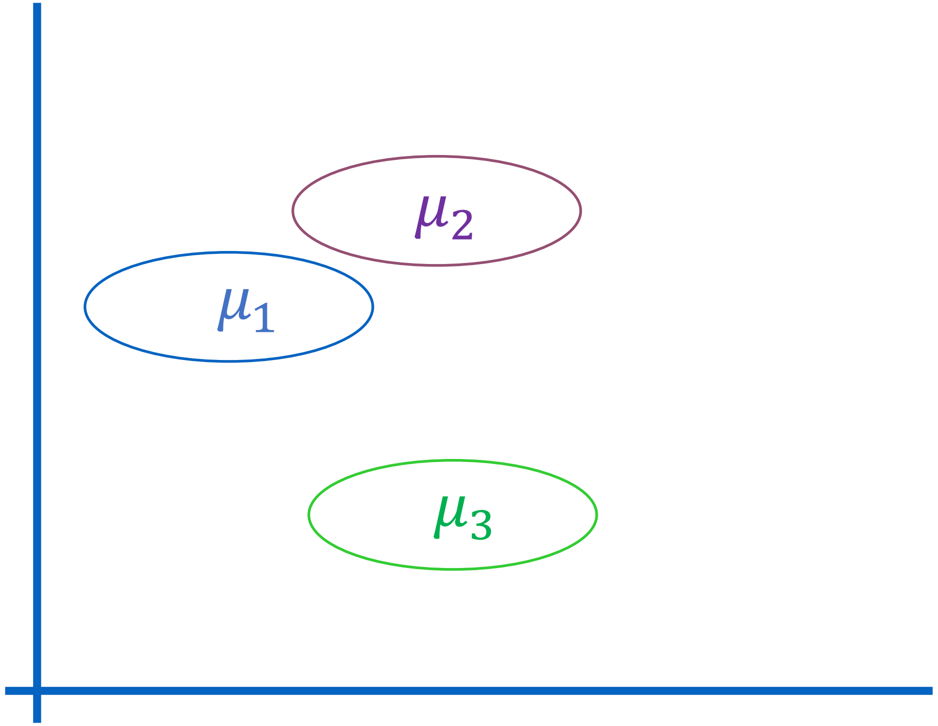
Mixture of K Gaussian distributions (multi-modal distribution)

(for simplicity: fixed covariance, Σ, across all three Gaussians)

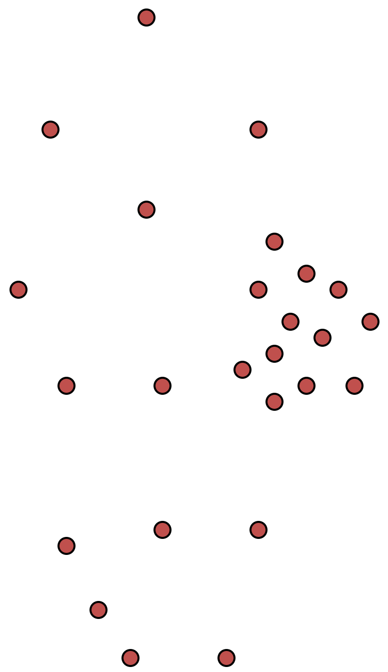
$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid z_k = 1) p(z_k = 1)$$

Mixture component — Mixture proportion



(One) bad case for K-means



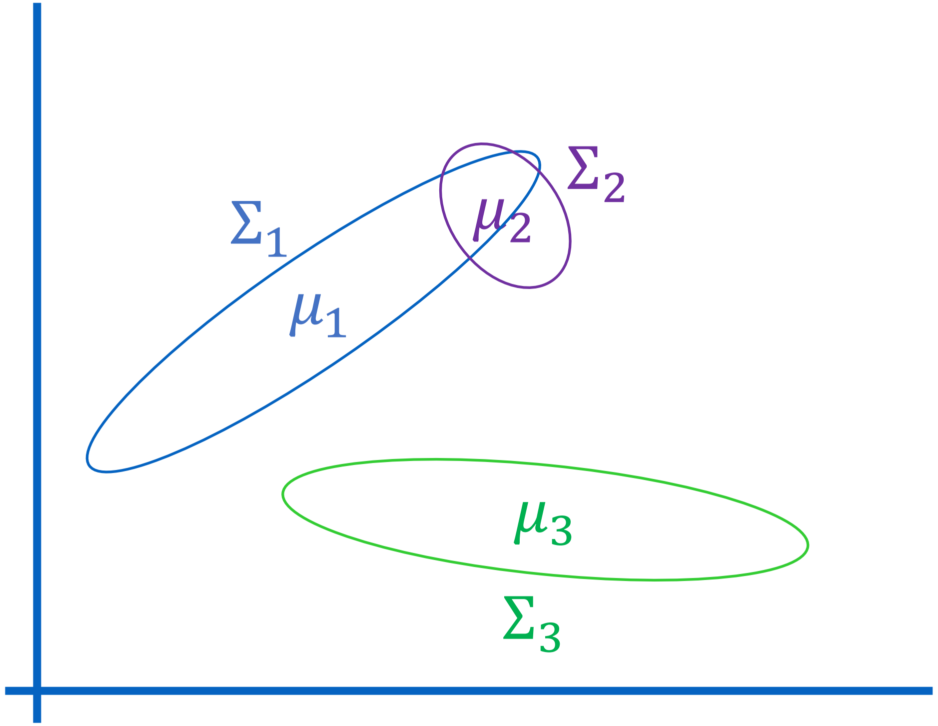
- Clusters may overlap
- Some clusters may be “wider” than others
- Clusters may not be linearly separable

Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid z_k = 1)p(z_k = 1)$$



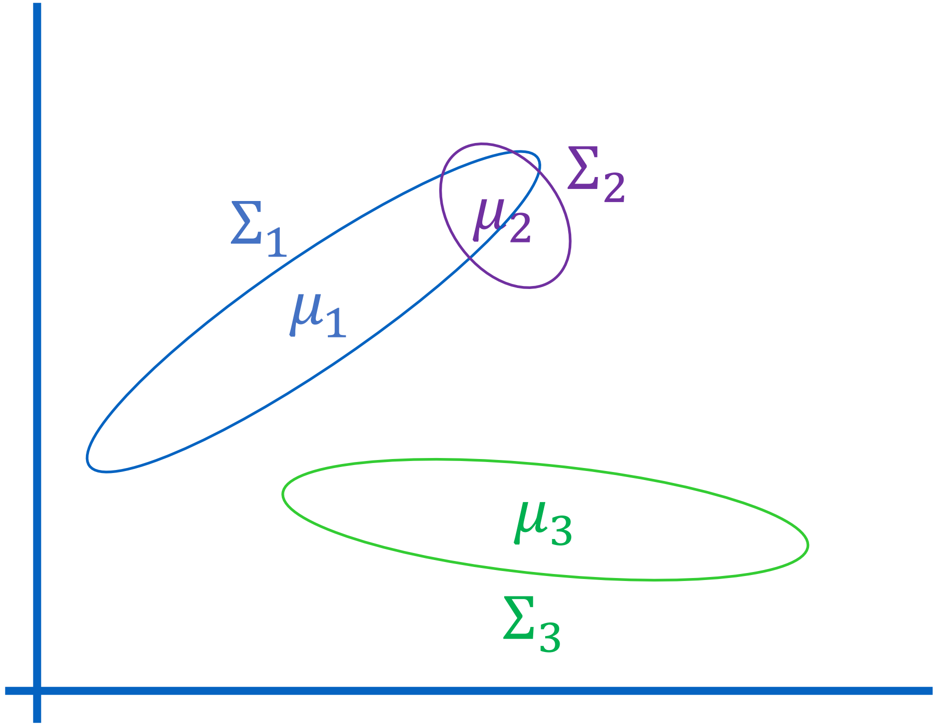
Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

- There are K components
- Component k generates data from a Gaussian with mean vector μ_k and covariance matrix Σ_k

Each data point is generated according to the follow recipe:

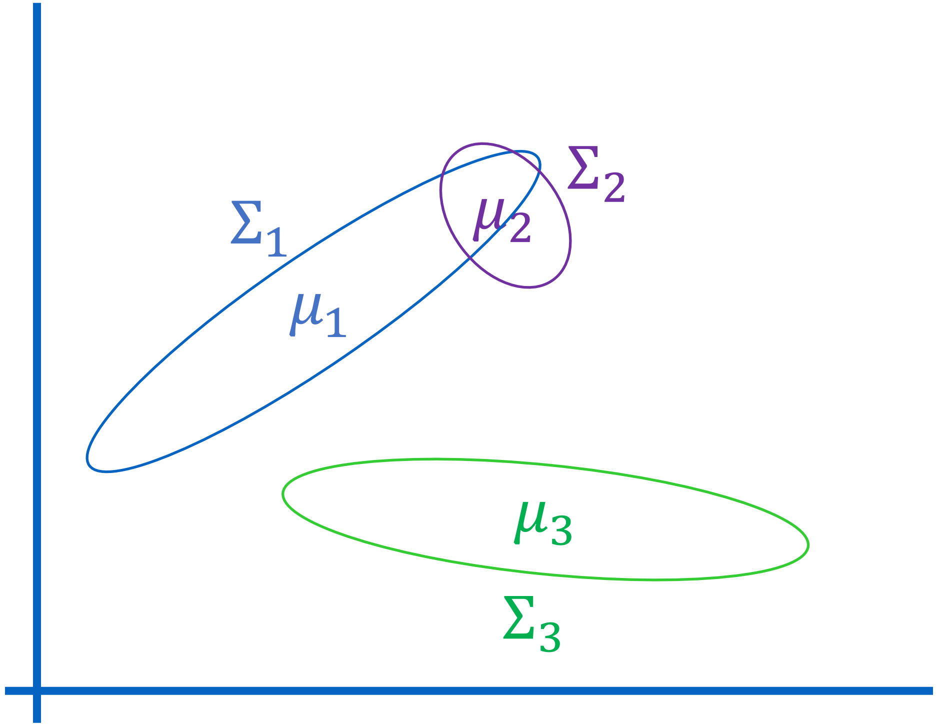
1) Pick a component at random: Choose component k with probability $p(z_k = 1)$
2) Data point $\mathbf{x} \sim \mathcal{N}(\mu_k, \Sigma_k)$



Learning General GMM

Mixture of K Gaussian distributions (multi-modal distribution)

$$x_1, \dots, x_M \sim p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid z_k = 1) p(z_k = 1)$$



Mixture: $\pi_k \overset{\text{def}}{=} p(z_k = 1)$

Gaussian components:

$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

Parameters:

$$\theta \overset{\text{def}}{=} \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$$

How to estimate parameters? Can we do MLE even without labels $\mathbf{z}$?

Learning General GMM

Maximize marginal likelihood:

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} p\left(\mathbf{x}^{(i)}, z_k^{(i)} = 1 \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} p\left(z_k^{(i)} = 1\right) \; p\left(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} \pi_k \;\; |\Sigma_k|^{-\frac{1}{2}} \, e^{-\frac{1}{2}\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k\right)^T \Sigma_k^{-1} \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k\right)}$$

Learning General GMM

Maximize marginal likelihood:

$$\underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i}^{N} \sum_{k=1}^{K} \pi_k \quad |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k\right)}$$

How do we find the π_k, μ_k, Σ_k which give the max. marginal likelihood?

a) Set $\frac{\partial}{\partial \mu_k} \ell(\theta; \mathcal{D}) = 0$ and solve for μ_k, etc. ? — No closed-form solution

b) Use gradient descent? — Doable, but complicated, often slow, and need to consider constraints on parameters

Log (Marginal) Likelihood for Missing Data

Marginalize over missing data, $\mathbf{z}^{(i)}$

$$\ell(\theta \mid \mathcal{D}) = \log \prod_{i}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

GMM vs K-means

Maximize marginal likelihood:

$$\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} \sum_{k=1}^{K} p\left(z^{(i)} = k\right) \; p\left(\mathbf{x}^{(i)} \mid z^{(i)} = k\right)$$

What happens if we assume a **hard-assignment**?

$p\left(z^{(i)} = k\right) = 1$ if point i belongs to the k-th cluster ← $p(z \mid x)$

(and assume variances are all the same) ←

$$\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} \sum_{k=1}^{K} p\left(z^{(i)} = k\right) \; \underline{p\left(\mathbf{x}^{(i)} \mid z^{(i)} = k\right)}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} e^{-\frac{1}{2}\left\|x^{(i)} - \mu_{z^{(i)}}\right\|_2^2}$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left\|x^{(i)} - \mu_{z^{(i)}}\right\|_2^2$$

Same as K-means!

K-means Optimization

Alternating minimization

a) $$\boldsymbol{z} = \underset{\boldsymbol{z}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{z^{(i)}} \right\|_2^2$$

b) $$\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K = \underset{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{z^{(i)}} \right\|_2^2$$

Expectation-Maximization for GMM

Log Likelihood vs Complete Log Likelihood

Log likelihood $\mathcal{D} = \{\mathbf{x}^{(i)}\}$

$$\ell(\theta \mid \mathcal{D}) = \log \prod_i^N p\left(\mathbf{x}^{(i)} \mid \theta\right)$$

Complete Log likelihood $\mathcal{D}_c = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\}$

$$\ell_c(\theta \mid \mathcal{D}_c) = \log \prod_i^N p\left(\mathbf{x}^{(i)}, \mathbf{z}^{(i)} \mid \theta\right)$$

Expected Value of Complete Log Likelihood

Replace know value of z with

$E_{Z|X,\theta}[\ell_c(\theta \mid \mathcal{D}_c)]$

Complete Log likelihood $\mathcal{D}_c = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\}$

$\ell_c(\theta \mid \mathcal{D}_c) = \log \prod_i^N p\left(\mathbf{x}^{(i)}, \mathbf{z}^{(i)} \mid \theta\right)$

Notes on EM

- EM is an optimization strategy for objective functions that can be interpreted as likelihoods in the presence of missing data.

- It is much simpler than gradient methods:
  - No need to choose step size.
  - Enforces constraints.
  - Calls inference and fully observed learning as subroutines.

- EM is an Iterative algorithm with two linked steps:
  - E-step: fill-in hidden values using inference, $p(z|x, \theta^t)$.
  - M-step: update parameters t+1 using standard MLE/MAP method applied to completed data

- This procedure monotonically improves (or leaves it unchanged). Thus, it always converges to a local optimum of the likelihood.

EM for GMMS

Initialize parameters

For $t = 0$, $\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \Sigma_k^{(0)}$

E-step

For a fixed set of Gaussian mixture model parameters, $\theta^{(t)}$, update the probability that each point, $\boldsymbol{x}^{(i)}$, belongs to cluster k, $p\left(z_k^{(i)} = 1 \mid \boldsymbol{x}^{(i)}, \theta^{(t)} \right)$

M-step

For a fixed $p\left(z_k^{(i)} = 1 \mid \boldsymbol{x}^{(i)}, \theta^{(t)} \right)$, update the estimate for each parameter, $\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \Sigma_k^{(t+1)}$

Iterate between E and M steps

EM for GMMS

Complete Log likelihood $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\}$

$$\ell_c(\theta \mid \mathcal{D}_c) = \log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)} \mid \theta)$$

E-step

$$E_{Z|X,\theta^{(t)}}\left[Z_k^{(i)}\right] = p\left(z_k^{(i)} = 1 \mid \boldsymbol{x}^{(i)}, \theta^{(t)}\right)$$

M-step

$$\left.\begin{matrix} \pi_k^{(t+1)} \\ \boldsymbol{\mu}_k^{(t+1)} \\ \Sigma_k^{(t+1)} \end{matrix}\right\} = \operatorname{argmax}_\theta E_{Z|X,\theta^{(t)}}[\ell_c(\theta \mid \mathcal{D}_c)]$$

EM for GMMS

E-step

$$p\left(z_k^{(i)}=1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \leftarrow \frac{\pi_k^{(t)} \mathcal{N}\left(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}\left(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}, \forall i, k$$

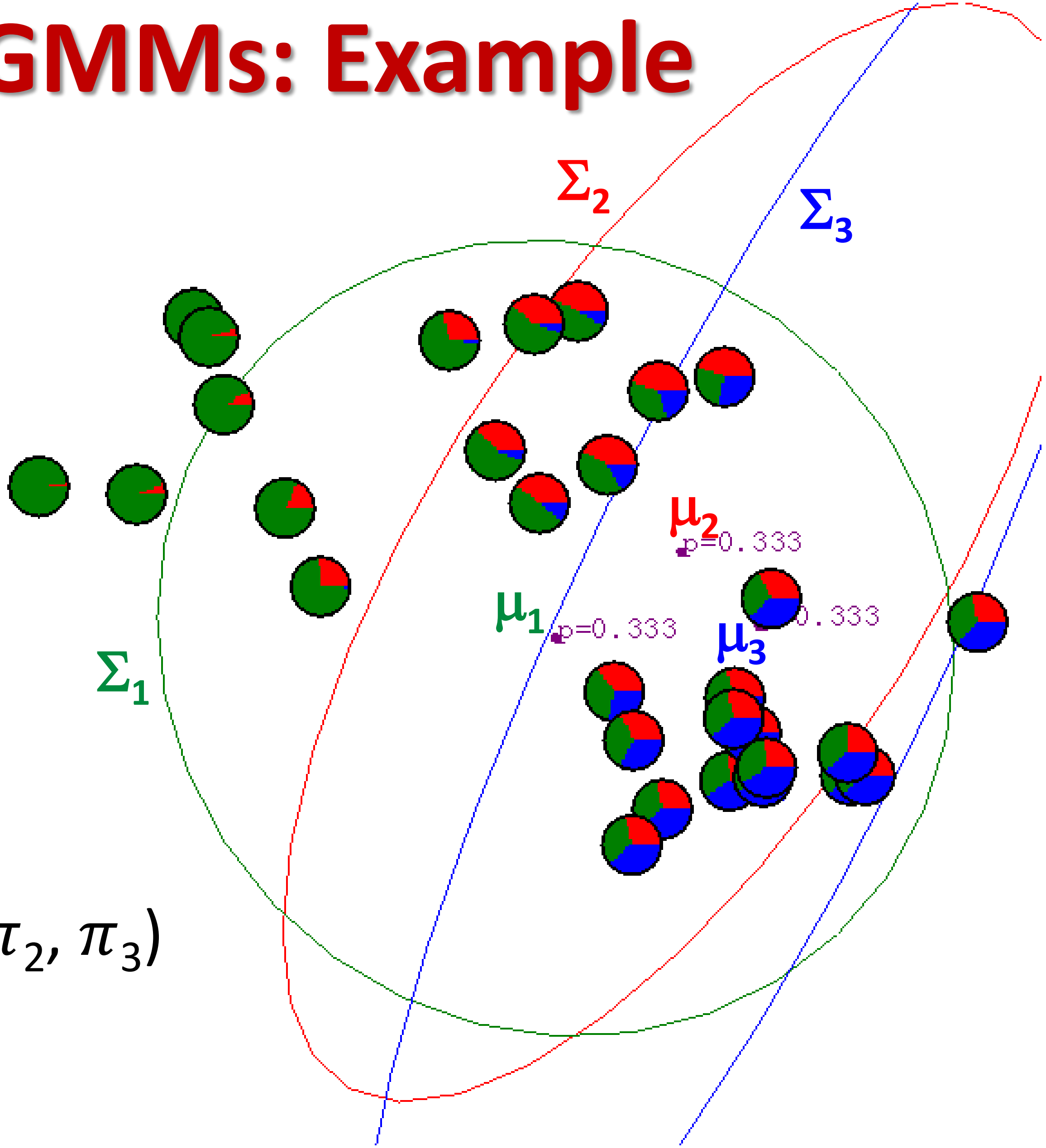
M-step

$$\pi_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^{N} P\left(z_k^{(i)}=1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}{N}, \forall k$$

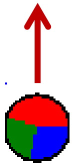
$$\boldsymbol{\mu}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^{N} P\left(z_k^{(i)}=1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \boldsymbol{x}^{(i)}}{\sum_{i=1}^{N} P\left(z_k^{(i)}=1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}, \forall k$$

$$\boldsymbol{\Sigma}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^{N} P\left(z_k^{(i)}=1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} P\left(z_k^{(i)}=1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}, \forall k$$

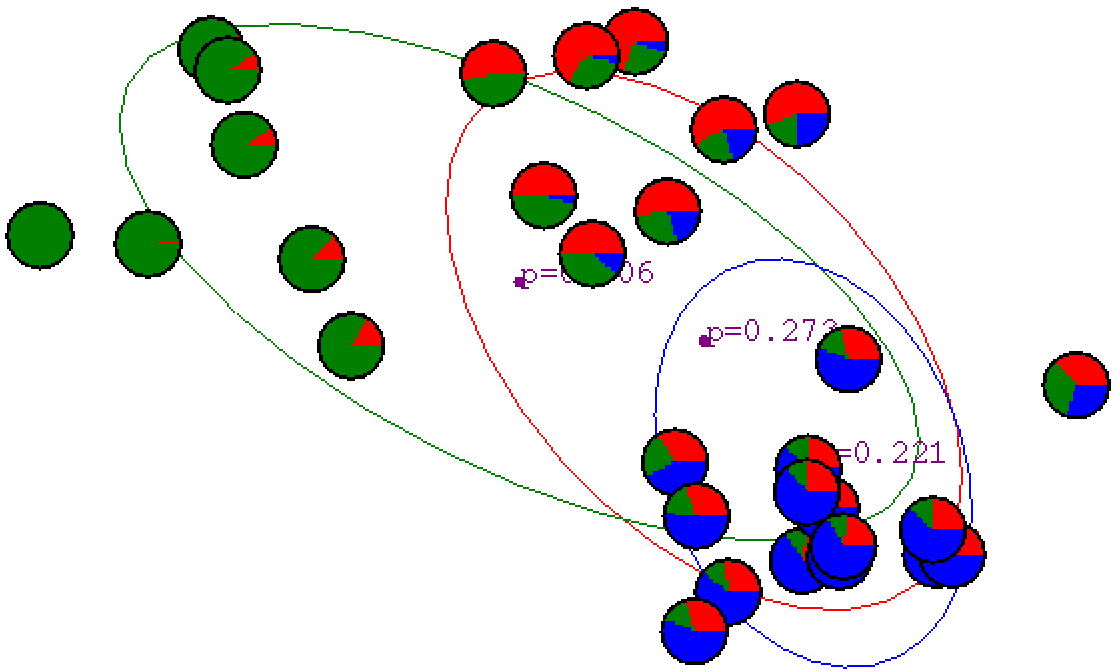
EM for GMMs: Example



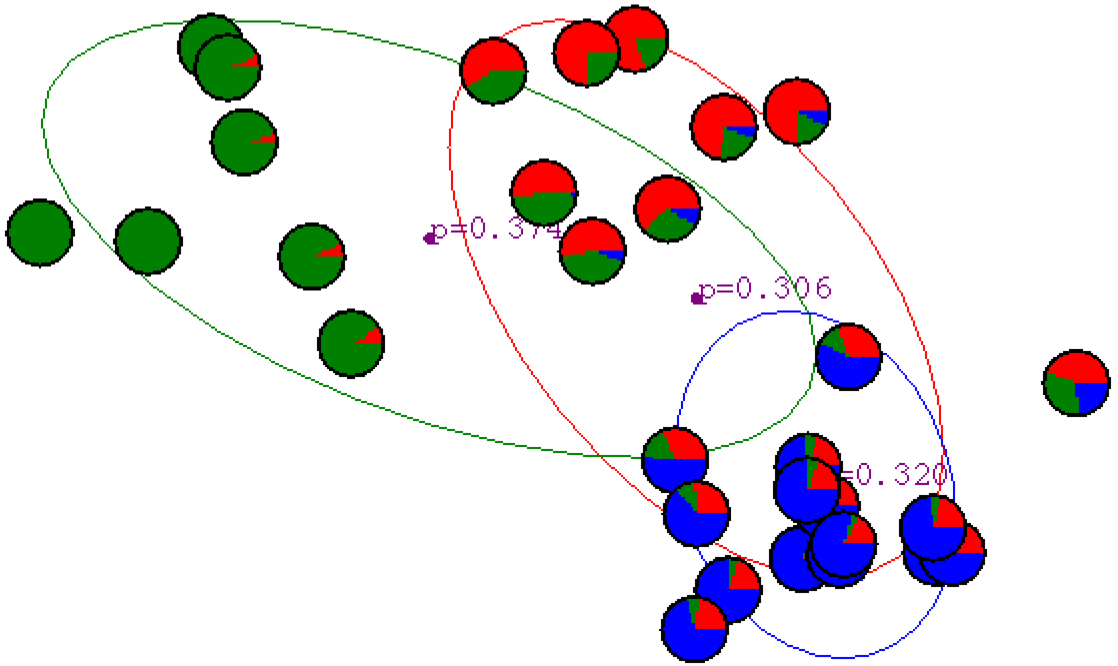
$P(z_2 = 1 \mid x_j, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3, \pi_1, \pi_2, \pi_3)$



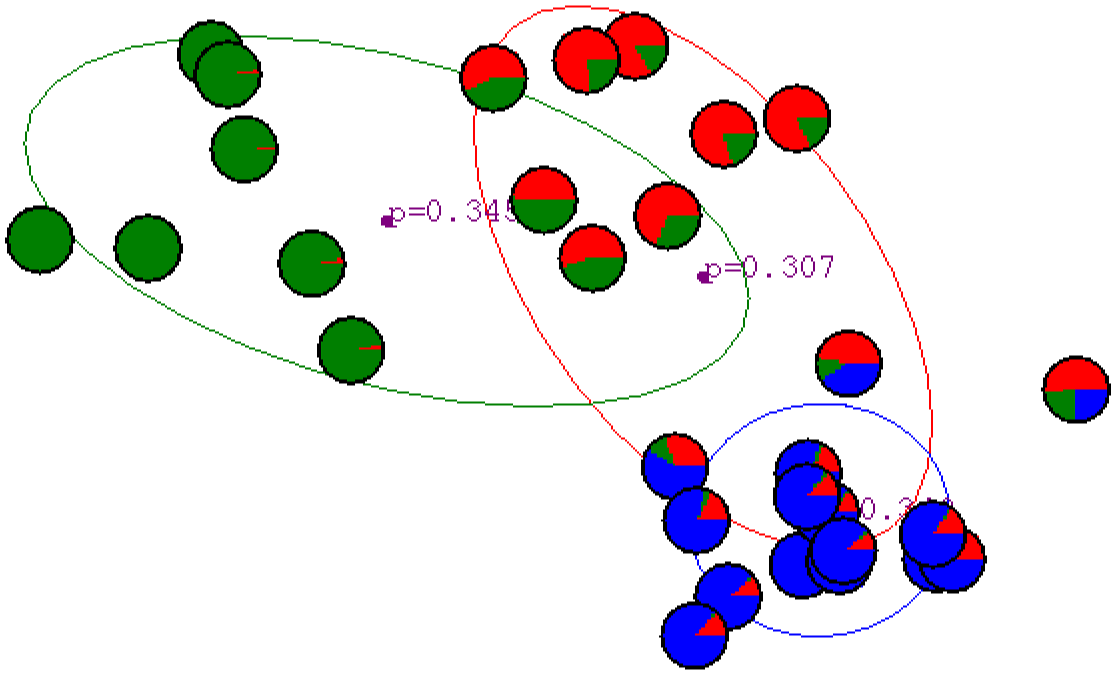
After 1st iteration



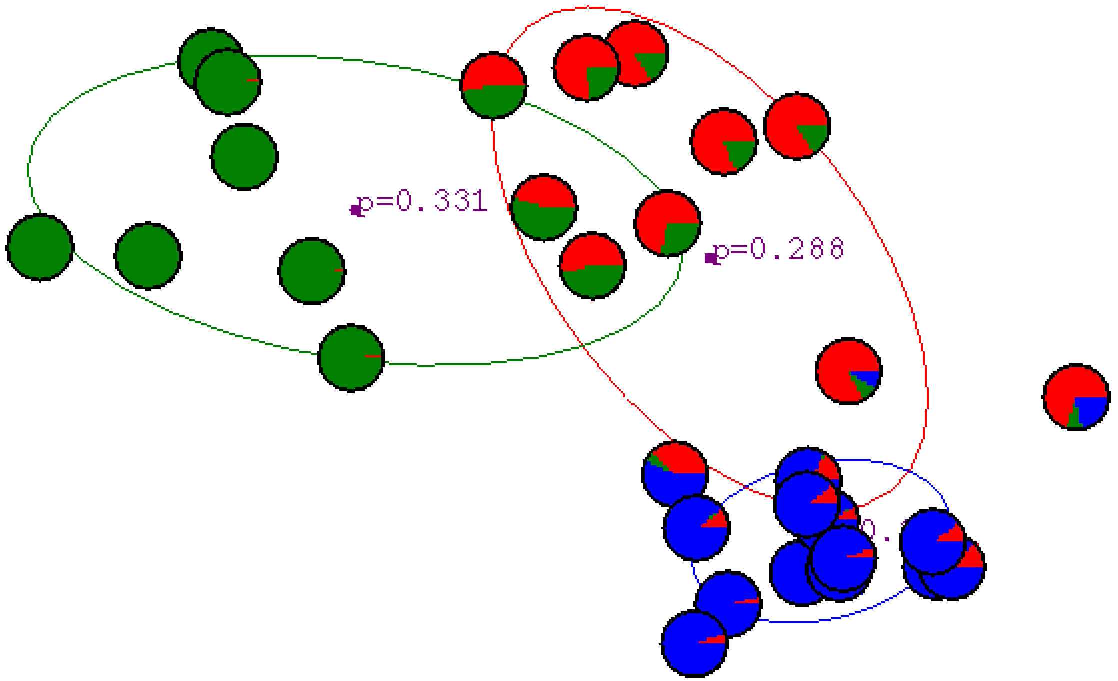
After 2nd iteration



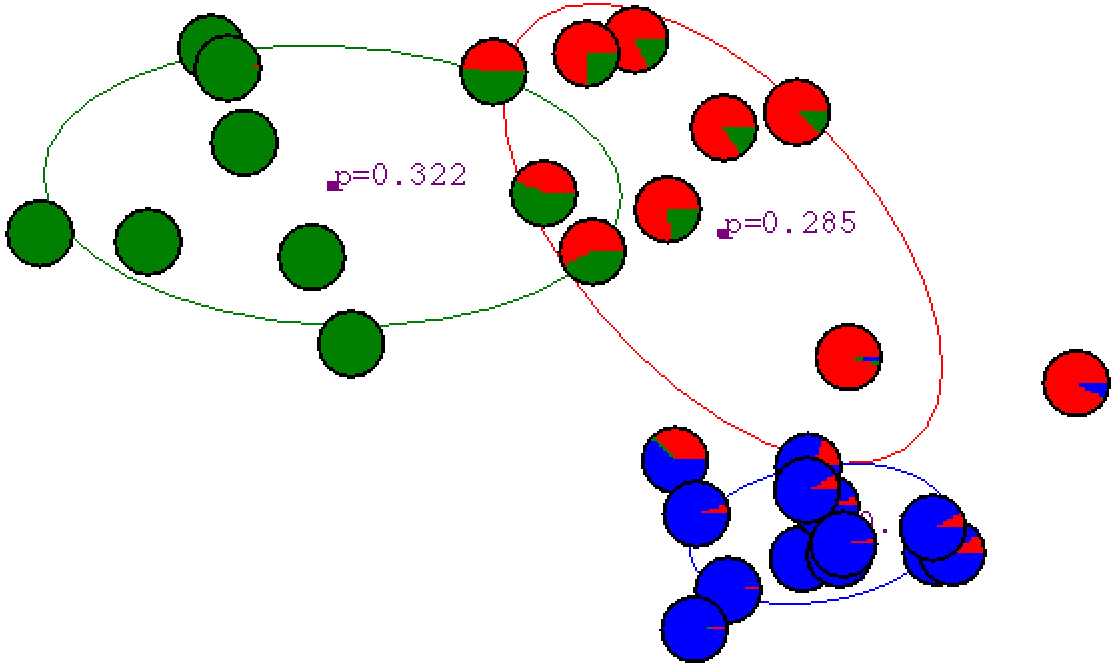
After 3rd iteration



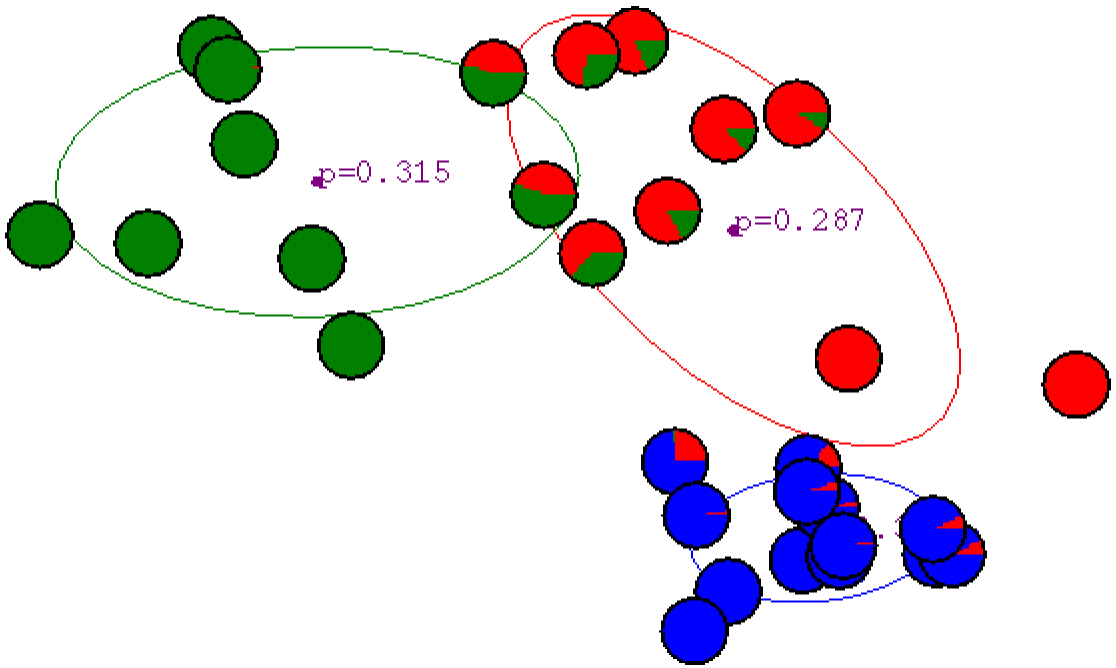
After 4^{th} iteration



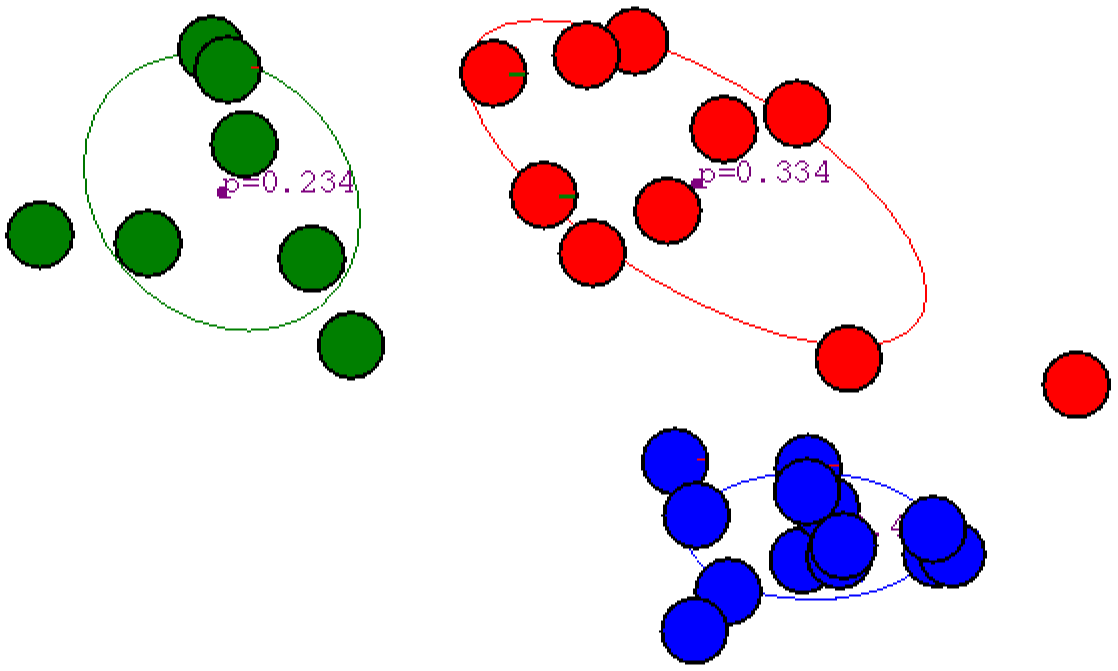
After 5^{th} iteration



After 6th iteration



After 20th iteration



Gaussian Mixture Model

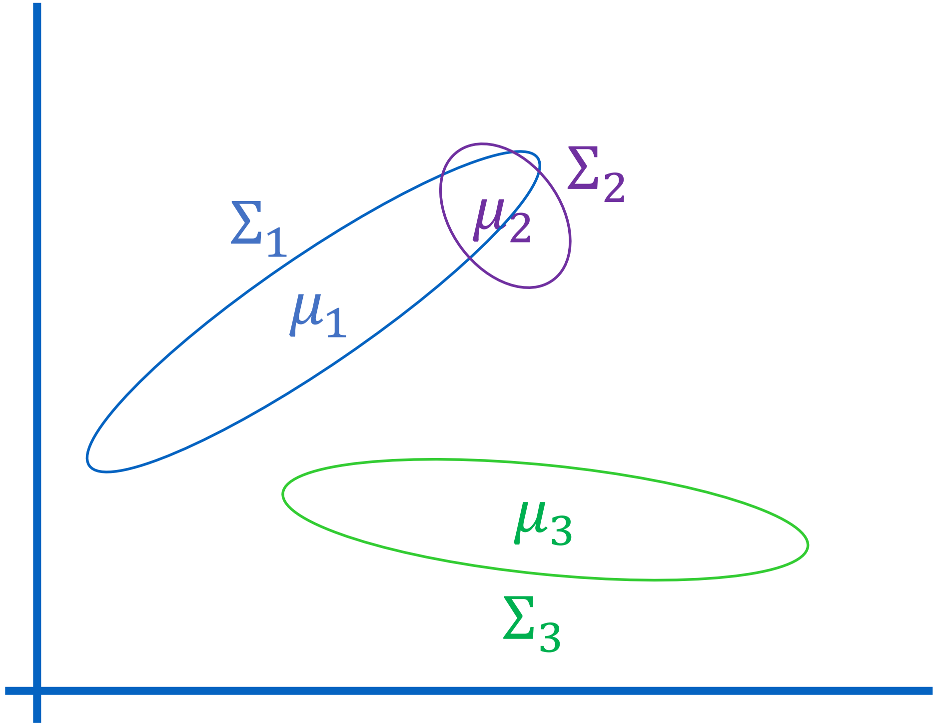
Mixture of K Gaussian distributions (multi-modal distribution)

$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_\mathrm{k})$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid z_k = 1)p(z_k = 1)$$

Mixture component

Mixture proportion



General EM Algorithm

Theory underlying EM

- What are we doing?

- Recall that according to MLE, we intend to learn the model parameter that would have maximize the likelihood of the data.

- But we do not observe z, so computing

$$\ell(\theta; D) = \log \sum_z p(x, z \mid \theta) = \log \sum_z p(z \mid \theta_z) p(x \mid z, \theta_x)$$

  is difficult!

- What shall we do?

Complete & Incomplete Log Likelihoods

- Complete log likelihood

  Let X denote the observable variable(s), and Z denote the latent variable(s).
  If Z could be observed, then

$$\ell_c(\theta; x, z) \overset{\text{def}}{=} \log p(x, z \mid \theta)$$

  - Usually, optimizing $\ell_c()$ given both z and x is straightforward (c.f. MLE for fully observed models).
  - Recalled that in this case the objective for, e.g., MLE, decomposes into a sum of factors, the parameter for each factor can be estimated separately.
  - But given that Z is not observed, $\ell_c()$ is a random quantity, cannot be maximized directly.

- Incomplete log likelihood

  With z unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; x) = \log p(x \mid \theta) = \log \sum_z p(x, z \mid \theta)$$

  - This objective won't decouple