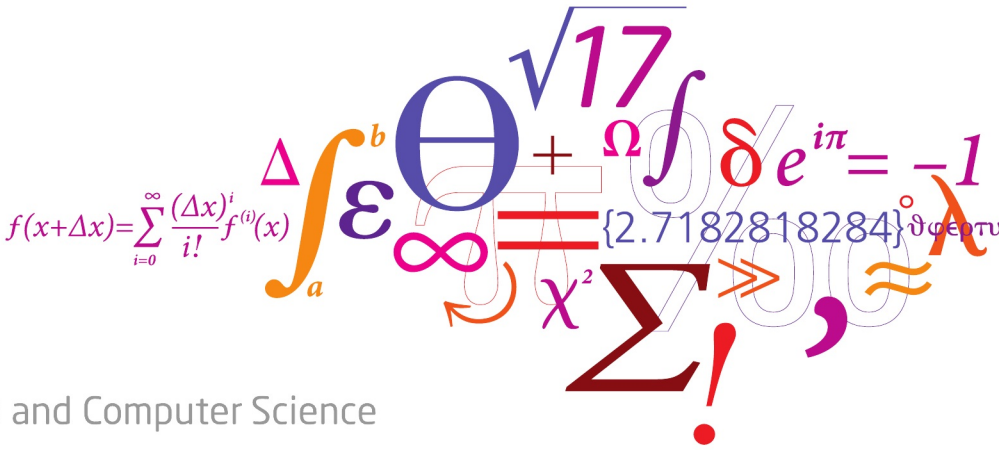
02450: Introduction to Machine Learning and Data Mining

Mixture models and density estimation

Tue Herlau

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

Remember you can evaluate the course on DTU inside.

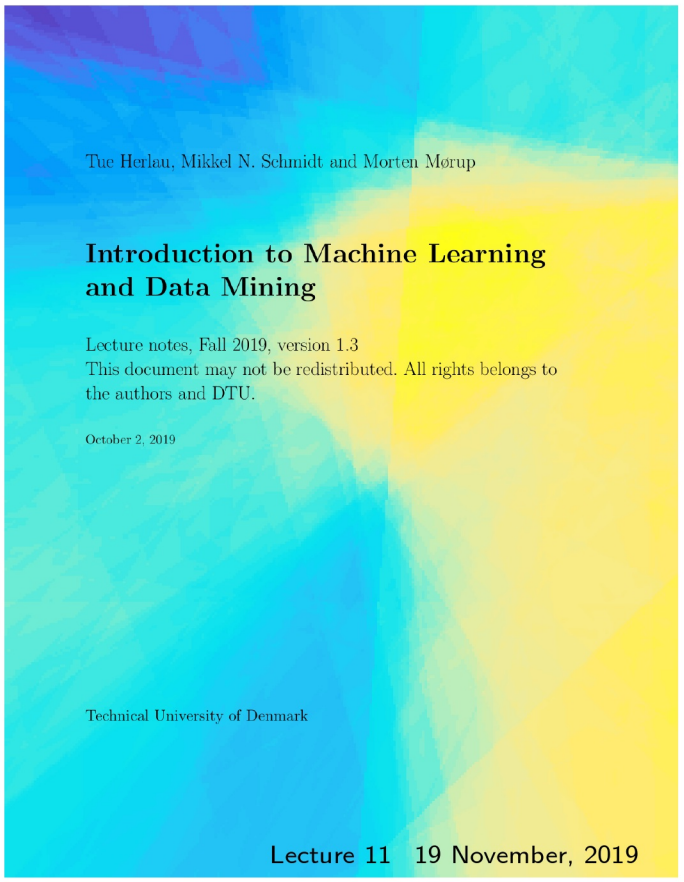
Today + Check exam on Inside.

Feedback Groups of the day:

Lau Halkier Wandall, Mabel-Brenda Ifeoma Okikaa, Jacob Pedersen, Carla Cristina Salazar Navalón, Silas El-Azm Stryhn, Yifei Xue, Anqi Yan, Mengge Hu, Naveen Karun Somasundaram, Carlos Parra, Ashwinth Mathivanan, Søren Bojer Jørgensen, Simon Majgaard, Joachim Secher, Michael Baand Severinsen, Sofie Betzer Rossen, Joakim Nilaus Hven, Frederik David Damsgaard Popp, Johannes Schou, Emma Dam Mortensen, Stefanie Wøhler Nielsen, Anna Bjerring Jensen, Sophia Troldborg Ohmann, Jens Damholt Richardt, Richard Thyssen Nørby Larsen, Niels Christian Dahl, Kasper Fibæk Schlosser, Lasse Pærgård Kristiansen, Jonas Weile, Evangelos Kerasidis, Allan Ozvan, Casper Chris Adriaan Bekkers, Michiel Goderie, Yannick Christopher Heijne, Maaike Elgersma, Maarten Marijn van Elst, Sean Christian Lindholm, Troels Lund, Michael Thorbøll Kristensen, Ali Jamal Jomeh, Jonathan Skovsholm Winther, Anthony Laye, Jimmy Xu, Andrew Nitu, Ruoyu Huang, Takuya Omori, Hui Jun Yap, Thor Larsen, Jakob Hammer Hedemann, Martin Illum, Hans Christian Bechsøfft Mikkelsen, Nicholas Rose, Elig Saraliev, Mikkel Nørgaard Schmidt, Mareva Ji Simon, Eskild Børsting Sørensen

Reading material:

Chapter 19, Chapter 20



Lecture Schedule

1. Introduction
   3 September: C1

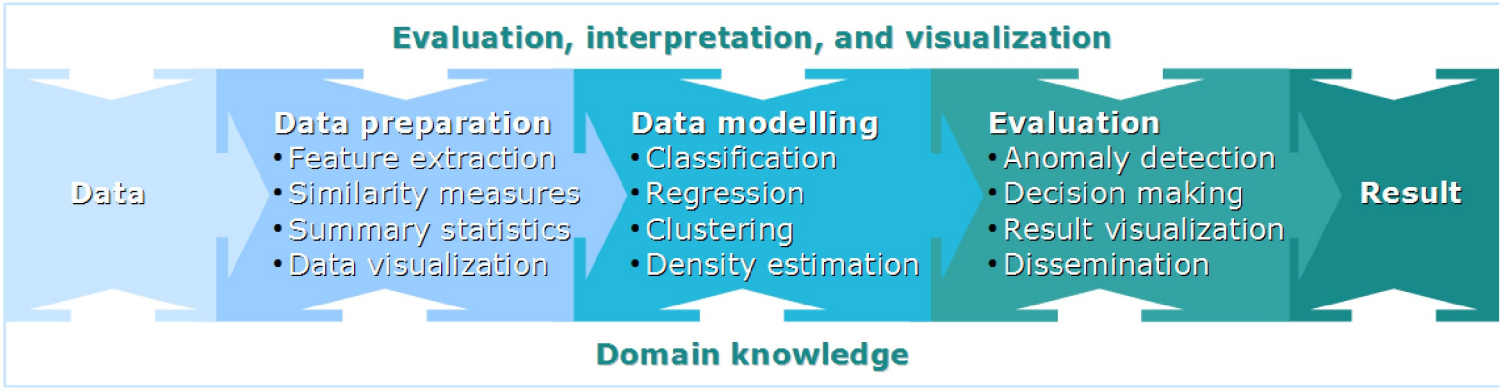
   Data: Feature extraction, and visualization
2. Data, feature extraction and PCA
   10 September: C2, C3
3. Measures of similarity, summary statistics and probabilities
   17 September: C4, C5
4. Probability densities and data Visualization
   24 September: C6, C7

   Supervised learning: Classification and regression
5. Decision trees and linear regression
   1 October: C8, C9 **(Project 1 due before 13:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   8 October: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   22 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   29 October: C14, C15
9. AUC and ensemble methods
   5 November: C16, C17

   Unsupervised learning: Clustering and density estimation
10. K-means and hierarchical clustering
    12 November: C18 **(Project 2 due before 13:00)**
11. **Mixture models and density estimation**
    19 November: C19, C20
12. Association mining
    26 November: C21

    Recap
13. Recap and discussion of the exam
    3 December: C1-C21 **(Project 3 due before 13:00)**

Piazza online help: `https://piazza.com/dtu.dk/fall2019/02450`
Videos/streaming of lectures: `https://video.dtu.dk`

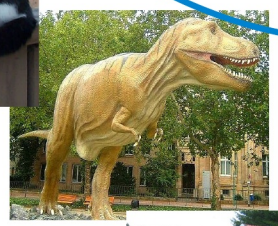


Learning Objectives

- Explain the role of the parameters in the Gaussian Mixture Model (GMM) and how the parameters are updated using the EM-algorithm
- Explain how cross-validation can be used for GMM
- Understand and apply kernel density, K-nearest neighbour density and average relative density estimation for outlier detection

Imagine you observe the world for the first time!

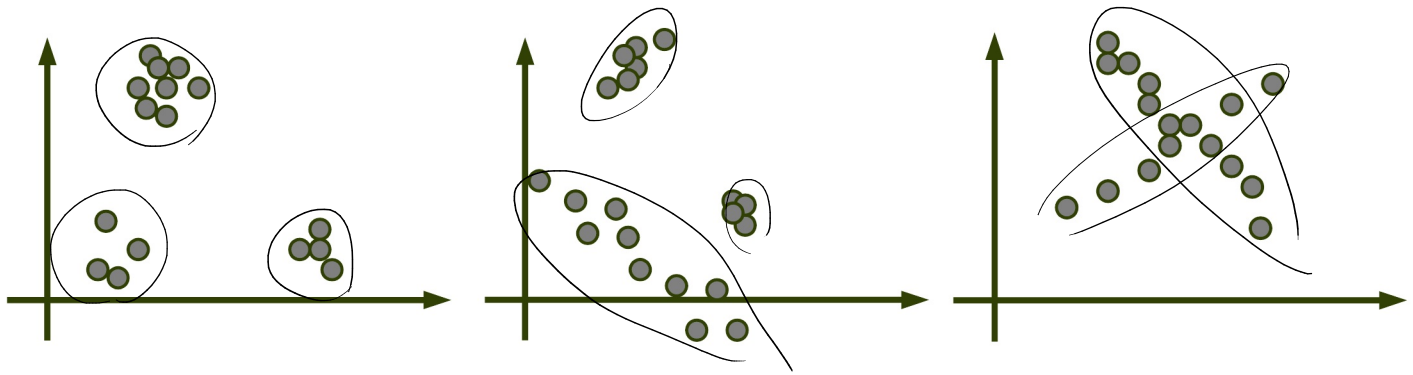
http://www.clipartlord.com/category/baby-clip-art/



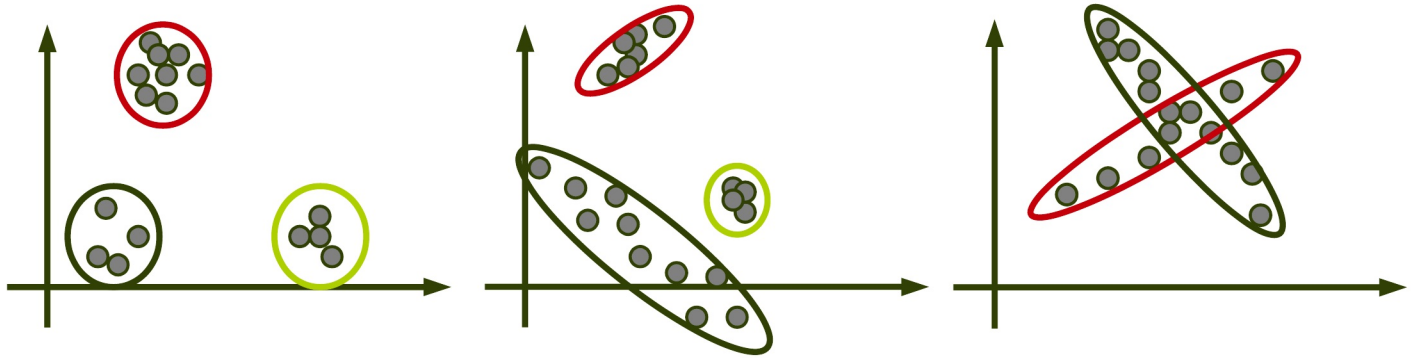
We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
http://commons.wikimedia.org/wiki/File:MasPiri-Astro-SVE.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg
http://commons.wikimedia.org/wiki/File:Cat002.jpg
https://commons.wikimedia.org/wiki/Dog#/media/File:BluetickCoonhound.jpg
http://commons.wikimedia.org/wiki/File:Saurier2.jpg

- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
  - Will it always find the optimum solution?
  - Can it model the sizes of the clusters?
  - Can it model the shape of the clusters?
  - How can we determine the number of clusters?



- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
  - Will it always find the optimum solution?
  - Can it model the sizes of the clusters?
  - Can it model the shape of the clusters?
  - How can we determine the number of clusters?

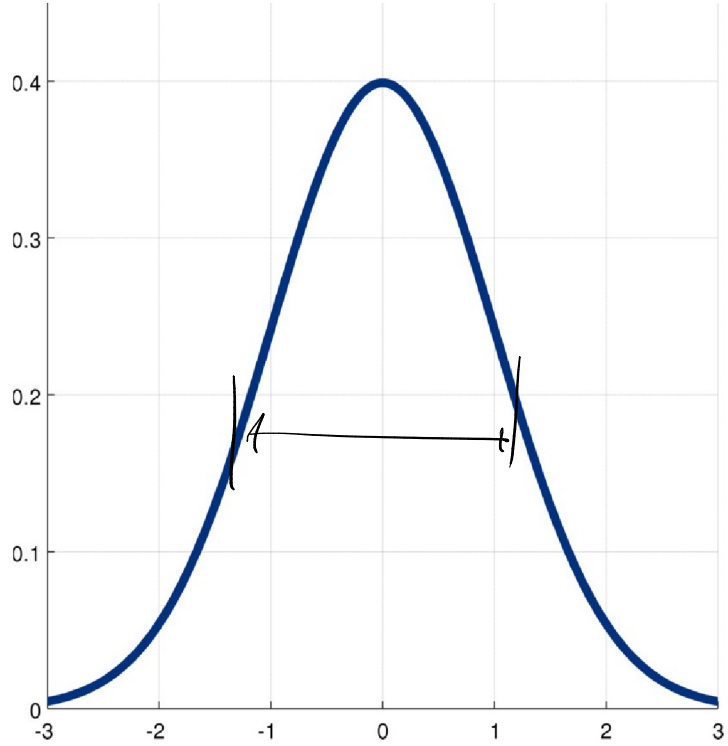


Normal distribution

- Probability density function describes the relative chance of a given value to occur

- Normal distribution characterized by
  - Mean
  - Variance

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



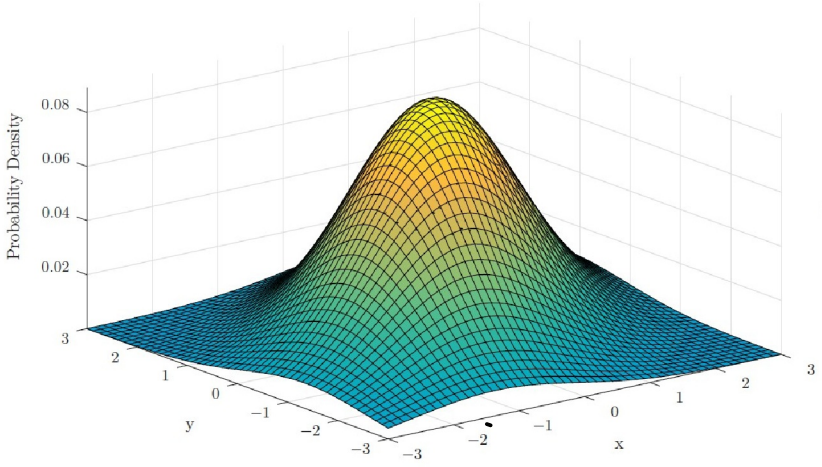
Multivariate Normal distribution

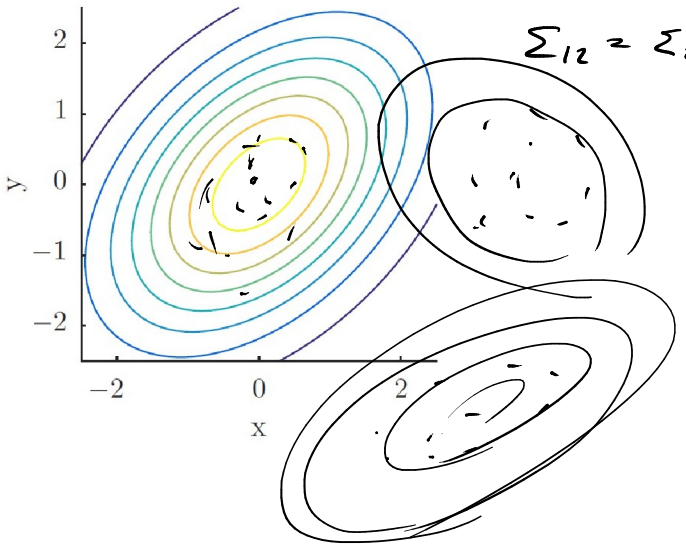
$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$p(x_1, x_2)$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$\Sigma_{12} = \Sigma_{21} > 0$





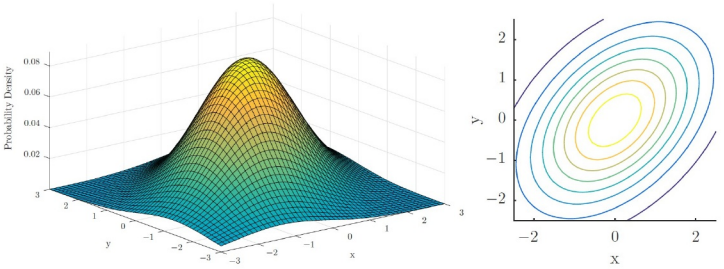
Multivariate Normal distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

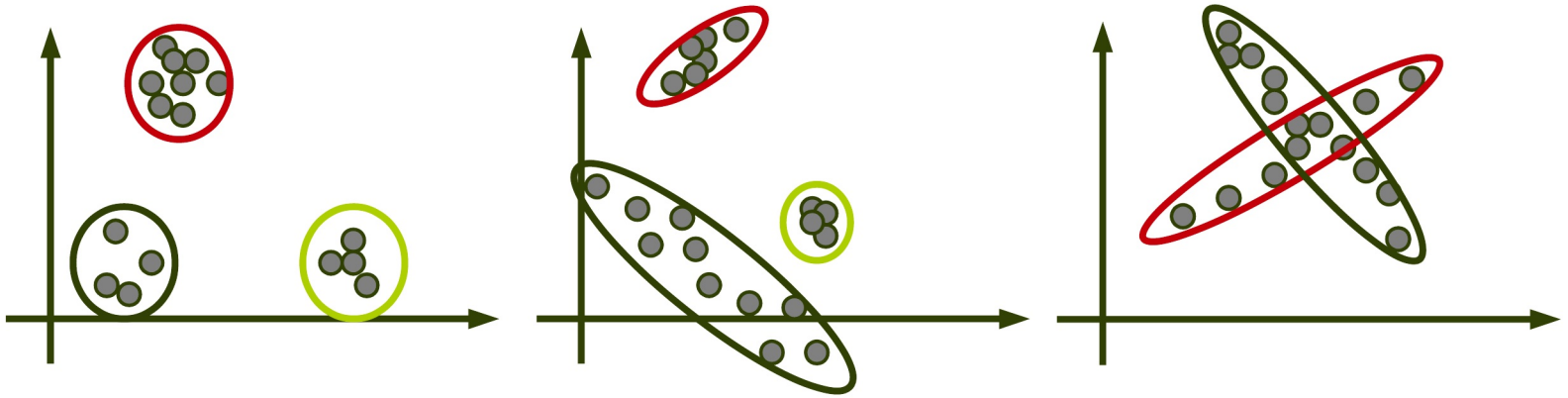
- Example: 2-dimensional Normal distribution

$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{bmatrix}$$



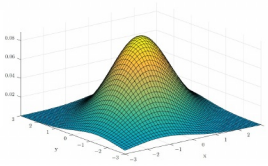
Prototypical mixture model



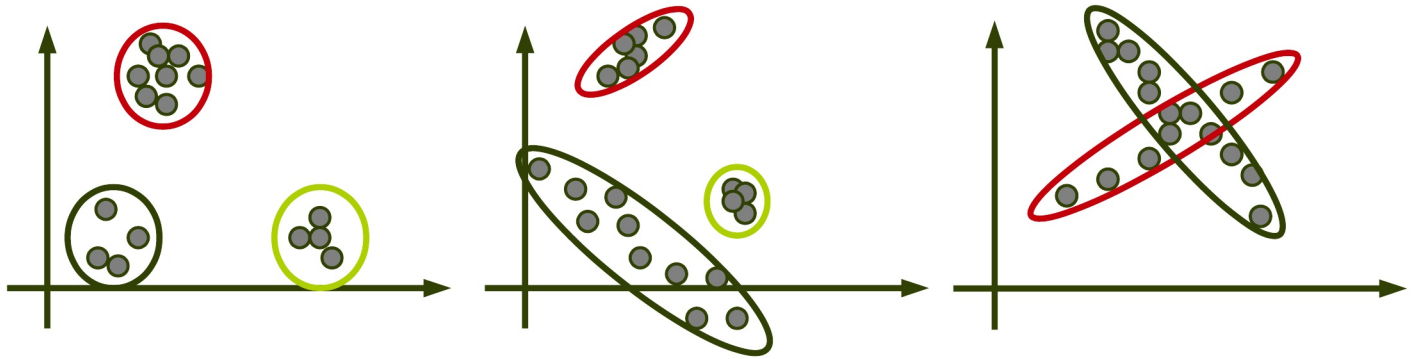
- We want a **density** $p(\boldsymbol{x})$ of our observations $\boldsymbol{x} \in \mathbb{R}^M$
- Suppose we have K clusters and let $z = k$ if $\boldsymbol{x}$ belongs to cluster k
- According to the the basic rules of probability:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}, z = k) = \sum_{k=1}^{K} p(\boldsymbol{x}|z = k)p(z = k)$$

- If we specify $p(\boldsymbol{x}|z = k)$ and $p(z = k) = w_k$ we have a model



The Gaussian Mixture Model (GMM)



- Different locations $\boldsymbol{\mu}_{(k)}$
- Different shape $\boldsymbol{\Sigma}_{(k)}$
- Different sizes w_k

Data density

Sum of cluster specific densities assumed normal distributed

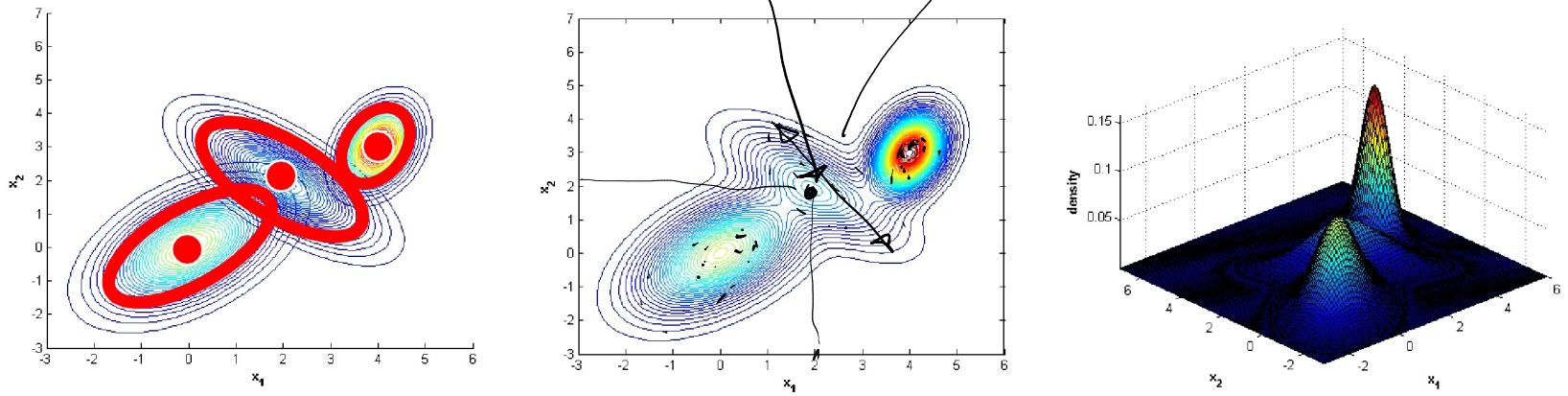
$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x} | \mu_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

$$\text{s.t.} \sum_{k=1}^{K} w_k = 1, \; w_k \geq 0$$

GMM example

$k = 3$, $w_1 = 0.5$, $w_2 = 0.2$, $w_3 = 0.3$, $\mu_{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, ...

$$p(\boldsymbol{x}) = 0.5\mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}) + 0.2\mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}) + 0.3\mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.5 \end{bmatrix})$$



$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative size/density of the cluster

Quiz 01 (please answer on Piazza): GMM

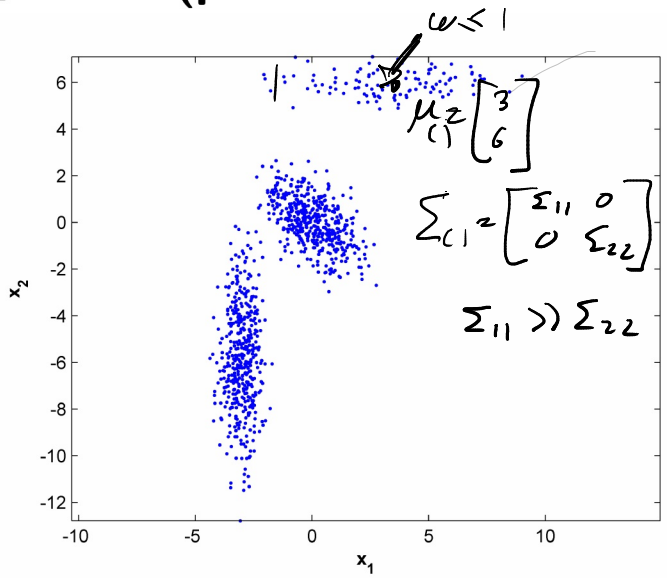


Figure 1: 1000 data observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

In Figure 1 is shown 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Suppose $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is the multivariate normal distribution, which one of the following GMM densities was used to generate the data?

A
$$p(x) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

B
$$p(x) = 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

C
$$p(x) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

D
$$p(x) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$$

E Don't know.

Solution:

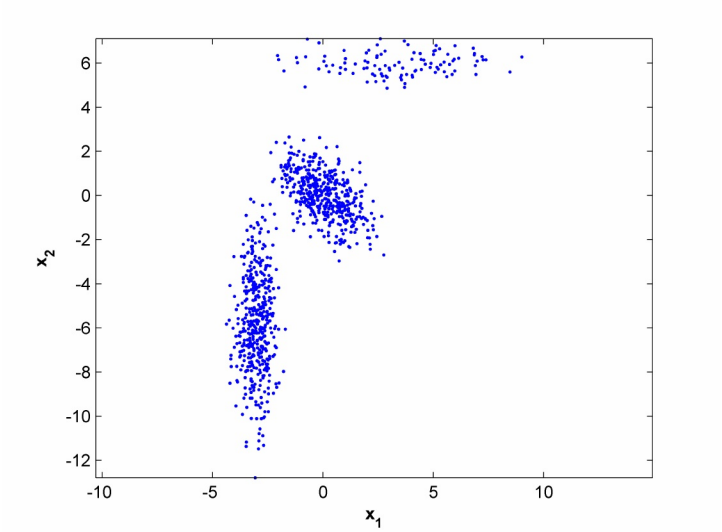


Figure 1: 1000 data observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

The centroids of the clusters are $\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The cluster at $\begin{bmatrix} 3 \\ 6 \end{bmatrix}$ is not very dense and should therefore have the coefficient 0.1, it further has a large spread in the x_1 direction and small spread in the x_2 direction corresponding to a covariance matrix of $\begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}$, thus, answer option one is the only correct answer.

Sanity check time:

- Consider the Gaussian mixture model (GMM)

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\, \mu_{(k)}, \mathbf{\Sigma}_{(k)}) \quad \text{s.t.} \sum_{k=1}^{K} w_k = 1,\ w_k \geq 0$$

- What is the value of the integral?

$$1 = \int p(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \sum_{k=1}^{K} w_k \, N(x \mid \mu_k, \Sigma_k) \, dx = \sum_k w_k \int N(x \mid \mu_k, \Sigma_k) \, dx.$$

$$= \sum w_k = 1$$

Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

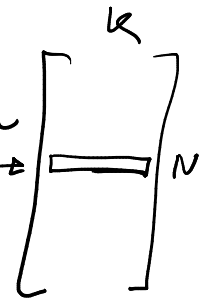
Until the parameters do not change

$n = 1 \ldots N$: observation nr.
x_n : observation n
z_n : which cluster x_n is assigned to.

E-step

$$p(z_n = k|x_n) = \frac{w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}{\sum_{K=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}$$

$p(z_n = k|x_n) \sim$
$p(z_n = \ldots | x_n) \rightarrow$



$n = 1 \ldots N$.

M-step

$$N_k = \sum_{n=1}^{N} p(z_n = k|x_n)$$

$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} \boldsymbol{x}_n p(z_n = k|\boldsymbol{x}_n)$$

$$w_k = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})(\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})^\top p(z_n = k|\boldsymbol{x}_n)$$

$k = 1 \ldots K$.

The GMM update rules approximately implements

$w_k \mathcal{N}(x \mid \mu_k, \Sigma_k) = p(z=k)\, p(x \mid z=k) = p(z=k, x)$

$$w_k = p(z = k)$$
$$\mu_{(k)} = \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[\boldsymbol{x}\right]$$
$$\boldsymbol{\Sigma}_{(k)} = \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\top\right]$$

E-step

$$p(\boldsymbol{z}_n = k|\boldsymbol{x}_n) = \frac{w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}{\sum_{K=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})} \approx \frac{p(z=k, x)}{p(x)} = p(z=k \mid x)$$

$= p(z)$

M-step

$$N_k = \sum_{n=1}^{N} p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} \boldsymbol{x}_n p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

$$w_k = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})(\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})^\top p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

$$w_k = \frac{N_k}{N} = \frac{1}{N} \sum_{n=1}^{N} p(z_n = k \mid x_n)$$
$$= \sum_{n=1}^{N} p(x_n)\, p(z_n = k \mid x_n)$$
$$= \sum_{n=1}^{N} p(x_n, z_n = k)$$
$$= p(z = k)$$

The GMM update rules approximately implements

$$w_k = p(z = k)$$
$$\mu_{(k)} = \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[\boldsymbol{x}\right]$$
$$\boldsymbol{\Sigma}_{(k)} = \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\top\right]$$

E-step

$$p(\boldsymbol{z}_n = k|\boldsymbol{x}_n) = \frac{w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}{\sum_{K=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}$$

M-step

$$N_k = \sum_{n=1}^{N} p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$
$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k}\sum_{n=1}^{N} \boldsymbol{x}_n p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$
$$w_k = \frac{N_k}{N}$$
$$\boldsymbol{\Sigma}_{(k)} = \frac{1}{N_k}\sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})(\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})^\top p(\boldsymbol{z}_n = k|\boldsymbol{x}_n)$$

$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k}\sum_{n=1}^{N} \boldsymbol{x}_n p(z_n = k|\boldsymbol{x}_n)$$

$$= \frac{1}{\frac{N_k}{N}} \frac{1}{N}\sum_{n=1}^{N} x_n\, p(z_n = k|x_n)$$

$$= \frac{1}{w_k}\sum_{n=1}^{N} p(x_n)\, x_n\, p(z_n = k|x_n)$$

$$= \sum_{n=1}^{N} x_n \frac{p(z_n = k|x_n)\, p(x_n)}{p(z_n = k)}$$

$$= \sum_{n=1}^{N} x_n\, p(x_n|z_n = k) = \mathbb{E}_{p(x|z_n=k)}[x]$$

Derivation

The GMM update rules approximately implements

$$
\begin{aligned}
w_k &= p(z = k) \\
\mu_{(k)} &= \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[\boldsymbol{x}\right] \\
\boldsymbol{\Sigma}_{(k)} &= \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\top\right]
\end{aligned}
$$

Derivation:

$$
\begin{aligned}
w_k &= \frac{N_k}{N} \\
&= \sum_{n=1}^{N} p(z_n = k|\boldsymbol{x}_n)\frac{1}{N} \\
&= \sum_{n=1}^{N} p(z_n = k|\boldsymbol{x}_n)p(\boldsymbol{x}_n) \\
&= \sum_{n=1}^{N} p(z_n = k, \boldsymbol{x}_n) \\
&= p(z = k)
\end{aligned}
$$

Derivation

The GMM update rules approximately implements

$$
\begin{aligned}
w_k &= p(z = k) \\
\mu_{(k)} &= \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[\boldsymbol{x}\right] \\
\boldsymbol{\Sigma}_{(k)} &= \mathbb{E}_{p(\boldsymbol{x}|z=k)}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\top\right]
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\mu}_{(k)} &= \frac{1}{N_k}\sum_{n=1}^{N} \boldsymbol{x}_n p(z_n = k|\boldsymbol{x}_n) \\
&= \frac{1}{\frac{N_k}{N}}\sum_{n=1}^{N} \boldsymbol{x}_n p(z_n = k|\boldsymbol{x}_n)\frac{1}{N} \\
&= \sum_{n=1}^{N} \boldsymbol{x}_n \frac{p(z = k)p(\boldsymbol{x}_n)}{p(z = k)} \\
&= \sum_{n=1}^{N} \boldsymbol{x}_n \frac{p(z = k, \boldsymbol{x}_n)}{p(z = k)} \\
&= \sum_{n=1}^{N} \boldsymbol{x}_n p(\boldsymbol{x}_n|z = k)
\end{aligned}
$$

Quiz 02 (please answer on Piazza): GMM

14:05

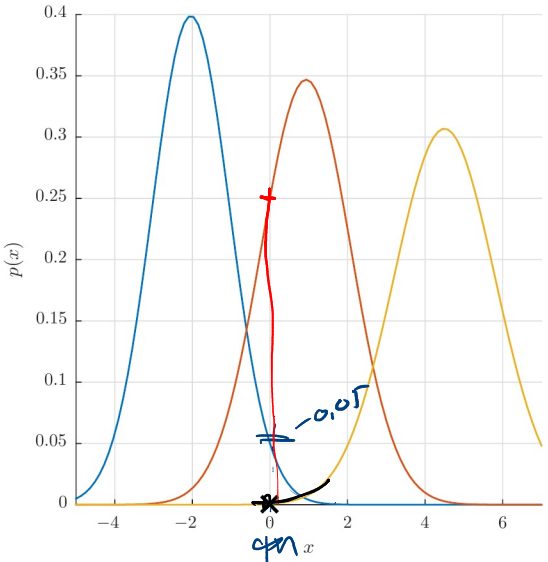


Figure 1: Mixture components in a GMM mixture model with $K = 3$

Consider a 1D GMM mixture model where each of the $K = 3$ (Gaussian) mixture components are illustrated in Figure 1 as the colored curves and the figure also shows a new observation indicated by the cross. Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step (i.e. assuming the mixture components has the means and variances indicated by Figure 1 and equal weights). According to the EM algorithm, what is the (approximate) probability the black cross is assigned to the blue (left-most) mixture component?

A. 0.05

B. 0.17

C. 0.25

D. 0.02

E. Don't know.

$r_n \approx 0$.

$N(x_n \mid z_n = k)$

$p(z_n = \text{"blue"} \mid x_n)$ ✓ $\quad \omega_n = \frac{1}{3}$

$= \dfrac{p(x_n \mid z_n = \text{"blue"})\, p(z_n = \text{"blue"})}{\left(p(x_n \mid z_n = \text{"blue"})\, p(z_n = \text{"blue"}) + p(x_n \mid z_n = \text{red})\, p(\text{red}) + p(x_n \mid z_n = y)\, p(y)\right)}$

$= \dfrac{0.05 \cdot \frac{1}{3}}{0.05 \cdot \frac{1}{3} + 0.25 \cdot \frac{1}{3} + 0} = \dfrac{1}{1+5}$

$= \frac{1}{6}$.

Solution:

The probability of the black cross under each of the three mixture components can be read of as approximately $p(x_0|\mu_1, \sigma_1) \approx 0.05$, $p(x_0|\mu_2, \sigma_2) \approx 0.25$, $p(x_0|\mu_3, \sigma_3) \approx 0$. Since they are weighted equally the assignment to the left-most component is

$$p(z = 1|x_0) = \frac{\frac{1}{3}p(x_0|\mu_1, \sigma_1)}{\sum_{i=1}^{3} \frac{1}{3}p(x_0|\mu_i, \sigma_i)} \approx \frac{0.05}{0.05 + 0.25} = 0.17$$

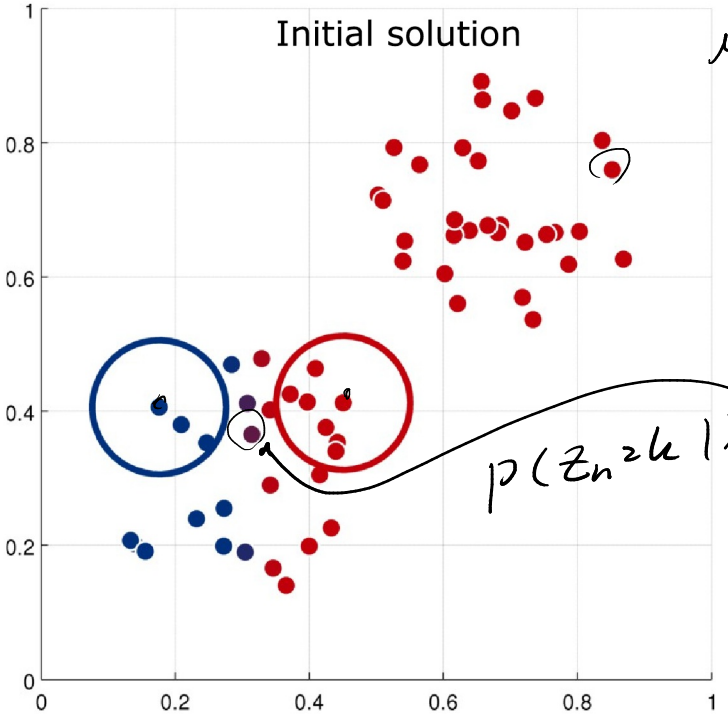
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



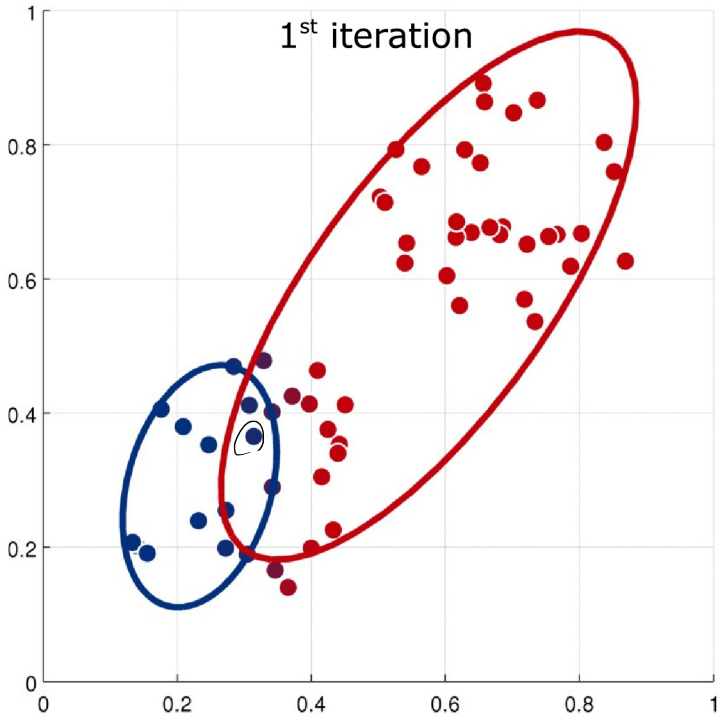
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



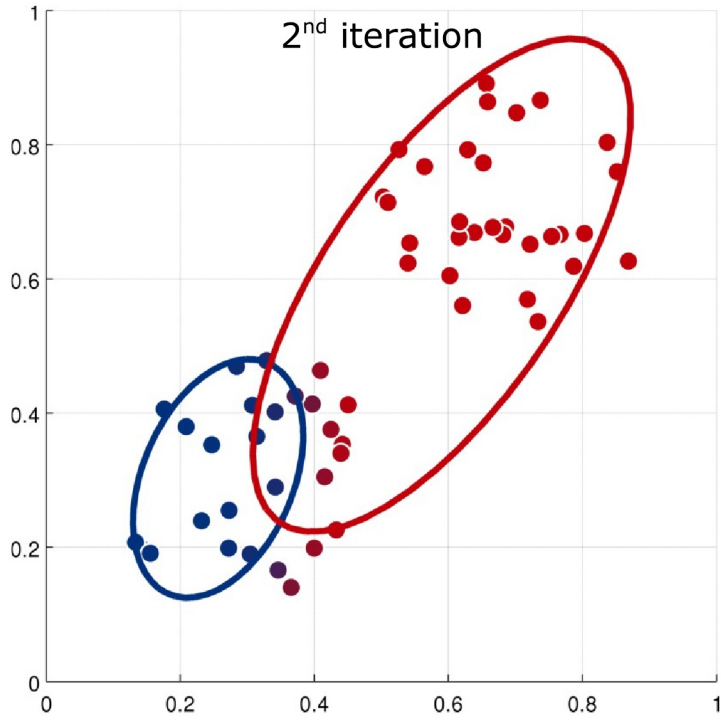
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



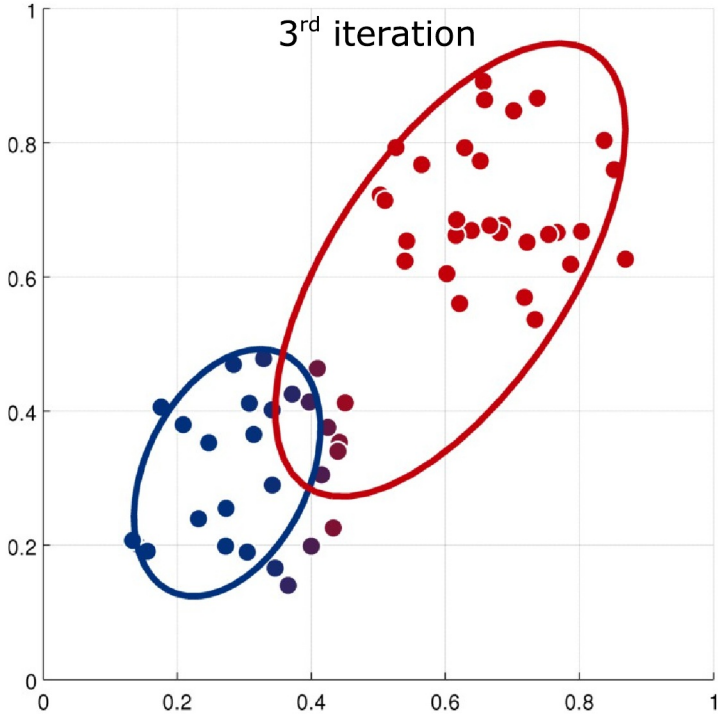
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



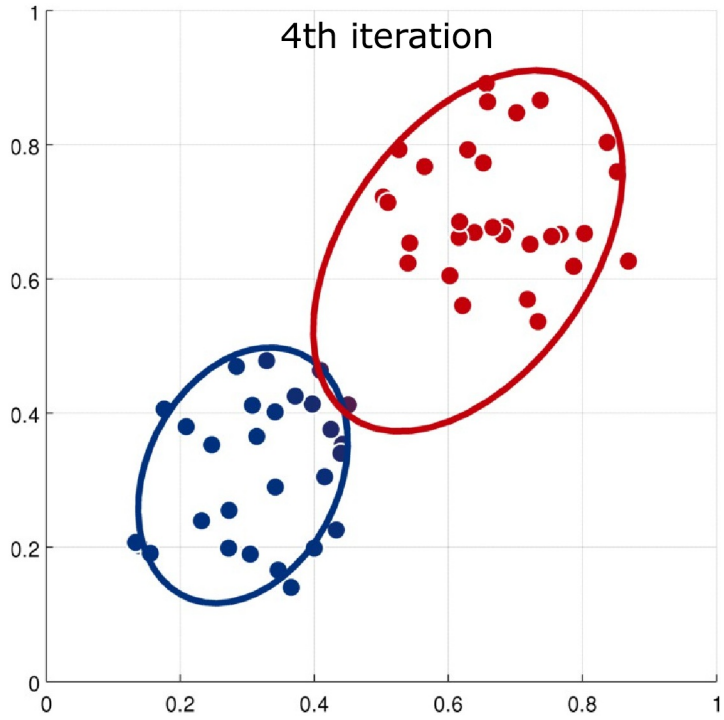
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



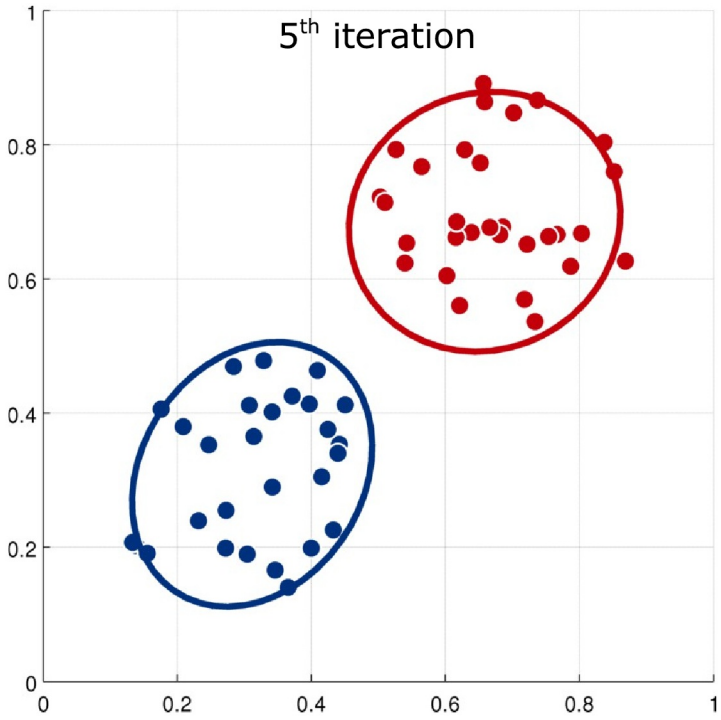
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



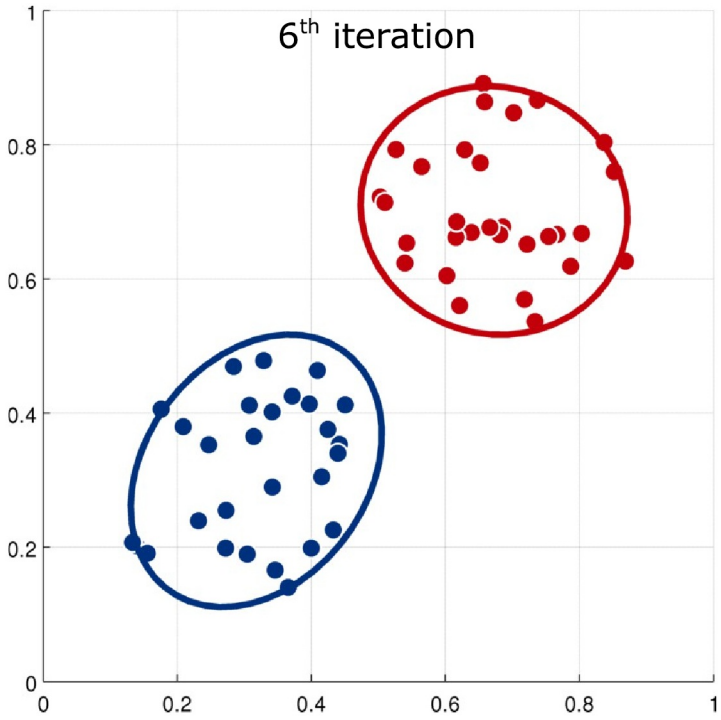
Gaussian mixture models, EM algorithm

Select an initial set of model parameters (mean and covariance for each cluster)

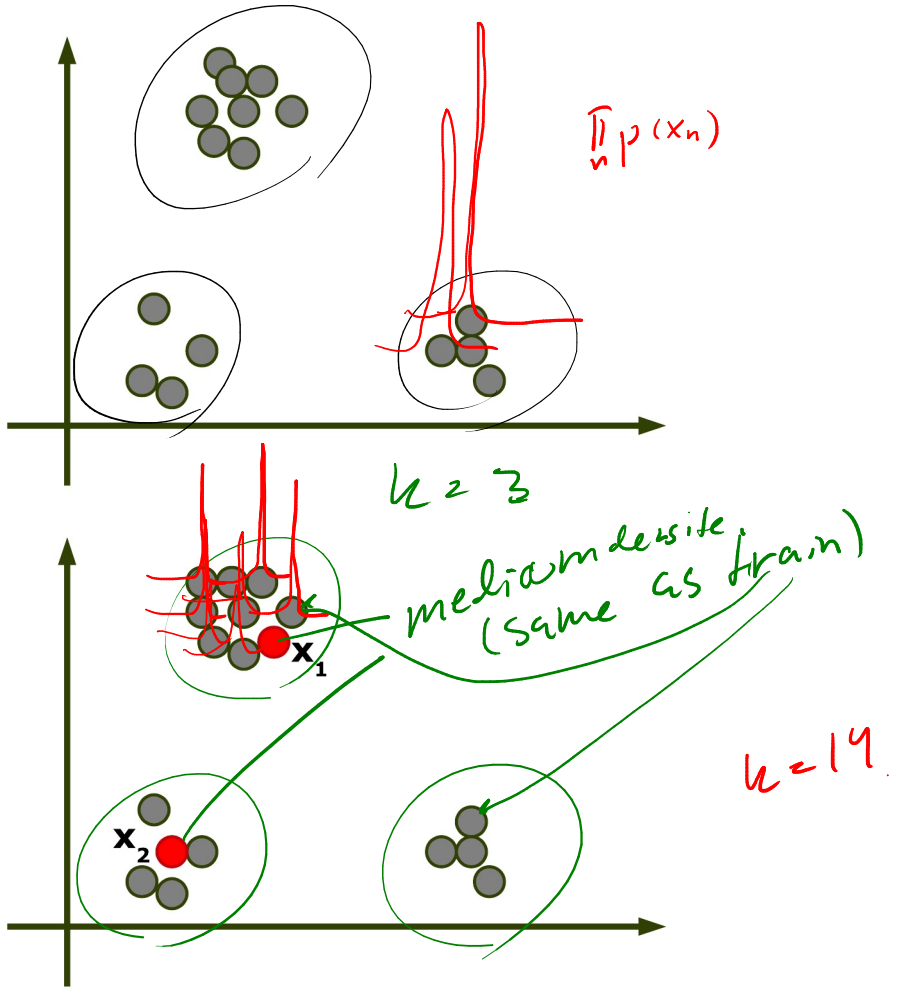
Repeat

- **Expectation**
  - For each object, calculate the probability of belonging to each distribution
- **Maximization**
  - For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



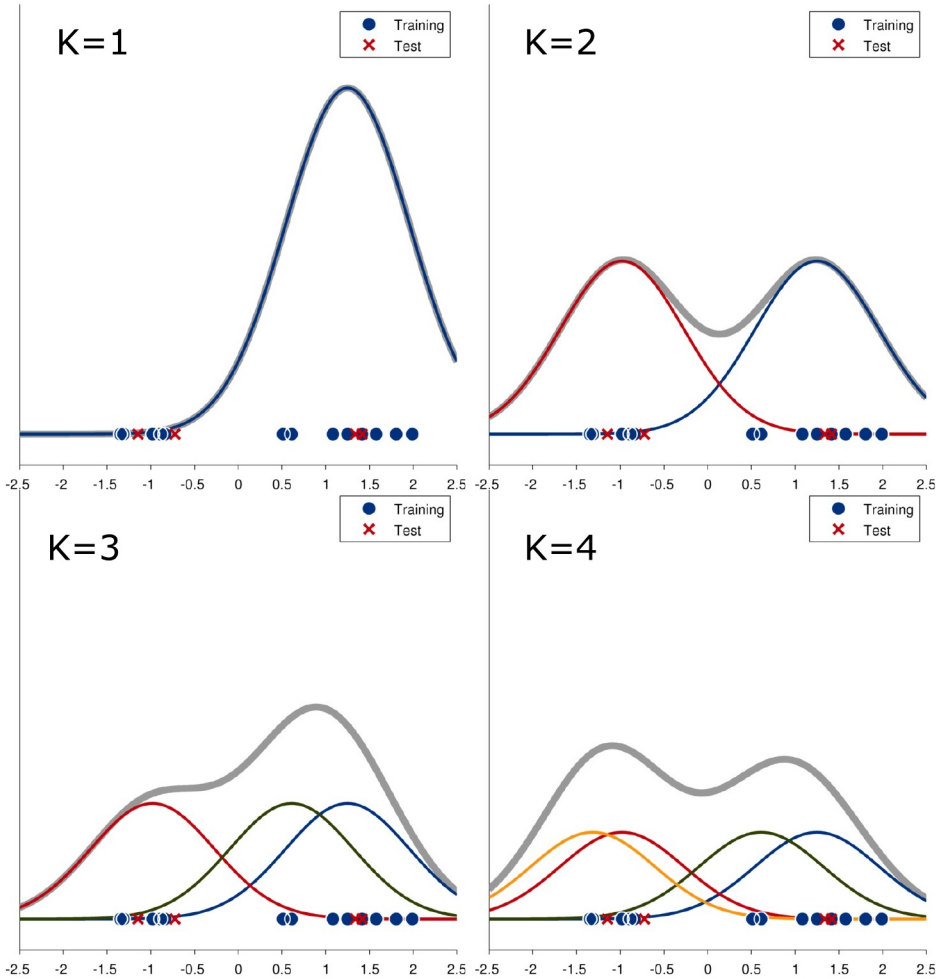
- Consider the data to the right with 16 observations.
  - What would ideally happen if we used a GMM with K=16 clusters to model the data?
- Imagine we have two **test observations** denoted $\mathbf{x}_1$ and $\mathbf{x}_2$ (red points) that are not used for training.
  - What happens to $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2)$ if we use K=3 and K=16 clusters?



Mixture models

- Selecting complexity using crossvalidation

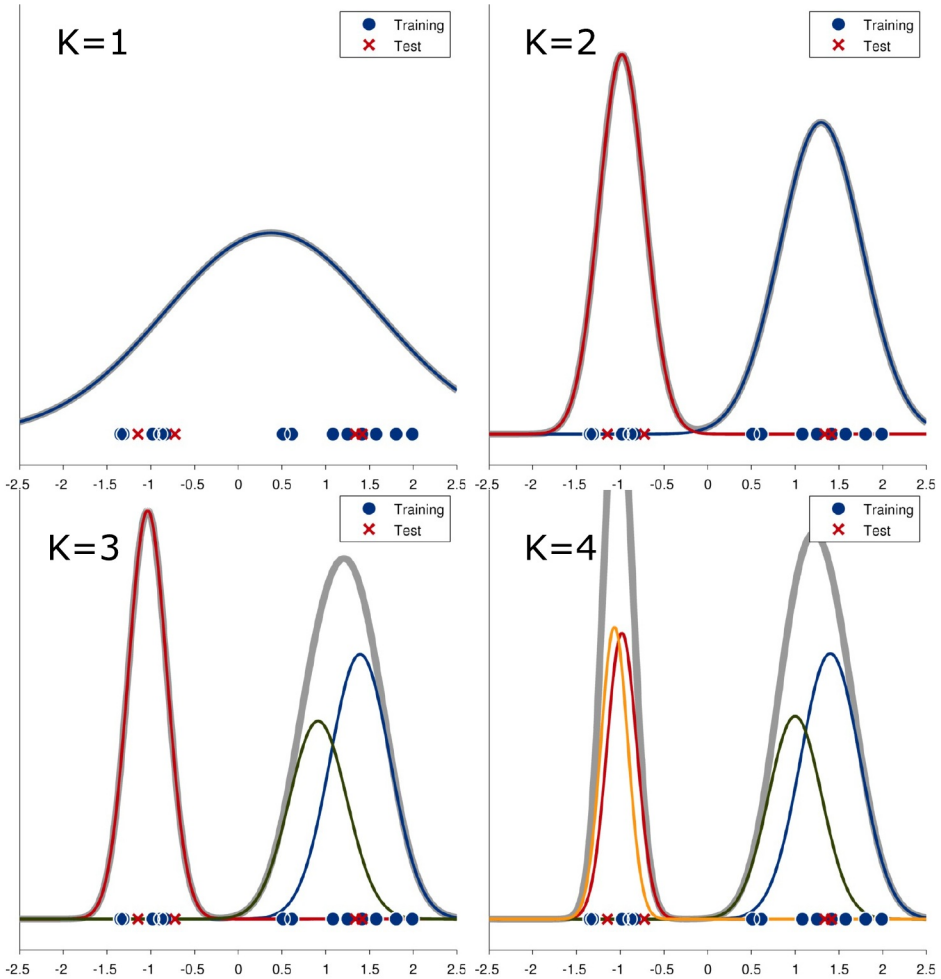
EM Initial solution



Mixture models

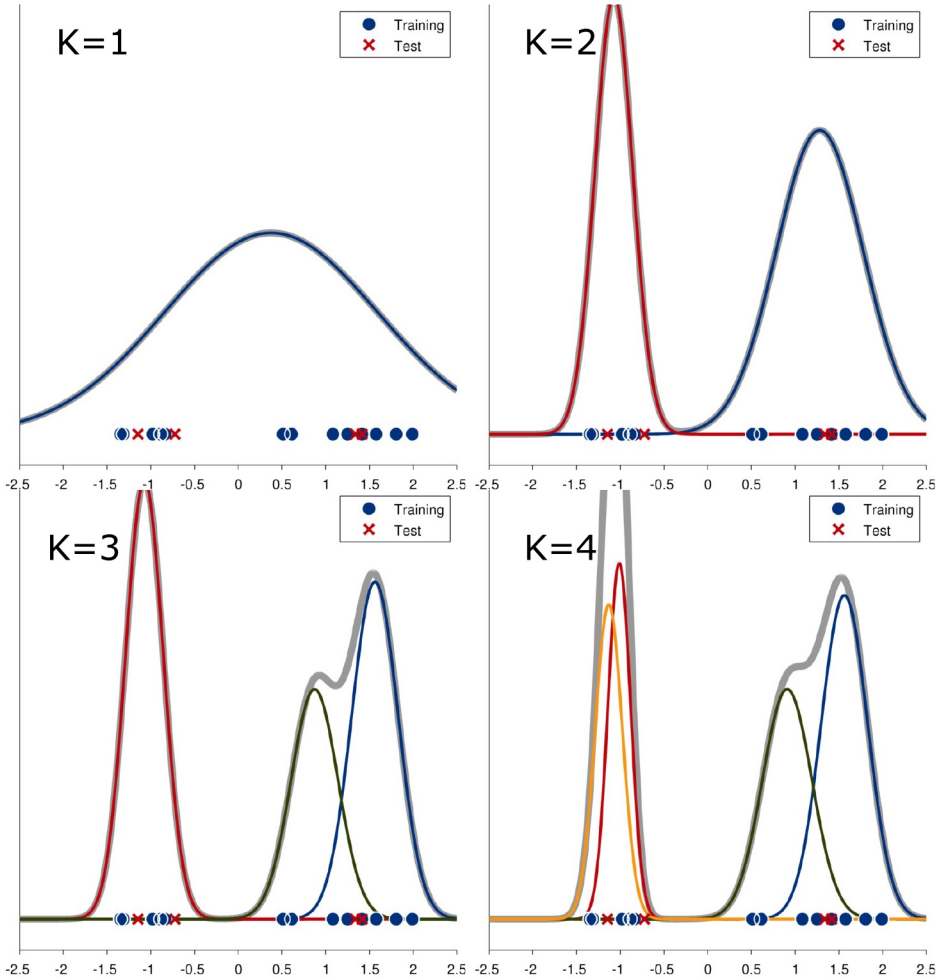
- Selecting complexity using crossvalidation

EM 1st iteration



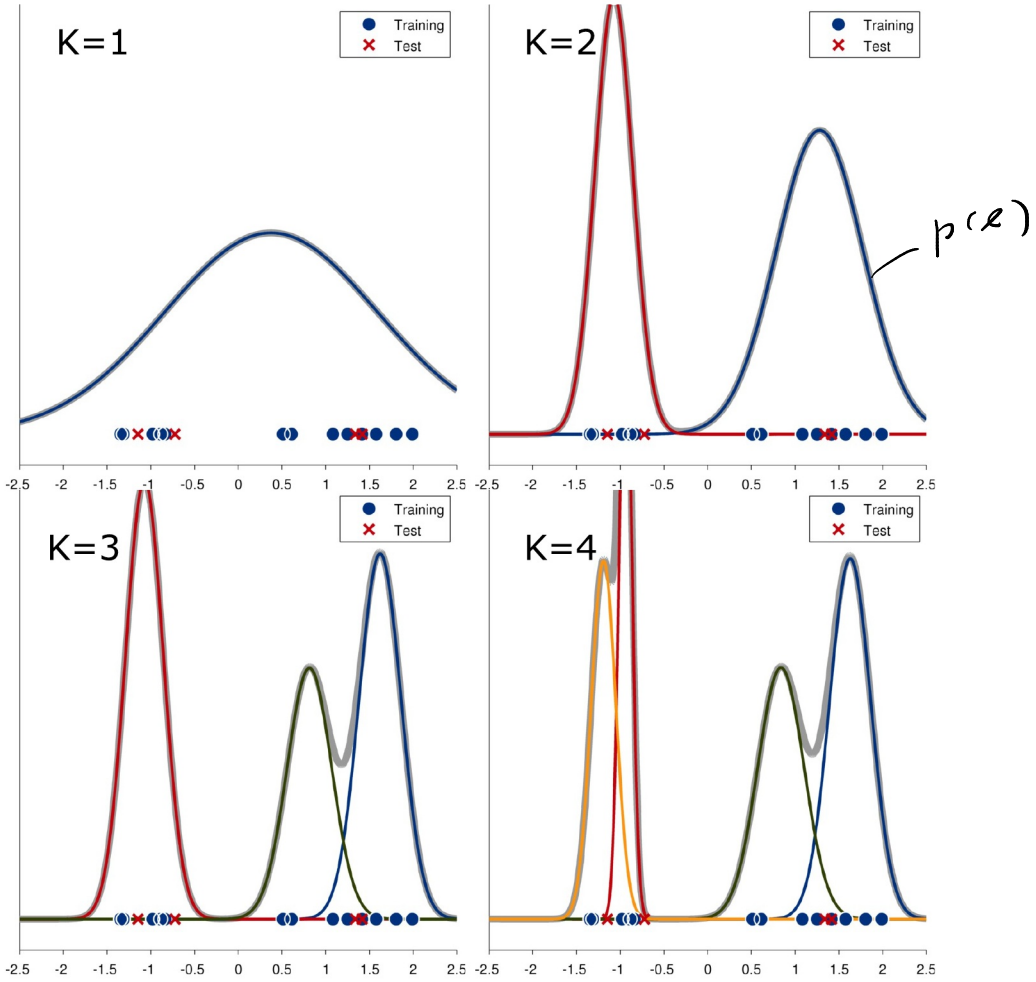
Mixture models

- Selecting complexity using crossvalidation



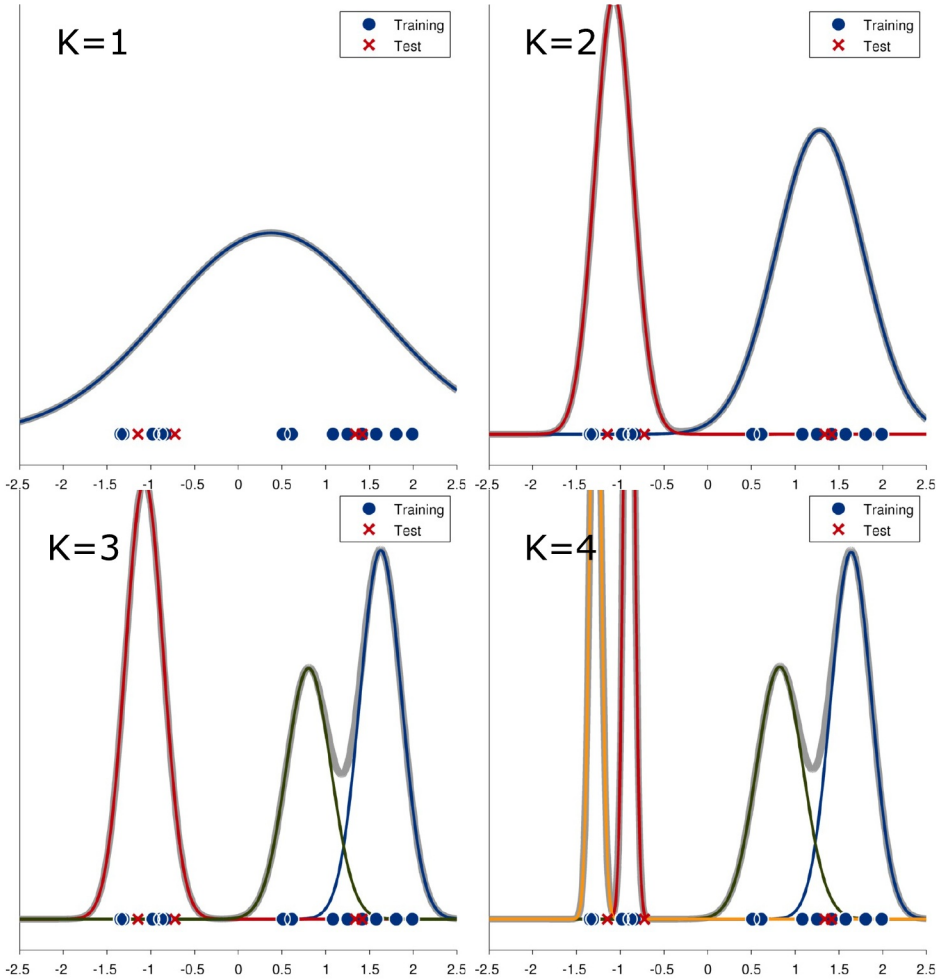
Mixture models

- Selecting complexity using crossvalidation



Mixture models

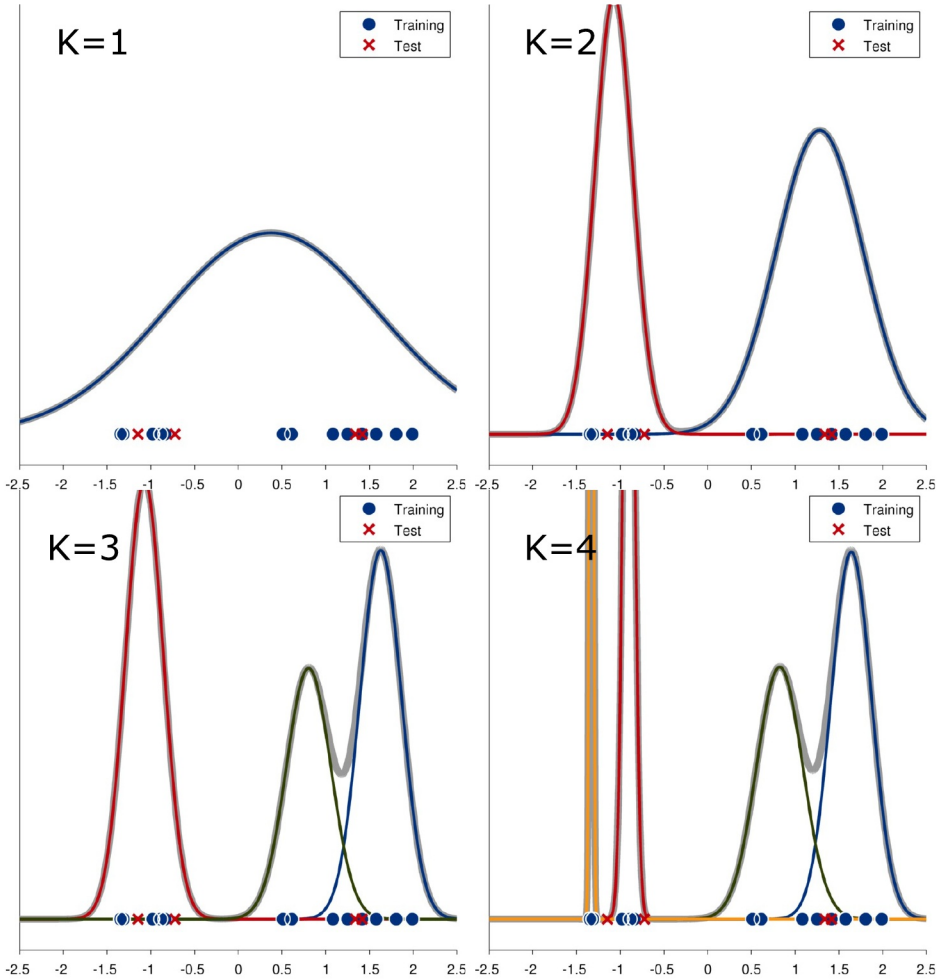
- Selecting complexity using crossvalidation



Mixture models

- Selecting complexity using crossvalidation

EM 5th iteration



Mixture models

- Selecting complexity using crossvalidation

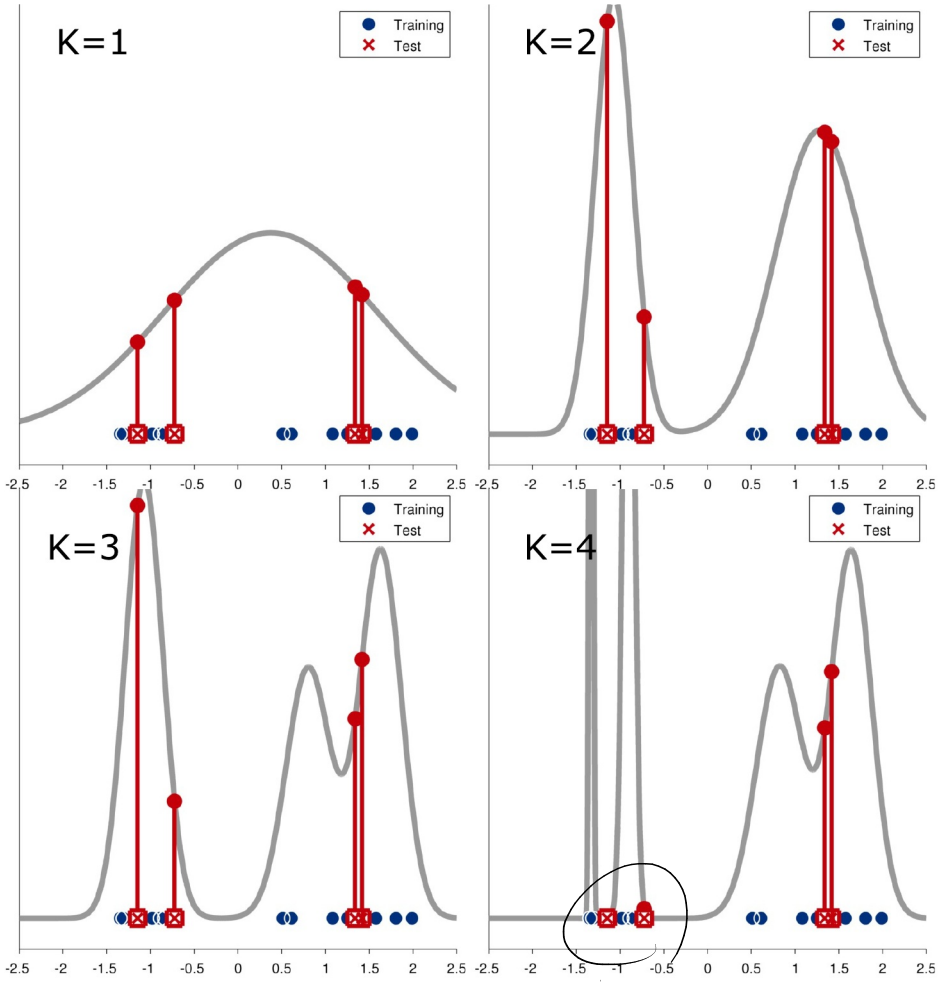
$$p(X^{test}) = \prod_{i=1}^{N^{test}} p(x_i^{test})$$

$$\log p(X^{test}) = \sum_{i=1}^{N^{test}} \log p(x_i^{test})$$

$\log 2$

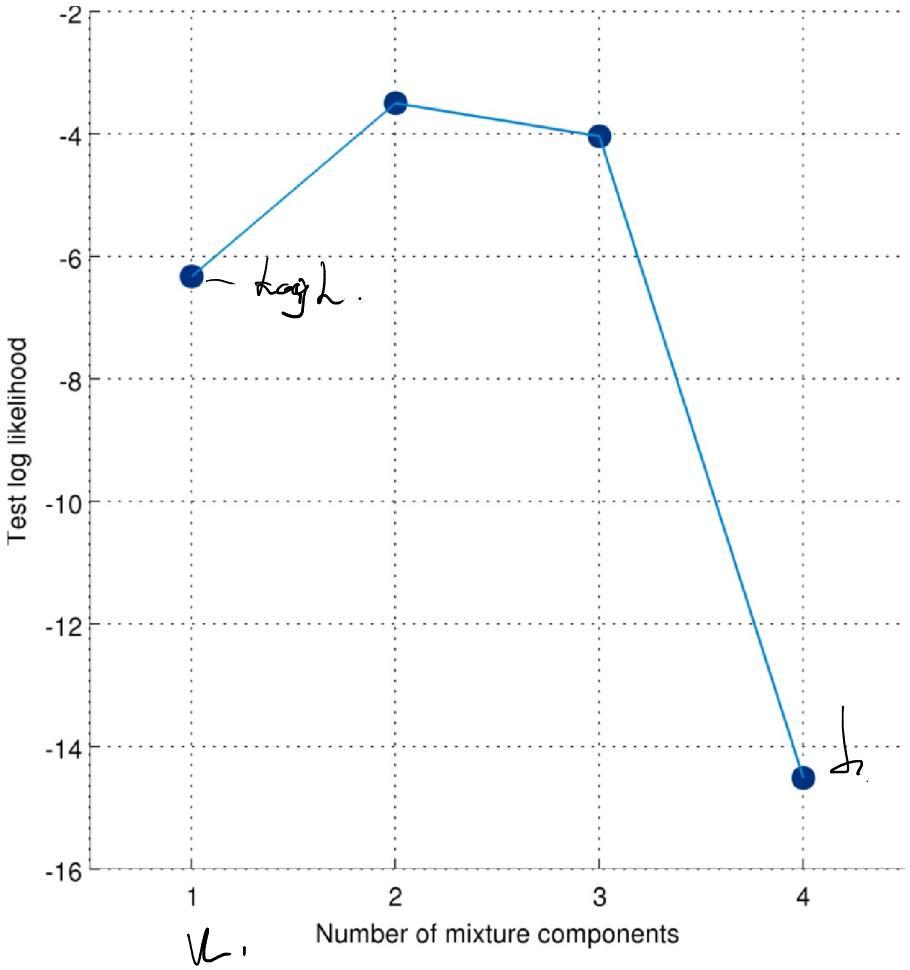
$(E = -\log L.)$

Test data evaluation



Mixture models

- Selecting complexity using crossvalidation

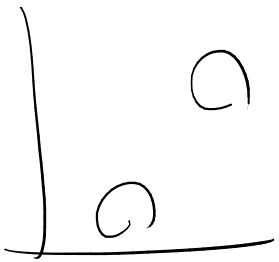


K-means versus GMM

K-means

- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model the size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

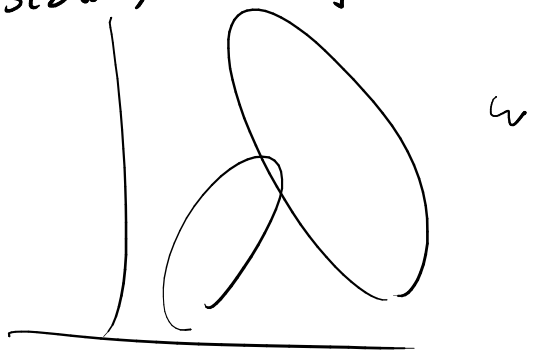
+ fast.



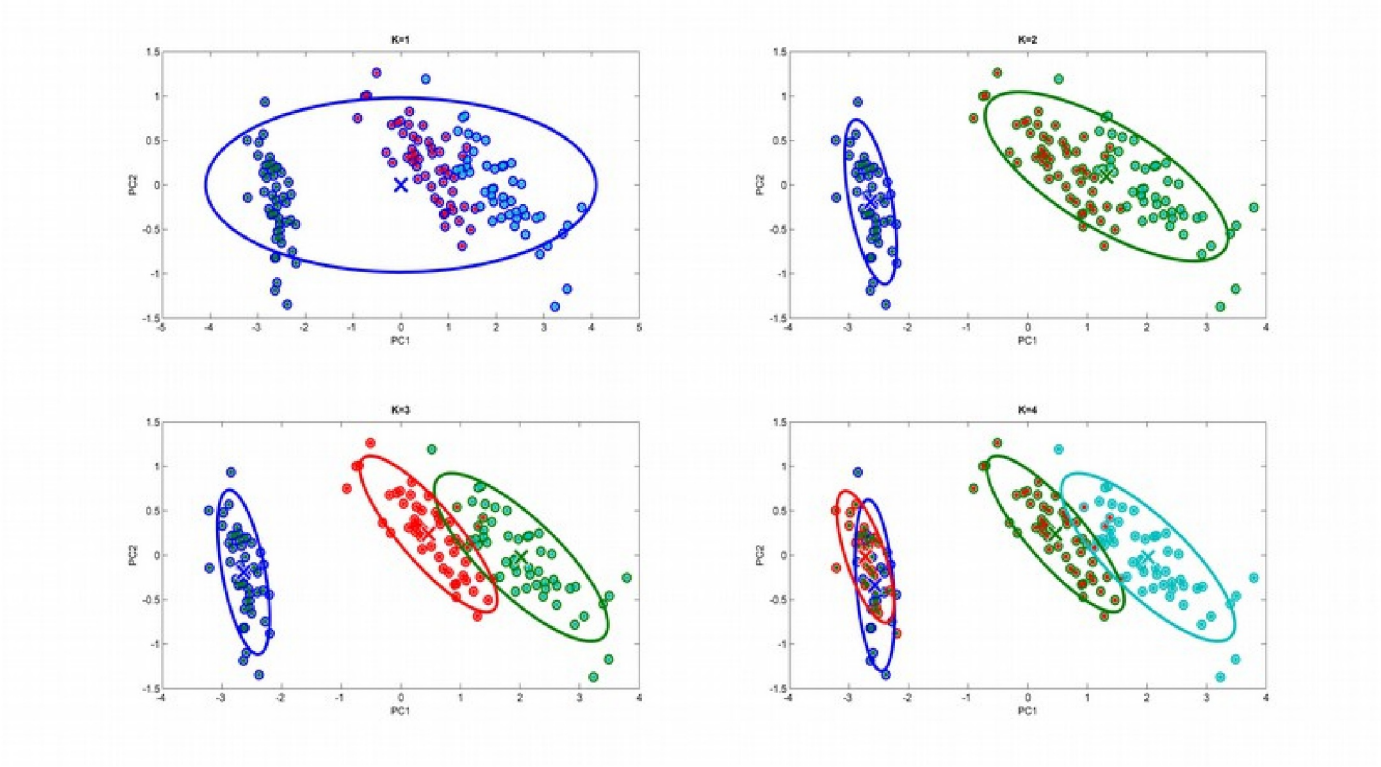
Gaussian mixture model (GMM)

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation

- slow, memory intensive.

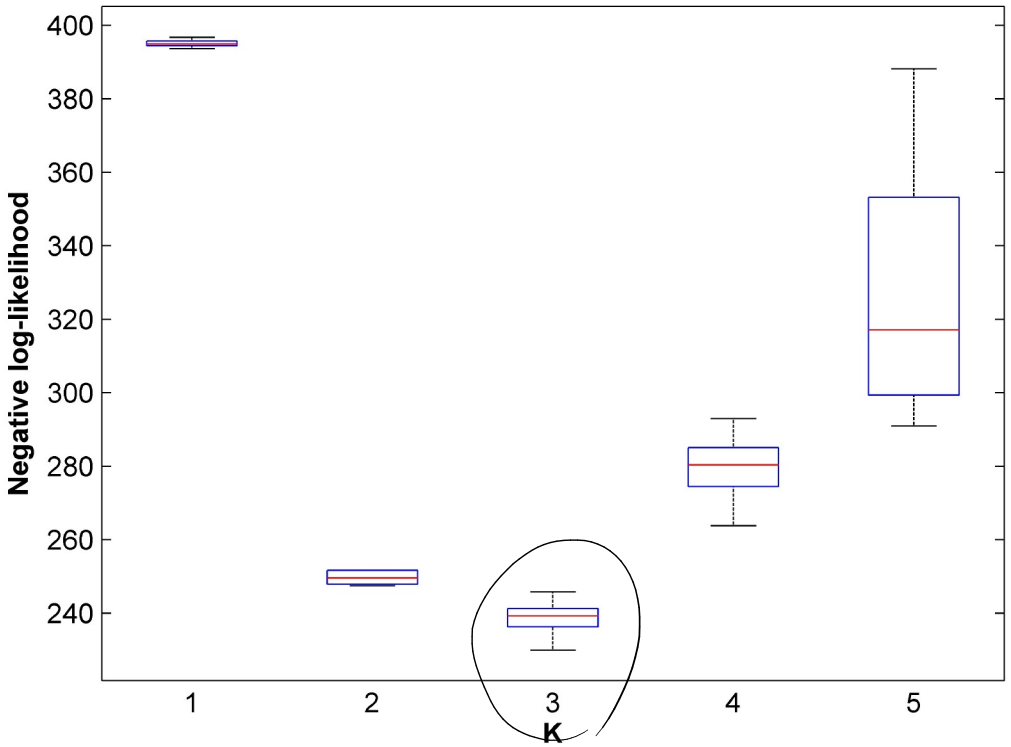


GMM on Iris data using 1,2,3 and 4 components



Recap of GMM on Iris data

GMM 10 fold cross-validation on Iris data repeated five times where the five runs are plotted using box-plots.



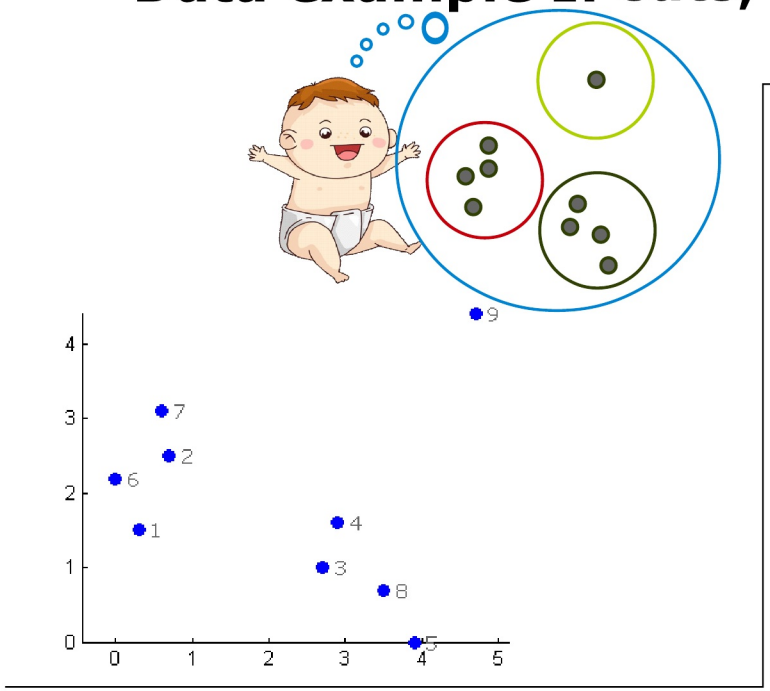
Anomaly detection: Definition

- Given a collection of data objects
  - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour

Anomaly detection: Example

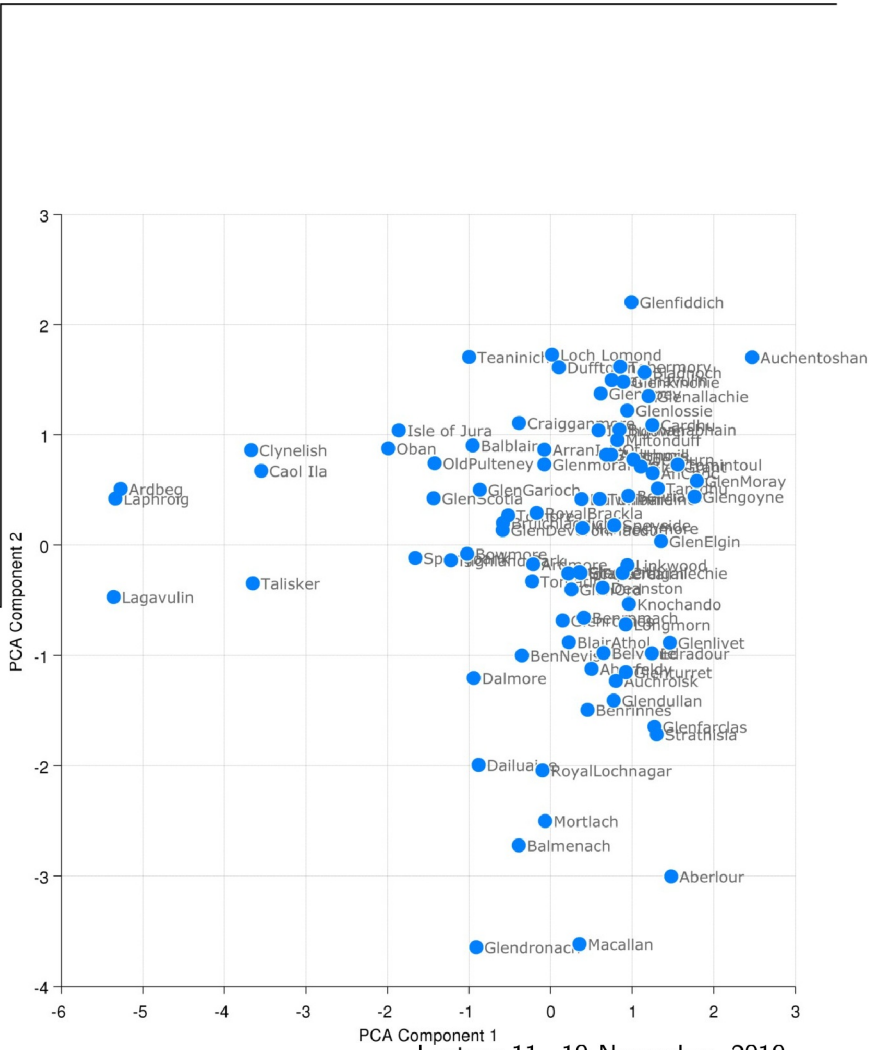
- Credit card **fraud detection**
  - Recognize dubious credit card transactions based on the transaction history of the card holder
- Network **intrusion detection**
  - Detect hacker attacks, web crawlers etc.
- **Ecosystem disturbances**
  - Detect hurricanes, floods droughts, heat waves and fires
- **Health and medicine monitoring**
  - Detect abnormal behaviour in populations and patients
- **Fault detection in industry systems**
  - Detect when a wind turbine performs poorly due to ice coating on blades
- Detection of **outliers** in data measurements
  - Remove erroneous measurements due to misreading from an instrument

Data example I: Cats, Dogs and Dinosaurs

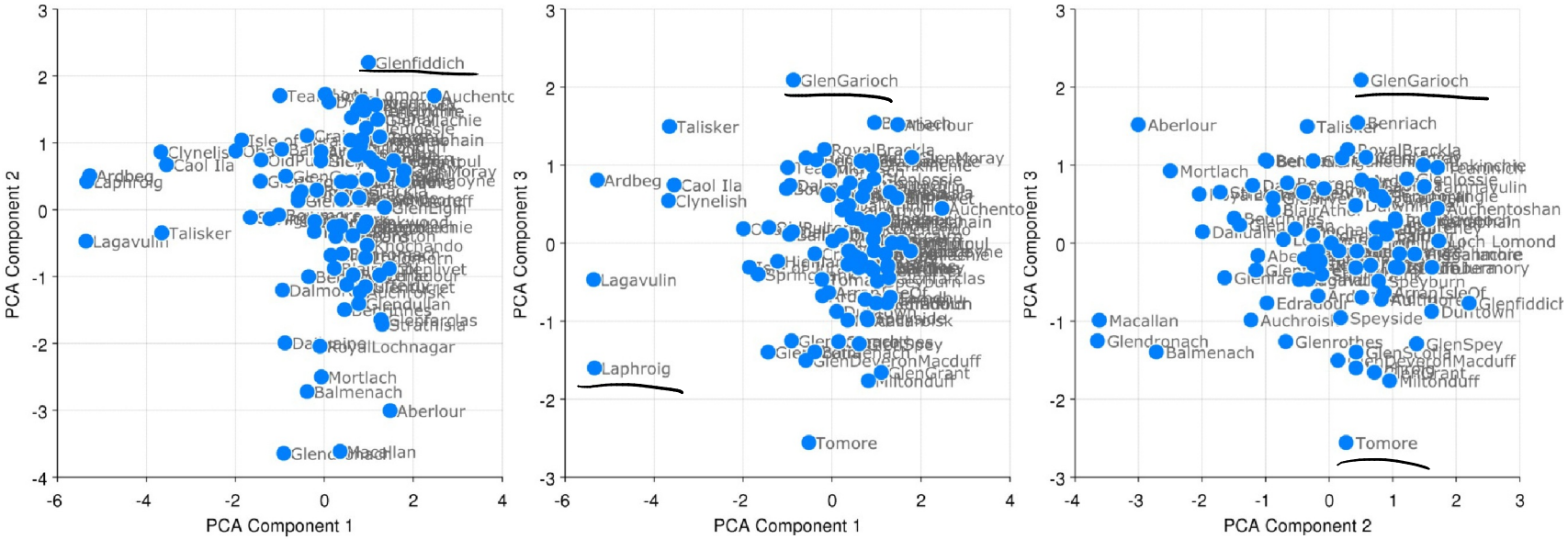


Data example II: Whisky

- 86 types of Scotch whisky
- Human ratings 1-5
- 12 taste categories
  - body, sweetness, smoky, medicinal, tobacco, honey, spicy, winey, nutty, malty, fruity, floral



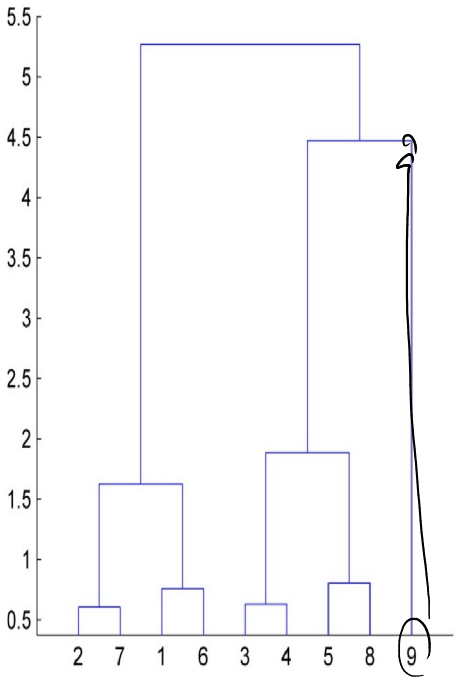
PCA plot


PCA Component 2
PCA Component 1
Glenfiddich
Ardbeg
Caol Ila
Talisker
Lagavulin
Mortlach
Balmenach
Aberlour
PCA Component 3
PCA Component 1
GlenGarioch
Talisker
Ardbeg
Caol Ila
Clynelish
Lagavulin
Laphroig
Tomore
PCA Component 3
PCA Component 2
GlenGarioch
Aberlour
Benriach
Mortlach
Macallan
Glendronach
Balmenach
Glenrothes
Speyside
Tomore

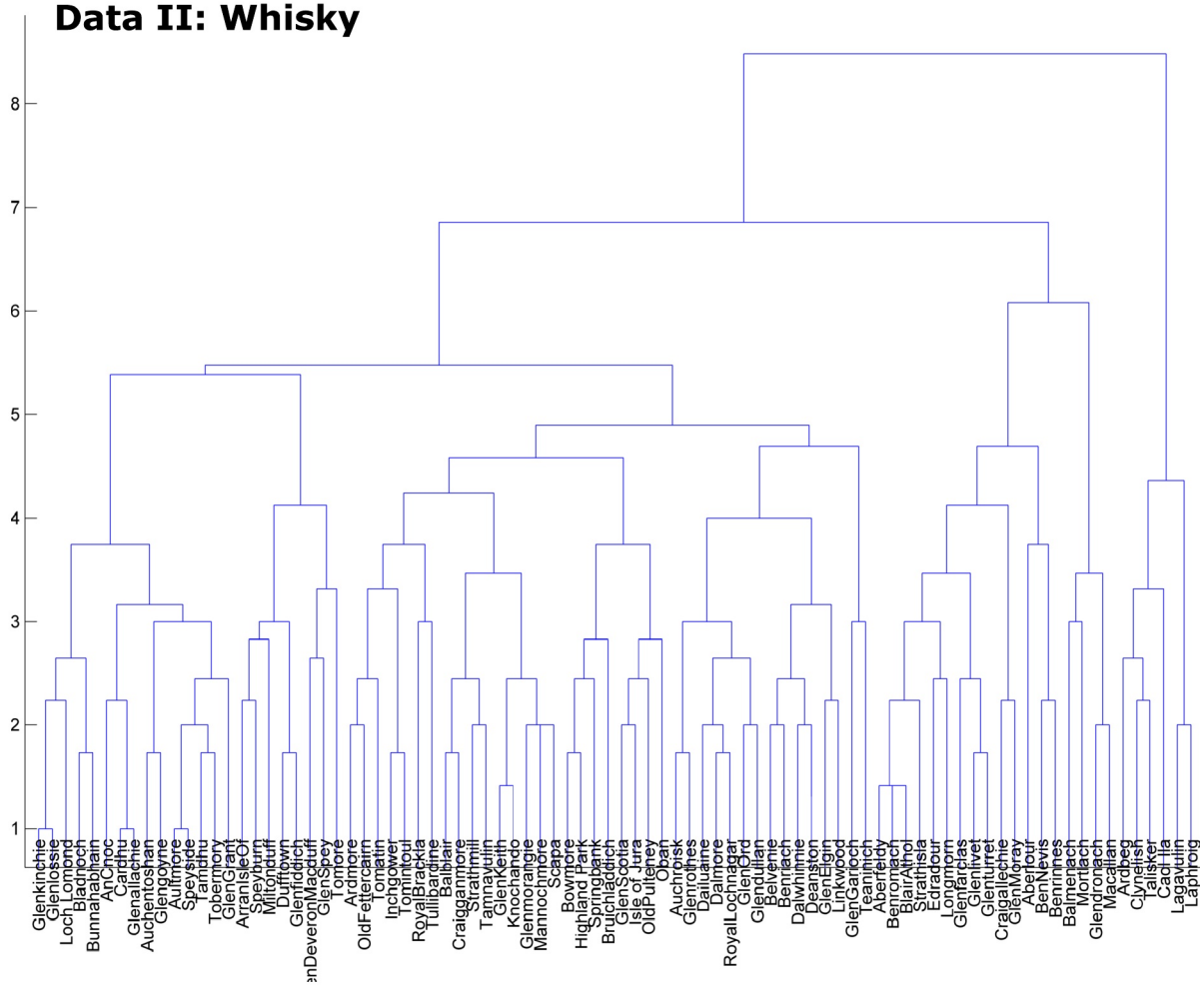
Dendrogram

- Dendrograms can be used to visualize relative distances between the observations

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



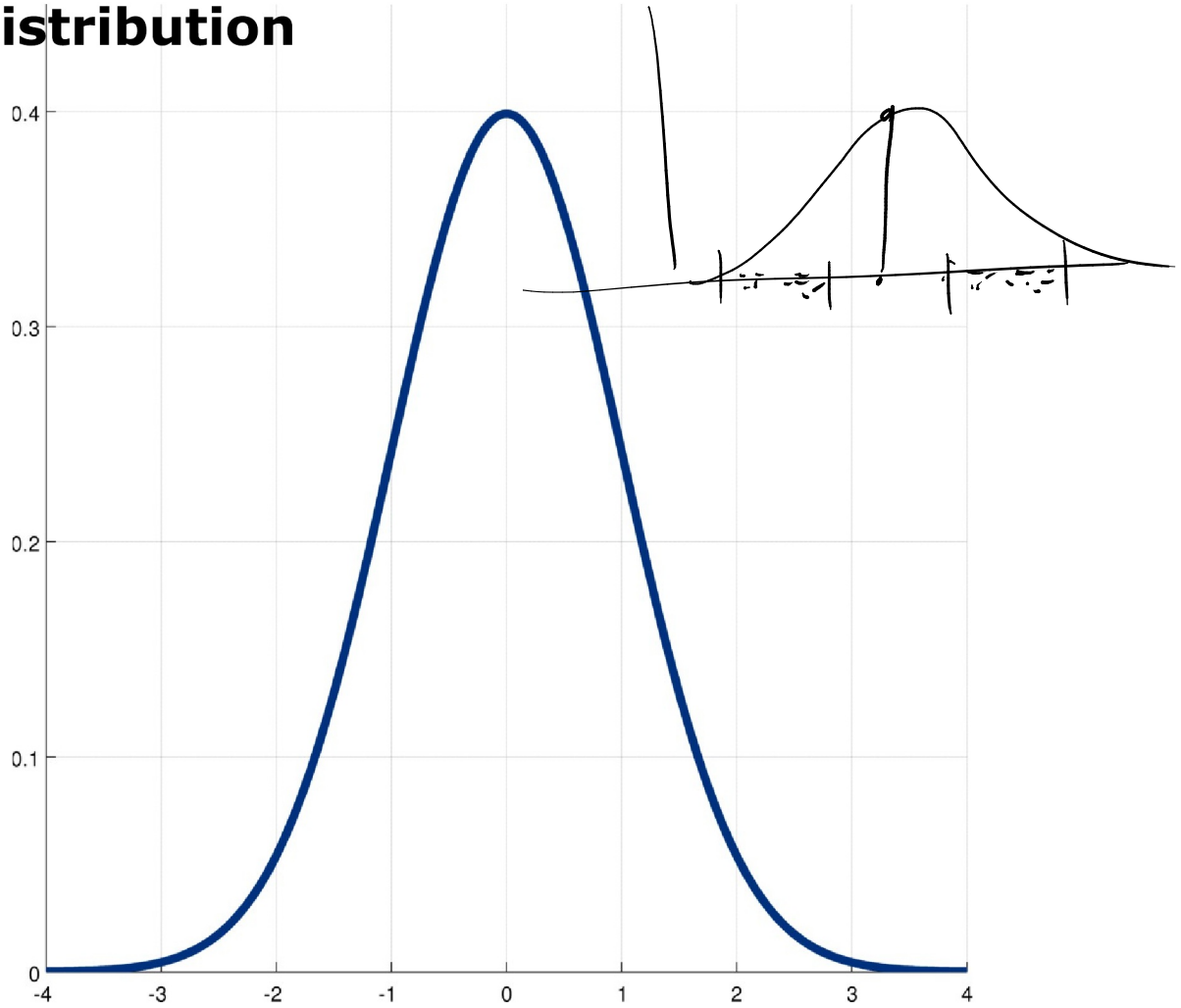
Density based techniques: Univariate normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



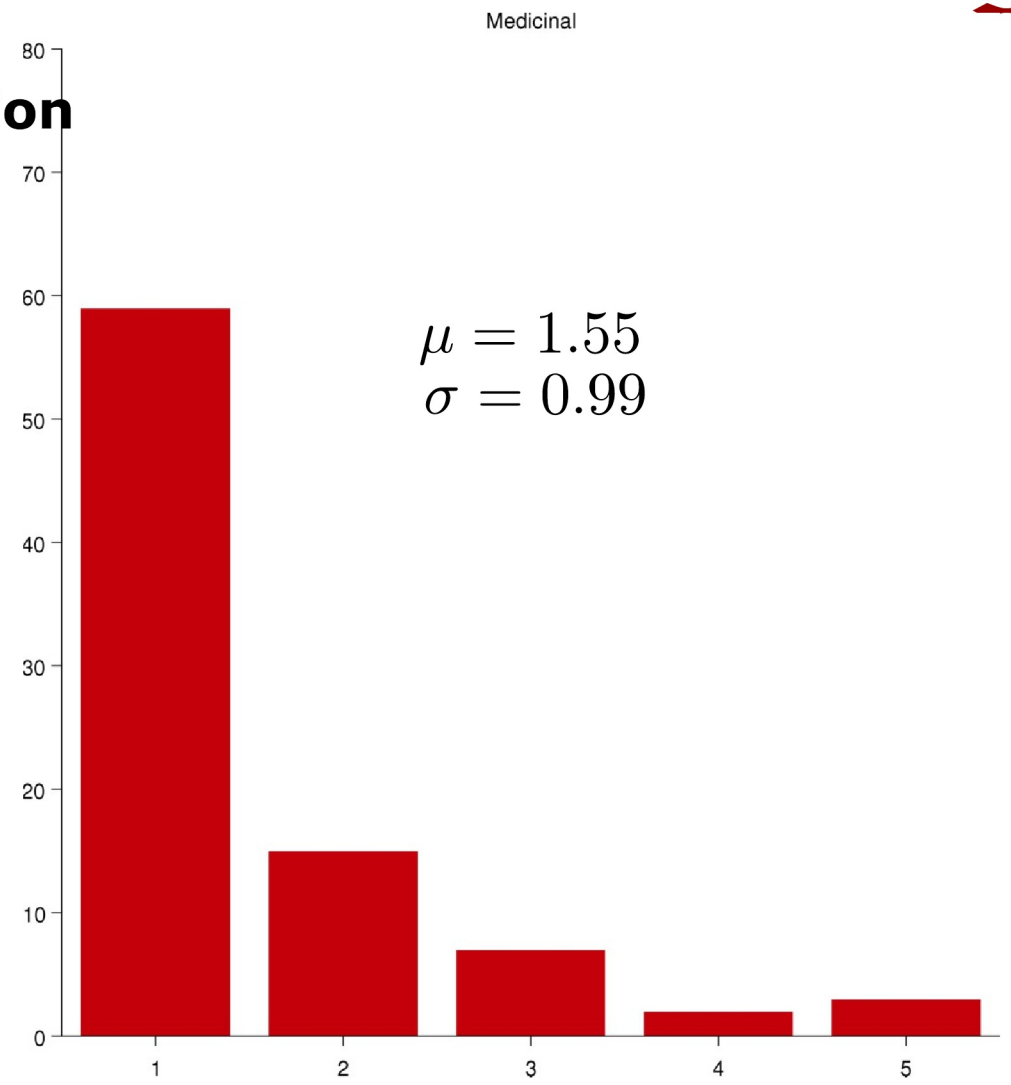
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



$$\mu = 1.55$$
$$\sigma = 0.99$$

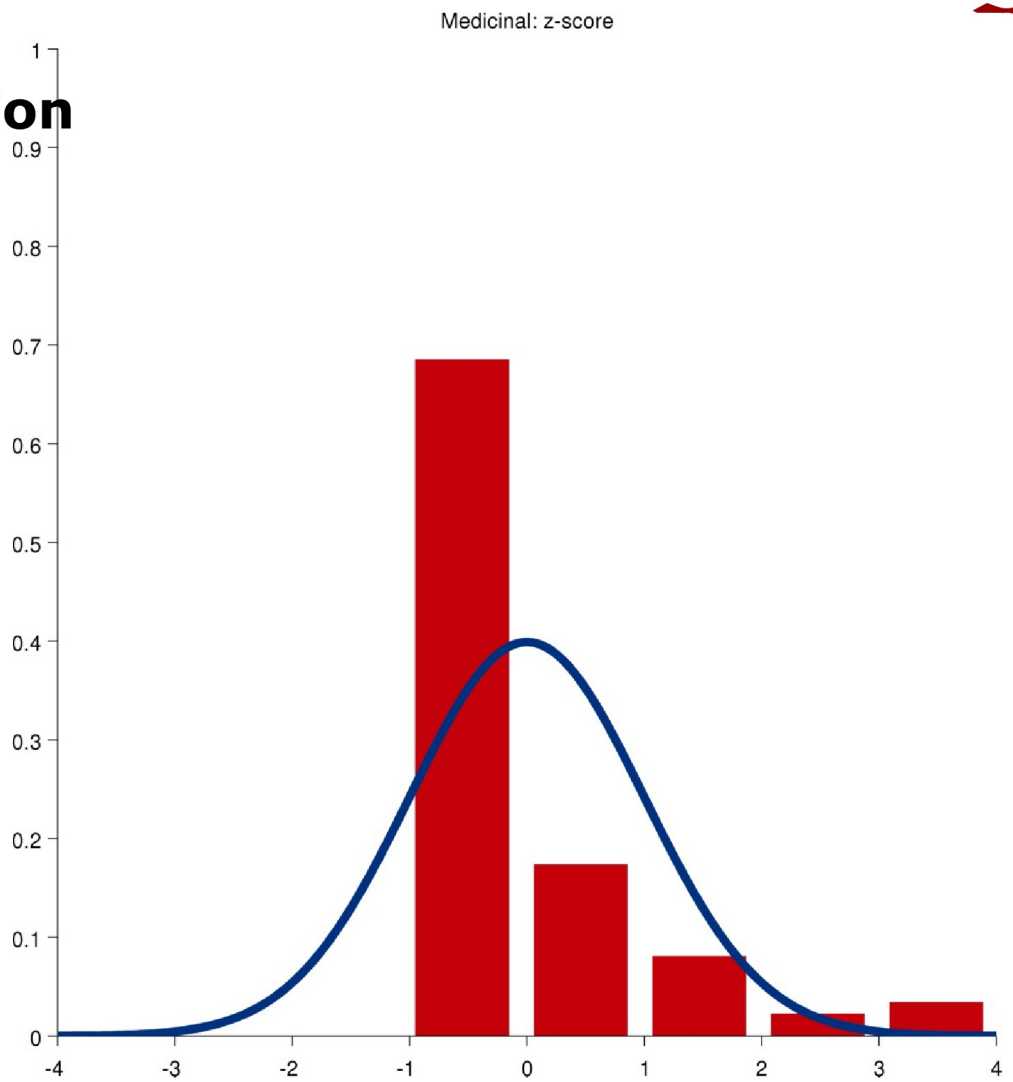
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



Normal distribution

- Map attribute to standard Normal variable

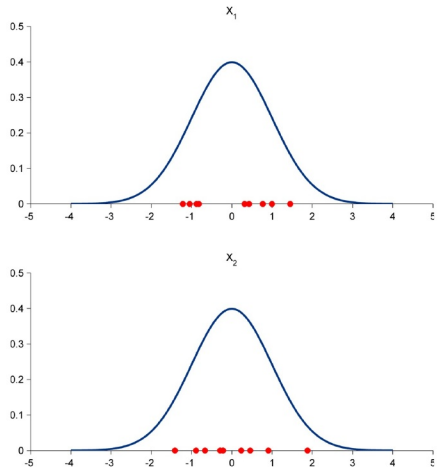
- Choose a threshold

$$z = \frac{x - \mu}{\sigma}$$

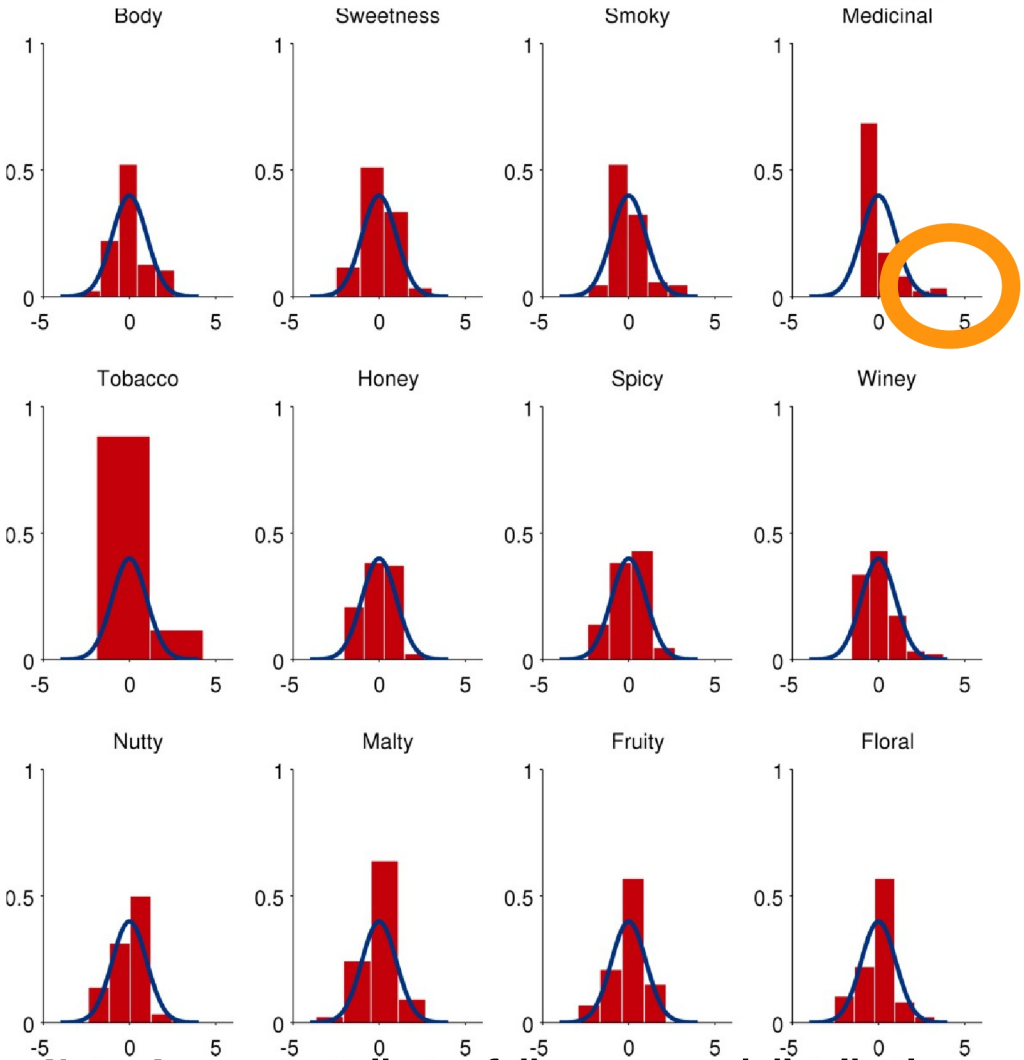
$$p(|z| > c) = 0.001$$

$$c = 3.2905$$

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



Note: Assumes attributes follow a normal distribution which may not be a valid assumption!

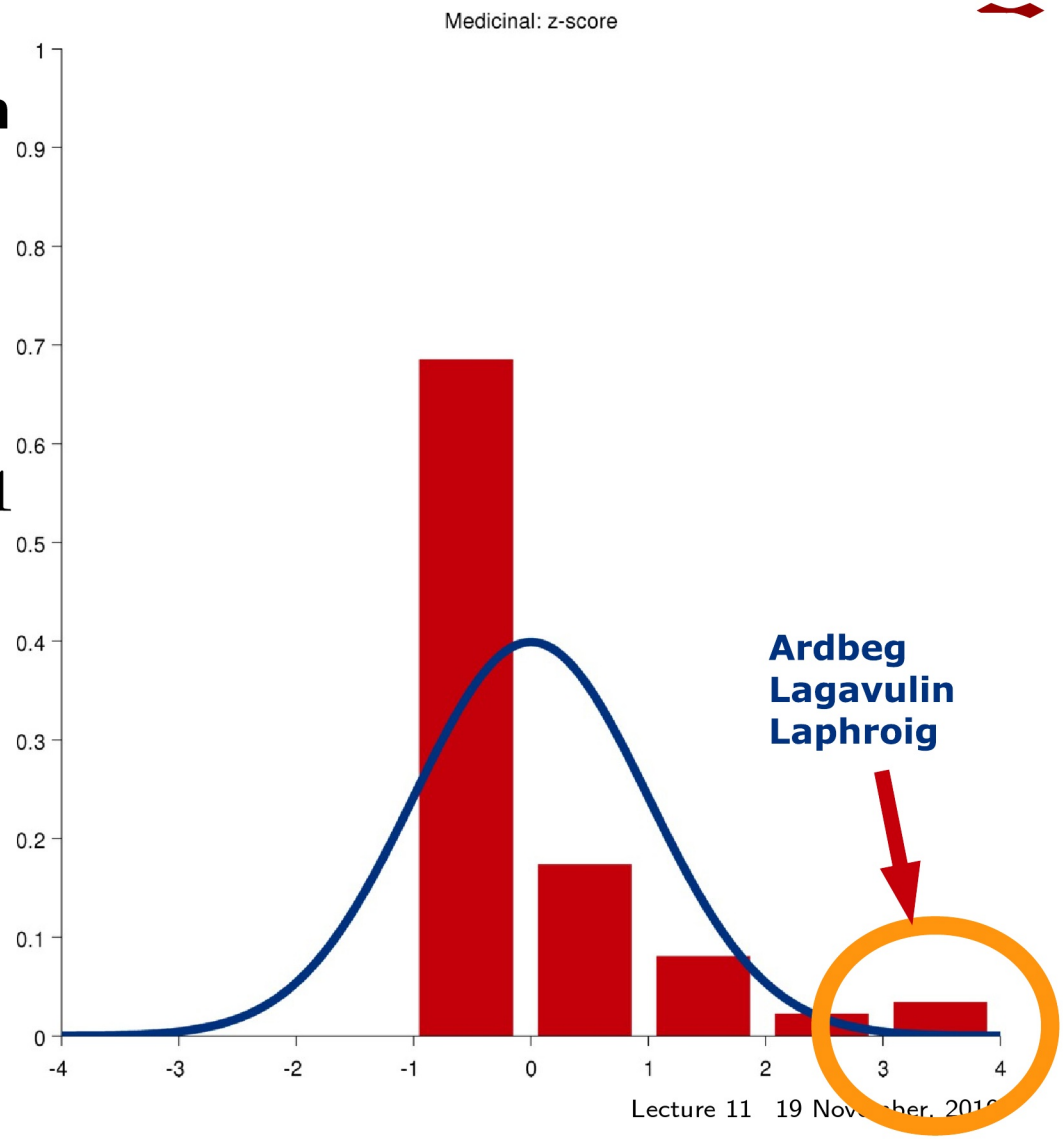
Normal distribution

- Map attribute to standard Normal variable

$$z = \frac{x - \mu}{\sigma}$$

- Choose a threshold

$$p(|z| > c) = 0.001$$
$$c = 3.2905$$



Approaches to anomaly detection

- **Density-based techniques**
  - Estimate the density of data objects
  - Outliers are:
    - Data objects in low density area

- **We can of course use the GMM to evaluate the density of test data.**
  - why not on the training data?

- **Approaches we will presently also consider**:
  - Kernel density estimation
  - Inverse average distance to K nearest neighbours (KNN density)
  - Average relative KNN density

Density based techniques: Kernel Density Estimator

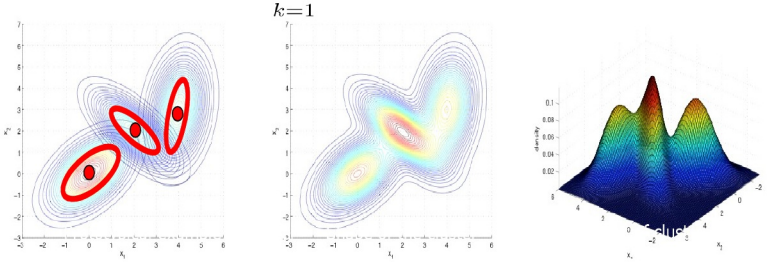
Recall the Gaussian Mixture Model (GMM)

Data density

Sum of cluster specific densities assumed normal distributed

$$p(\boldsymbol{x}) = \sum_{K=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

$$(s.t. \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0)$$

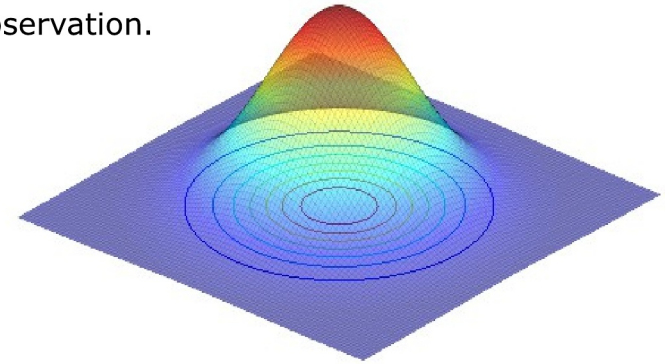


$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative density of the cluster

Kernel Density estimation based on Gaussian Kernel: Consider the GMM and define a Gaussian with mean $\boldsymbol{x}_n$ and co-variance $\sigma^2 I$ around each Observation.



$\mu_k = x_k$.
$\Sigma_k = \sigma^2 I$

Let all observation weight the same, i.e. $w_n = 1/N$

$$p(\mathbf{x}) = \sum_{n=1}^{N} \frac{1}{N} \mathcal{N}(\mathbf{x}|\mathbf{x}_n, \sigma^2 \mathbf{I})$$

Only free parameter σ^2!

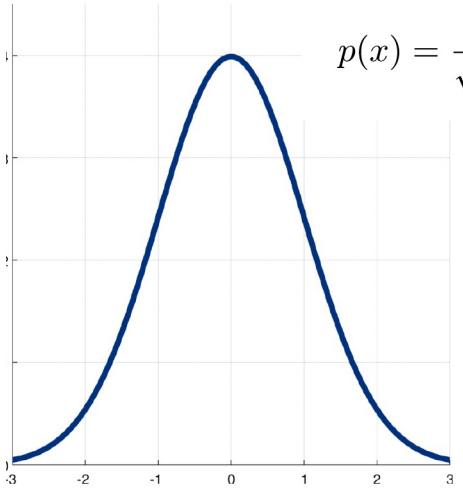
There is nothing special about the normal distribution. For a general mixture distribution p the general form of kernel density estimator is:

$$p(\mathbf{x}) = \sum_{n=1}^{N} \frac{1}{N} p(\mathbf{x}|\mathbf{x}_n, \theta)$$

This may be useful if x is discrete or non-negative.

Piazza quiz 03: Kernel density (Spring 2013)

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



Consider five observations of an attribute x given by

$$\boldsymbol{X} = \{2, 3, 5, 10, 12\}.$$

Based on the five observations, what is the Gaussian kernel density estimate at $x = 4$ using $\sigma^2 = 4$?

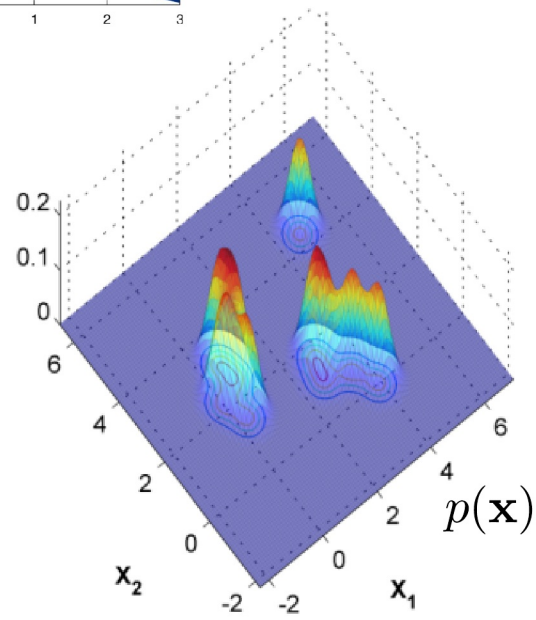
A. $\frac{1}{\sqrt{8\pi}^5} \exp(-\frac{53}{4})$

B. $\frac{1}{5\sqrt{8\pi}} \exp(-\frac{53}{4})$

C. $\frac{1}{5\sqrt{8\pi}}(\exp(-\frac{1}{2}) + 2 \cdot \exp(-\frac{1}{8}) + \exp(-\frac{9}{2}) + \exp(-8))$

D. $\frac{1}{5\sqrt{8\pi}}(\exp(-1) + 2 \cdot \exp(-\frac{1}{4}) + \exp(-9) + \exp(-16))$

E. Don't know.



$$p(\mathbf{x}) = \sum_{n=1}^{N} \frac{1}{N} \mathcal{N}(\mathbf{x}|\mathbf{x}_n, \sigma^2\mathbf{I})$$

n

Solution:

When inserting $\sigma^2 = 4$ the Gaussian kernel density is given by

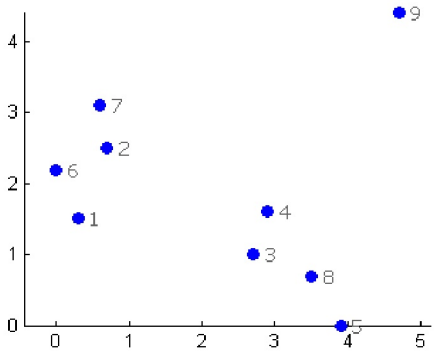
$$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\tfrac{(x-x_n)^2}{2\sigma^2}) = \frac{1}{5\sqrt{8\pi}}(\exp(-\tfrac{(x-2)^2}{2\cdot 4}) + \exp(-\tfrac{(x-3)^2}{2\cdot 4}) + \exp(-\tfrac{(x-5)^2}{2\cdot 4}) + \exp(-\tfrac{(x-10)^2}{2\cdot 4}) + \exp(-\tfrac{(x-12)^2}{2\cdot 4})).$$

Evaluating the density at $x = 4$ we obtain

$$p(x=4) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\tfrac{(x-x_n)^2}{2\sigma^2}) = \frac{1}{5\sqrt{8\pi}}(\exp(-\tfrac{(2)^2}{2\cdot 4}) + \exp(-\tfrac{(1)^2}{2\cdot 4}) + \exp(-\tfrac{(-1)^2}{2\cdot 4}) + \exp(-\tfrac{(-6)^2}{2\cdot 4}) + \exp(-\tfrac{(-8)^2}{2\cdot 4})) = \frac{1}{5\sqrt{8\pi}}(\exp(-\tfrac{1}{2}) + 2\cdot(\exp(-\tfrac{1}{8}) + \exp(-\tfrac{9}{2}) + \exp(-8)).$$

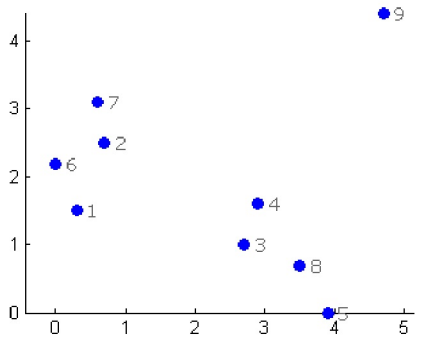
How do we determine σ^2?

Data I: Cats, Dogs and Dinosaurs

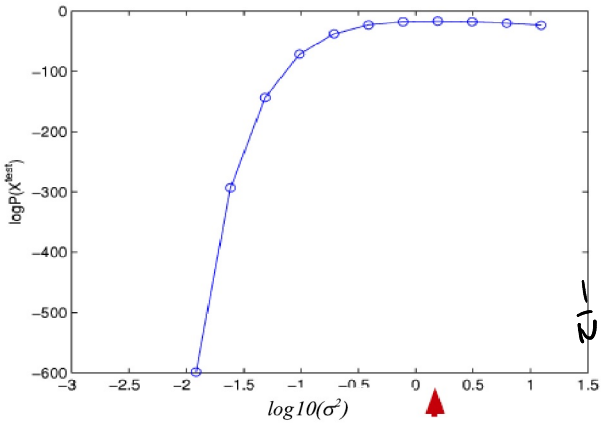


How do we determine σ^2? Crossvalidation!

Data I: Cats, Dogs and Dinosaurs

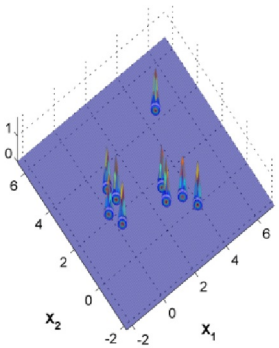


Density of test set based on leave-one-out cross validation

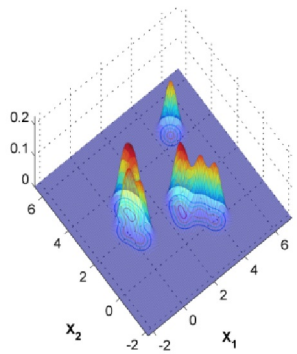


Optimal $\sigma^2=1.55$

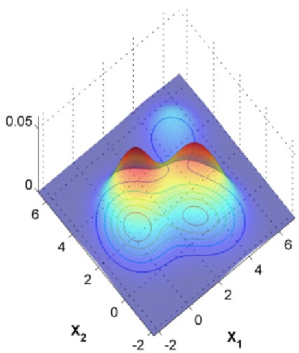
$\sigma^2=0.01$



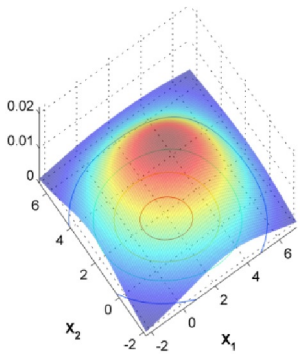
$\sigma^2=0.1$



$\sigma^2=1$



$\sigma^2=5$

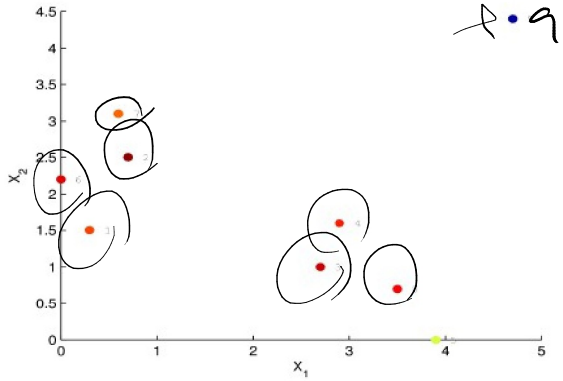


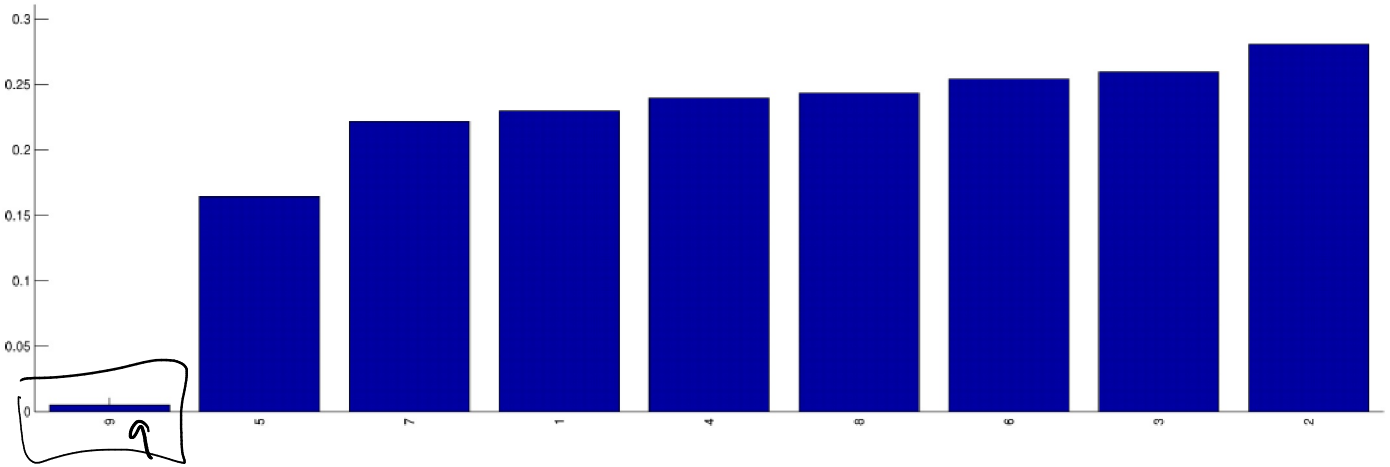
$\sigma = 0.01, 0.1, 1, 5$.

$p^{\sigma}_{N\setminus i}(x)$

$\frac{1}{N}\sum_{i=1}^{N} -\log p^{\sigma}_{N\setminus i}(x_i) = L^{avg.}(\sigma)$

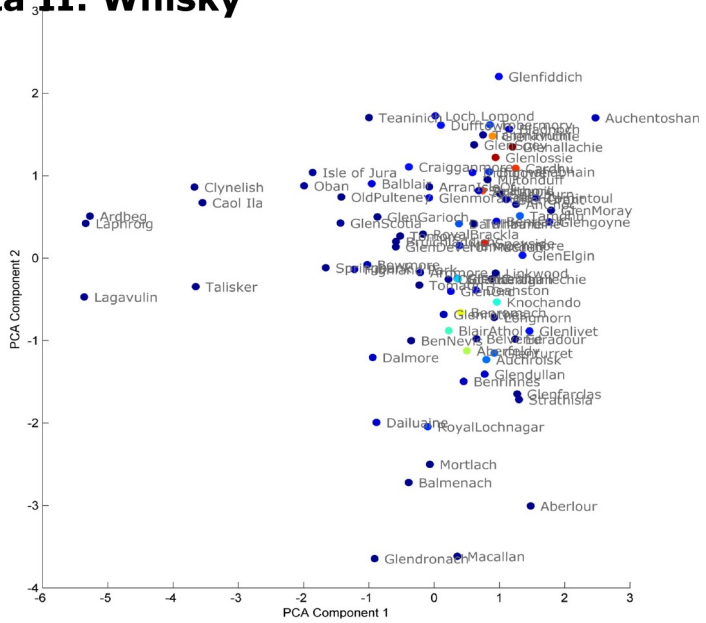
Etimated leave-one-out density evaluated at each observation



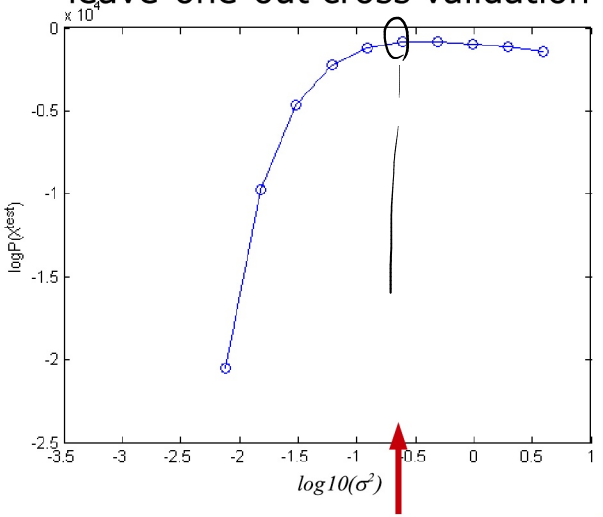


Etimated density evaluated at each observation

Data II: Whisky

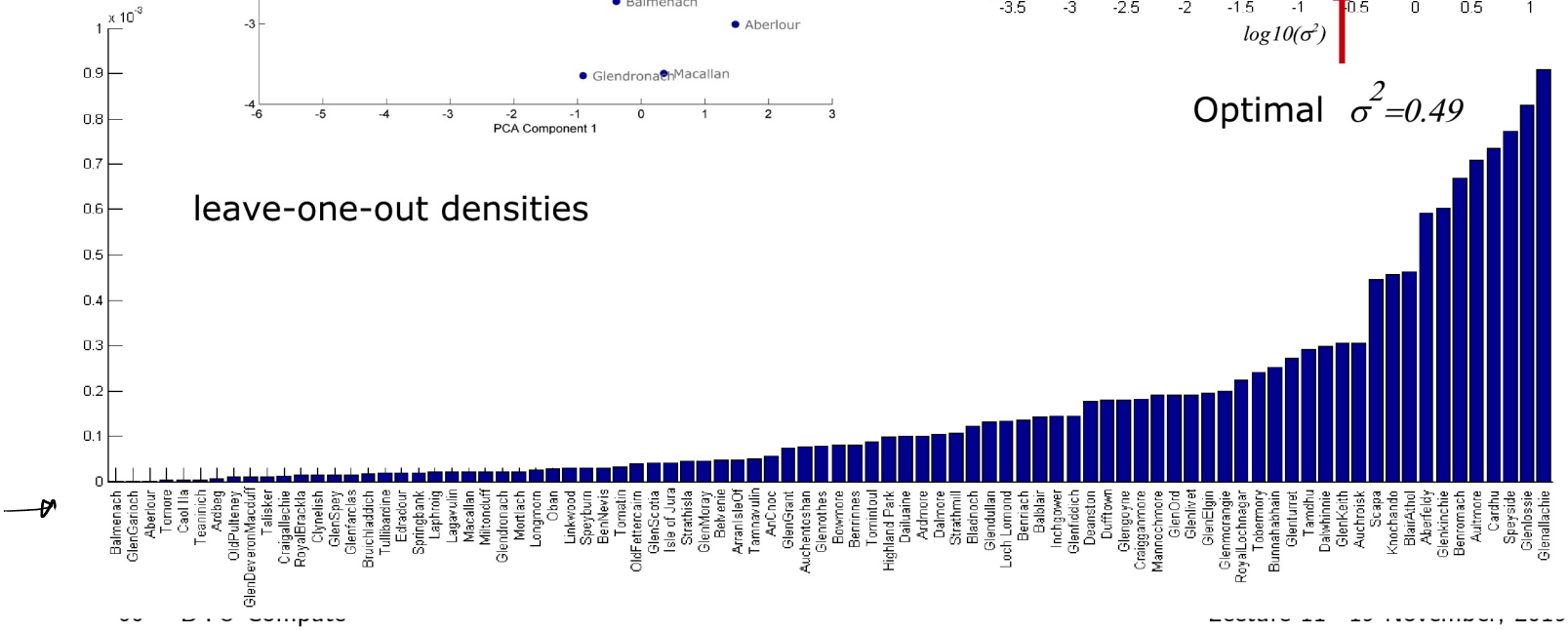


Density of test set based on leave-one-out cross validation

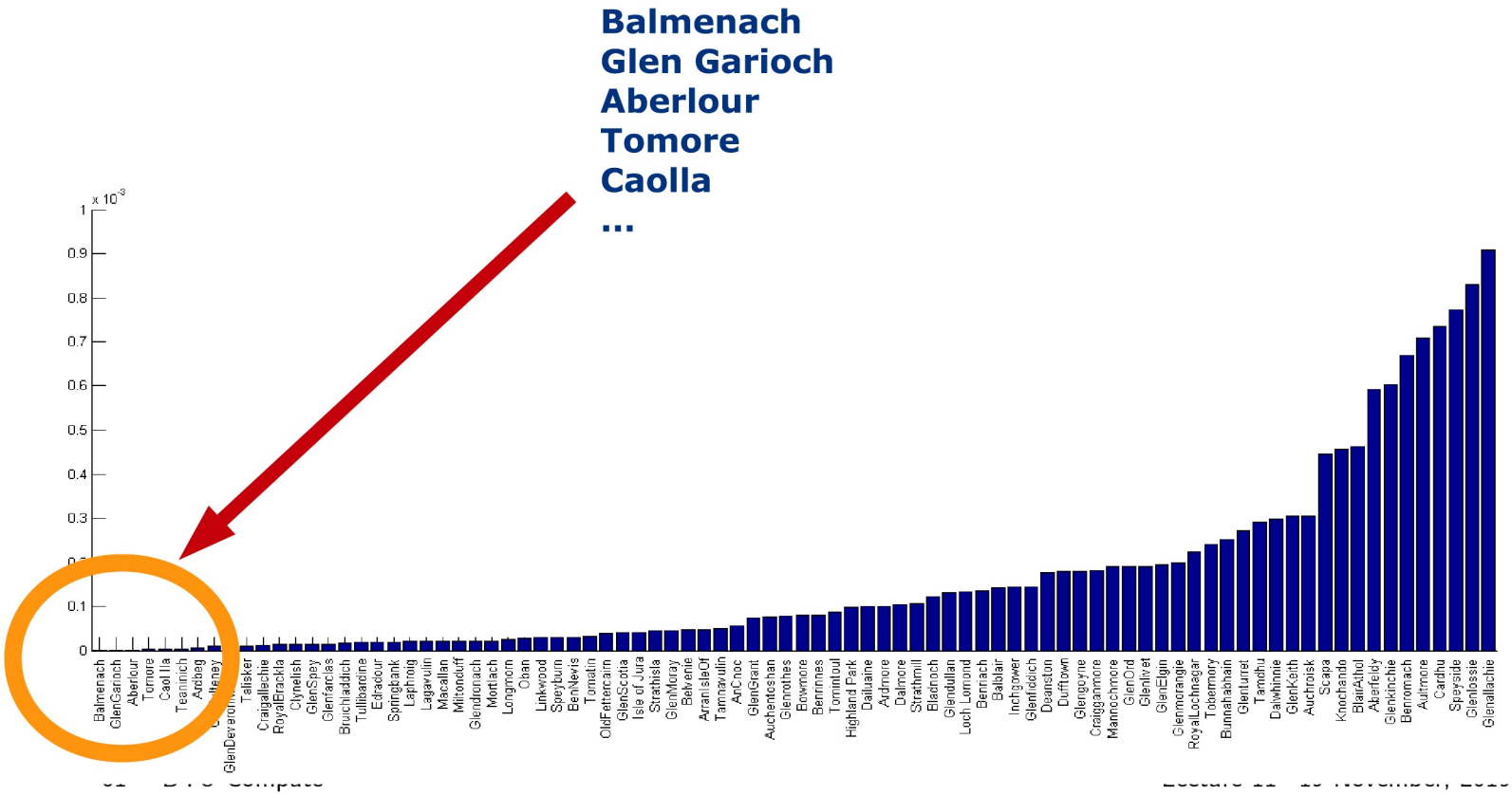


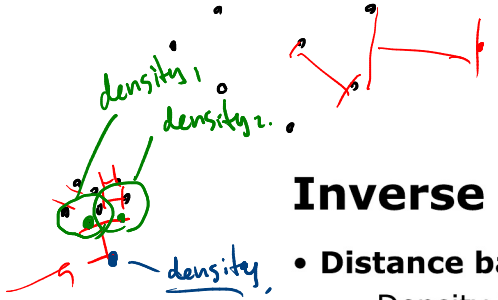
Optimal $\sigma^2 = 0.49$

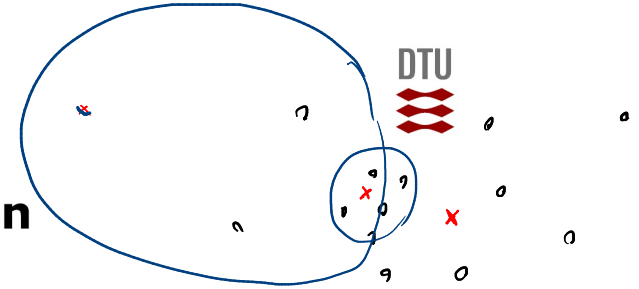
leave-one-out densities



Data II: Whisky







Inverse distance density estimation

- **Distance based measure of density**
  - Density is inverse proportional to avereage distance to k nearest neighbors
  - Density is low if nearest neighbors are far away

$$\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K) = \frac{1}{\frac{1}{K}\sum_{\boldsymbol{x}' \in N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)} d(\boldsymbol{x}_i, \boldsymbol{x}')}$$

average distance to k nn. (not including x_i)

nearest k observations to x_i who are not x_i

- **Relative density**
  - Density compared to density at nearest neighbors

$$\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K) = \frac{\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)}{\frac{1}{K}\sum_{\boldsymbol{x}_j \in N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)} \text{density}_{\boldsymbol{X}_{\setminus j}}(\boldsymbol{x}_j, K)}$$

$k = 2$

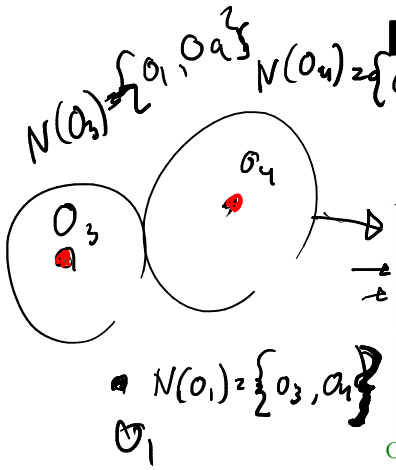
average of density$_1$, density$_2$.

$$N_{\boldsymbol{X}}(\boldsymbol{x}, K) = \{\text{The } K \text{ observations in } \boldsymbol{X} \text{ which are nearest to } \boldsymbol{x}\}$$

$$\boldsymbol{X}_{\setminus i}^T = [\boldsymbol{x}_1\ \boldsymbol{x}_2 \cdots \boldsymbol{x}_{i-2}\ \boldsymbol{x}_{i-1}\ \boldsymbol{x}_{i+1}\ \boldsymbol{x}_{i+2} \cdots \boldsymbol{x}_N]$$

These measures are taken from:
"Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, Vipin Kumar

Piazza quiz 4: ARD (Spring 2015)



	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0	393.5	68.1	165.4	271.8	200.6	210.9	206.1	166.3	365.0
O2	393.5	0	411.3	361.8	478.6	490.9	409.2	382.3	391.1	37.4
O3	68.1	411.3	0	119.8	208.4	136.6	152.8	154.3	111.1	387.1
O4	165.4	361.8	119.8	0	137.5	130.8	62.1	44.7	32.5	346.2
O5	271.8	478.6	208.4	137.5	0	99.0	76.8	101.0	116.4	468.5
O6	200.6	490.9	136.6	130.8	99.0	0	100.1	124.0	100.5	473.8
O7	210.9	409.2	152.8	62.1	76.8	100.1	0	29.5	45.2	396.8
O8	206.1	382.3	154.3	44.7	101.0	124.0	29.5	0	44.6	370.1
O9	166.3	391.1	111.1	32.5	116.4	100.5	45.2	44.6	0	375.1
O10	365.0	37.4	387.1	346.2	468.5	473.8	396.8	370.1	375.1	0

Table 1: Pairwise Euclidean distance between the 10 first observations in the PM10 (air polution) dataset. Red/green observations corresponds to high/low pollution levels.

$$\text{density}(O_1) = \frac{1}{\frac{1}{2}(68 + 165)}$$

$$\text{density}(O_3) = \frac{1}{\frac{1}{2}(68 + 111)}$$

$$\text{density}(O_4) = \frac{1}{\frac{1}{2}(45 + 33)}$$

$$\text{A r.d.}(O_1) = \frac{\text{density}(O_1)}{\frac{1}{2}(\text{density}(O_3) + \text{density}(O_4))} =$$

We suspect that observation O1 in Table 1 may be an outlier. In order to assess if this is the case we will calculate the average relative density (ARD) based on the distances in the table using the definitions:

$$\text{density}(\boldsymbol{x}, K) = \left(\frac{1}{K}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},K)} distance(\boldsymbol{x}, \boldsymbol{y})\right)^{-1},$$

$$\text{a.r.d.}(\boldsymbol{x}, K) = \frac{\text{density}(\boldsymbol{x}, K)}{\frac{1}{K}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},K)} \text{density}(\boldsymbol{y}, K)},$$

where $N(\boldsymbol{x}, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}$ and a.r.d$(\boldsymbol{x}, K)$ is the average relative density of $\boldsymbol{x}$ using K nearest neighbors. What is ARD for observation O1 for $K = 2$ nearest neighbors?

A. 0.01

B. 0.02

C. 0.23

D. 0.46

E. Don't know.

Solution:

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0	393.5	68.1	165.4	271.8	200.6	210.9	206.1	166.3	365.0
O2	393.5	0	411.3	361.8	478.6	490.9	409.2	382.3	391.1	37.4
O3	68.1	411.3	0	119.8	208.4	136.6	152.8	154.3	111.1	387.1
O4	165.4	361.8	119.8	0	137.5	130.8	62.1	44.7	32.5	346.2
O5	271.8	478.6	208.4	137.5	0	99.0	76.8	101.0	116.4	468.5
O6	200.6	490.9	136.6	130.8	99.0	0	100.1	124.0	100.5	473.8
O7	210.9	409.2	152.8	62.1	76.8	100.1	0	29.5	45.2	396.8
O8	206.1	382.3	154.3	44.7	101.0	124.0	29.5	0	44.6	370.1
O9	166.3	391.1	111.1	32.5	116.4	100.5	45.2	44.6	0	375.1
O10	365.0	37.4	387.1	346.2	468.5	473.8	396.8	370.1	375.1	0

Table 1: Pairwise Euclidean distance between the 10 first observations in the PM10 (air polution) dataset. Red/green observations corresponds to high/low pollution levels.

The intermediate computations are:

$$\text{density}(\boldsymbol{x}_{O1}, 2) = (\frac{1}{2} \cdot (68.1 + 165.4))^{-1} = 0.0086$$

$$\text{density}(\boldsymbol{x}_{O3}, 2) = (\frac{1}{2} \cdot (68.1 + 111.1))^{-1} = 0.0112$$

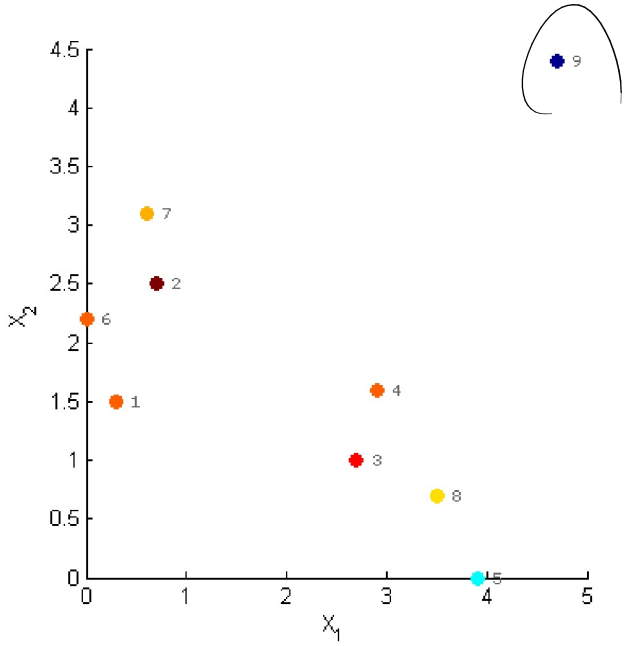
$$\text{density}(\boldsymbol{x}_{O4}, 2) = (\frac{1}{2} \cdot (32.5 + 44.7))^{-1} = 0.0259$$

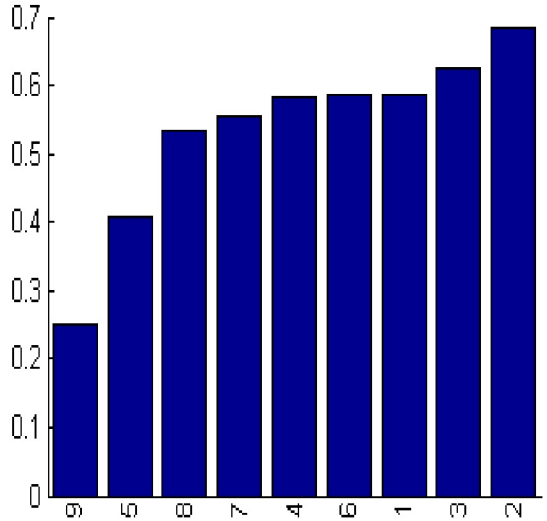
$$\begin{aligned} \text{a.r.d.}(\boldsymbol{x}, K) &= \frac{\text{density}(\boldsymbol{x}_{O1}, 2)}{\frac{1}{2}(\text{density}(\boldsymbol{x}_{O3}, 2) + \text{density}(\boldsymbol{x}_{O4}, 2))} \\ &= \frac{0.0086}{\frac{1}{2} \cdot (0.0112 + 0.0259)} = 0.46 \end{aligned}$$

Inverse distance density estimation

- KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs

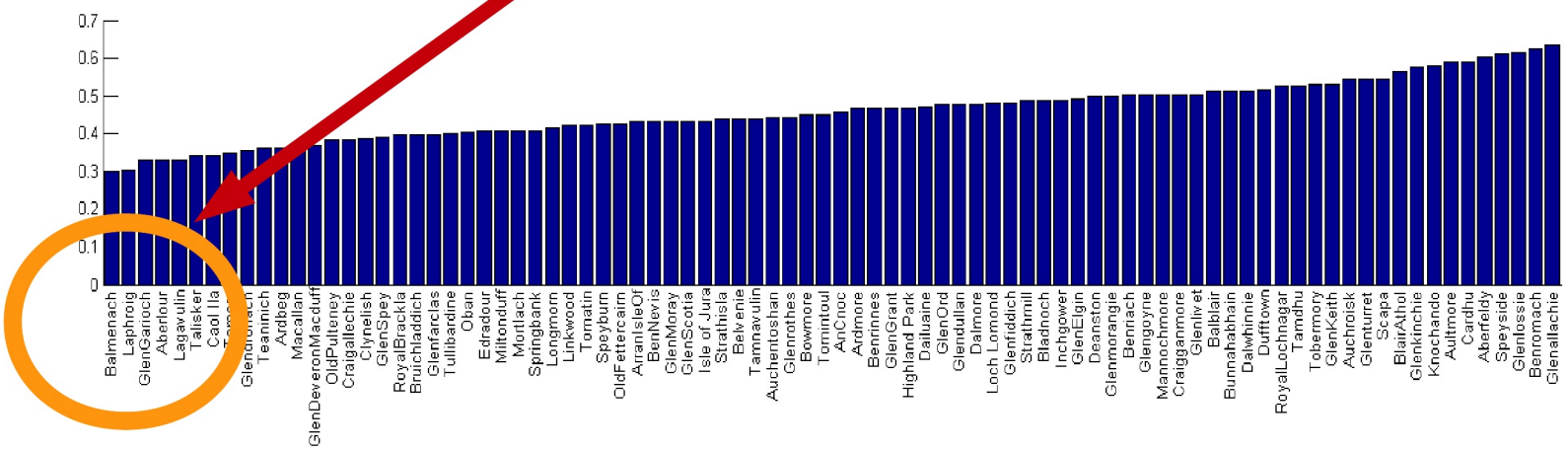




Inverse distance density estimation

- KNN density (5 nearest neighbors)

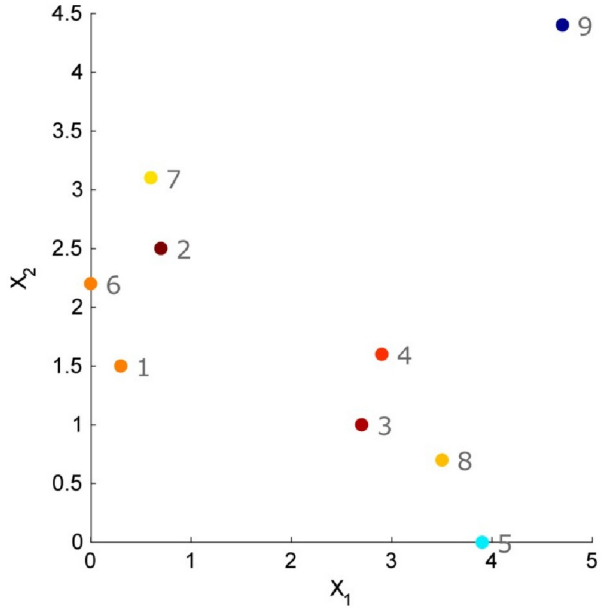
Balmenach
Laphroig
GlenGarioch
Aberlour
Lagevulin
...

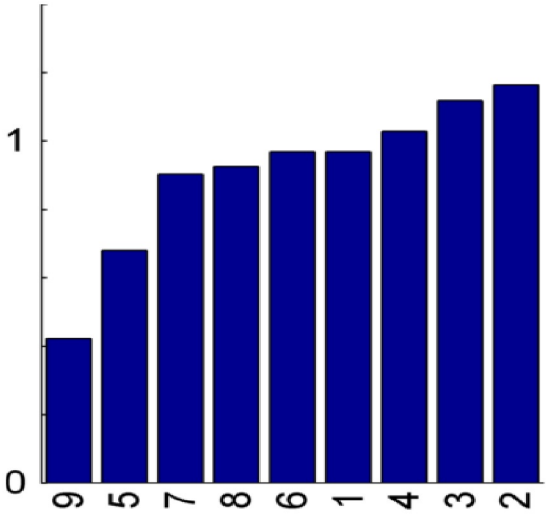


Average Relative density

- Average Relative KNN density (5 nearest neighbors)

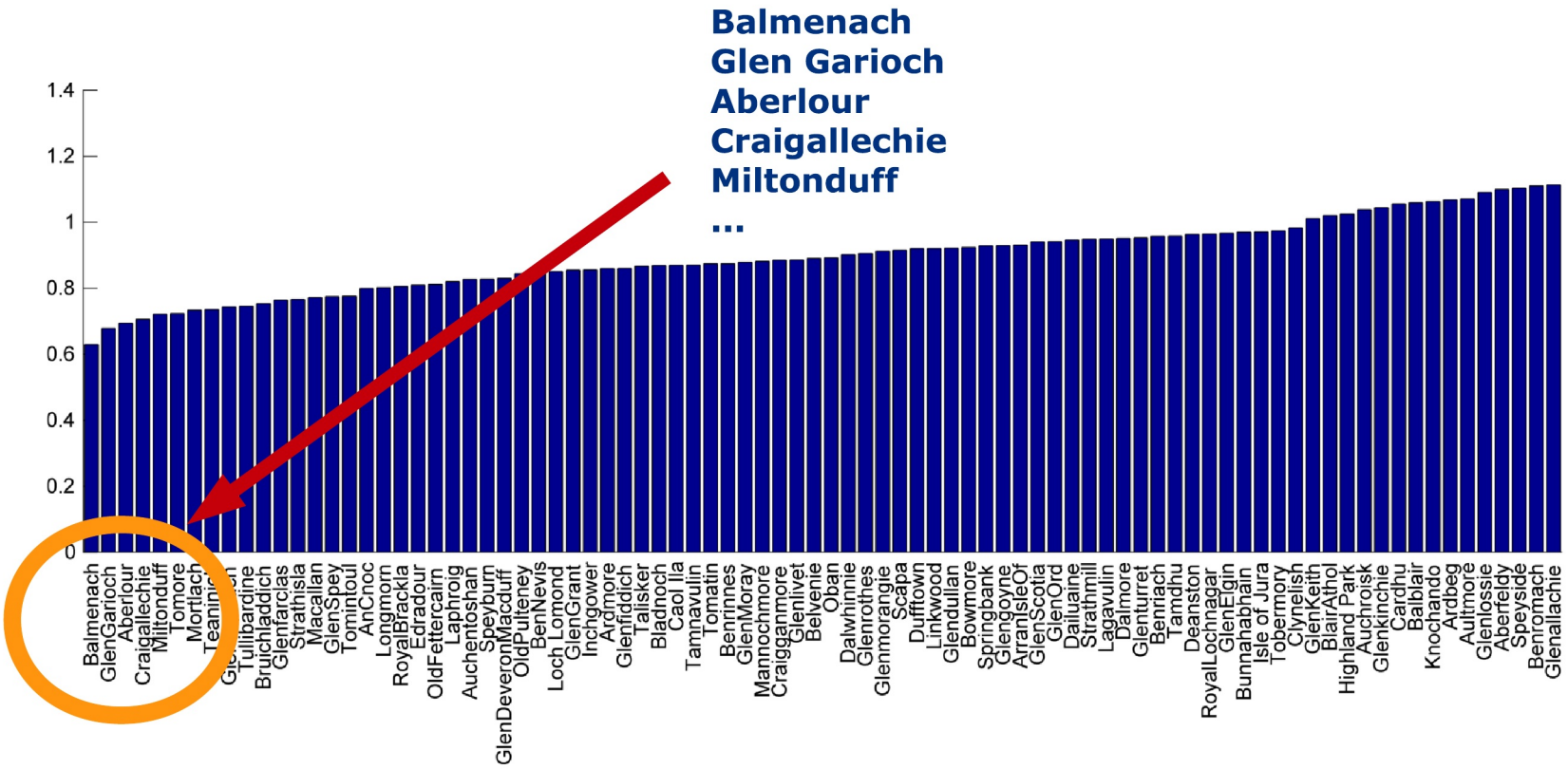
Data I: Cats , dogs and dinosaurs





Average relative density

- Average relative KNN density (5 nearest neighbors)



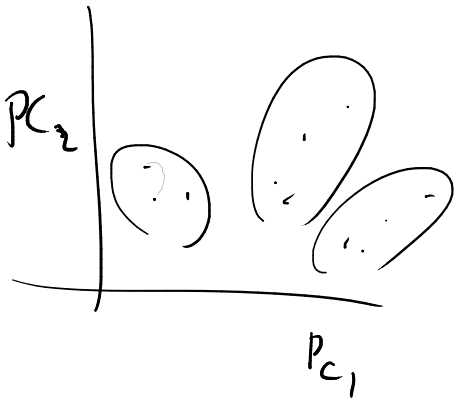
Results using different methods

- **Kernel Density Estimation**
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Tomore
  - Caolla
- **KNN density**
  - Balmenach
  - Laphroig
  - Glen Garioch
  - Aberlour
  - Lagavulin
- **KNN average relative density**
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Craigallechie
  - Miltonduff

Common: Balmenach, Glen Garioch, Aberlour

About project 3

- Assuming your (true) class labels are $\boldsymbol{y}$ and your data is $\boldsymbol{X}$
- Cluster your data in the original M-dimensional space. Remember to exclude the y-variable.
- Evaluate performance by computing similarity between clustering and $\boldsymbol{y}$ (last week)
- Plot data and clusters by projecting onto e.g. 2D space using PCA



Resources

https://www.youtube.com Nice explanation of expectation maximization for the Guassian Mixture Model (https://www.youtube.com/watch?v=WaKNSBeDLTw)