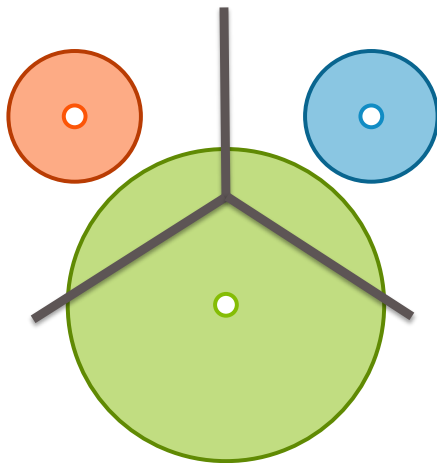


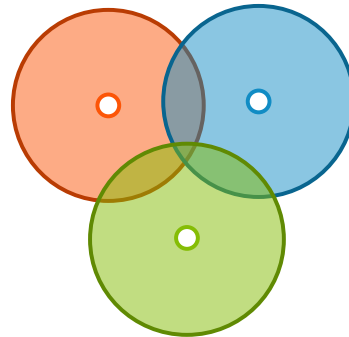
Mixture of Gaussians

CS229: Machine Learning
Carlos Guestrin
Stanford University

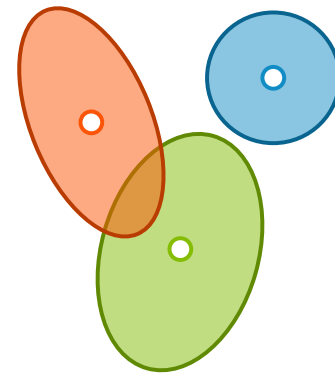
Failure modes of k-means



disparate cluster sizes

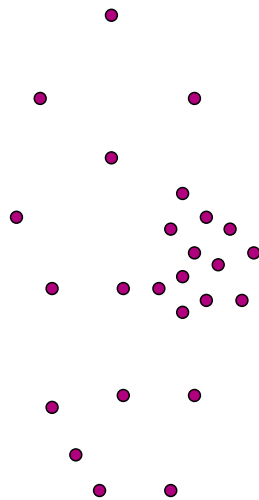


overlapping clusters



different
shaped/oriented
clusters

(One) bad case for k-means



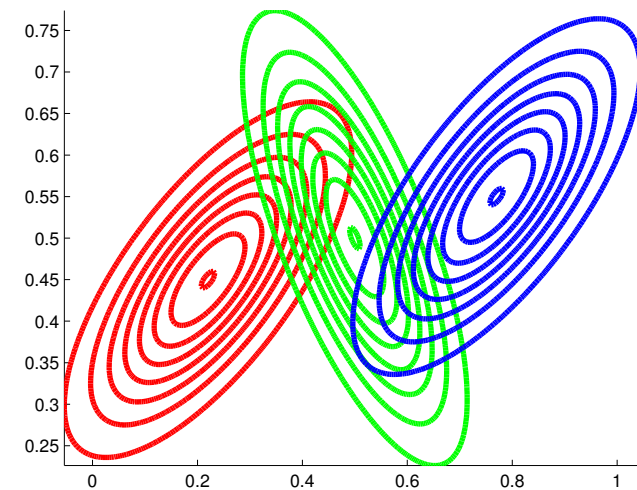
- Clusters may overlap
- Some clusters may be “wider” than others

Gaussians in m Dimensions

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right]$$

Suppose You Have a Gaussian For Each Class

$$\frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right]$$



Gaussian Bayes Classifier

- You have a Gaussian over \mathbf{x} for each class $y=i$:

$$\frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right]$$

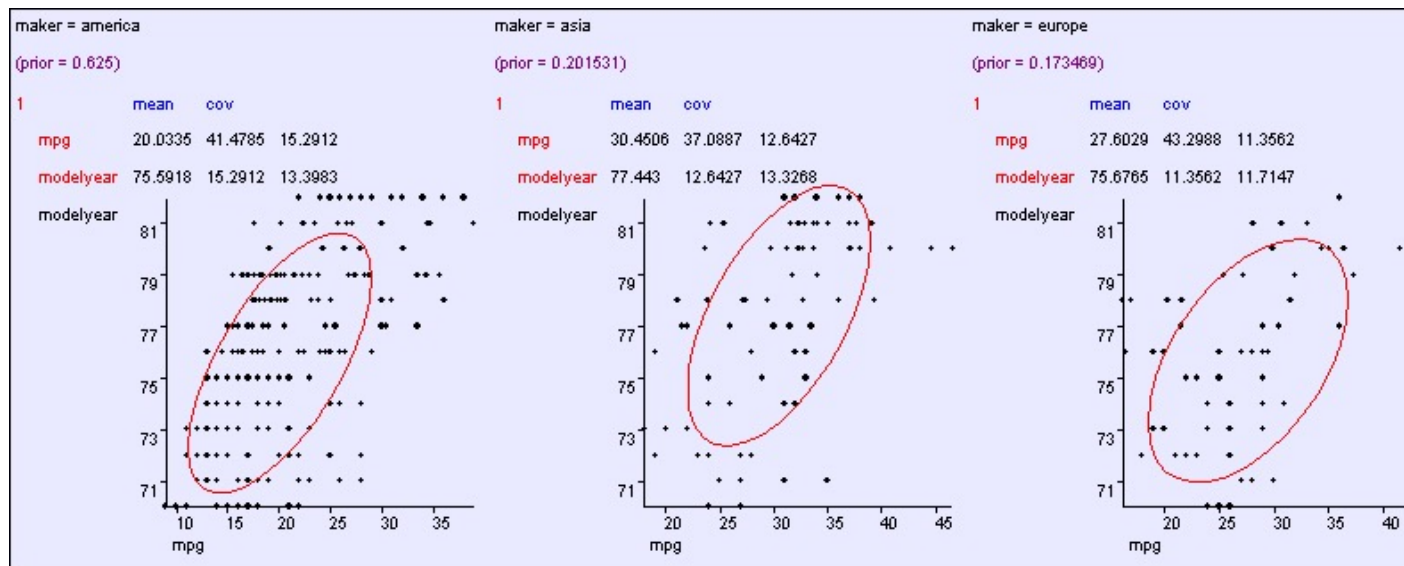
- But you need probability of class $y=i$ given \mathbf{x} :

- Thank you Bayes Rule!!

$$P(y = i | \mathbf{x}) = \frac{p(\mathbf{x} | y = i)P(y = i)}{p(\mathbf{x})}$$

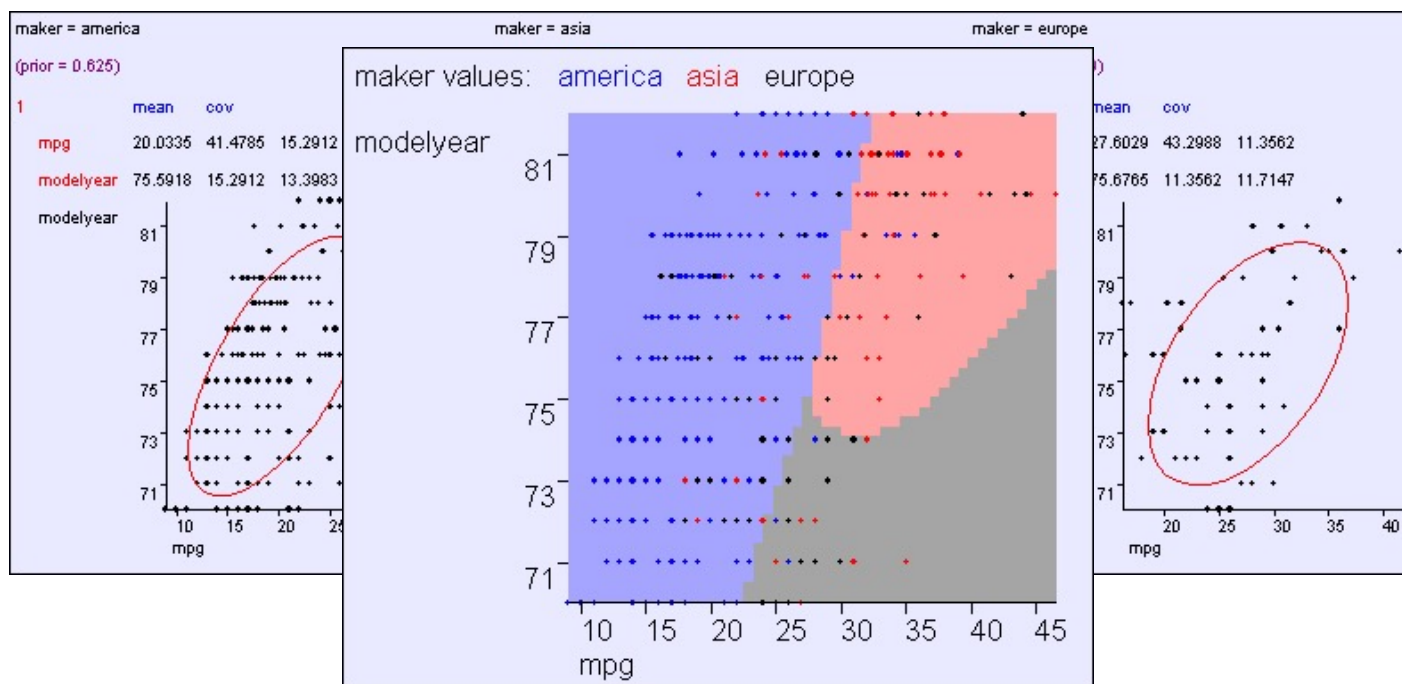
Learning modelyear , mpg
 ---> maker

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$



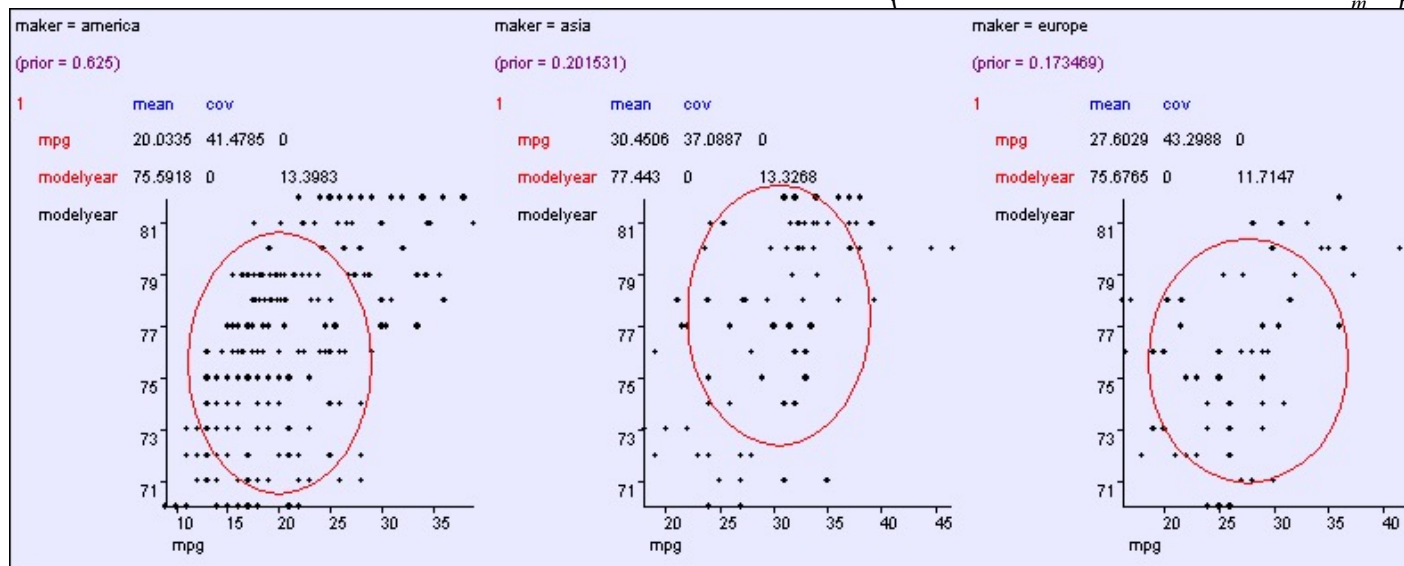
General: $O(m^2)$
parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$



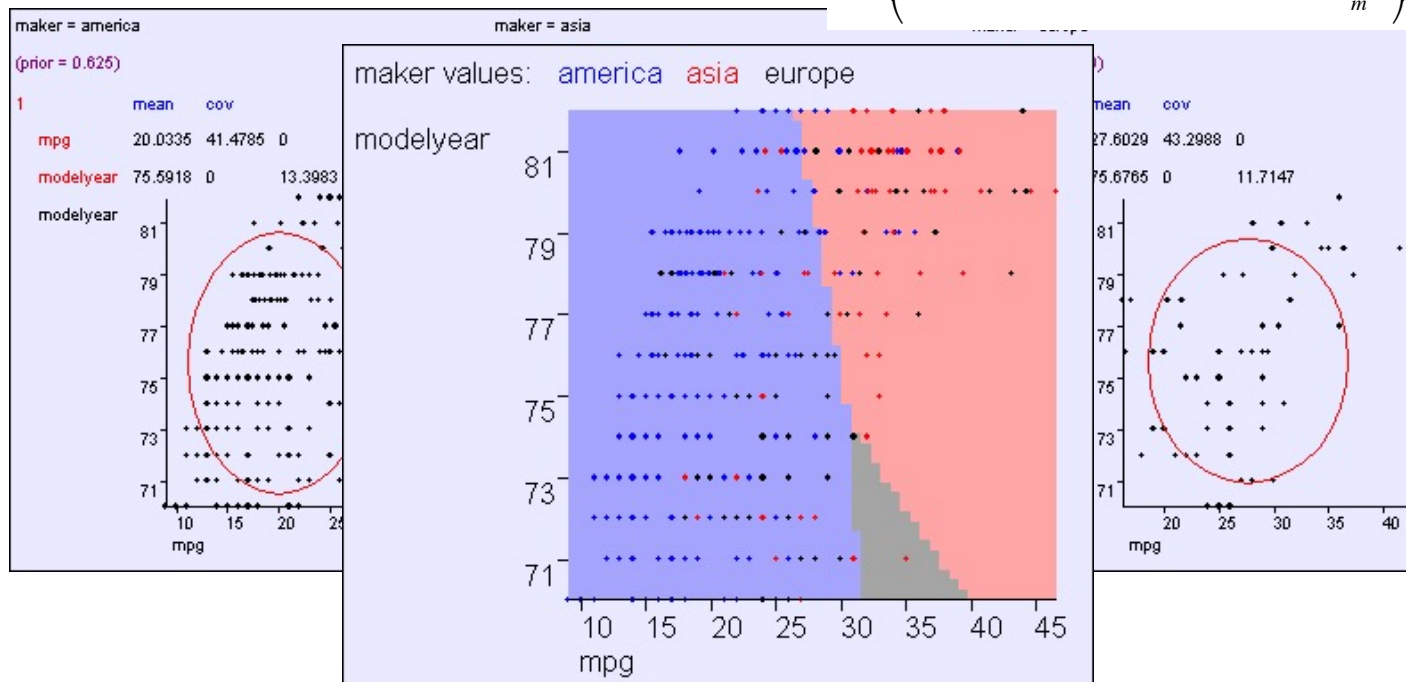
Aligned: $O(m)$
parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_{m-1}^2 & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma_m^2 \end{pmatrix}$$



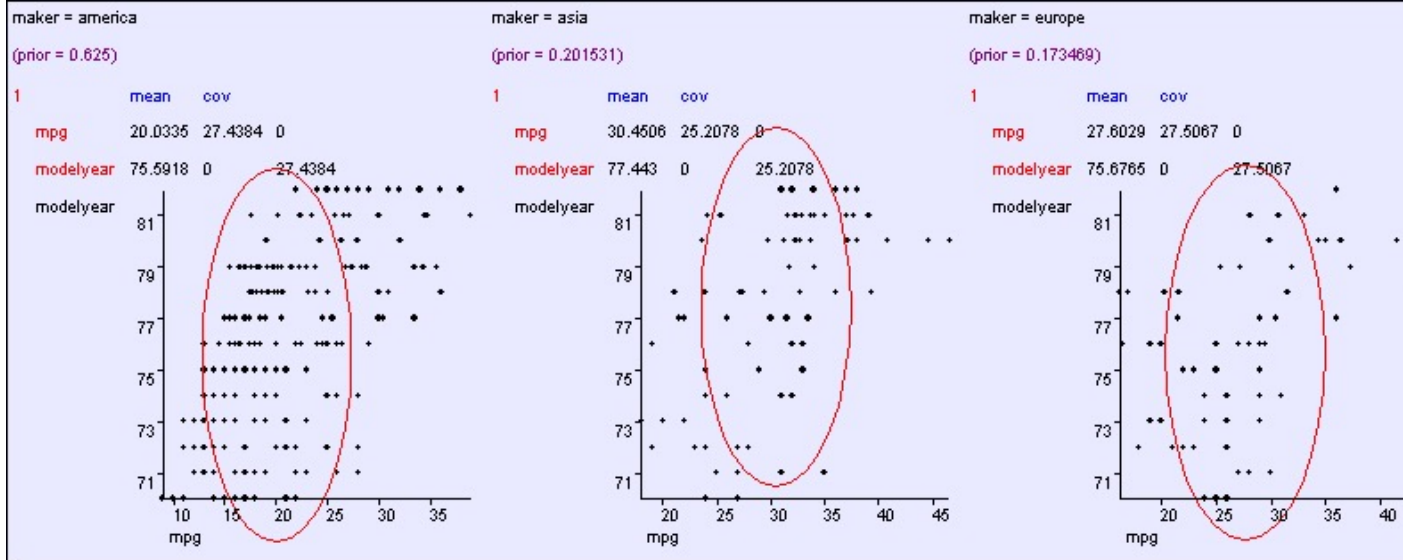
Aligned: $O(m)$
parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{m-1}^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma_m^2 \end{pmatrix}$$



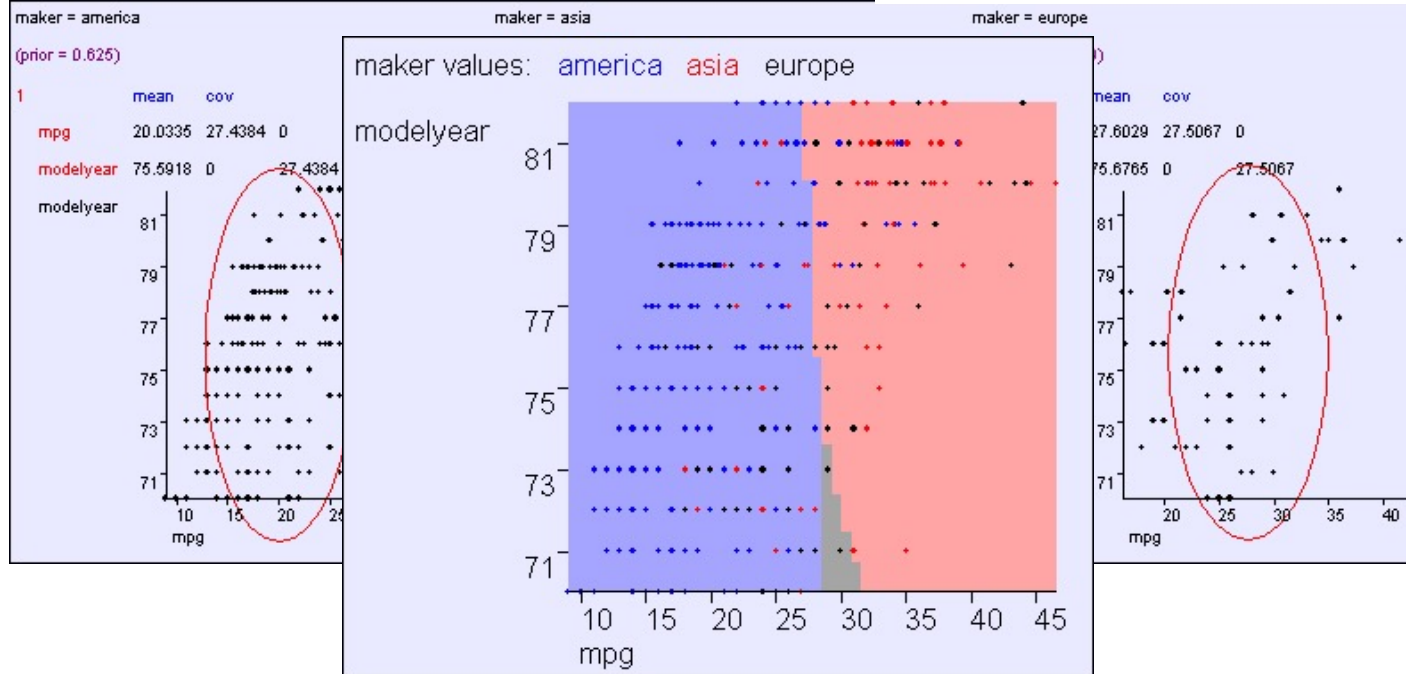
Spherical: $O(1)$ cov parameters

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix}$$



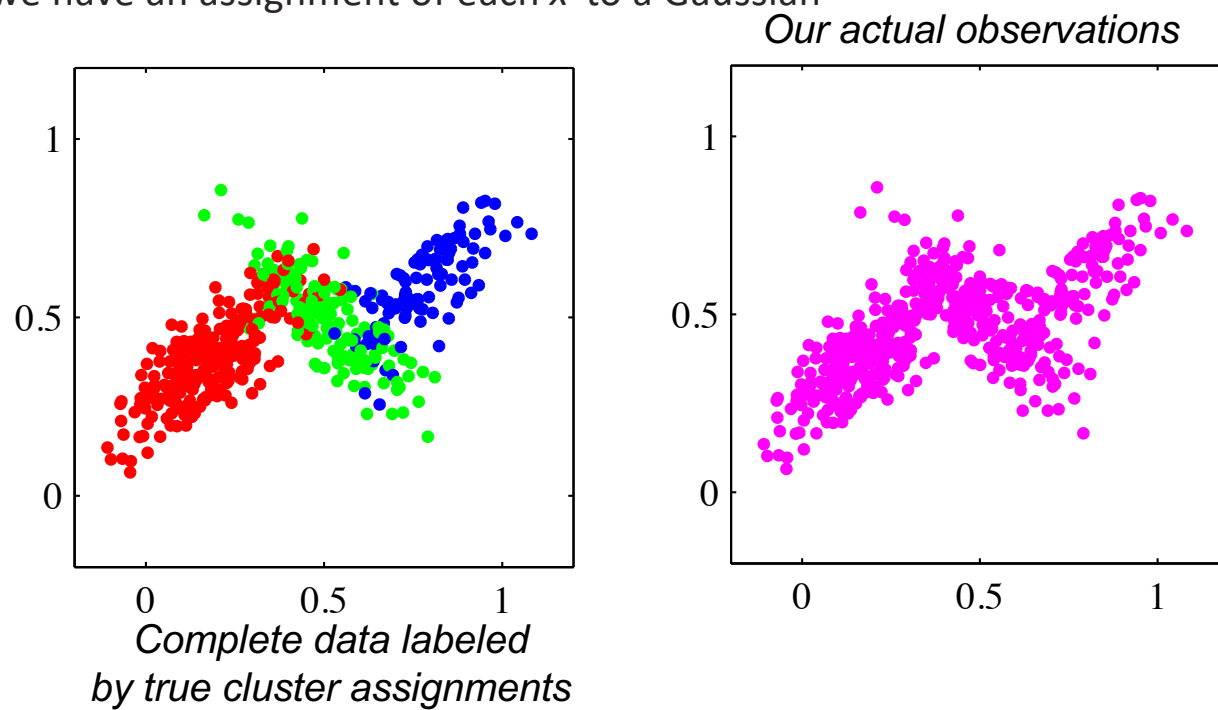
Spherical: $O(1)$ cov parameters

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma^2 \end{pmatrix}$$



Clustering our Observations

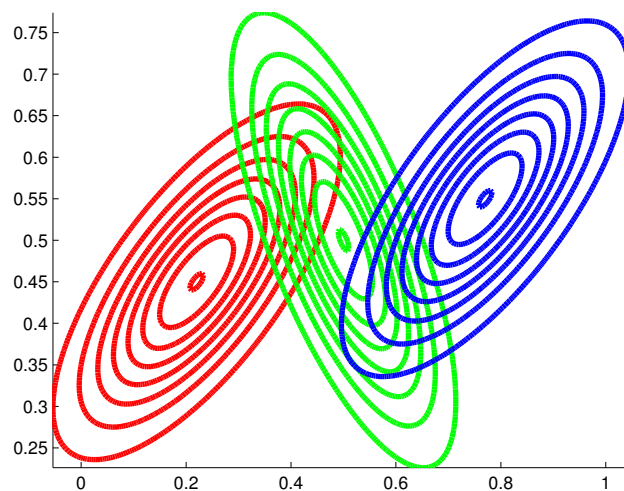
- Imagine we have an assignment of each x^i to a Gaussian



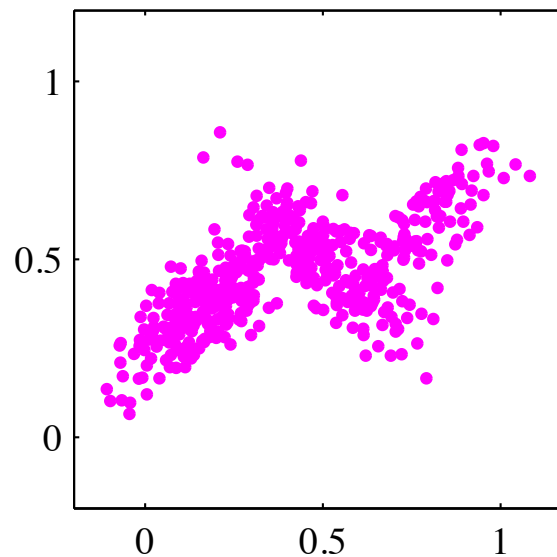
Density as Mixture of Gaussians

- Approximate with density with a mixture of Gaussians

Mixture of 3 Gaussians



Our actual observations

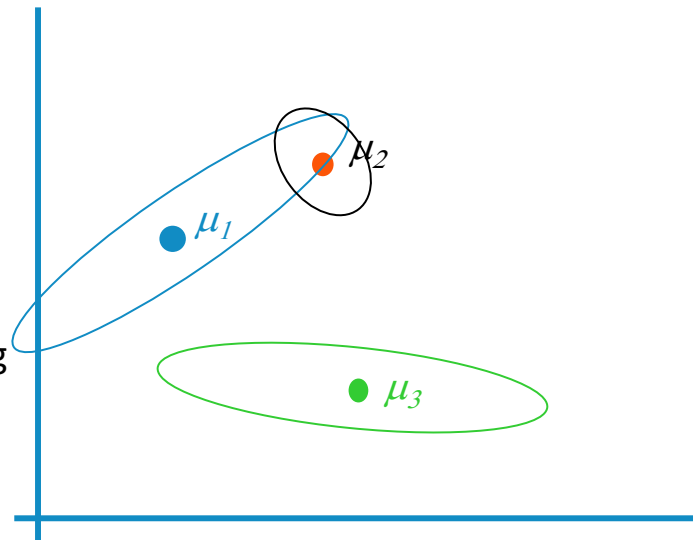


The General GMM assumption

- There are k components
- Component i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix Σ_i

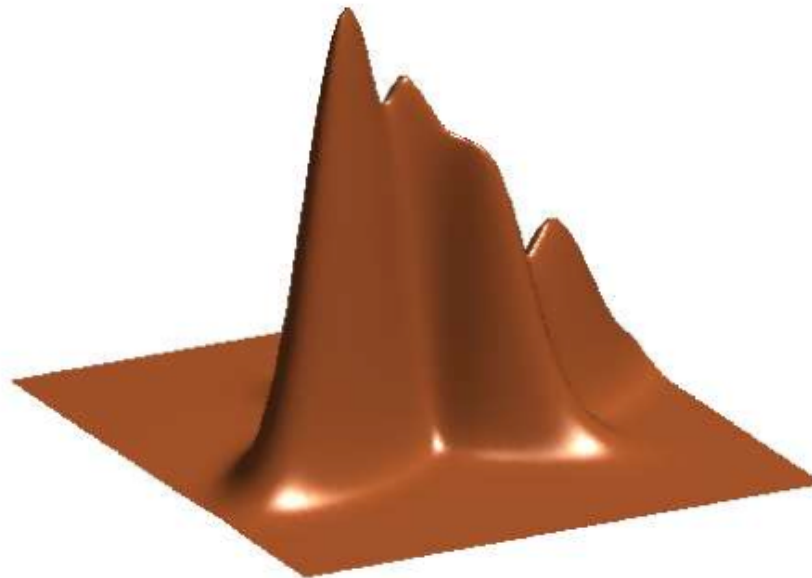
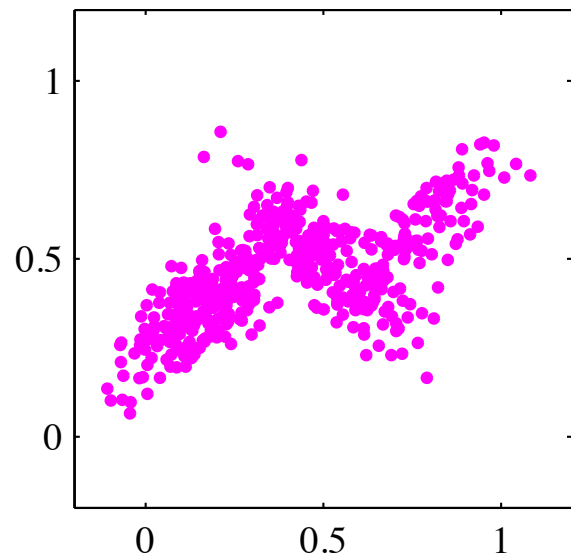
Each data point is generated according to the following recipe:

1. Pick a component at random:
Choose component i with probability $P(y=i)$
2. Datapoint $\sim N(\mu_i, \Sigma_i)$



Density Estimation

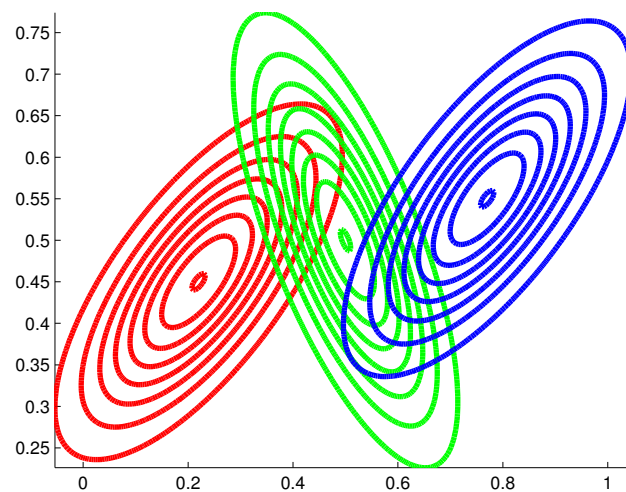
- Estimate a density based on x^1, \dots, x^N



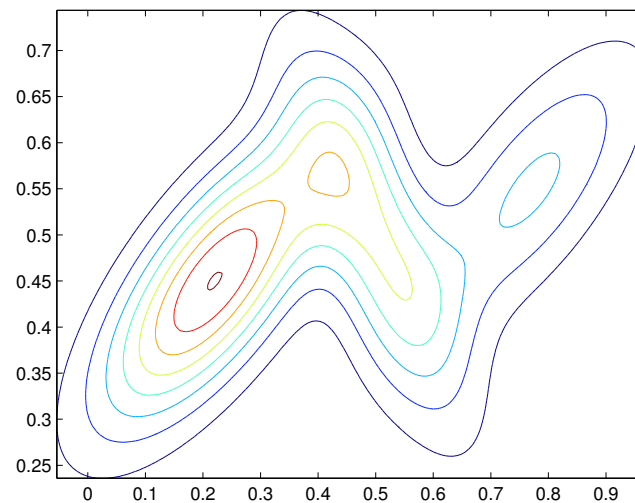
Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



Contour Plot of Joint Density



Summary of GMM Components

- Observations $x^i \in \mathbb{R}^d, \quad i = 1, 2, \dots, N$
- Hidden cluster labels $z_i \in \{1, 2, \dots, K\}, \quad i = 1, 2, \dots, N$
- Hidden mixture means $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \dots, K$
- Hidden mixture covariances $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \dots, K$
- Hidden mixture probabilities $\pi_k, \quad \sum_{k=1}^K \pi_k = 1$

Gaussian mixture marginal and conditional likelihood :

$$p(x^i | \pi, \mu, \Sigma) = \sum_{z^i=1}^K \pi_{z^i} p(x^i | z^i, \mu, \Sigma)$$

$$p(x^i | z^i, \mu, \Sigma) = \mathcal{N}(x^i | \mu_{z^i}, \Sigma_{z^i})$$