TEXAS A&M UNIVERSITY
Engineering

ECEN 758 Data Mining and Analysis: Lecture 8, Gaussian Mixture Models

Joshua Peeples, Ph.D.
Assistant Professor
Department of Electrical and Computer Engineering

Announcements

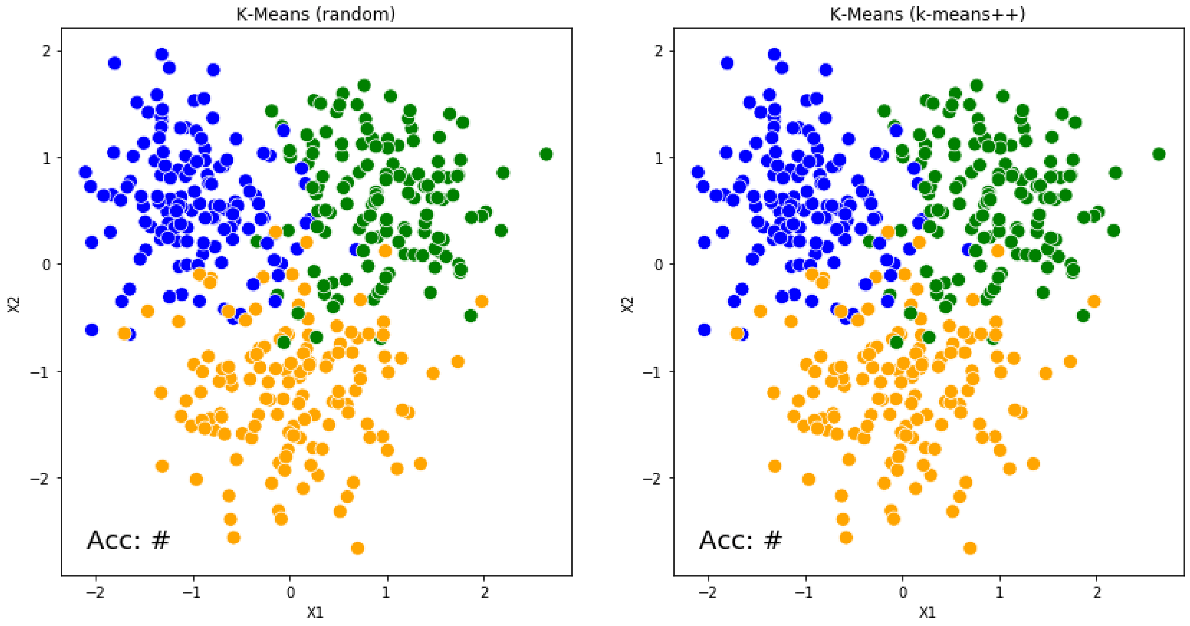
- Assignment #1 grades available
  - Please revise any grade discrepancies within a week (COB, 09/23)
  - Email Dr. Peeples (do not contact Grader) and/or stop by office hours
- Assignment #2 will be released this Wednesday (09/18)
  - Please upload submission as single PDF
  - Please share Python code (e.g., Jupyter Notebooks, Google Colab)

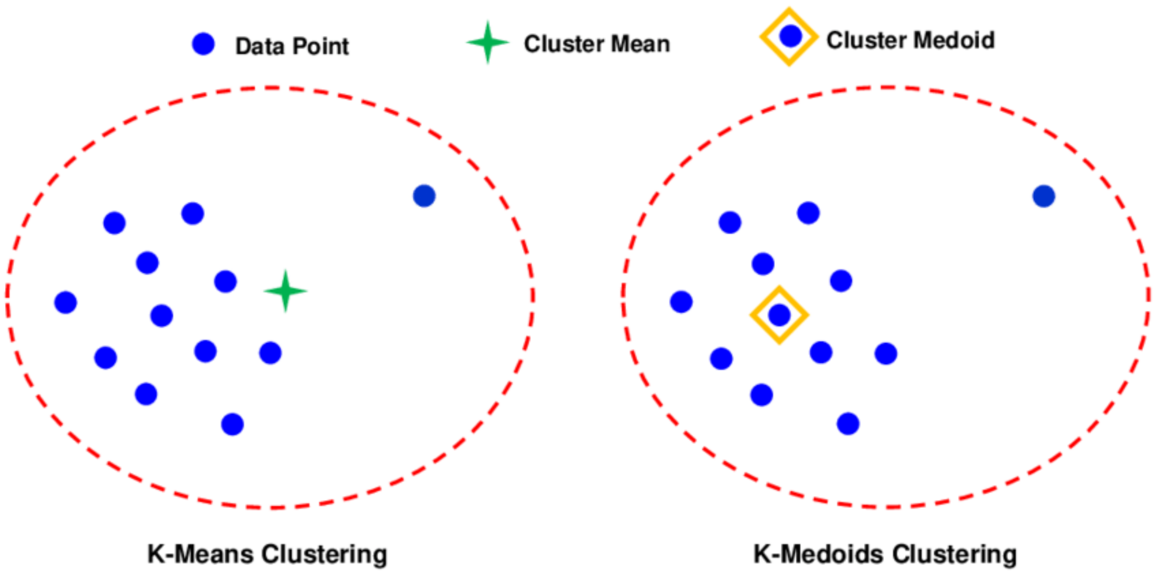
Assignment 1 Observations

- Make sure to clearly label your figures and tables
  - Communication is important
- Please have a clear discussion
  - Formal writing (i.e., no contractions)
- Show your work
  - State equations and show your steps
- Disclose if you use AI (e.g., ChatGPT, Copilot) for code development
  - Do not use for your discussions
- Ask questions if clarification is needed

Last Lecture

- Representative Clustering II





Gif from: D. Sheehan, Clustering with Scikit with GIFs

Today

- Gaussian Mixture Models
- Reading: ZM Chapter 13

Clustering Overview

TEXAS A&M UNIVERSITY Engineering

- We will discuss several variants of clustering
  - **Representative-based Clustering**
  - Hierarchical Clustering
  - Density-Based Clustering

What disadvantages of k-means?

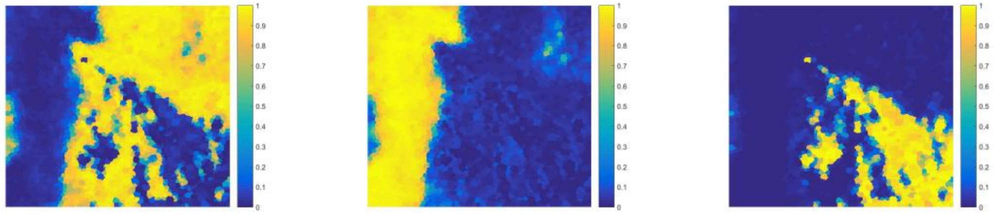
k-Means Disadvantages

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- Batch processing
- Selecting number of clusters (*k*)
- **“Crisp”/Hard clustering**

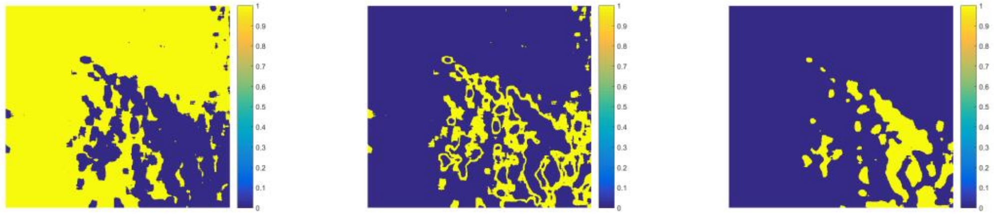
k-Means Disadvantage: “Crisp”/Hard Clustering

- Points can only “belong” to one cluster
- Different applications may require “soft” clustering
  - Points may belong to more than one group

Input Image



(h) FLICM Cluster 1 (i) FLICM Cluster 2 (j) FLICM Cluster 3



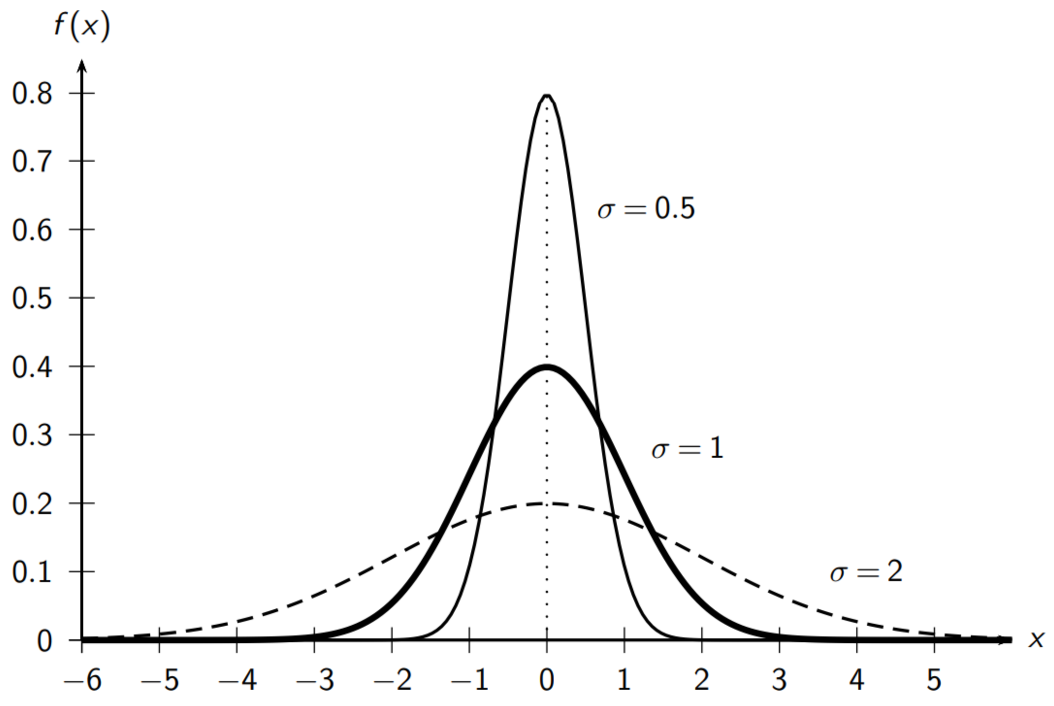
(k) K-Means Cluster 1 (l) K-Means Cluster 2 (m) K-Means Cluster 3

J. Peeples, et al. Possibilistic fuzzy local information C-means with automated feature selection for seafloor segmentation. SPIE, 2018.

Gaussian Mixture Models

Gaussian/Normal Distribution (1D)

- Two parameters, mean (μ) and variance (σ^2)
- Probability density decreases exponentially as a function of the distance from mean
- Maximum value when x = μ



$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Gaussian Mixture Models

- Model clusters as Gaussians
- "Soft" clustering approach
  - Assign probability of belonging to clustering
- Generative model

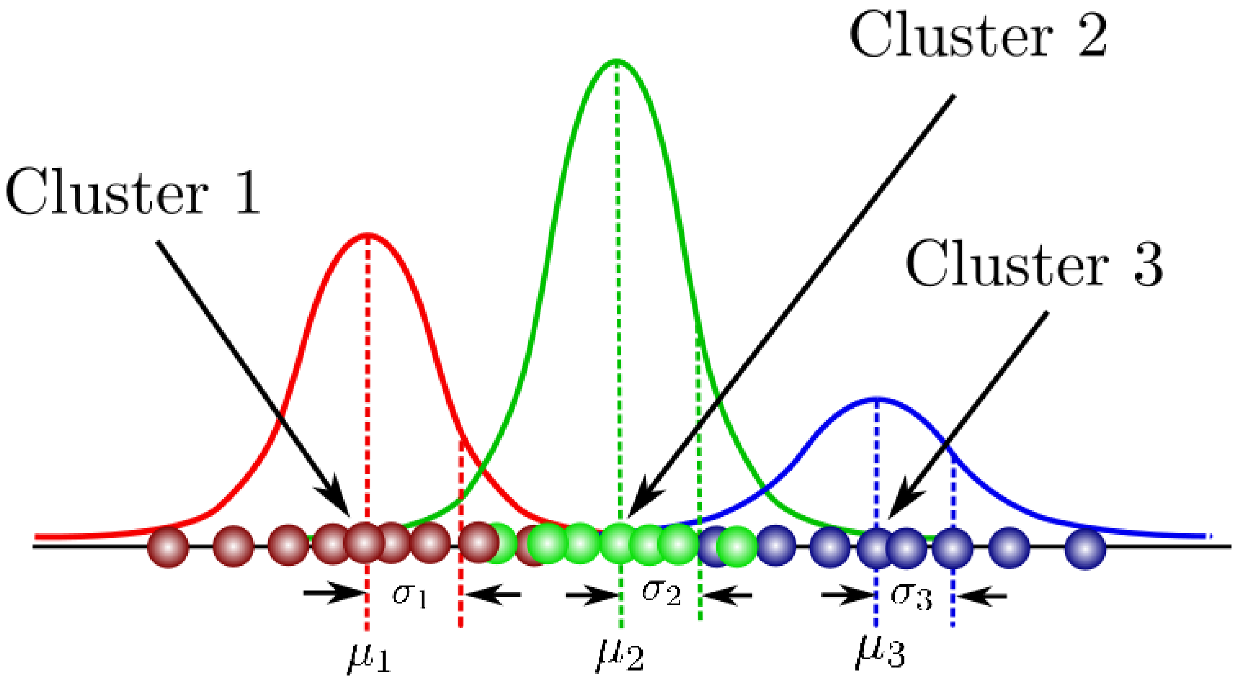
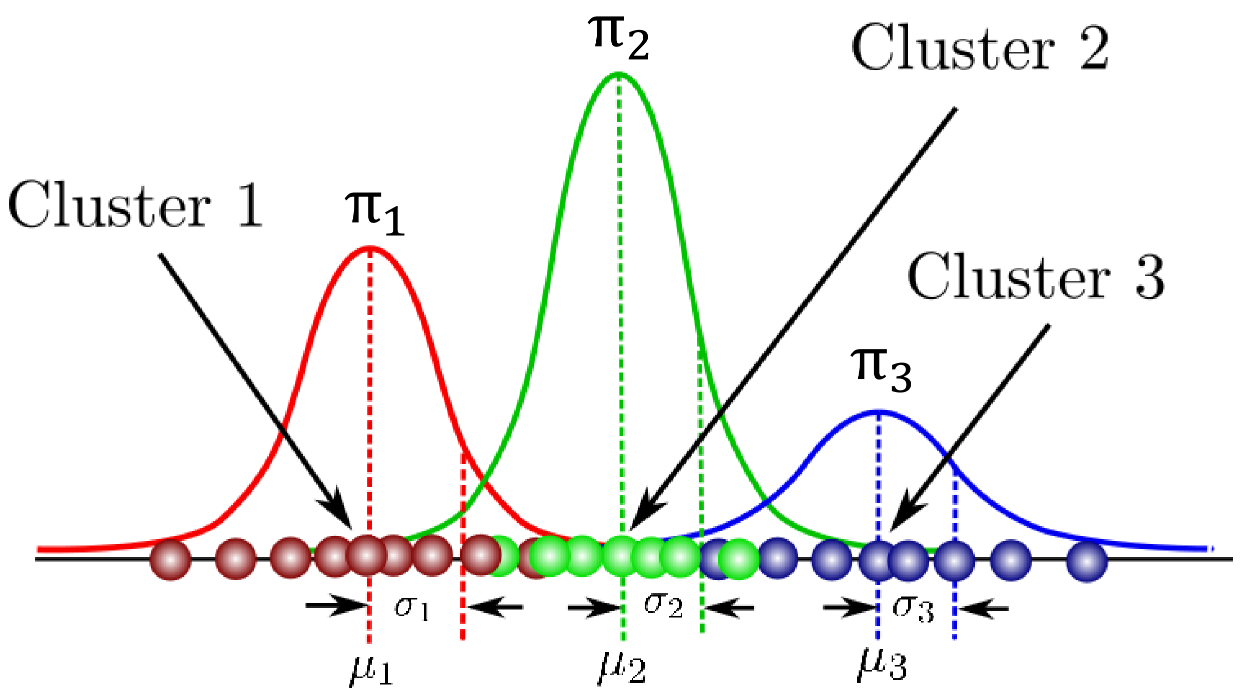


Image from: O. Carrasco, Gaussian Mixture Models Explained.

Mixtures of Gaussians (1D)

TEXAS A&M UNIVERSITY Engineering

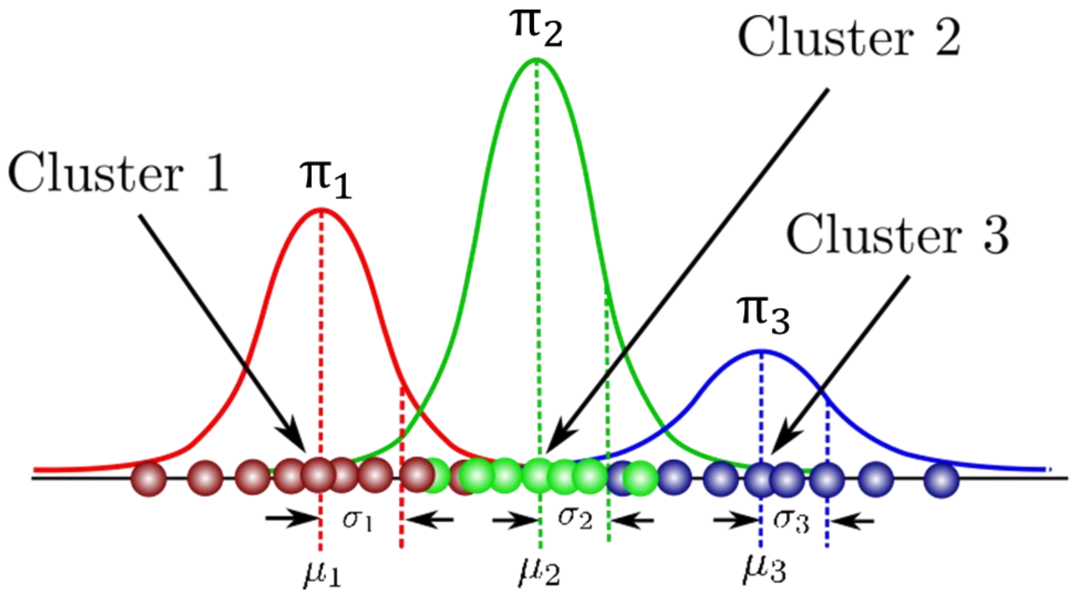
- Three parameters to describe clusters:
  - Mean (μ_k)
  - Variance (σ_k^2)
  - Mixture parameters (π_k)
    - Weights, “size”, prior probability
    - Sum to one constraint



Mixtures of Gaussians (1D)

- Three parameters to describe clusters:
  - Mean (μ_k)
  - Variance (σ_k^2)
  - Mixture parameters (π_k)
    - Weights, "size", prior probability
- Probability distribution:

$$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \sigma_i)$$



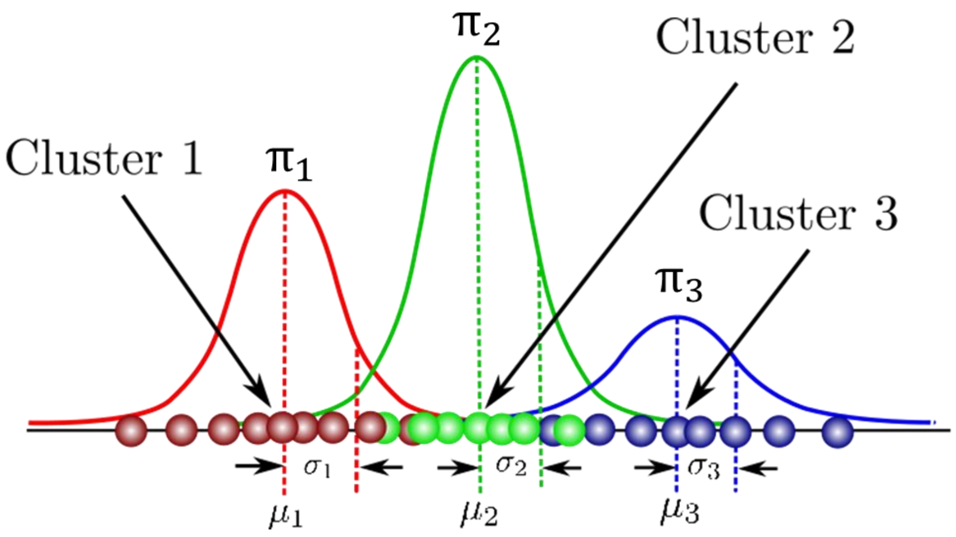
Mixtures of Gaussians (1D)

- Probability distribution:

$$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \sigma_i)$$

- Select mixture component with probability π_k

$$p(z = k) = \pi_k$$



Mixtures of Gaussians (1D)

- Probability distribution:

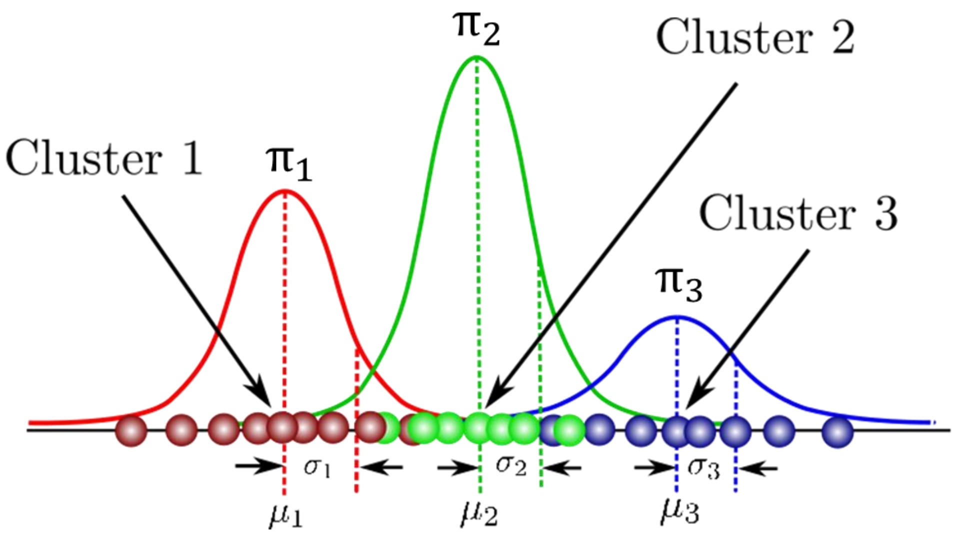
$$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \sigma_i)$$

- Select mixture component with probability π_k

$$p(z = k) = \pi_k$$

- Sample from that component's Gaussian

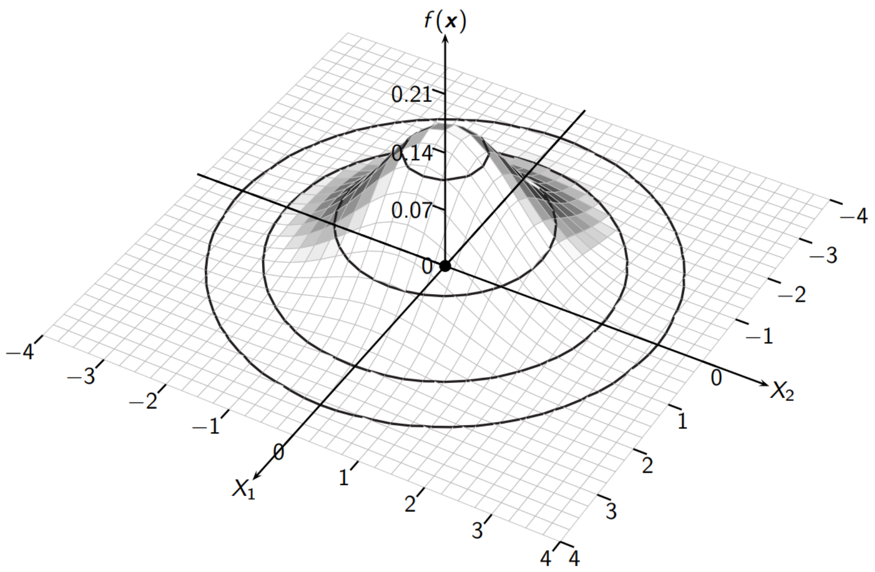
$$p(x|z = k) = \mathcal{N}(x|\mu_k, \sigma_k)$$



Gaussian Mixture Models: Multivariate

Multivariate Gaussian Distribution

- Parameters: mean vector (**μ**) and covariance matrix (Σ)
- |Σ| determinant of covariance matrix
- Numerator in exponential referred to as **Mahalanobis distance**



$$f(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(\sqrt{2\pi})^d\sqrt{|\Sigma|}}\exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right\}$$

Mixtures of Gaussians

- Three parameters to describe clusters:
  - Mean vector ($\boldsymbol{\mu}_i$)
  - Covariance matrix ($\boldsymbol{\Sigma}_i^2$)
  - Mixture parameters (π_i *or* $P(C_i)$)
    - Weights, “size”, prior probability
    - Sum to one constraint

$$\sum_{i=1}^{k} P(C_i) = 1$$

ith Cluster:

$$f_i(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{ -\frac{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)}{2} \right\}$$

Probability Density function of **x** as GMM:

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} f_i(\boldsymbol{x}) P(C_i) = \sum_{i=1}^{k} f(\boldsymbol{x}|\boldsymbol{\mu}_i, \Sigma_i) P(C_i)$$

Gaussian Mixture Models Algorithm

GMM Algorithm: Objective

TEXAS A&M UNIVERSITY Engineering

- Parameters of model represented as **ϴ**

$$\theta = \{\mu_1, \Sigma_1, P(C_1) \ldots, \mu_k, \Sigma_k, P(C_k)\}$$

- Maximum likelihood estimation (MLE)
- Usually maximize log-likelihood function

Likelihood:

$$P(\mathbf{D}|\theta) = \prod_{j=1}^{n} f(\mathbf{x}_j)$$

MLE:

$$\theta^* = \arg\max_{\theta}\{\ln P(\mathbf{D}|\theta)\}$$

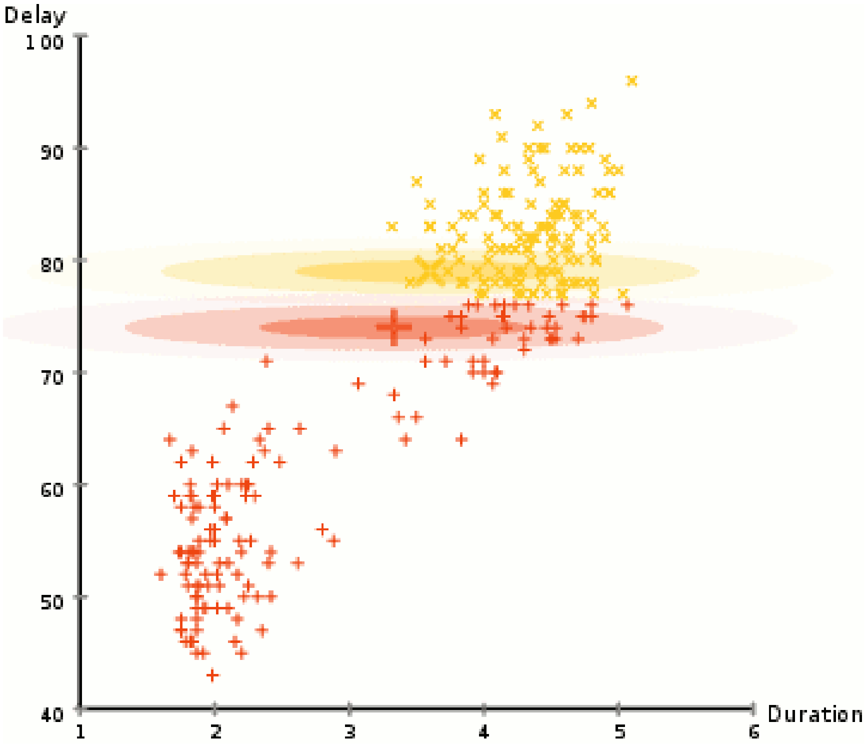
Log-likelihood:

$$\ln P(\mathbf{D}|\theta) = \sum_{j=1}^{n} \ln f(\mathbf{x}_j) = \sum_{j=1}^{n} \ln\left(\sum_{i=1}^{k} f(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)P(C_i)\right)$$

GMM Algorithm: Objective

TEXAS A&M UNIVERSITY Engineering

- Directly maximizing log-likelihood over Θ is hard
- Alternative approach: Expectation-Maximization (EM)
- Two steps:
  - Expectation: Assignment of points
  - Maximization: Estimation of parameters
- We will do a deep dive into EM next lecture!



Gif from: Expectation-maximization algorithm, Wikipedia

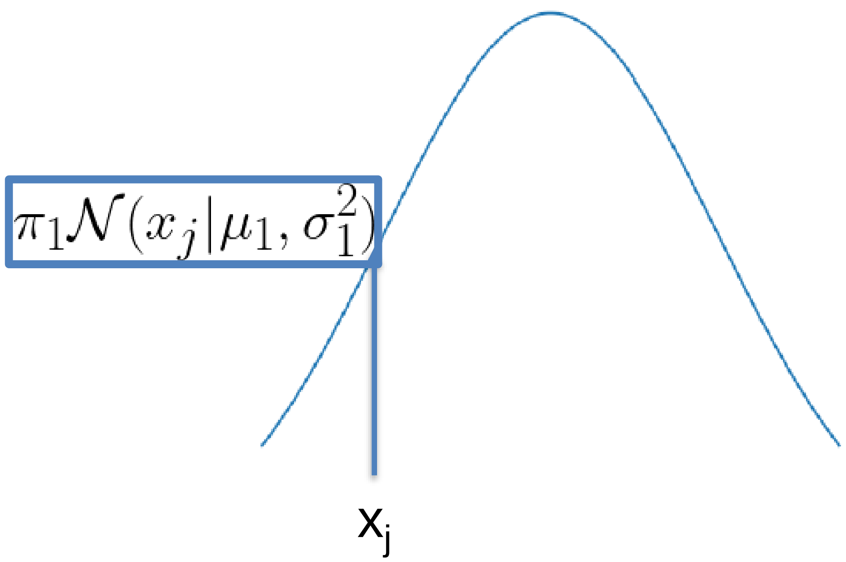
GMM Expectation-Maximization (1D)

GMM Expectation-Maximization (1D): E-step

- **Initialize cluster parameters**
- Expectation (E-Step)
  - For each data point, x_j
  - Compute cluster posterior probability
    - Compute probability with respect to C_i
    - Normalize to sum to one over clusters

For each cluster:

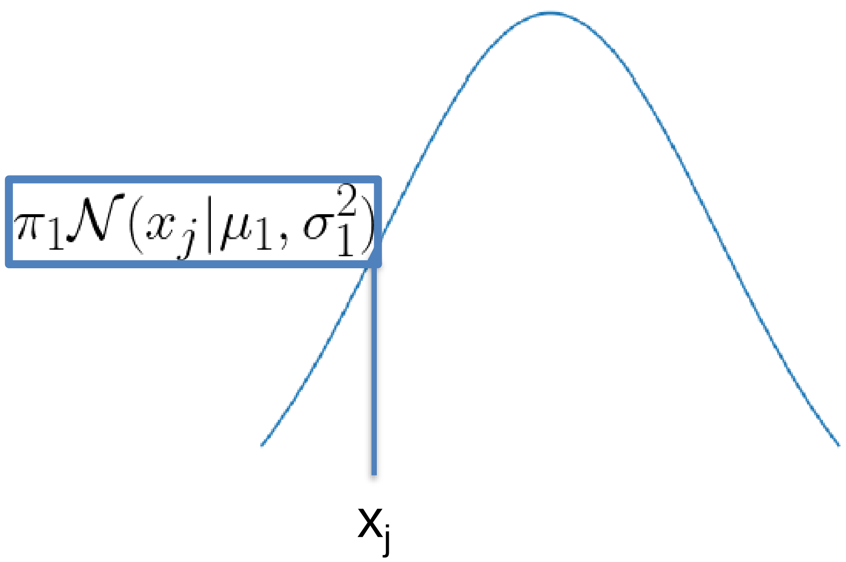
$$f_i(x) = f(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\}$$



GMM Expectation-Maximization (1D): E-step

- Initialize cluster parameters
- Expectation (E-Step)
  - For each data point, x_j
  - **Compute cluster posterior probability**
    - Compute probability with respect to C_i
    - Normalize to sum to one over clusters

$$w_{ij} = P(C_i | x_j) = \frac{f(x_j | \mu_i, \sigma_i^2) \cdot P(C_i)}{\sum_{a=1}^{k} f(x_j | \mu_a, \sigma_a^2) \cdot P(C_a)}$$

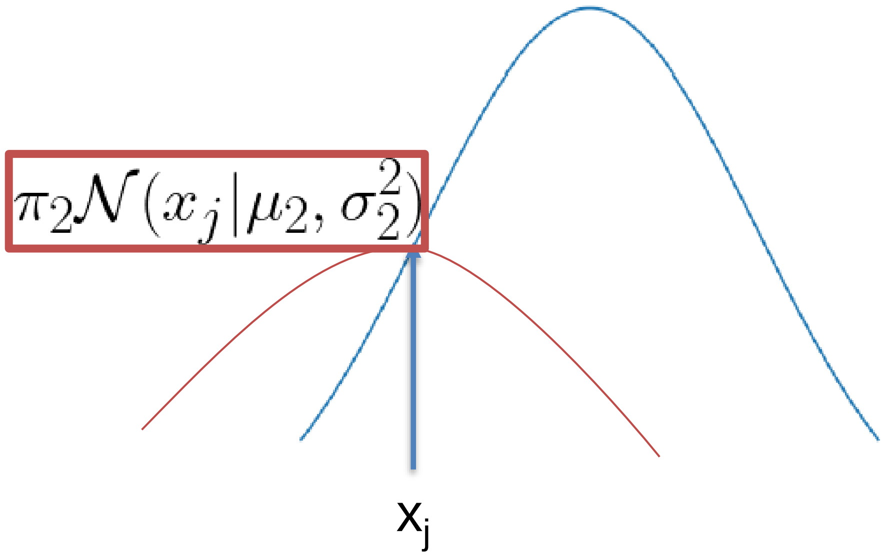


GMM Expectation-Maximization (1D): E-step

TEXAS A&M UNIVERSITY Engineering

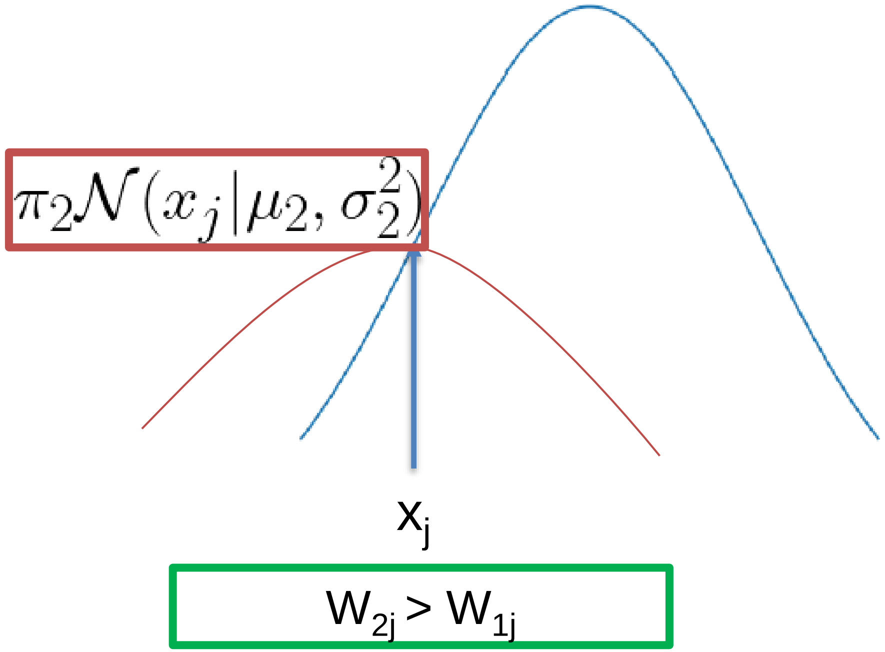
- Expectation (E-Step)
  - For each data point, x_j
  - **Compute cluster posterior probability**
    - Compute probability with respect to C_i
    - Normalize to sum to one over clusters

$$w_{ij} = P(C_i|x_j) = \frac{f(x_j|\mu_i, \sigma_i^2) \cdot P(C_i)}{\sum_{a=1}^{k} f(x_j|\mu_a, \sigma_a^2) \cdot P(C_a)}$$



- Expectation (E-Step)
  - For each data point, x_j
  - Compute cluster posterior probability
    - Compute probability with respect to C_i
    - Normalize to sum to one over clusters
- **Higher probability will be assigned to Gaussian that is more likely**

$$w_{ij} = P(C_i|x_j) = \frac{f(x_j|\mu_i, \sigma_i^2) \cdot P(C_i)}{\sum_{a=1}^{k} f(x_j|\mu_a, \sigma_a^2) \cdot P(C_a)}$$



GMM Expectation-Maximization (1D): M-step

- Maximization (M-Step)
  - Update parameters using (weighted) data points

$$w_{ij} = P(C_i|x_j) = \frac{f(x_j|\mu_i, \sigma_i^2) \cdot P(C_i)}{\sum_{a=1}^{k} f(x_j|\mu_a, \sigma_a^2) \cdot P(C_a)}$$

Mean:

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot x_j}{\sum_{j=1}^{n} w_{ij}}$$

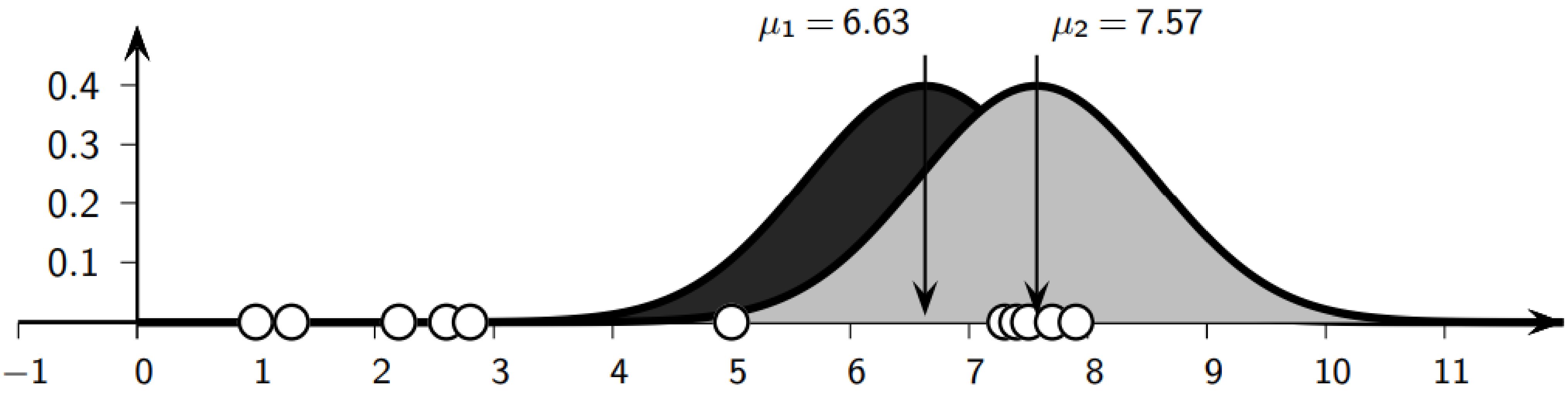
Variance:

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} w_{ij}(x_j - \mu_i)^2}{\sum_{j=1}^{n} w_{ij}}$$

Mixture Weight/Prior Probability:

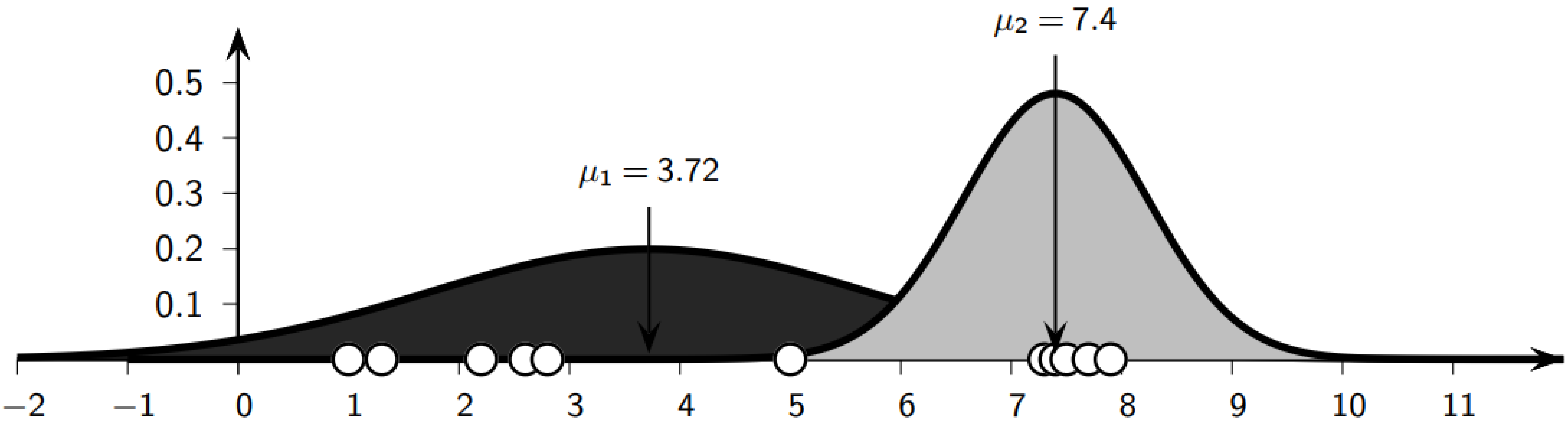
$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

GMM EM 1D Example



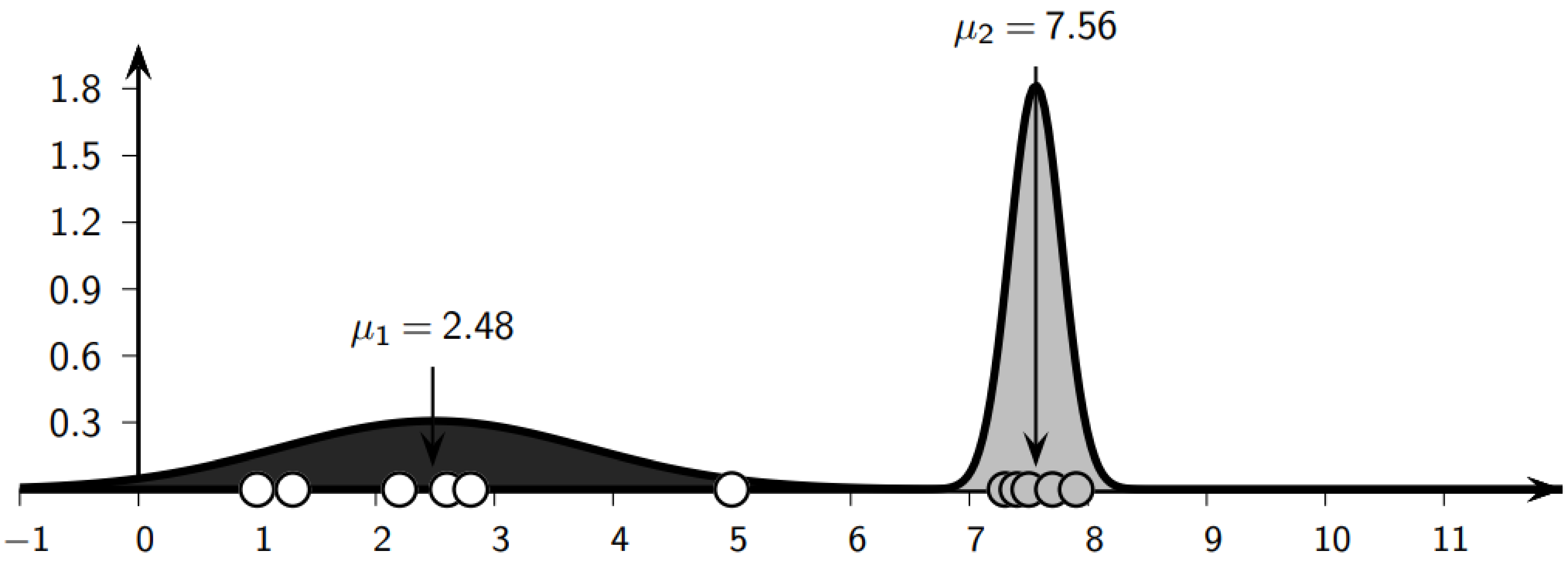
(a) Initialization: $t = 0$

GMM EM 1D Example



(b) Iteration: $t = 1$

GMM EM 1D Example



(c) Iteration: $t = 5$ (converged)

GMM Expectation-Maximization (d-dimensions)

EM in d Dimensions

- Each cluster will have d x d covariance matrix
- Expensive to calculate and may be unreliable estimation
- Can use diagonal covariance
  - Assumes dimensions are independent

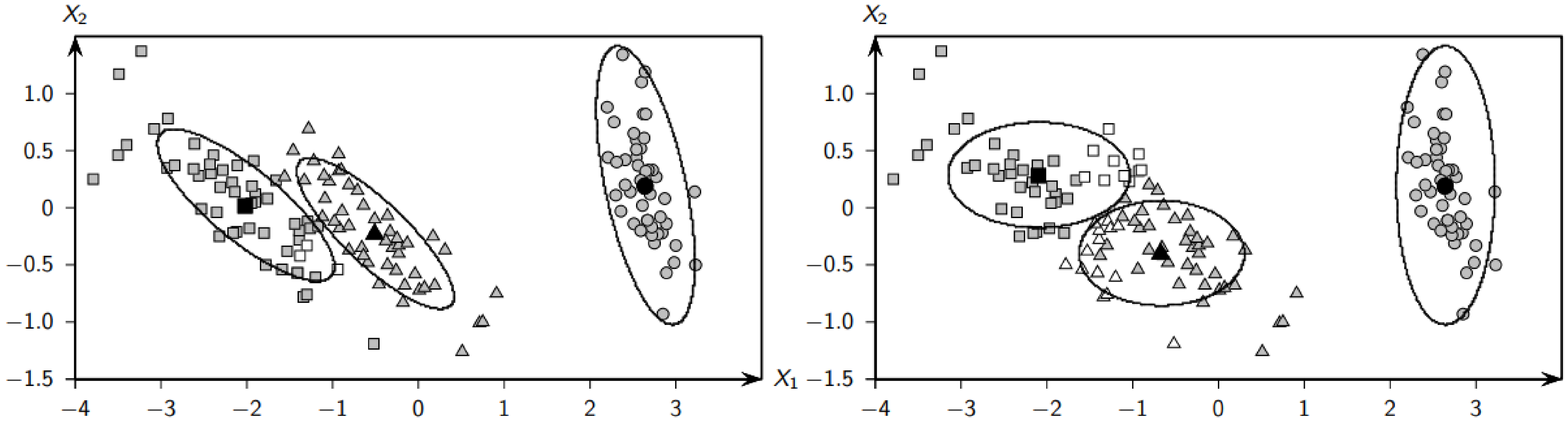
Full Covariance:

$$\Sigma_i = \begin{pmatrix} (\sigma_1^i)^2 & \sigma_{12}^i & \cdots & \sigma_{1d}^i \\ \sigma_{21}^i & (\sigma_2^i)^2 & \cdots & \sigma_{2d}^i \\ \vdots & \vdots & \ddots & \\ \sigma_{d1}^i & \sigma_{d2}^i & \cdots & (\sigma_d^i)^2 \end{pmatrix}$$

Diagonal Covariance:

$$\Sigma_i = \begin{pmatrix} (\sigma_1^i)^2 & 0 & \cdots & 0 \\ 0 & (\sigma_2^i)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & (\sigma_d^i)^2 \end{pmatrix}$$

Full vs Diagonal



(a) Full covariance matrix ($t = 36$)   (b) Diagonal covariance matrix ($t = 29$)

GMM in Scikit-learn

- [Additional options]() for covariance matrices include:
  - Spherical: Each cluster has a single variance (isotropic covariance)
  - Tied: All clusters share same covariance matrix

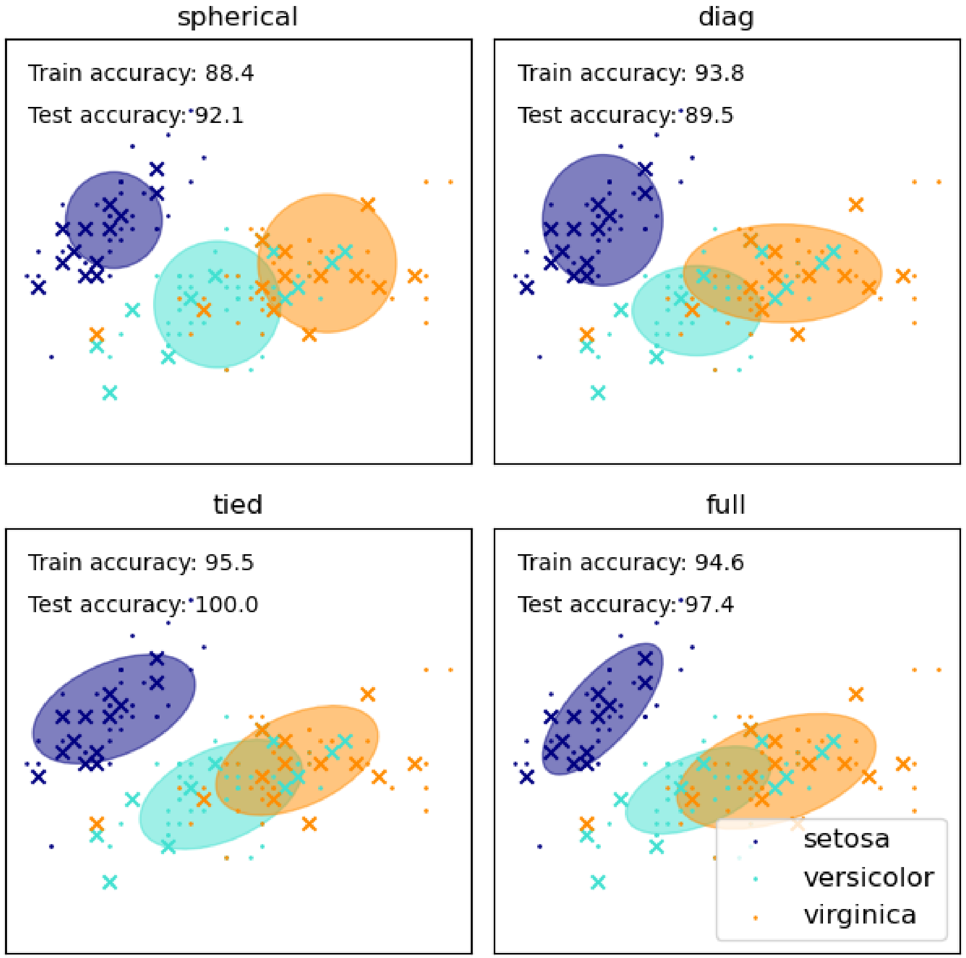


Image from: Scikit-learn, 2.1.1. Gaussian Mixture.

GMM Number of Parameters

- Given *k* clusters, *n* samples, and *d* features, what are the total number of parameters for a **tied covariance matrix** (i.e., all clusters share the same covariance matrix) GMM?
- Break into pairs
- 5 minutes for activity

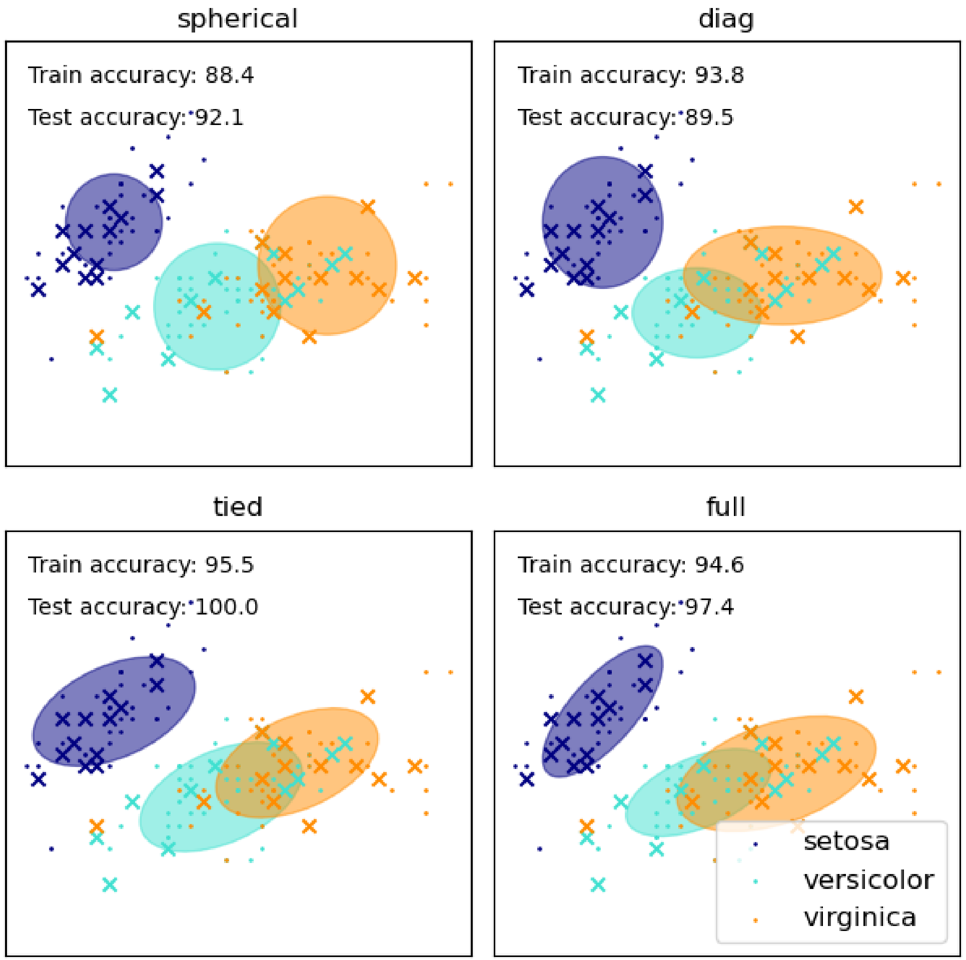


Image from: Scikit-learn, 2.1.1. Gaussian Mixture.

GMM Number of Parameters

TEXAS A&M UNIVERSITY Engineering

- Given *k* clusters, *n* samples, and *d* features, what are the total number of parameters for a **tied covariance matrix** (i.e., all clusters share the same covariance matrix) GMM?
- Solution:
  - $k*d + d^2 + k$
  - *k* mean vectors (d by 1), single covariance (d by d), and *k* mixture parameters (1x1, scalers)

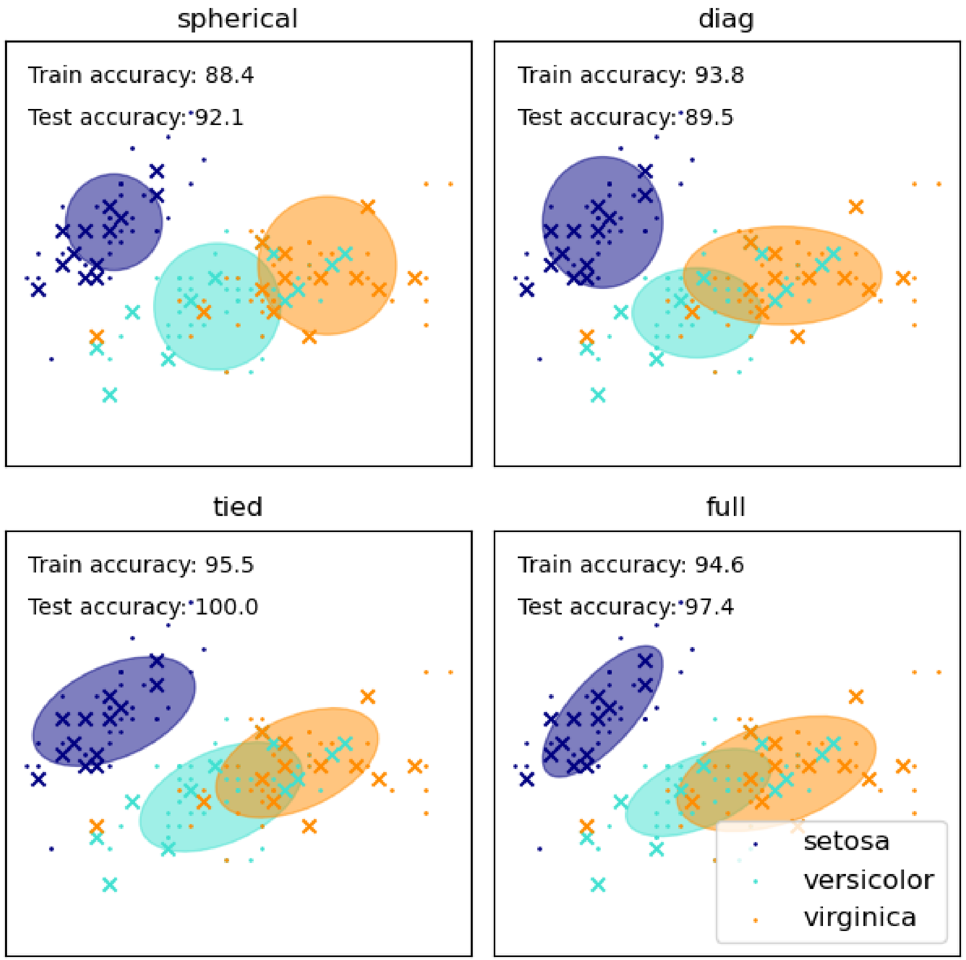


Image from: Scikit-learn, 2.1.1. Gaussian Mixture.

EM in d Dimensions

- Expectation step:

$$w_{ij} = P(C_i|\mathbf{x}_j) = \frac{f_i(\mathbf{x}_j) \cdot P(C_i)}{\sum_{a=1}^{k} f_a(\mathbf{x}_j) \cdot P(C_a)}$$

- Maximization step:

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}} \qquad \Sigma_i = \frac{\sum_{j=1}^{n} w_{ij}(\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}} \qquad P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

GMM EM Algorithm

- Each step maximizes log-likelihood
- Iterate until convergence
  - Set maximum iterations or set threshold for changes in parameters
  - May converge to local optima

MLE:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}}\{\ln P(\boldsymbol{D}|\boldsymbol{\theta})\}$$

Log-likelihood:

$$\ln P(\boldsymbol{D}|\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln f(\mathbf{x}_j) = \sum_{j=1}^{n} \ln\left(\sum_{i=1}^{k} f(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)P(C_i)\right)$$

GMM EM Algorithm Pseudocode

Expectation-Maximization $(\boldsymbol{D}, k, \epsilon)$**:**

$t \leftarrow 0$

Randomly initialize $\boldsymbol{\mu}_1^t, \ldots, \boldsymbol{\mu}_k^t$

$\Sigma_i^t \leftarrow \boldsymbol{I}, \ \forall i = 1, \ldots, k$

repeat

  $t \leftarrow t + 1$

  for $i = 1, \ldots, k$ *and* $j = 1, \ldots, n$ **do**

    $w_{ij} \leftarrow \frac{f(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i) \cdot P(C_i)}{\sum_{a=1}^{k} f(\mathbf{x}_j|\boldsymbol{\mu}_a, \Sigma_a) \cdot P(C_a)}$ `// posterior probability` $P^t(C_i|\mathbf{x}_j)$

  for $i = 1, \ldots, k$ **do**

    $\boldsymbol{\mu}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$ `// re-estimate mean`

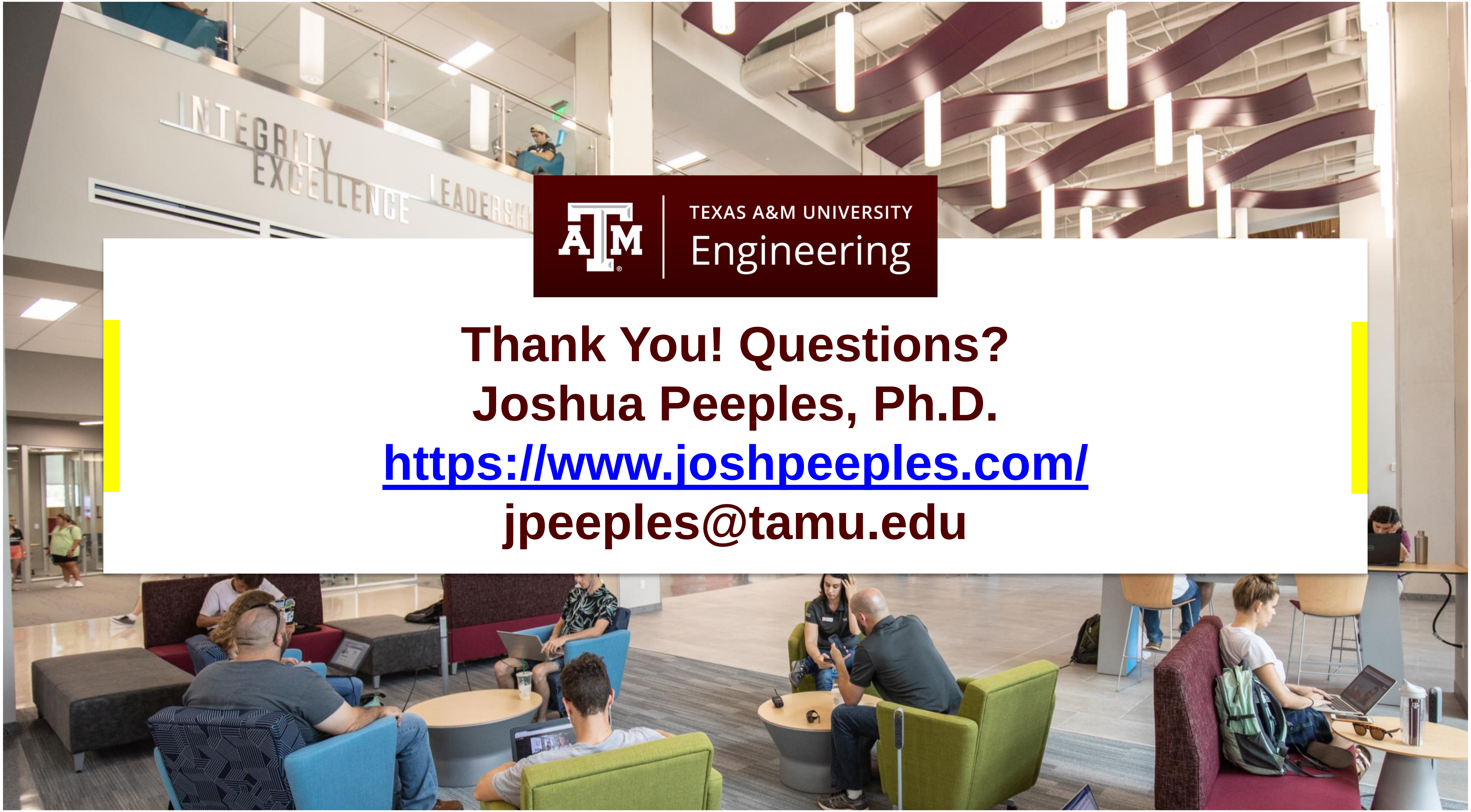
    $\Sigma_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$ `// re-estimate covariance matrix`

    $P^t(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{n}$ `// re-estimate priors`

until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

Next class

- Expectation-Maximization



TEXAS A&M UNIVERSITY
Engineering

Thank You! Questions?

Joshua Peeples, Ph.D.
https://www.joshpeeples.com/
jpeeples@tamu.edu

TEXAS A&M UNIVERSITY
Engineering

Supplemental Slides

Useful Links

TEXAS A&M UNIVERSITY Engineering

- Gaussian Mixture Models and EM
- Gaussian Mixture Models Google Colab