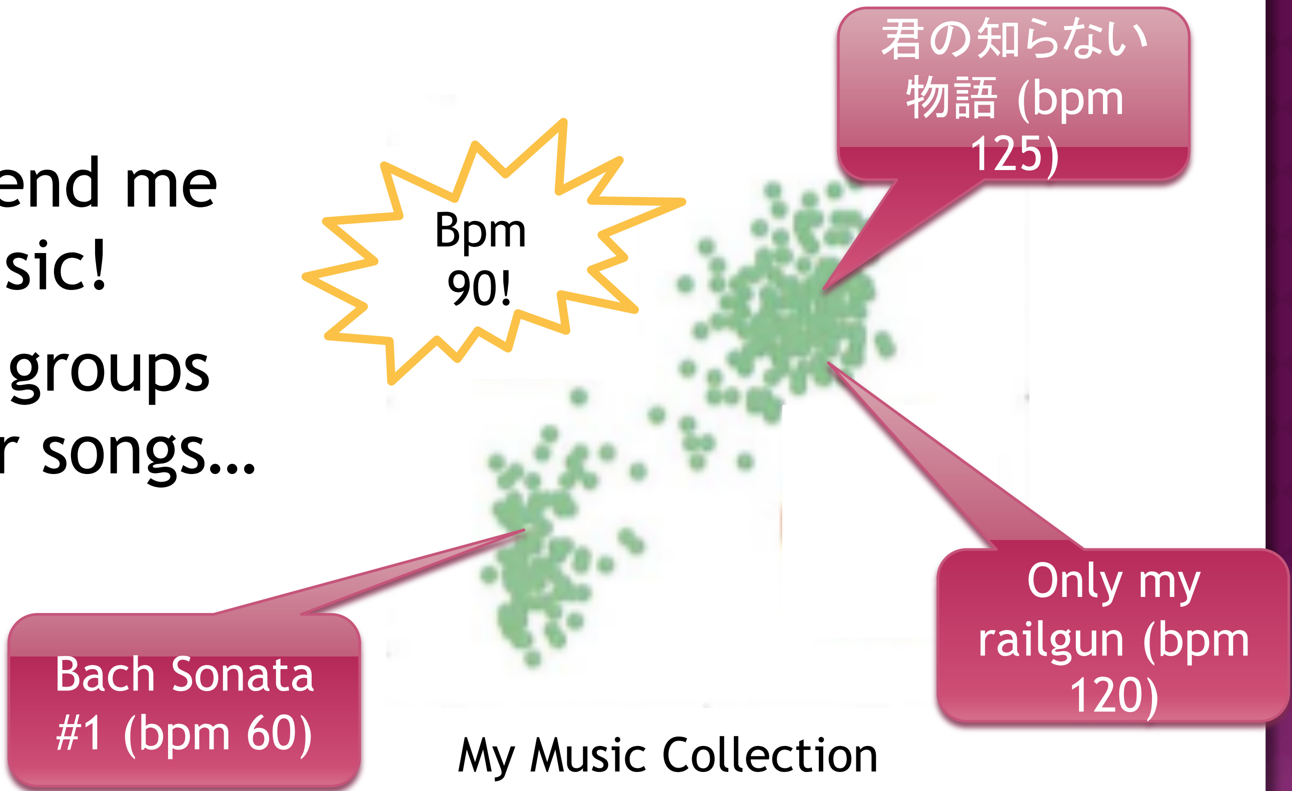
PRML: CHAPTER 9

Expectation Maximization and

Mixture of Gaussians

A CLUSTERING PROBLEM

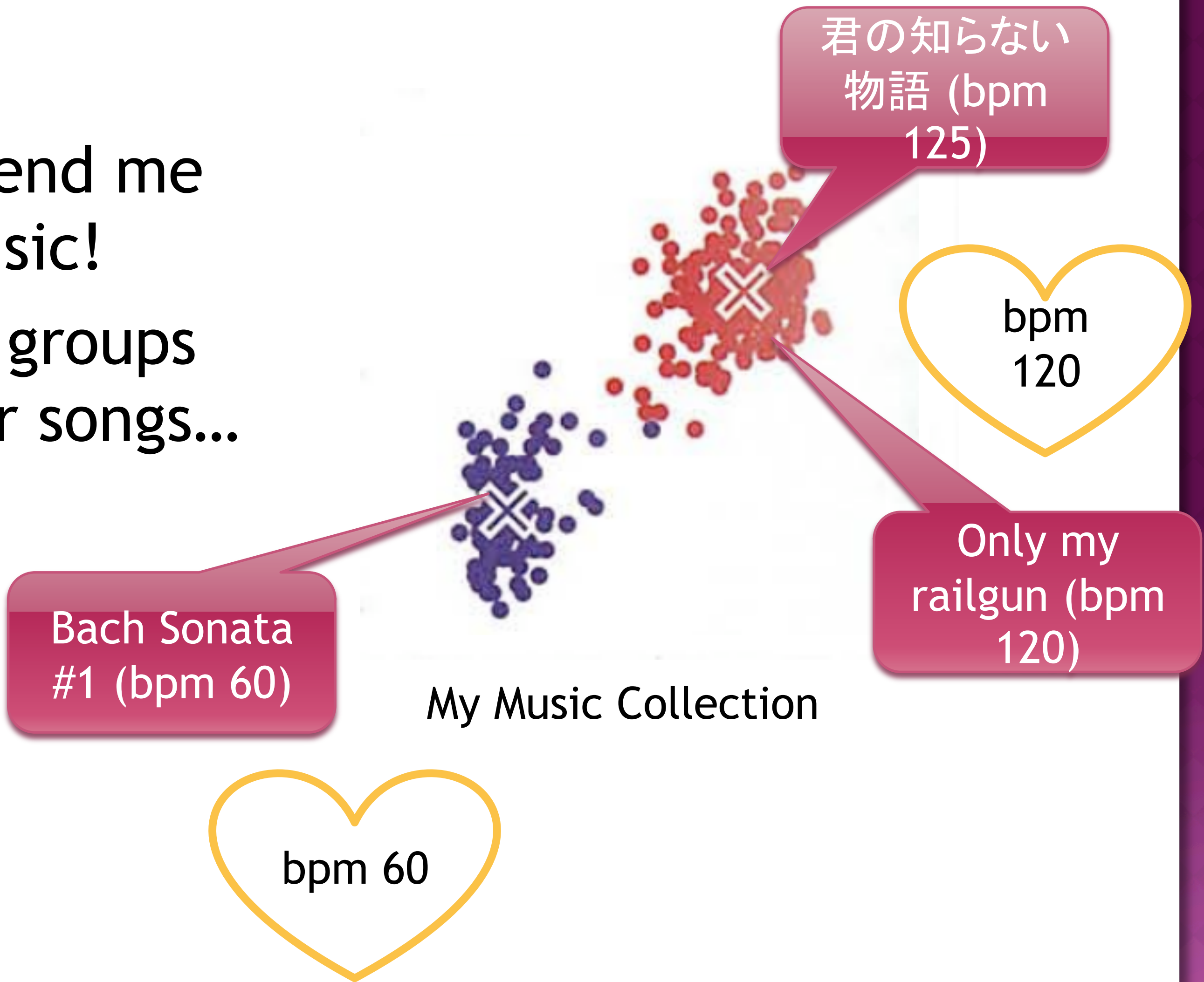
- Recommend me some music!
- Discover groups of similar songs…



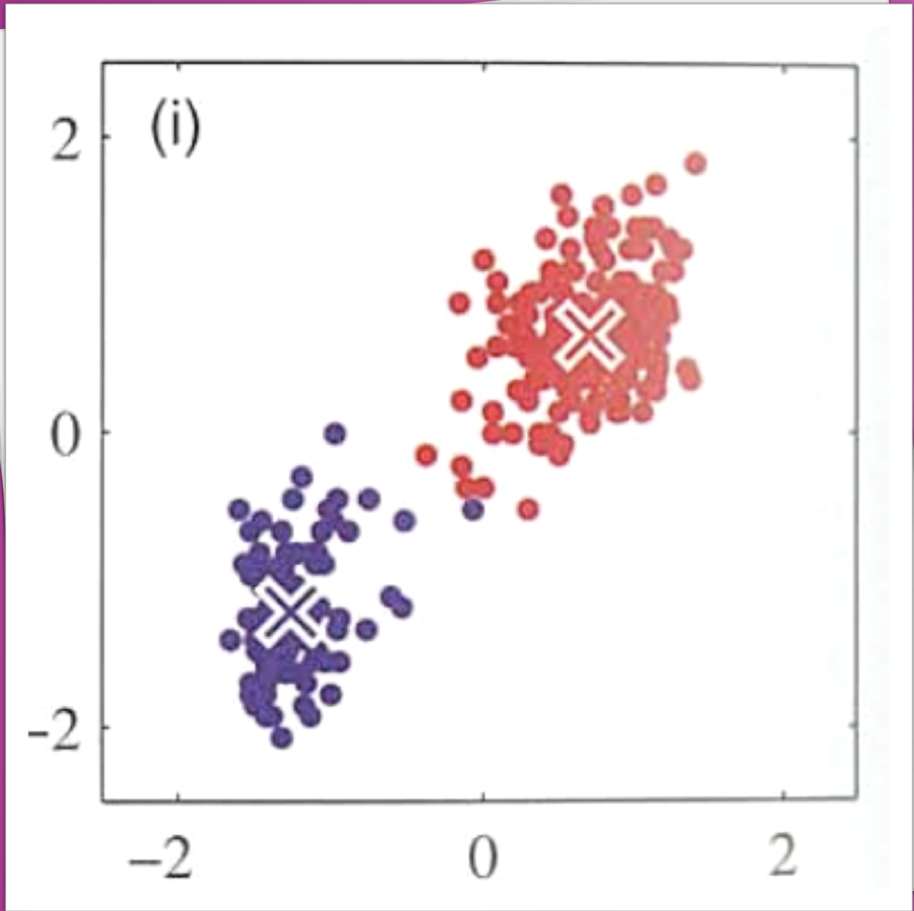
My Music Collection

A CLUSTERING PROBLEM

- Recommend me some music!
- Discover groups of similar songs…



My Music Collection



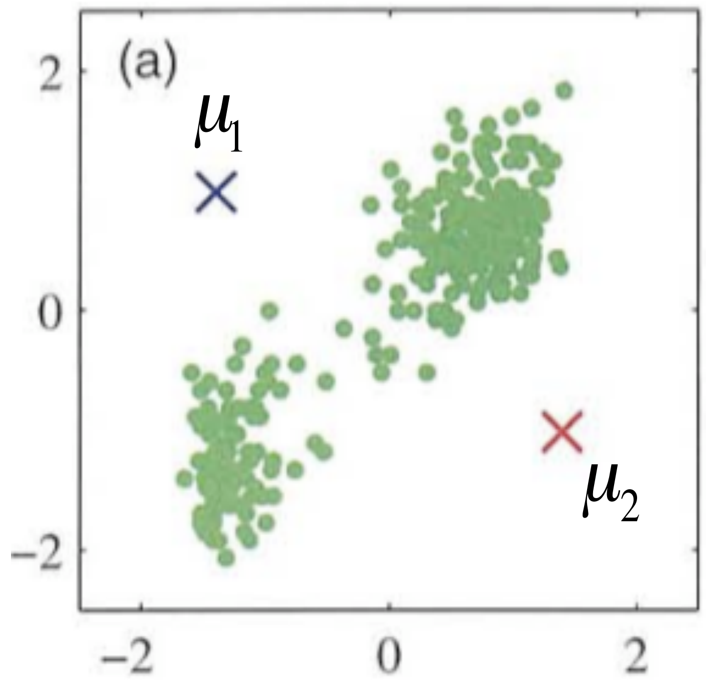
9.1 K-MEANS CLUSTERING

An unsupervised classifying method

K-MEANS ALGORITHM

1. Initialize K "means" μ_k , one for each class

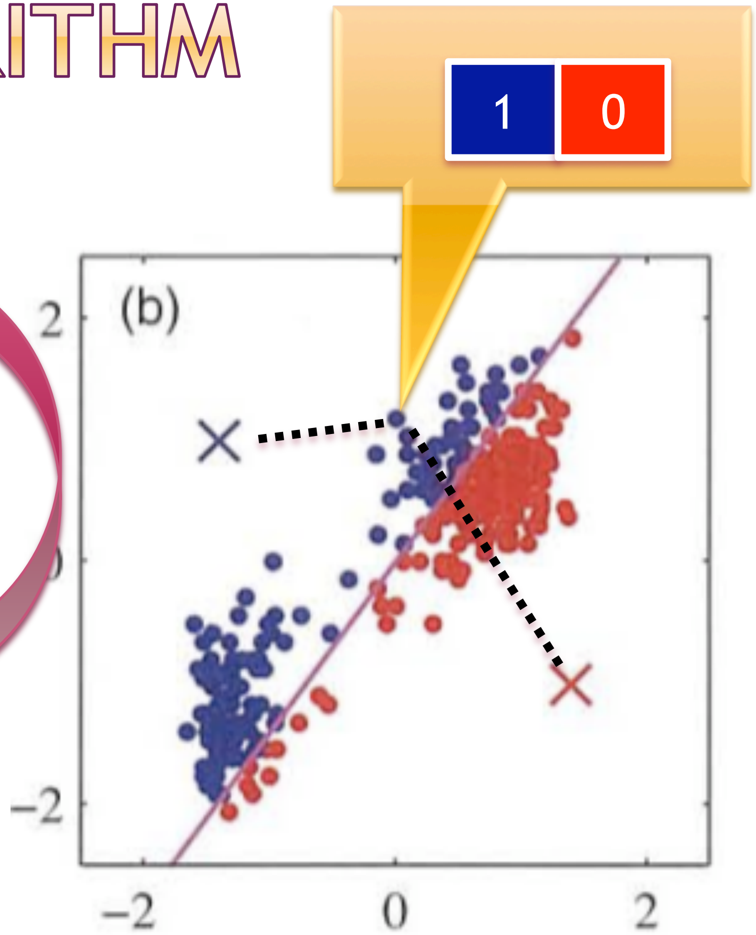
   - Eg. Use random starting points, or choose k random points from the set



K=2

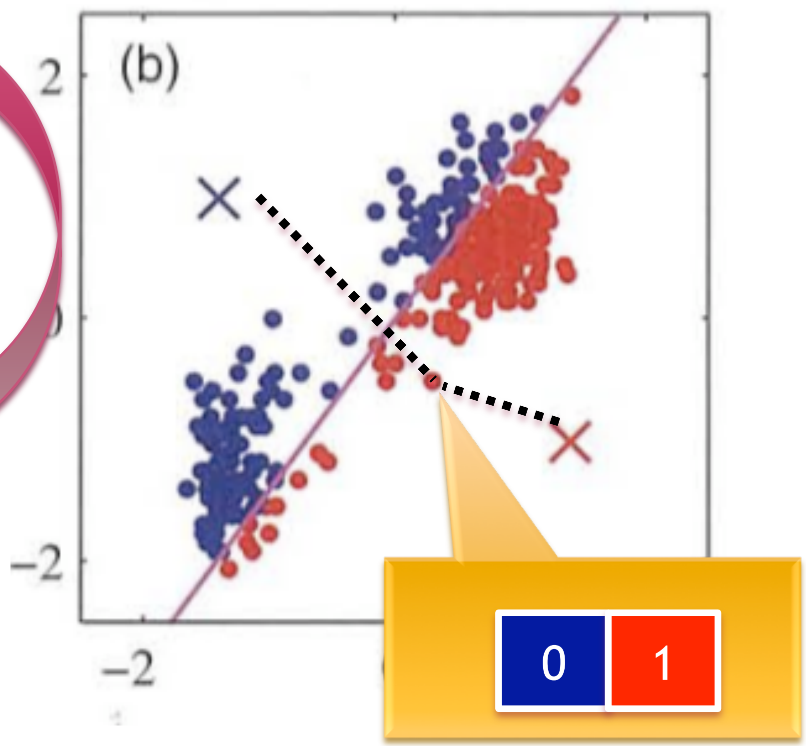
K-MEANS ALGORITHM

2. **Phase 1: Assign** each point to closest mean μ_k
3. **Phase 2: Update** means of the new clusters



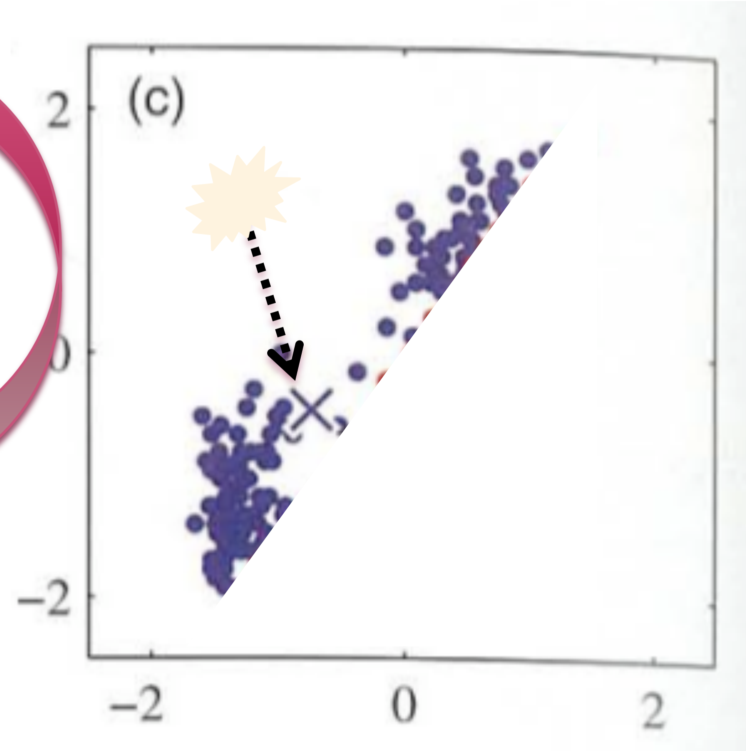
K-MEANS ALGORITHM

2. **Phase 1: Assign** each point to closest mean μ_k
3. **Phase 2:** Update means of the new clusters



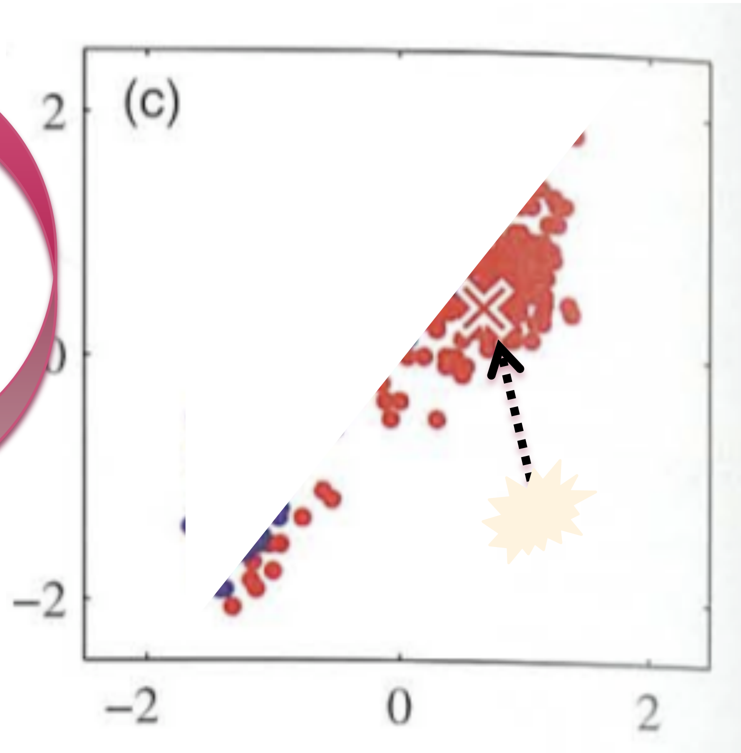
K-MEANS ALGORITHM

2. Phase 1: Assign each point to closest mean
3. **Phase 2:** Update means of the new clusters



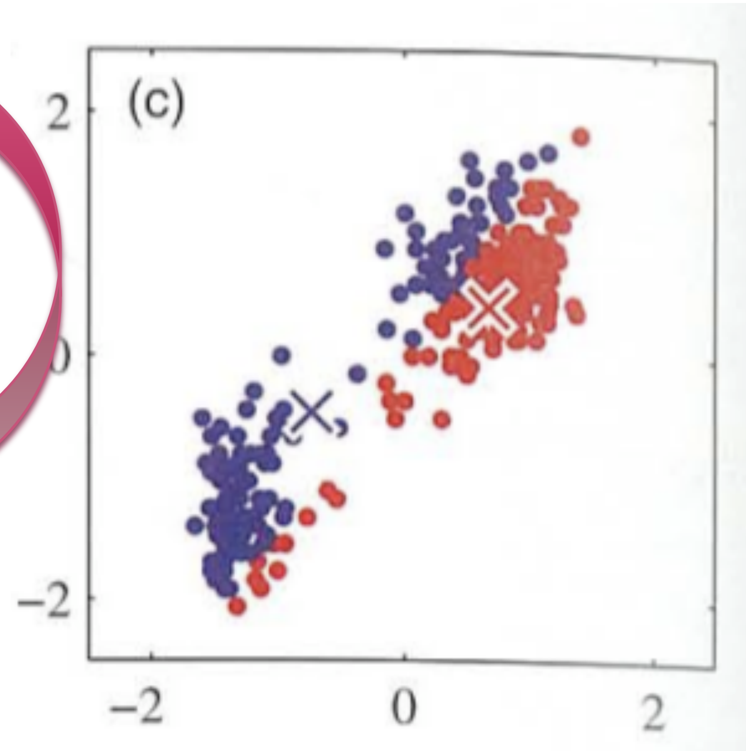
K-MEANS ALGORITHM

2. Phase 1: Assign each point to closest mean
3. **Phase 2:** Update means of the new clusters



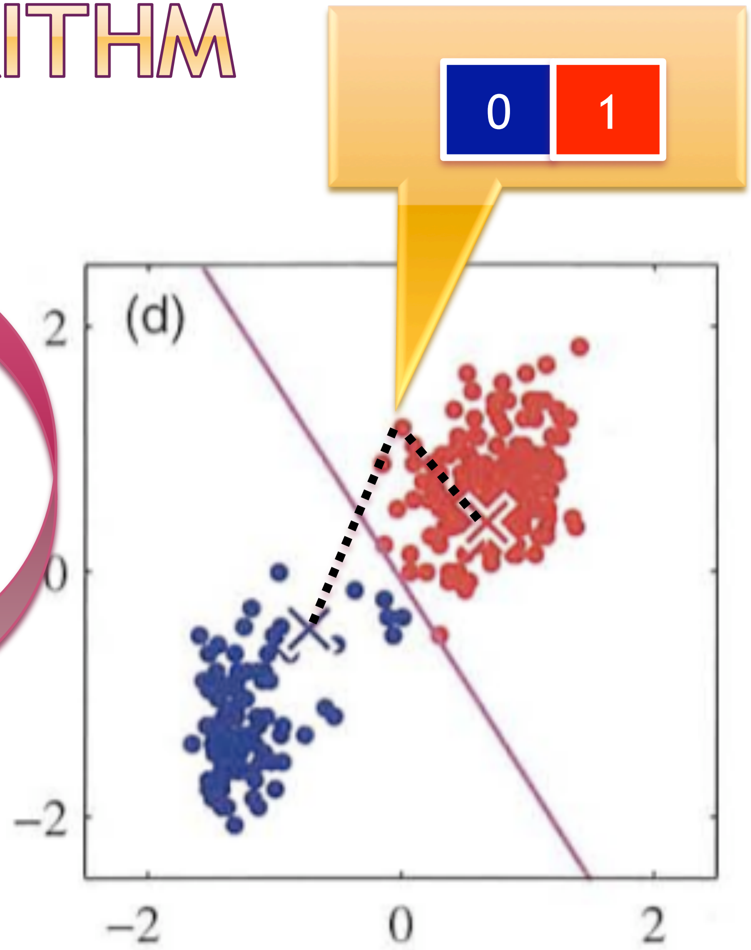
K-MEANS ALGORITHM

2. Phase 1: Assign each point to closest mean
3. **Phase 2:** Update means of the new clusters



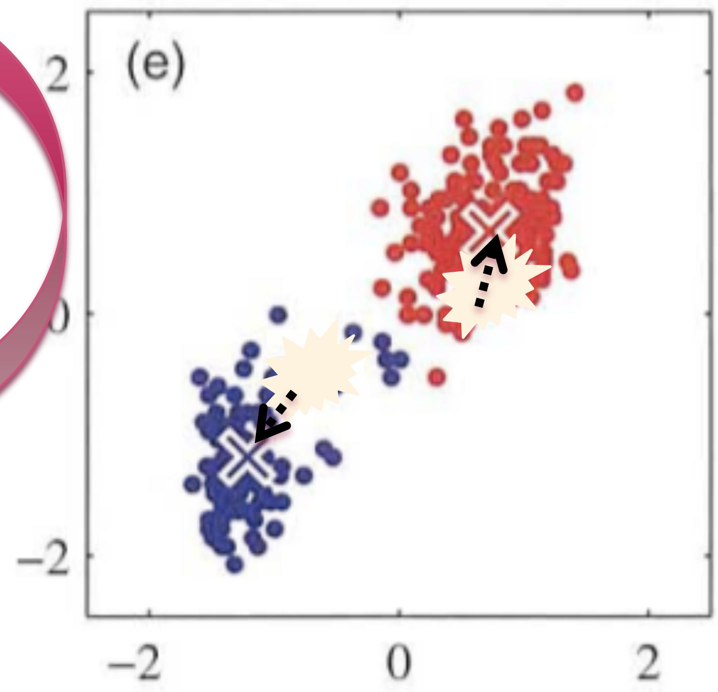
K-MEANS ALGORITHM

2. **Phase 1: Assign each point to closest mean μ_k**
3. Phase 2: Update means of the new clusters



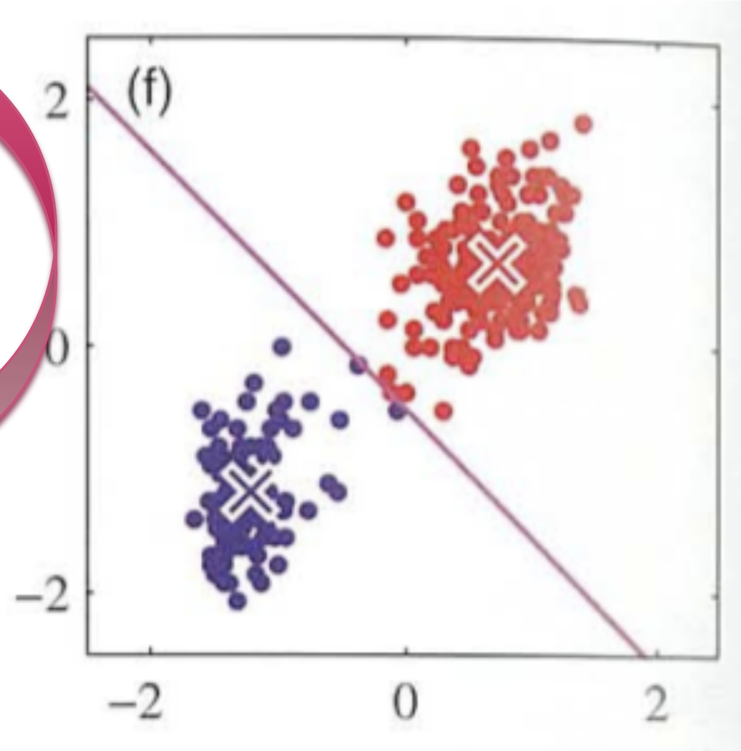
K-MEANS ALGORITHM

2. Phase 1: Assign each point to closest mean
3. **Phase 2:** Update means of the new clusters



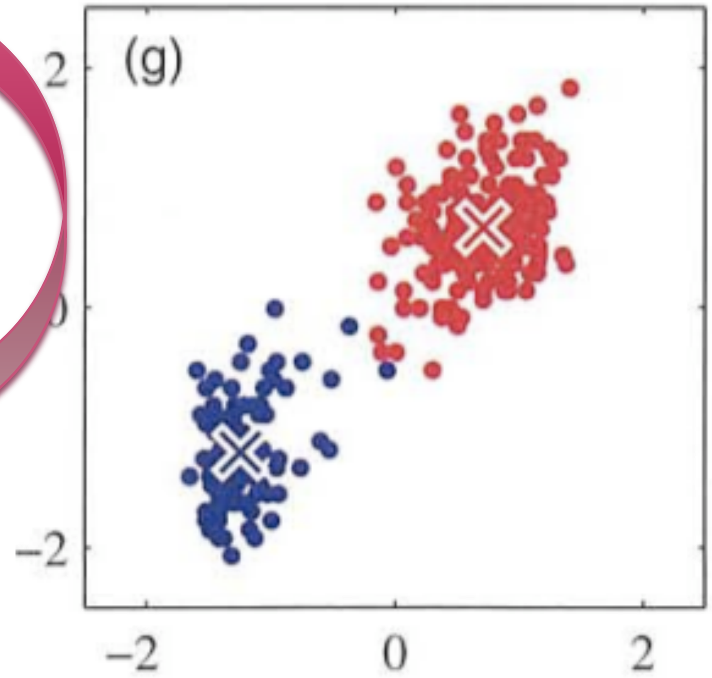
K-MEANS ALGORITHM

2. **Phase 1:** Assign each point to closest mean μ_k
3. **Phase 2:** Update means of the new clusters



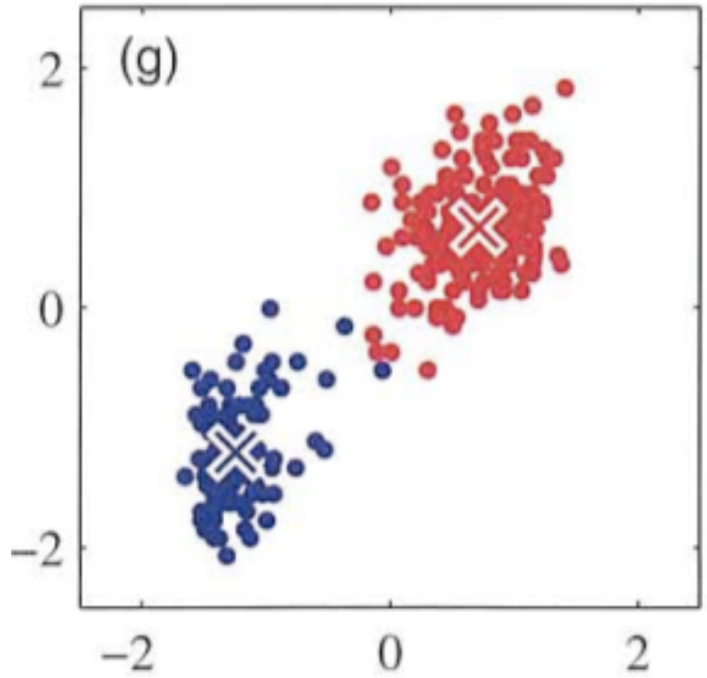
K-MEANS ALGORITHM

2. Phase 1: Assign each point to closest mean
3. **Phase 2:** Update means of the new clusters



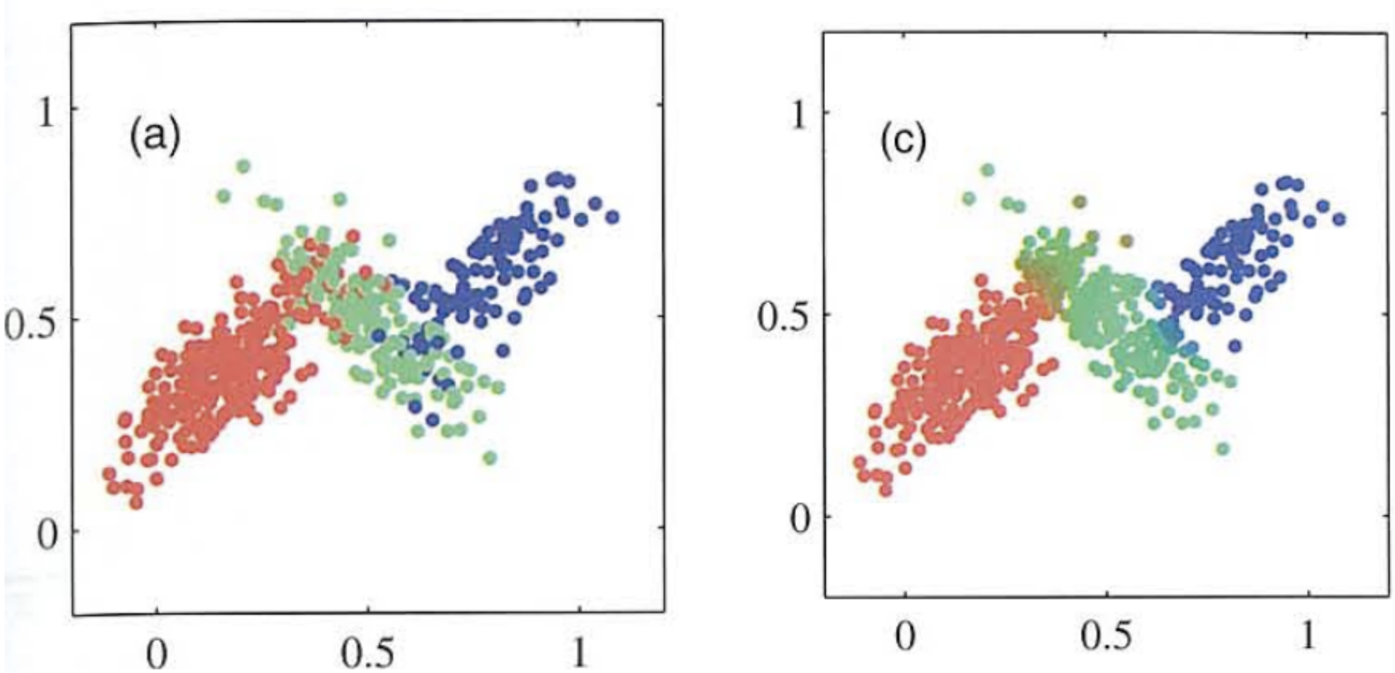
K-MEANS ALGORITHM

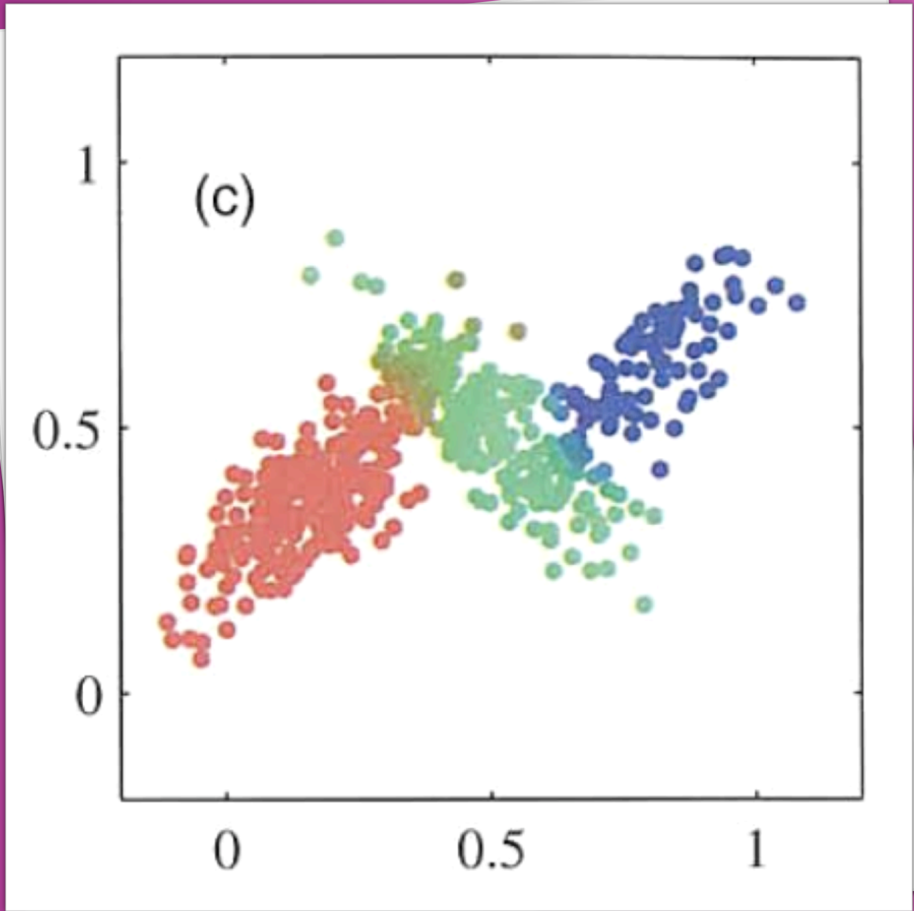
4. When means do not change anymore → clustering DONE.



PROBLEM WITH K-MEANS

- In K-means, a point can only have 1 class
- But what about points that lie in between groups? eg. Jazz + Classical





9.2 MIXTURE OF GAUSSIANS

The Famous “GMM”:
Gaussian Mixture Model

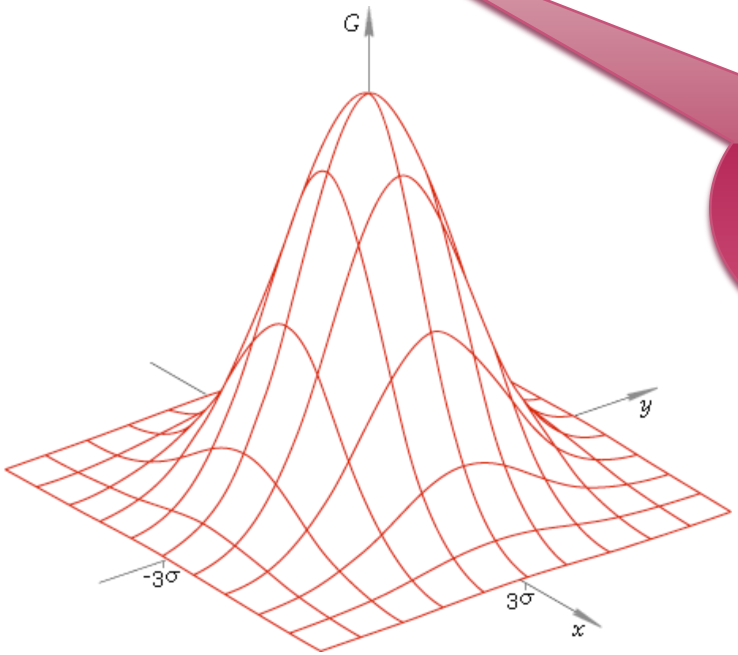
WHAT'S A GAUSSIAN?

$$p(X) = N(X \mid \mu, \Sigma)$$

Mean

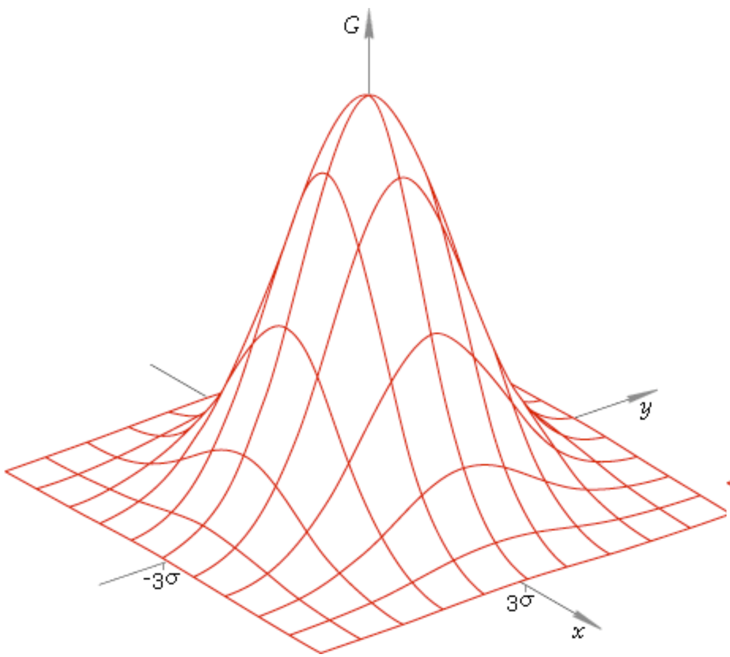
Variance

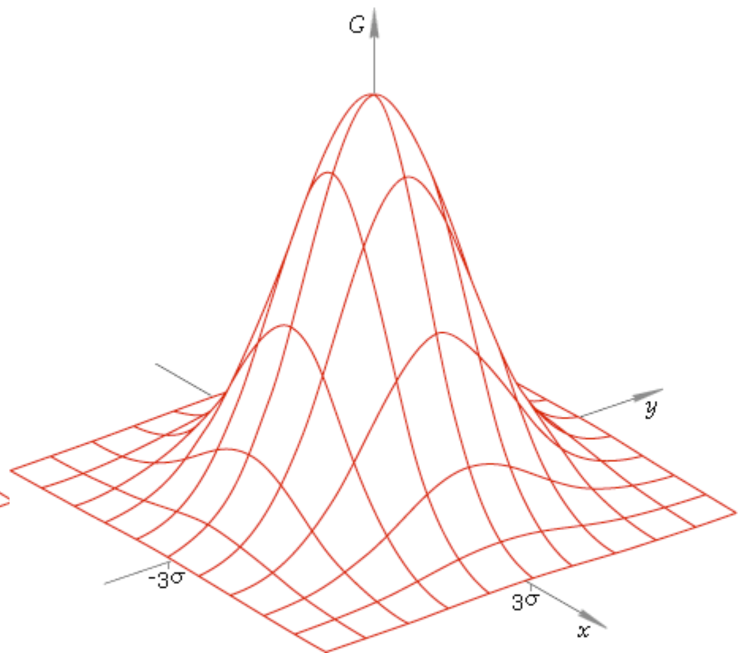
Gaussian == "Normal" distribution



WHAT'S A GAUSSIAN MIXTURE?

$$p(X) = N(X \mid \mu, \Sigma) + N(X \mid \mu, \Sigma)$$



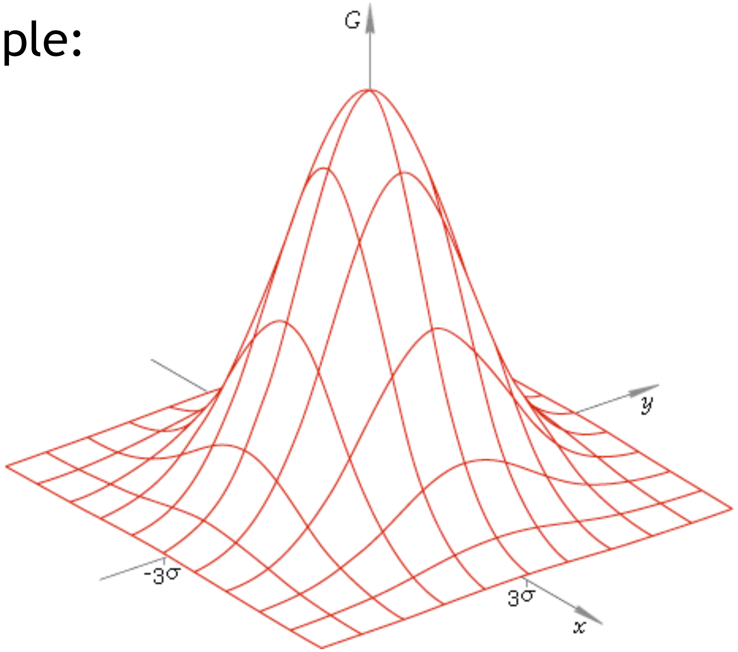


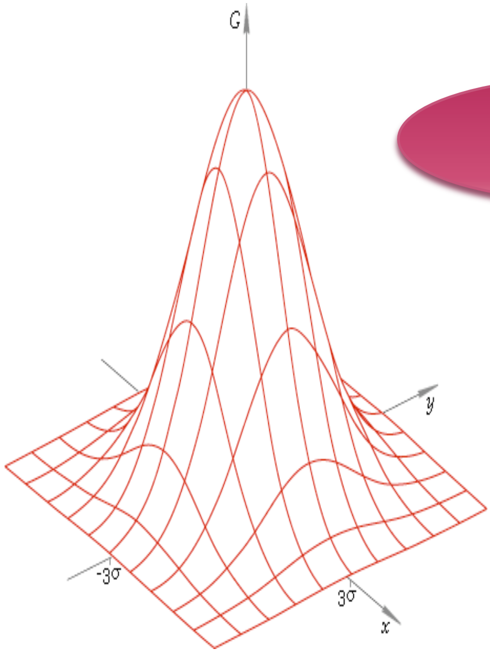
WHAT'S A GAUSSIAN MIXTURE?

$$p(X) = N(X \mid \mu_1, \Sigma_1) + N(X \mid \mu_2, \Sigma_2)$$

Variance

Example:





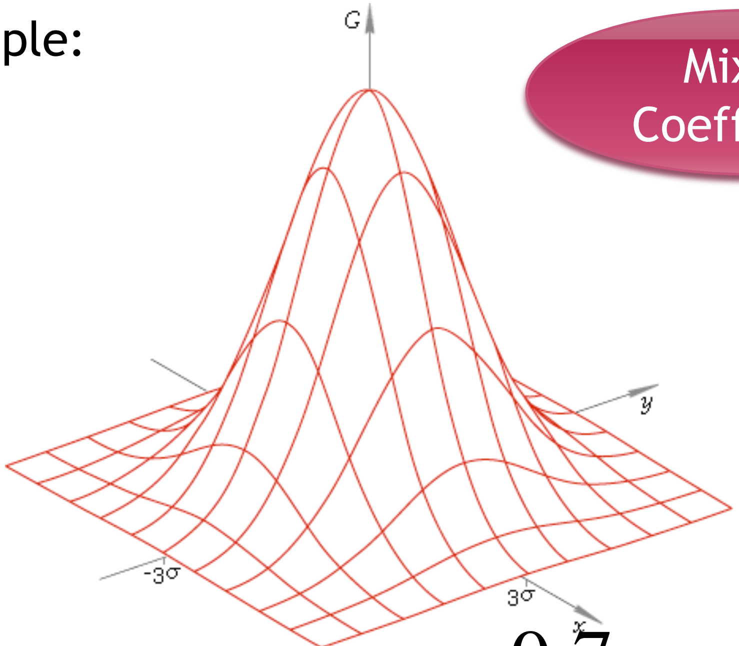
WHAT'S A GAUSSIAN MIXTURE?

$$p(X) = \pi_1 N(X \mid \mu_1, \Sigma_1) + \pi_2 N(X \mid \mu_2, \Sigma_2)$$

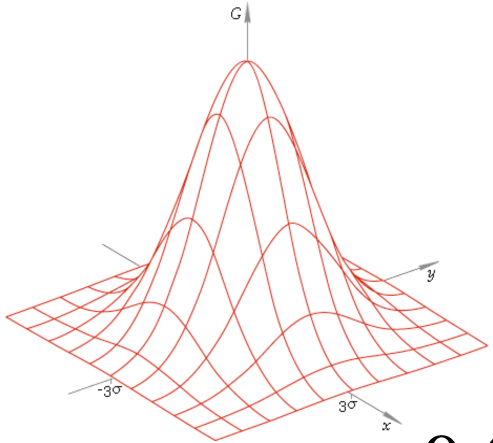
Mixing Coefficient

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

Example:



$$\pi_1 = 0.7$$



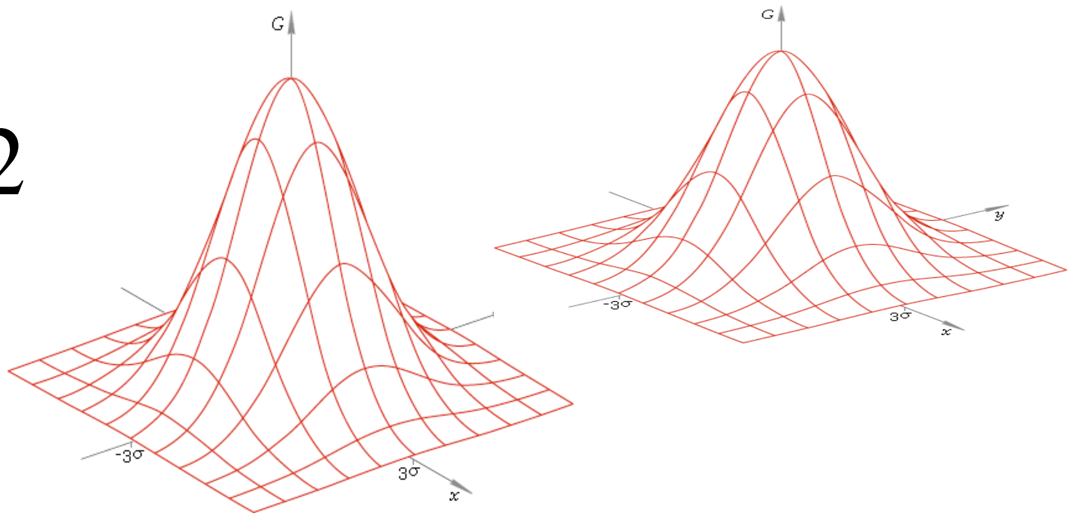
$$\pi_2 = 0.3$$

GAUSSIAN MIXTURE DEFINITION

$$p(X) = \sum_{k=1}^{K} \pi_k N(X \mid \mu_k, \Sigma_k)$$

Example:

$$K = 2$$



K-MEANS → GAUSSIAN MIXTURE

- K-means is a classifier

- Mixture of Gaussians is a probability model
- We can USE it as a "soft" classifier

K-MEANS → GAUSSIAN MIXTURE

- K-means is a classifier

- Mixture of Gaussians is a probability model
- We can USE it as a "soft" classifier

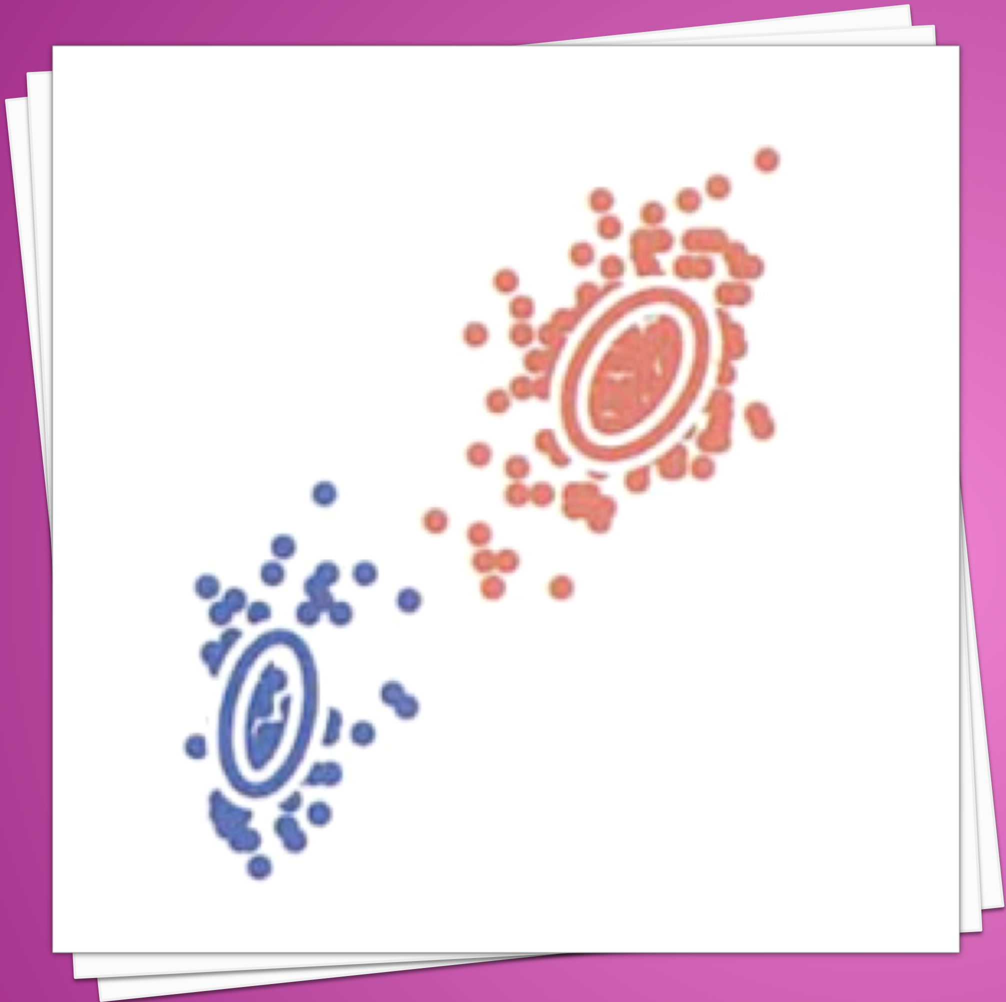
K-MEANS → GAUSSIAN MIXTURE

- K-means is a classifier

- Mixture of Gaussians is a probability model
- We can USE it as a "soft" classifier

Parameter to fit to data:
- Mean μ_k

Parameters to fit to data:
- Mean μ_k
- Covariance Σ_k
- Mixing coefficient π_k



9.2.2 EXPECTATION MAXIMIZATION FOR GAUSSIAN MIXTURES

EM for GMM

K-MEANS ALGORITHM REMINDER

1. Initialize means μ_k
   2. E Step: Assign each point to a cluster 

   3. M Step: Given clusters, refine mean μ_k of each cluster k
4. Stop when change in means is small

EXPECTATION MAXIMIZATION (EM) FOR GAUSSIAN MIXTURES

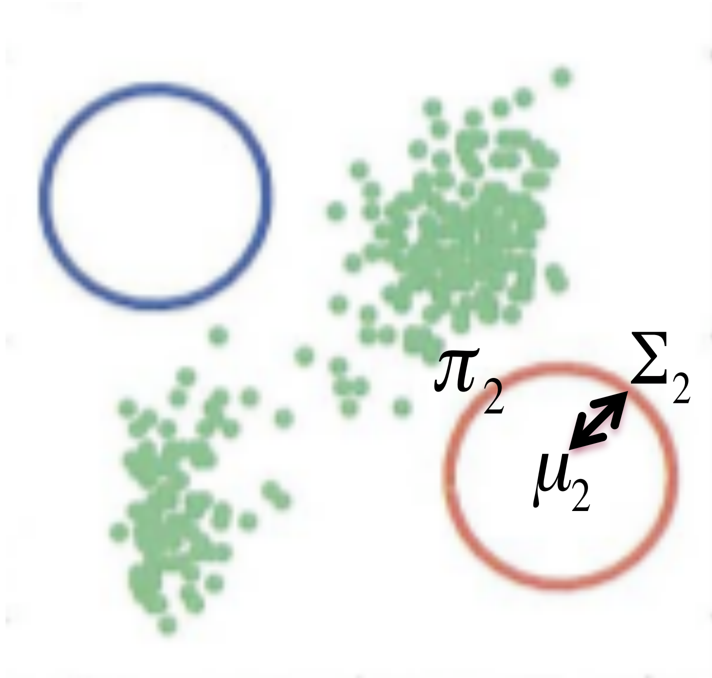
1. Initialize Gaussian* parameters: means μ_k , covariances Σ_k and mixing coefficients π_k
   2. **E Step:** Assign each point X_n an assignment score $\gamma(z_{nk})$ for each cluster k
   3. **M Step:** Given scores, adjust μ_k , π_k , Σ_k for each cluster k
4. Evaluate likelihood. If likelihood or parameters converge, stop.

0.5 | 0.5

*There are k Gaussians

EM FOR GAUSSIAN MIXTURES

1. Initialize μ_k , Σ_k π_k , one for each Gaussian k



- Tip! Use K-means result to initialize:

$$\mu_k \leftarrow \mu_k$$

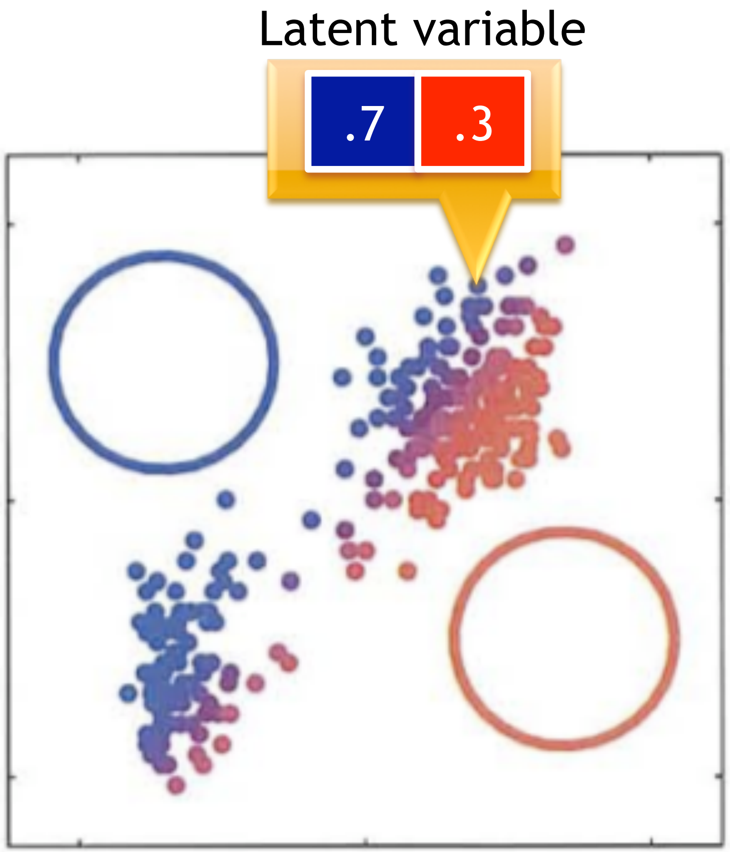
$$\Sigma_k \leftarrow \mathrm{cov}(cluster(K))$$

$$\pi_k \leftarrow \frac{\text{Number of points in k}}{\text{Total number of points}}$$

EM FOR GAUSSIAN MIXTURES

2. **E Step:** For each point X_n, determine its assignment score to each Gaussian k:

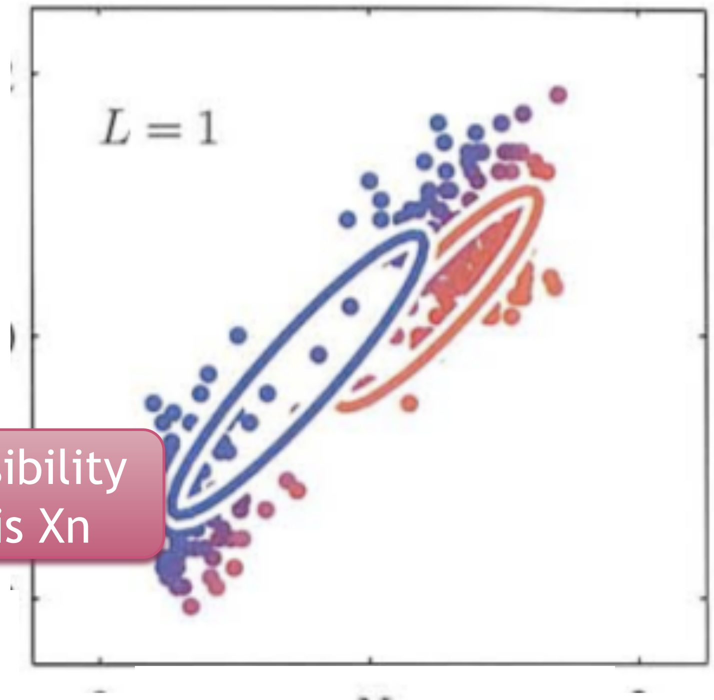
$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$



$\gamma(z_{nk})$ is called a “responsibility”: how much is this Gaussian k responsible for this point X_n?

EM FOR GAUSSIAN MIXTURES

3. **M Step:** For each Gaussian k, update parameters using new $\gamma(z_{nk})$



Responsibility for this Xn

Mean of Gaussian k

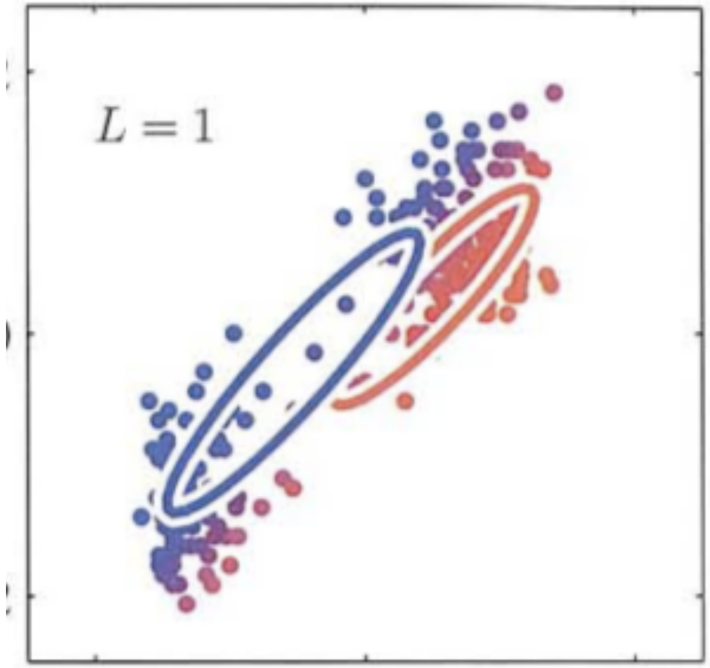
$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

Find the mean that "fits" the assignment scores best

EM FOR GAUSSIAN MIXTURES

3. **M Step:** For each Gaussian k, update parameters using new $\gamma(z_{nk})$



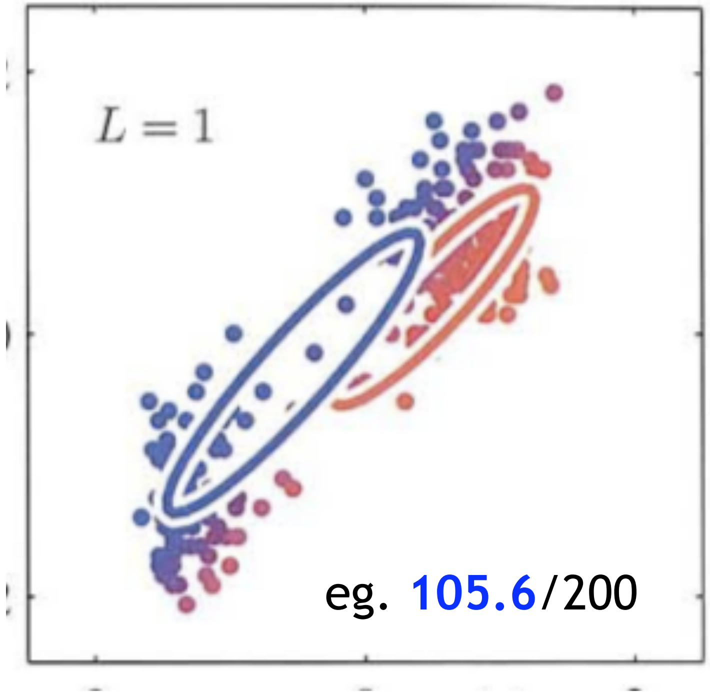
Covariance matrix of Gaussian k

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

Just calculated this!

EM FOR GAUSSIAN MIXTURES

3. **M Step:** For each Gaussian k, update parameters using new $\gamma(z_{nk})$



Mixing Coefficient for Gaussian k

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

Total # of points

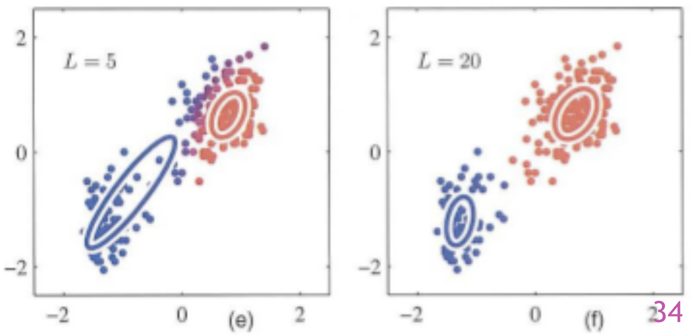
$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

EM FOR GAUSSIAN MIXTURES

4. Evaluate log **likelihood.** If likelihood or parameters converge, stop. Else go to Step 2 (**E step**).

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Likelihood is the probability that the data X was generated by the parameters you found. ie. Correctness!



9.4 THE GENERAL EM ALGORITHM

GENERAL EM ALGORITHM

1. Initialize parameters θ^{old}
   2. **E Step:** Evaluate $p(Z \mid X, \theta^{old})$
   3. **M Step:** Evaluate

Hidden variables

Observed variables

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}})$$

where

Likelihood

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

4. Evaluate log likelihood. If likelihood or parameters converge, stop. Else $\theta^{old} \leftarrow \theta^{new}$ and go to E Step.

EM IN MANY FORMS

- K-means can be formulated as EM
- EM for Gaussian Mixtures
- EM for Bernoulli Mixtures
- EM for Bayesian Linear Regression

EXPECTATION MAXIMIZATION SUMMARY

- "Expectation"

Calculated the fixed, data-dependent parameters of the function *Q*.

- "Maximization"

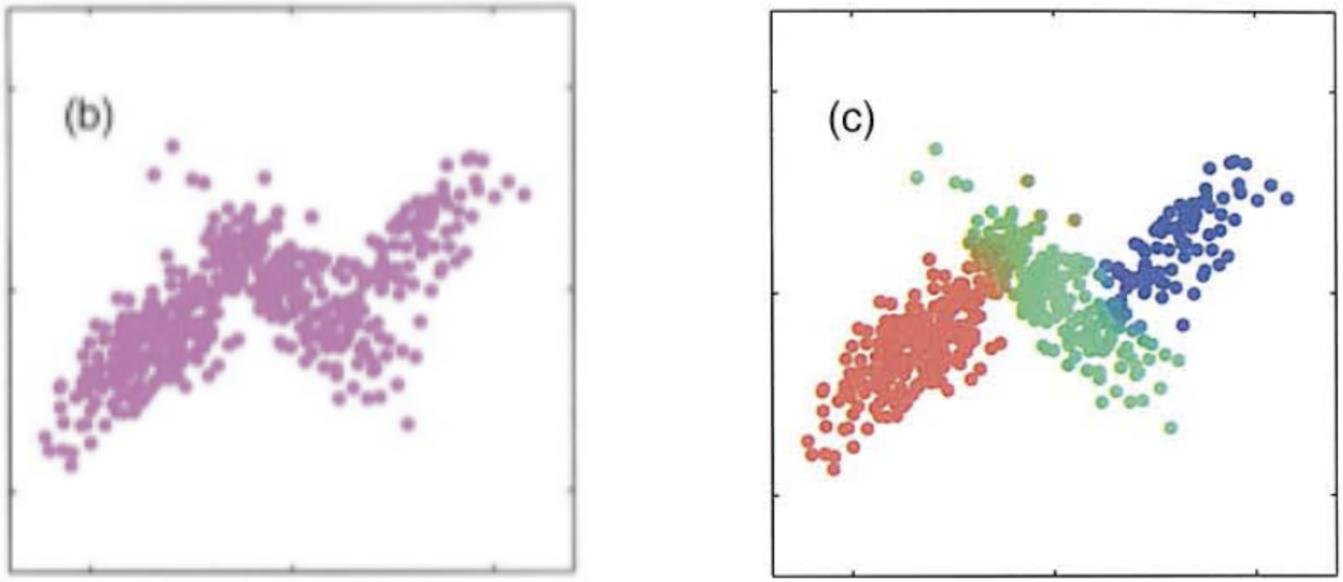
Once the parameters of *Q* are known, it is fully determined, so now we can maximize *Q*.

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

CHAPTER 9 SUMMARY

- We learned how to cluster data in an unsupervised manner
- Gaussian Mixture Models are useful for modeling data with "soft" cluster assignments
- Expectation Maximization is a method used when we have a model with latent variables (values we don't know, but estimate with each step)

QUESTIONS?



- My question: What other applications could use EM? How about EM of GMMs?