CS57300
PURDUE UNIVERSITY

NOVEMBER 10, 2021

DATA MINING

K-MEANS

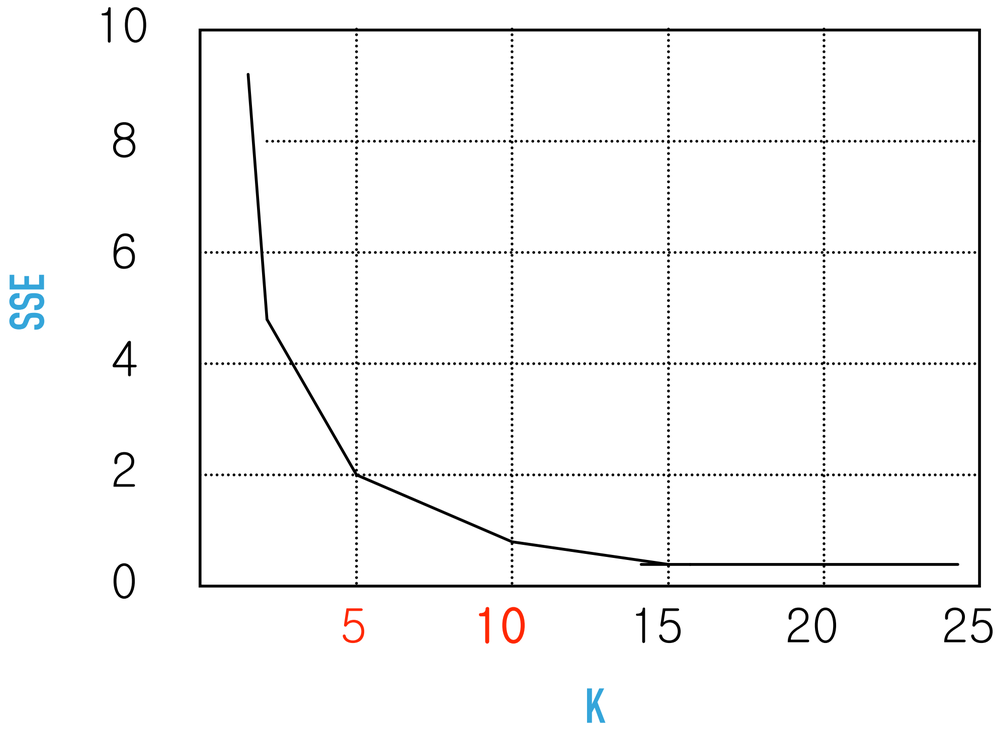
- Strengths:
  - Relatively efficient (time complexity is O(K·N·i), where i is the number of iterations)
  - Finds spherical clusters
- Weaknesses:
  - Terminates at local optimum (sensitive to initial seeds)
  - Applicable only when mean is defined
  - Need to specify K
  - Susceptible to outliers/noise

VARIATIONS

- Selection of initial centroids
  - Select first seed randomly and then pick successive points that are farthest away
  - Run with multiple random selections, pick result with best score
  - Use hierarchical clustering to identify likely clusters and pick seeds from distinct groups
- When mean is undefined
  - K-medoids: use one of the data points as cluster center
  - K-modes: uses categorical distance measure and frequency-based update method

HOW TO SELECT K?

- Plot objective function (i.e., within cluster SSE) as a function of K, and look for "elbow" in plot



K-MEANS SUMMARY

- Knowledge representation
  - K clusters are defined by canonical members (e.g., centroids)
- Model space the algorithm searches over?
  - All possible partitions of the examples into k groups
- Scoring function?
  - Minimize within-cluster Euclidean distance
- Search procedure?
  - Iterative refinement correspond to greedy hill-climbing

HIERARCHICAL CLUSTERING

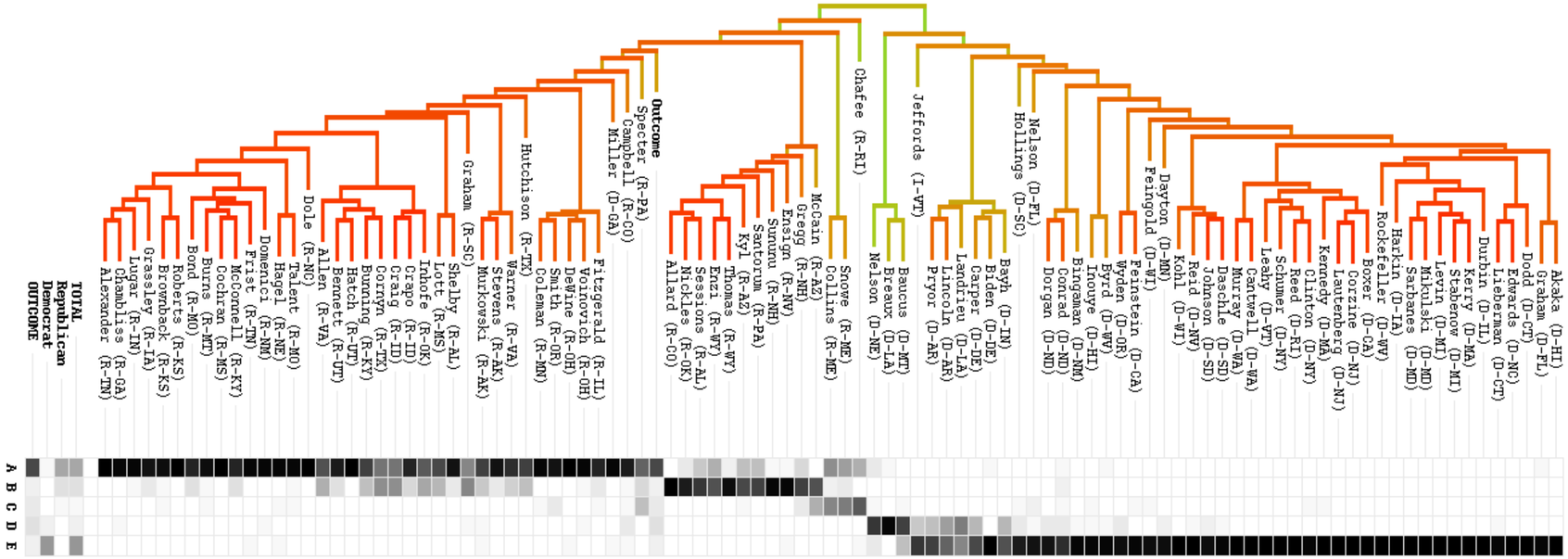
HIERARCHICAL METHODS

- Construct a hierarchy of nested clusters rather than picking K beforehand
- Approaches:
  - Agglomerative: merge clusters successively
  - Divisive: divided clusters successively
- Dendrogram depicts sequences of merges or splits and height indicates distance

AGGLOMERATIVE

- For i = 1 to n:
  - Let $C_i = \{x(i)\}$
- While $|C|>1$:
  - Let C_i and C_j be the pair of clusters with min $D(C_i,C_j)$
  - $C_i=C_i \cup C_j$
  - Remove C_j

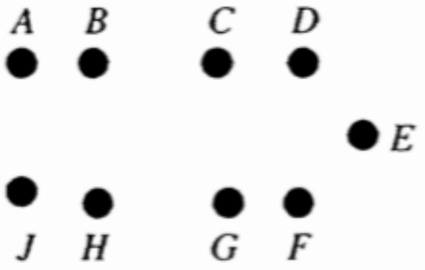
HIERARCHICAL CLUSTERING



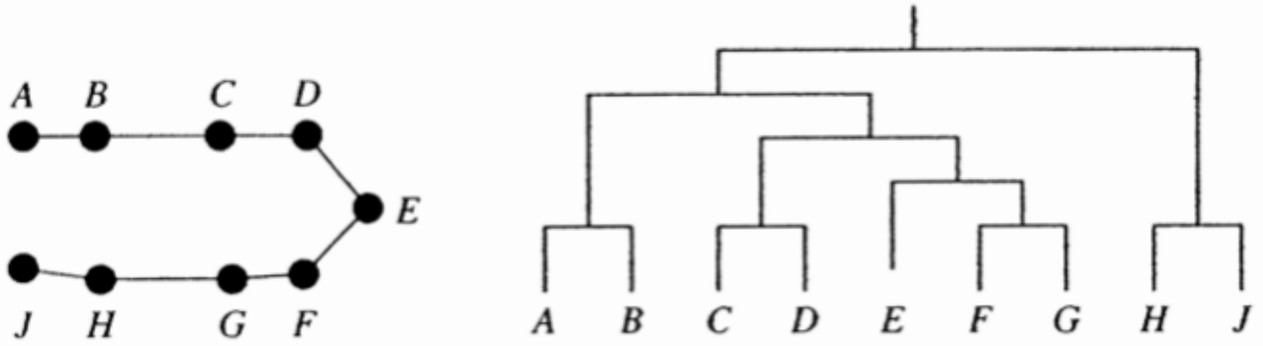
Clustering represented with dendrogram

DISTANCE MEASURES BETWEEN CLUSTERS

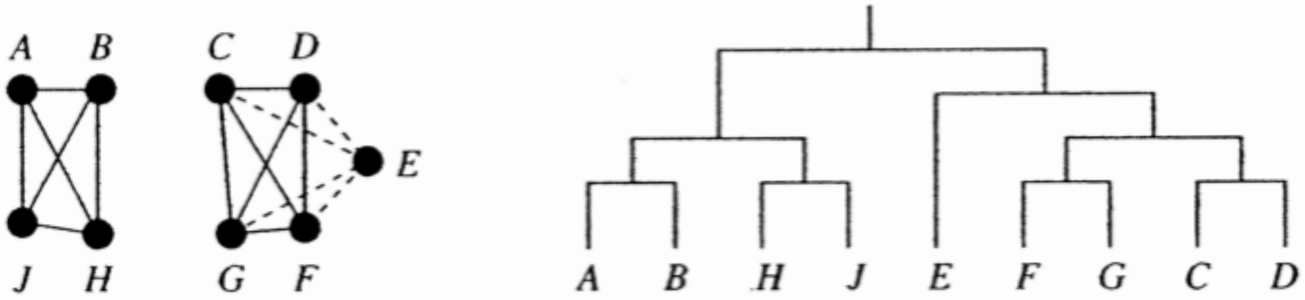
- Single-link/nearest neighbor:
  - $D(C_i,C_j) = \mathbf{min}\{ d(x,y) \mid x \in C_i, y \in C_j \}$ $\Rightarrow$ can produce long thin clusters
- Complete-link/furthest neighbor:
  - $D(C_i,C_j) = \mathbf{max}\{ d(x,y) \mid x \in C_i, y \in C_j \}$ $\Rightarrow$ is sensitive to outliers
- Average link:
  - $D(C_i,C_j) = \mathbf{avg}\{ d(x,y) \mid x \in C_i, y \in C_j \}$ $\Rightarrow$ compromise between the two



(a) Data set



(b) Clustering using single linkage



(c) Clustering using complete linkage

HIERARCHICAL CLUSTERING SUMMARY

- Knowledge representation
  - Dendrogram represents a hierarchy of clusterings
- Model space the algorithm searches over?
  - All possible dendrograms (i.e., hierarchies of partitions from 1 to N)
- Score function?
  - Locally minimize within-cluster distance (e.g., single link)
- Search procedure?
  - Local greedy search

DIVISIVE

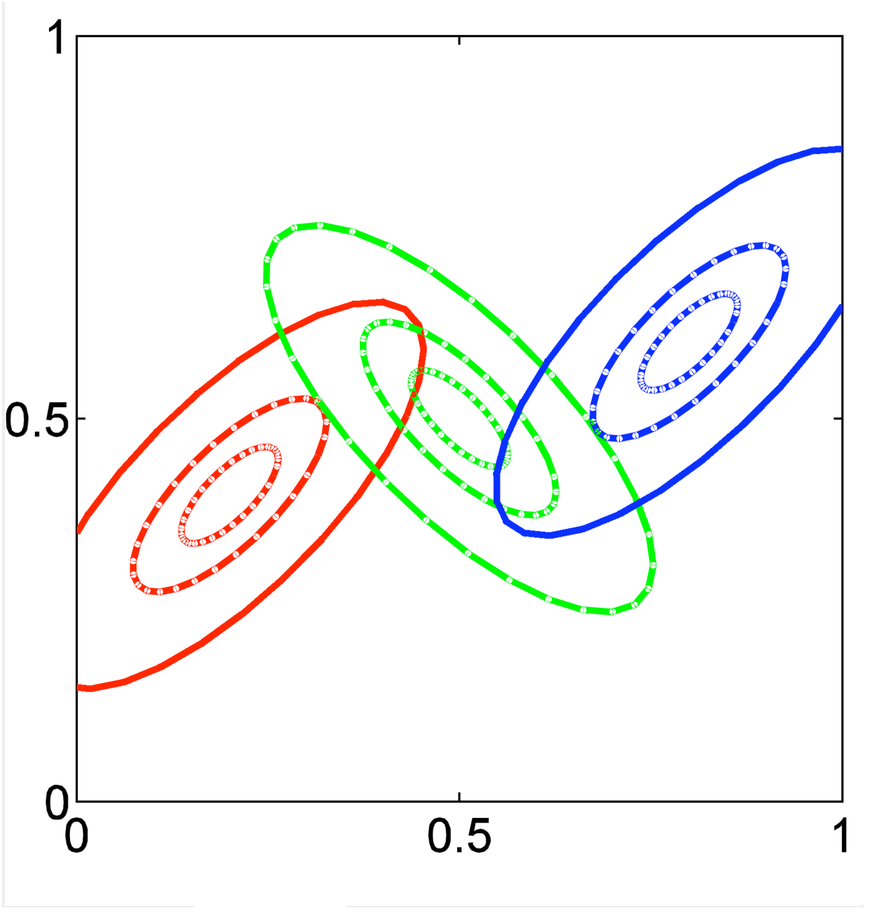
- While $|C| < n$:
  - For each C_i with more than 2 objects:
    - Apply partition-based clustering method to split C_i into two clusters C_j and C_k
    - $C = C - \{C_i\} \cup \{C_j, C_k\}$
- Example: spectral clustering

MODEL-BASED CLUSTERING

PROBABILISTIC MODEL-BASED CLUSTERING

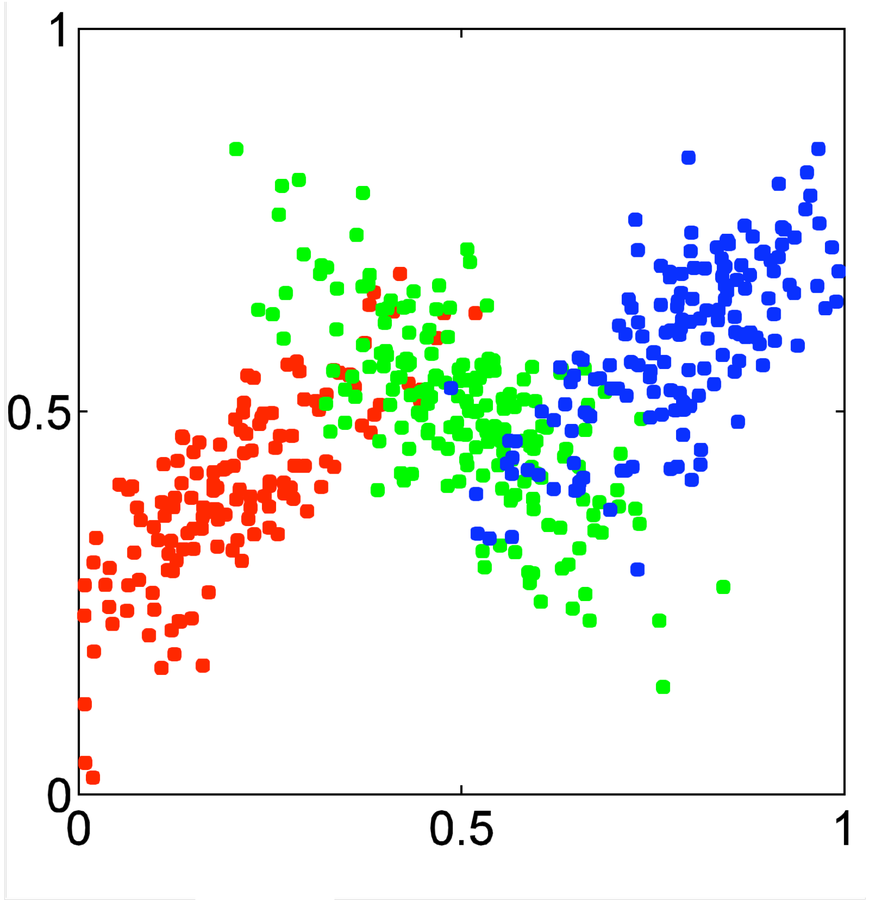
- Assumes a probabilistic model for each underlying cluster (component)
- Mixture model describes data as being generated from a weighted combination of component distributions (e.g., Gaussian)
- Generative process for data:
  - For each data point:
    - Select component *i* randomly based on component weights
    - Generate data point by sampling randomly from component *i*

GAUSSIAN MIXTURE MODEL

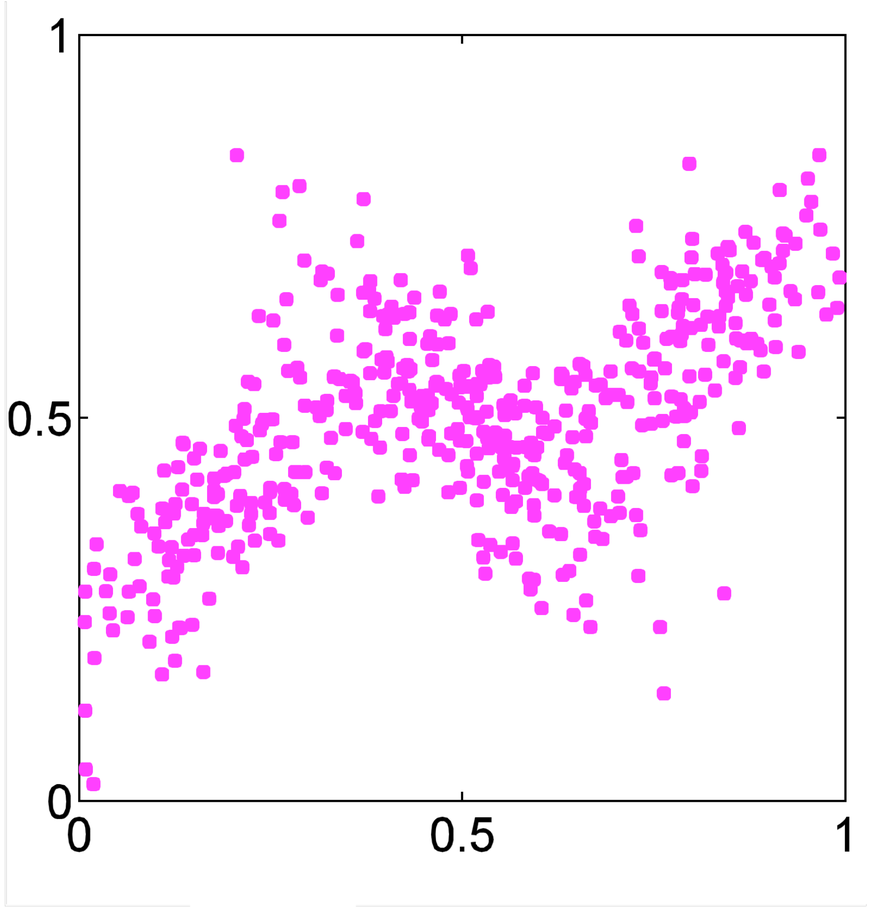


Mixture of three equi-probable Gaussians

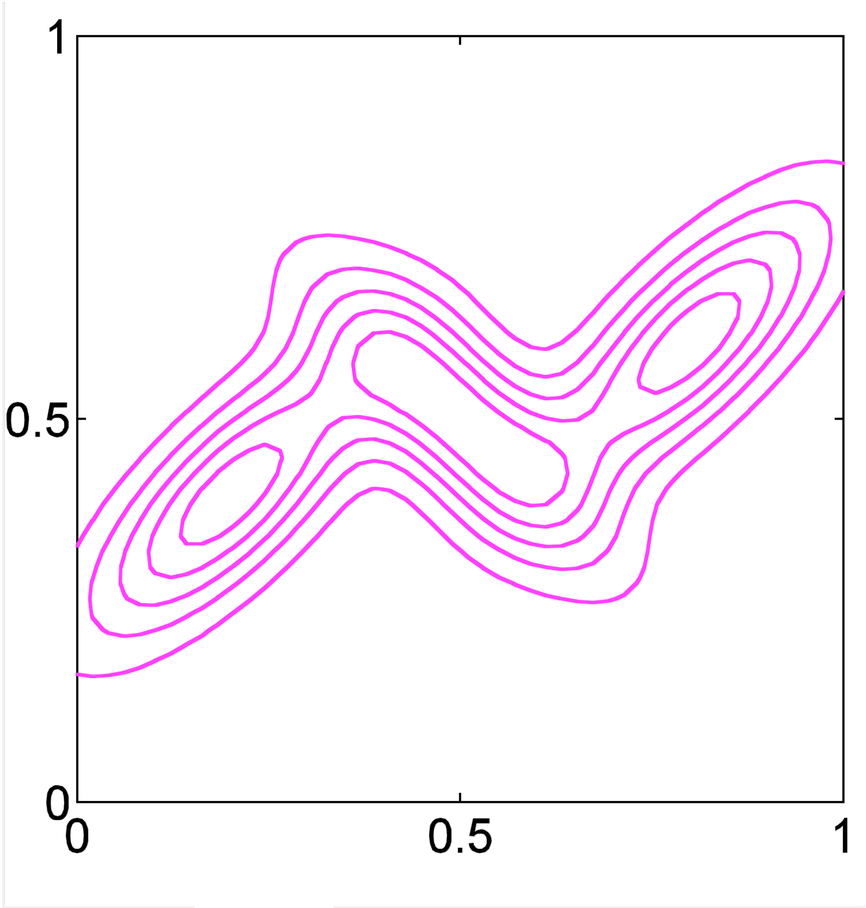
SAMPLE DATASET



UNLABELED DATASET



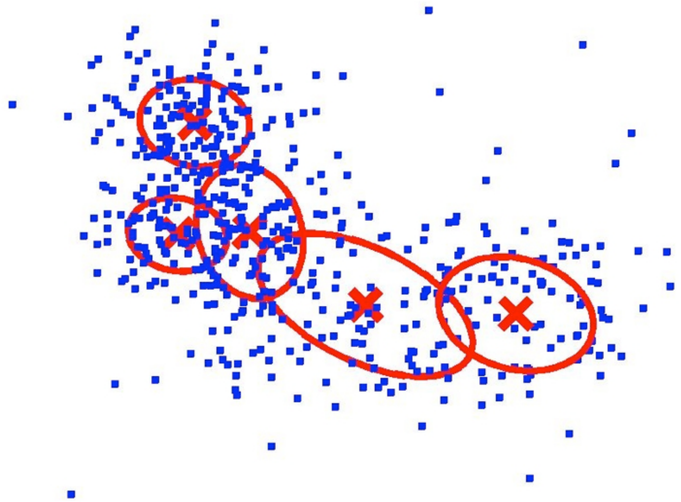
CONTOURS OF PROBABILITY DISTRIBUTION



Mixture of three Gaussians

PROBABILISTIC MIXTURE MODEL

- Instances represented as a weighted combination of *mixture* distributions



$$f(x) = \sum_{k=1}^{K} w_k f_k(x;\theta)$$

probability of observing x

likelihood of x being generated from cluster k

likelihood of point belonging to cluster k

GENERATIVE PROCESS (REVISITED)

- Assume that the data are generated from a mixture of K multi-dimensional Gaussians, where each component has parameters: $N_k(\mu_k, \Sigma_k)$
- For each data point:
  - Pick component Gaussian randomly with probability *p*(*k*)
  - Draw point from that Gaussian randomly by sampling from: $N_k(\mu_k, \Sigma_k)$

$$
\begin{aligned}
p(x) &= \sum_{k=1}^{K} p(k)p(x|k) \\
&= \sum_{k=1}^{K} p(k)p\Big(x|x \sim N(\mu_k, \Sigma_k)\Big)
\end{aligned}
$$

MULTIDIMENSIONAL GAUSSIAN

- A multi-dimensional Gaussians, for data with p dimensions is specified as follows

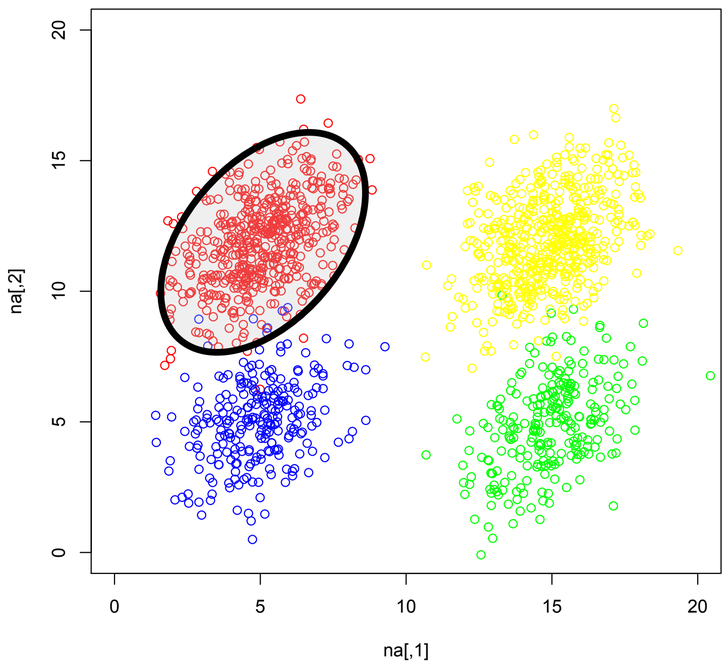
$$x \sim \mathcal{N}(\mu, \Sigma)$$

where:

$$\mu = \Big(E[X_1], ..., E[X_p]\Big)$$

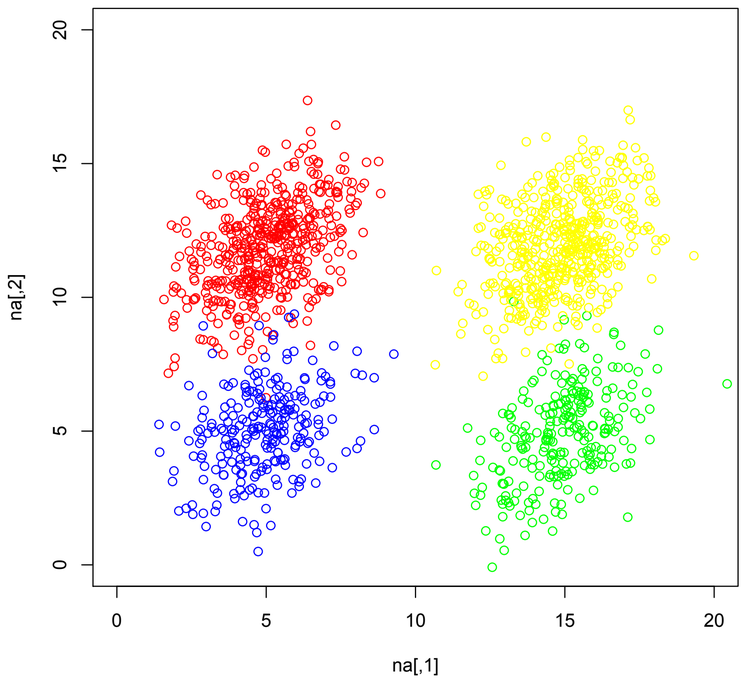
$$\Sigma = \begin{bmatrix} Var(X_1) & ... & Cov(X_1, X_p) \\ ... & ... & ... \\ Cov(X_1, X_p) & ... & Var(X_p) \end{bmatrix}$$

$$p(\mathbf{x}) = p(x_1, ..., x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right)$$



EXAMPLE GENERATIVE PROCESS

```
sigma <- matrix(c(2,1,1,3),2,2)
na=mvrnorm(n=500, c(5,12), sigma)
nb=mvrnorm(n=250, c(5,5), sigma)
nc=mvrnorm(n=250, c(15,5), sigma)
nd=mvrnorm(n=500, c(15,12), sigma)
d=rbind(na,nb,nc,nd)
plot(na,xlim=c(0,20),ylim=c(0,20),col='red')
points(nb,col='blue')
points(nc,col='green')
points(nd,col='yellow')
```



Parameters

$$p(k) = [0.333, 0.167, 0.167, 0.333]$$

$$\mu_1 = [5, 15], \mu_2 = [5, 5], \mu_3 = [15, 5], \mu_2 = [15, 12]$$

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{bmatrix} \qquad \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

LEARNING THE MODEL FROM DATA

- We want to invert this process
- Given the dataset, find the parameters
  - Mixing coefficients $p(k)$
  - Component means and covariance matrix $N_k(\mu_k, \Sigma_k)$
- Once the parameters are learned, we can decide the cluster membership of a data point using the Bayes rule
  - $p(c \mid x) = \dfrac{p(c)p(x \mid c)}{p(x)}$

HOW TO LEARN GMMS?

SCORE FUNCTION FOR GMM

- **Log likelihood** takes the following form (for model M={w,μ,Σ}):

$$
\begin{aligned}
log\, p(D|w,\mu,\Sigma) &= \sum_{n=1}^{N} log\, p(x_n|M) \\
&= \sum_{n=1}^{N} log\left[\sum_{k=1}^{K} p(x_n|k,M)P(k|M)\right] \\
&= \sum_{n=1}^{N} log\left[\sum_{k=1}^{K} w_k N(x_n|\mu_k,\Sigma_k)\right]
\end{aligned}
$$

- Note the sum over components is inside the log
- There is no closed form solution for the MLE

HIDDEN CLUSTER MEMBERSHIP VARIABLES

- Consider k cluster indicator variables for example x_n: $\mathbf{z_n} = [z_{n1}, \ldots, z_{nk}]$ which equals 1 for the cluster that x_n is a member of, and 0 otherwise
- If we knew the values of the hidden cluster membership variables (z) we could easily maximize the complete data log-likelihood, which has a closed form solution:

$$
\begin{aligned}
log\, p(D, \mathbf{z}|w, \mu, \Sigma) &= \sum_{n=1}^{N} log\Big[\sum_{k=1}^{K} z_{nk} \cdot w_k N(x_n|\mu_k, \Sigma_k)\Big] \\
&= \sum_{n=1}^{N} log\ \Big[w_{k'} N(x_n|\mu_{k'}, \Sigma_{k'})\Big] \quad \text{where } z_{nk'} \neq 0 \\
&= \sum_{n=1}^{N} log\ w_{k'} + \quad log\, N(x_n|\mu_{k'}, \Sigma_{k'}) \quad \text{where } z_{nk'} \neq 0
\end{aligned}
$$

- Unfortunately we don't know the values for the hidden variables!
- But, for given set of parameters we can compute the ***expected values*** of the hidden variables (cluster memberships)

POSTERIOR PROBABILITIES OF CLUSTER MEMBERSHIP

- We can think of the mixing coefficients as **prior** probabilities for cluster membership
- Then for a given example x_n, we can evaluate the corresponding **posterior** probabilities of **cluster membership** with Bayes theorem:

$$\gamma_k(x_n) \equiv p(z_{nk} = 1|x_n) = \frac{p(x_n|z_{nk} = 1)p(z_{nk} = 1)}{p(x_n)}$$

$$= \frac{w_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j N(x_n|\mu_j, \Sigma_j)}$$

cluster membership for x

EXPECTATION-MAXIMIZATION (EM) ALGORITHM

- Popular algorithm for parameter estimation in data with hidden/unobserved values
  - Hidden variables=cluster membership
- Basic idea
  - Initialize parameters
  - Predict values for hidden variables given current parameters
  - Estimate parameters given current prediction for hidden variables
  - Repeat

E STEP

M-STEP