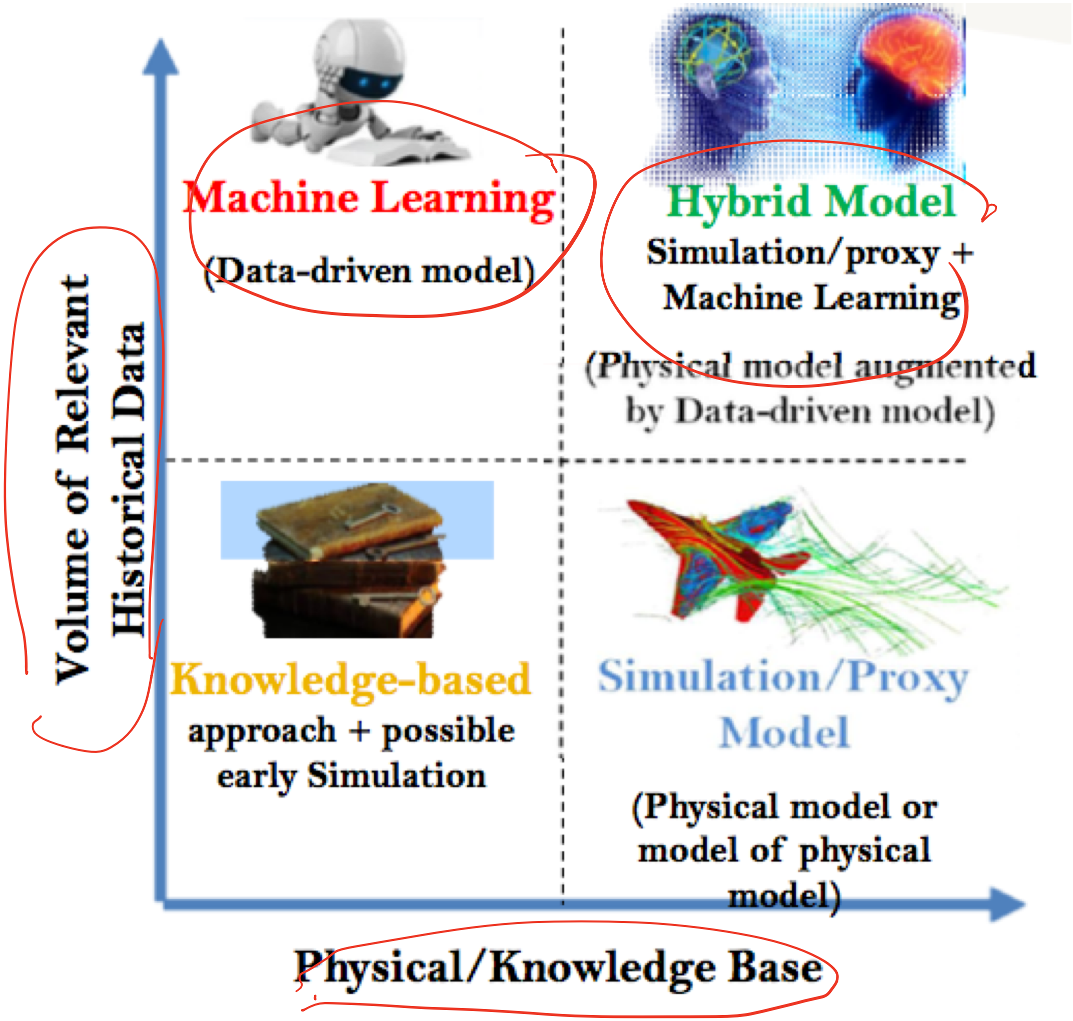
First principles vs Data driven

Data	Small data	Big data to train
Domain expertise	High reliance on domain expertise	Results with little domain knowledge
Fidelity/ Robustness	Universal link can handle non-linear complex relations	Limited by the range of values spanned by training data
Adaptability	Complex and time consuming derivation to use new relations	Rapidly adapt to new problems
Interpretability	Parameters are physical!	Physically agnostic, limited by the rigidity of the functional form

Perceived Importance. **SIO** **SP** **Peter** **Google**

Machine learning versus knowledge based



Machine Learning

(Data-driven model)

Hybrid Model

Simulation/proxy +
Machine Learning

(Physical model augmented
by Data-driven model)

Knowledge-based
approach + possible
early Simulation

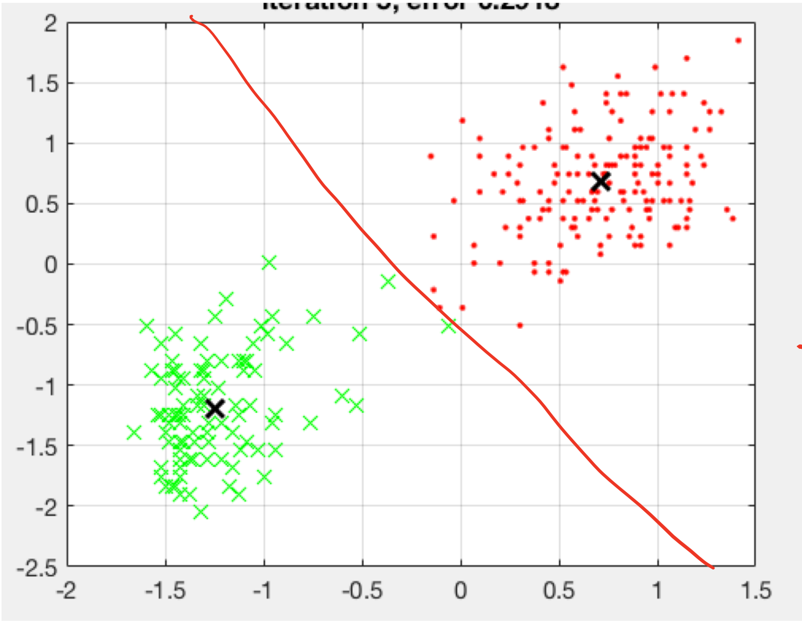
**Simulation/Proxy
Model**

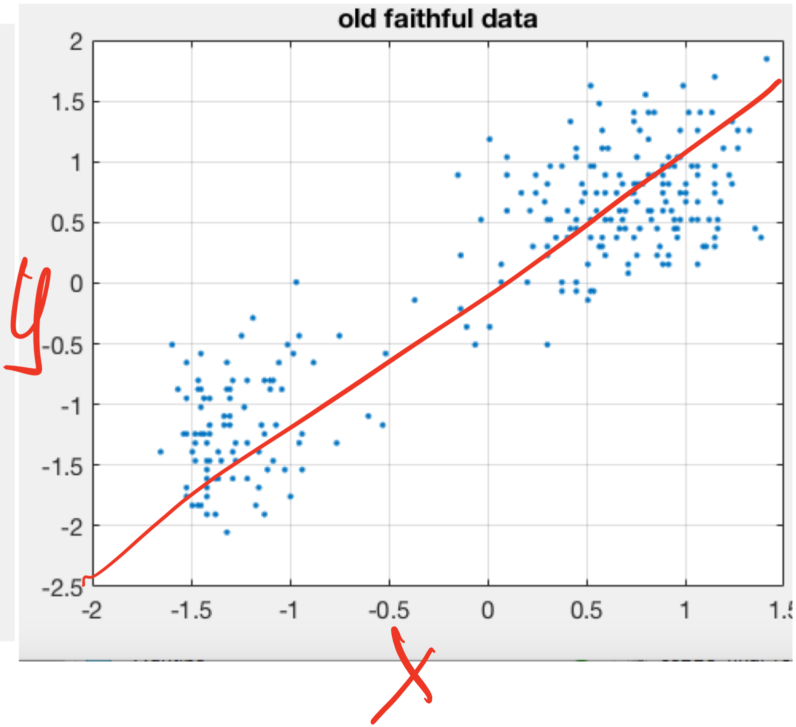
(Physical model or
model of physical
model)

Volume of Relevant
Historical Data

Physical/Knowledge Base

Supervised learning



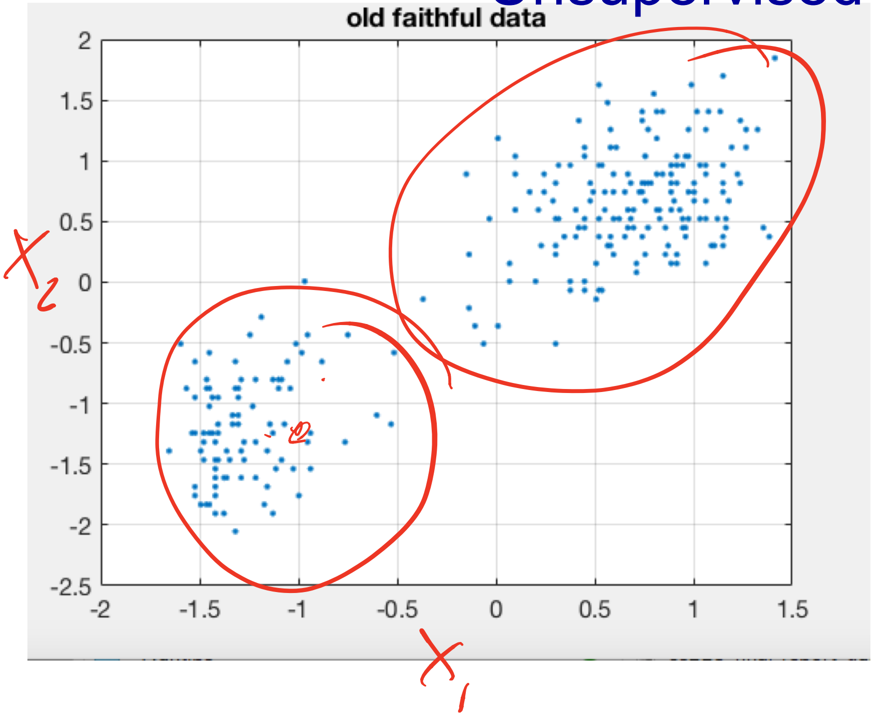


Supervised $\underline{\mathbf{y}=\mathbf{w}^T\mathbf{x}}$

Training set $\{(x^1, y^1), (x^2, y^2), (x^3, y^3)\}$

We are given the two classes.

Unsupervised learning



~~Supervised y=wx~~

Training set $\{(x_1^1, x_2^1), (x_1^2, x_2^2), (x_1^3, x_2^3)\}$

Unsupervised learning

Unsupervised machine learning is inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations). Since the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm—which is one way of distinguishing unsupervised learning from supervised learning.

We are not interested in prediction

Supervised learning: all classification and regression.

$$\boldsymbol{Y} = \boldsymbol{W}^T \boldsymbol{X}$$

Prediction is important.

10^{14} $\pi \cdot 10^{7}$

Unsupervised learning

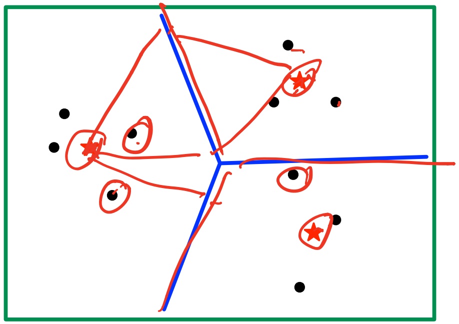
- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- But techniques for unsupervised learning are of growing importance in several fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.
- It is often easier to obtain unlabeled data — from a lab instrument or a computer — than labeled data, which can require human intervention.
  - For example it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?

Kmeans

- **Input**: Points $\mathbf{x}_1,\ldots,\mathbf{x}_N \in R^p$; integer $\underline{K}$
- **Output**: "Centers", or representatives, $\mu_1,\ldots,\mu_K \in R^p$
- Output also $\mathbf{z}_1,\ldots,\mathbf{z}_N \in R^K$

Goal: Minimize average squared distance between points and their nearest representatives:

- $cost(\mu_1,\ldots,\mu_K) = \sum_{n=1}^{N} \min_j \|\underline{x_n} - \underline{\mu_j}\|$



The centers carve $\mathbb{R}^p$ up into k convex regions: μ_j's region consists of points for which it is the closest center.

K-means

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \tag{9.1}$$

Solving for r$_{nk}$

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

Differentiating for μ_k

$$\frac{\partial J}{\partial \mu_k}$$

$$2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{9.3}$$

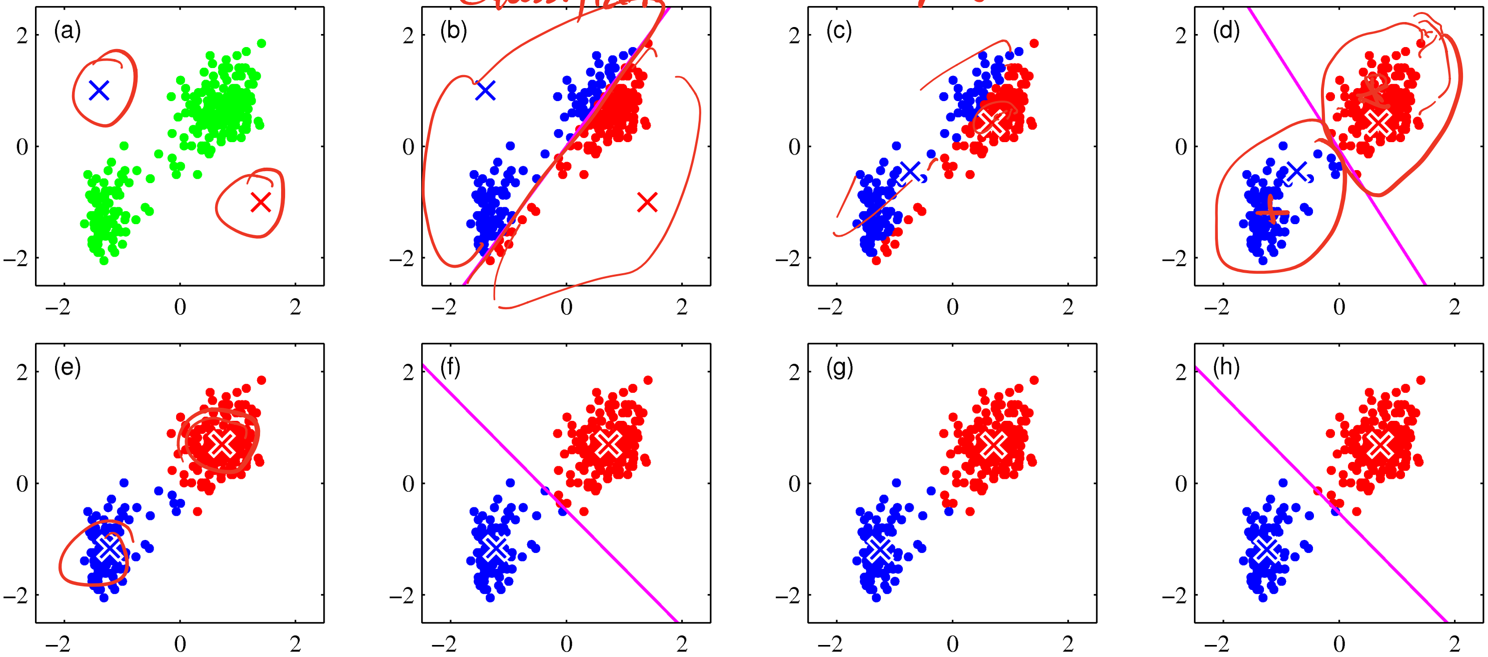
which we can easily solve for $\boldsymbol{\mu}_k$ to give

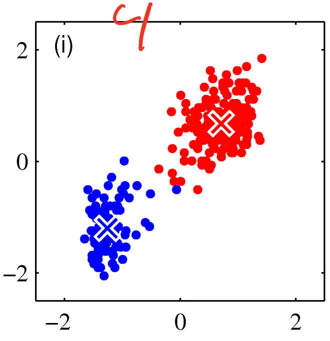
$$\underline{\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\boxed{\sum_n r_{nk}}}} \quad N_k \tag{9.4}$$

classification

K-means

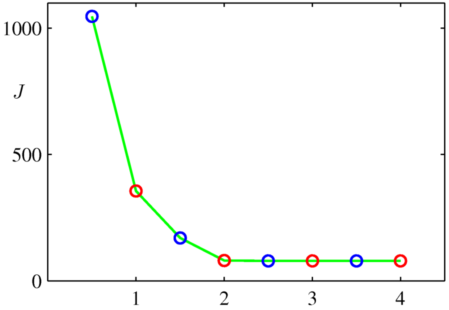
μ_n





$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$$



Old Faithful, Kmeans from Murphy

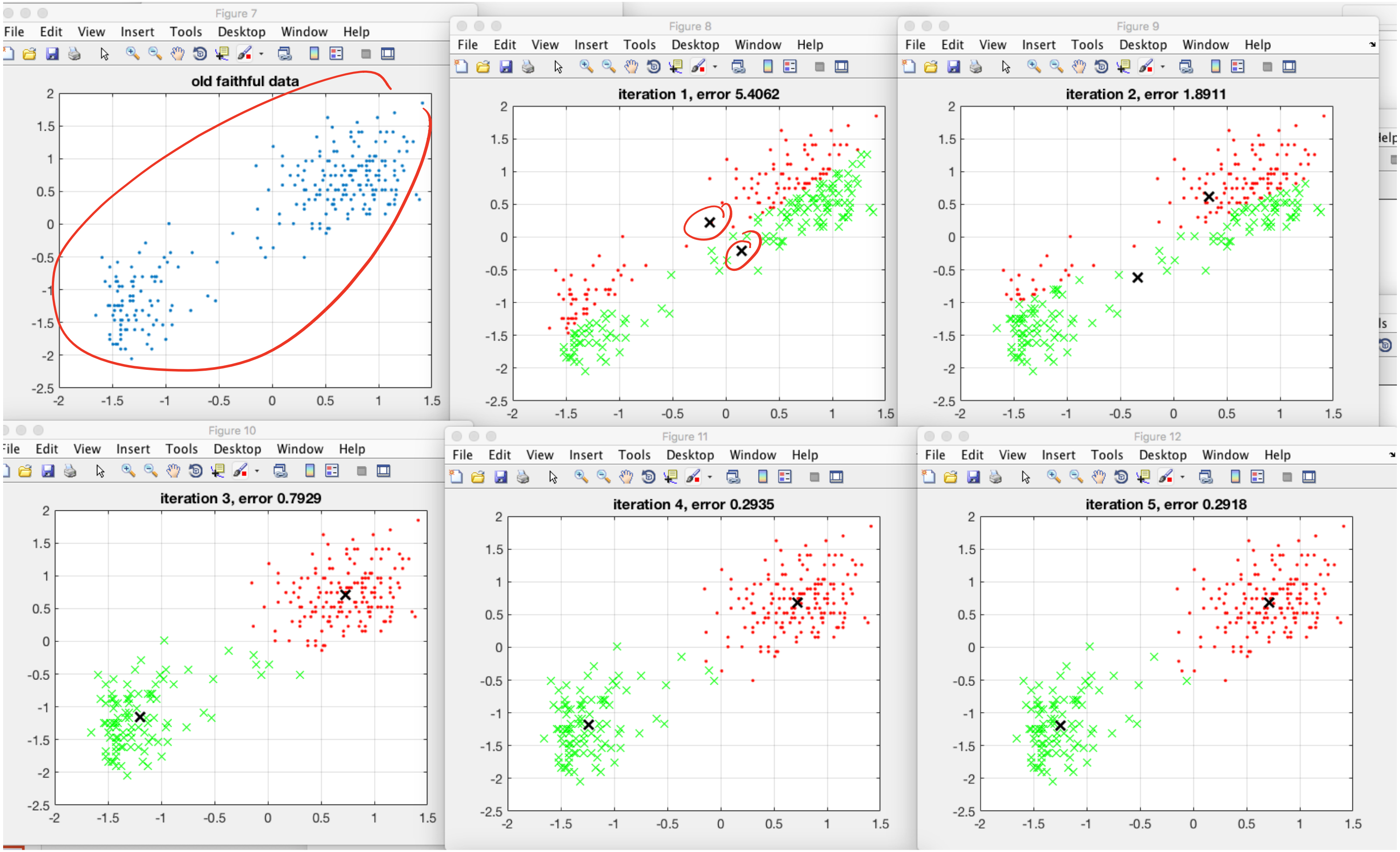
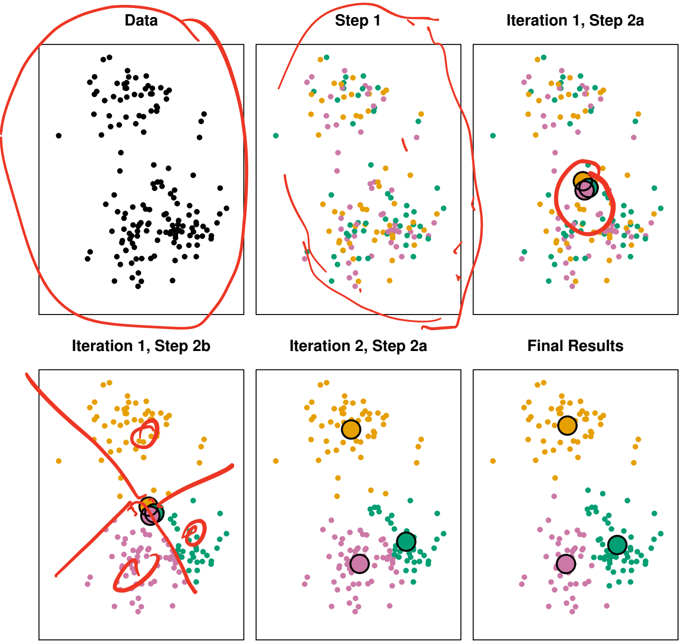


Figure 7
File Edit View Insert Tools Desktop Window Help
old faithful data
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
Figure 8
iteration 1, error 5.4062
Figure 9
iteration 2, error 1.8911
Figure 10
iteration 3, error 0.7929
Figure 11
iteration 4, error 0.2935
Figure 12
iteration 5, error 0.2918

Example

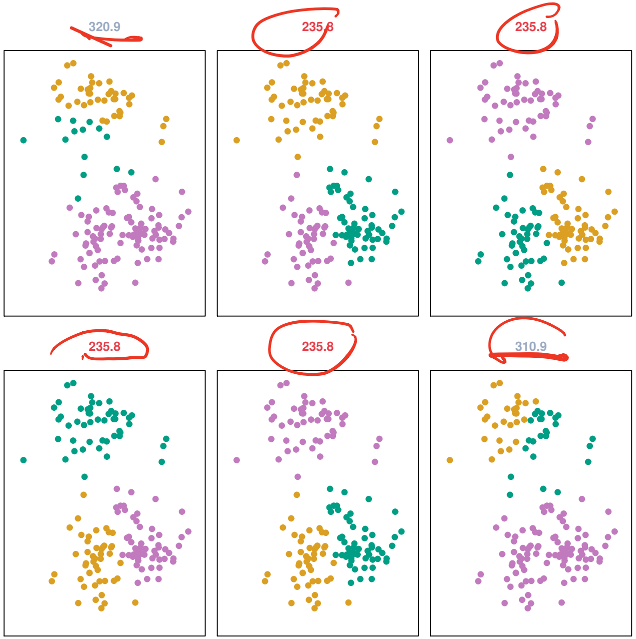


The progress of the K-means algorithm with $K=3$.

- *Top left:* The observations are shown.
- *Top center:* In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
- *Top right:* In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
- *Bottom left:* In Step 2(b), each observation is assigned to the nearest centroid.
- *Bottom center:* Step 2(a) is once again performed, leading to new cluster centroids.
- *Bottom right:* The results obtained after 10 iterations.

Likely From Hastie book

Different starting values



K-means clustering performed six times on the data from previous figure with $K = 3$, each time with a different random assignment of the observations in Step 1 of the K-means algorithm.

Above each plot is the value of the objective (4).

Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters.

Those labeled in red all achieved the same best solution, with an objective value of 235.8

Likely From Hastie book

Vector Quantization VQ

Murphy book Fig 11.12 vqdemo.m

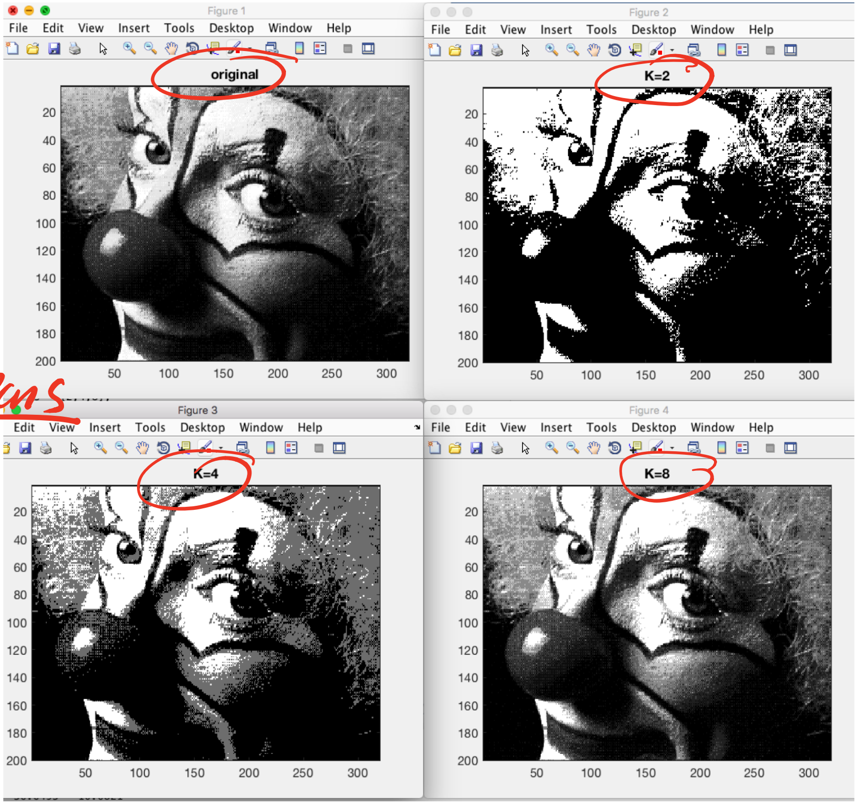
Each pixel $\mathbf{x}_i$ is represented
By codebook of K entries μ_k

Encode($\mathbf{x}_i$)=$\underset{k}{\operatorname{argmin}} \|x_i - \mu_k\|$ k-means

Consider N=64k observations, of
D=1 (b/w) dimension, C=8 bit

NC=513k

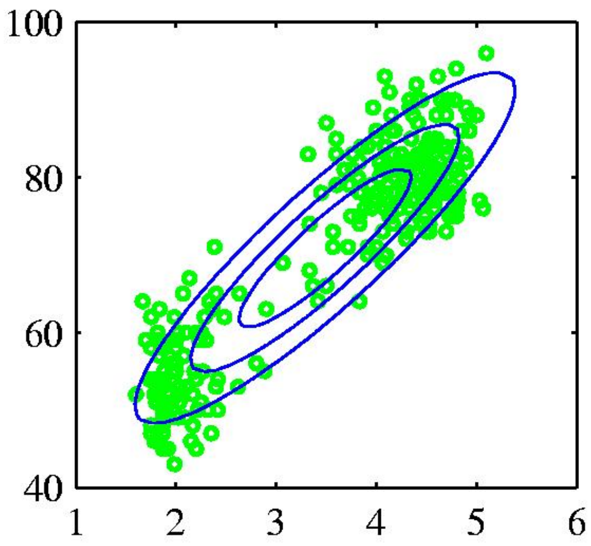
Nlog$_2$ K+KC bits is needed
K=4 gives 128k a factor 4.



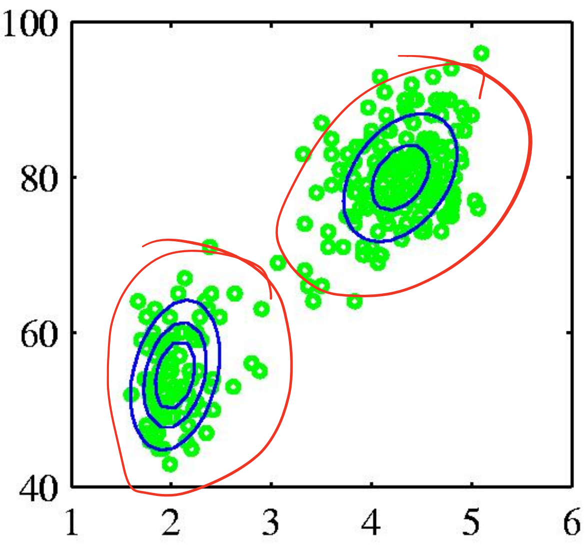
Mixtures of Gaussians (1)

Old Faithful geyser:

The time between eruptions has a bimodal distribution, with the mean interval being either 65 or 91 minutes, and is dependent on the length of the prior eruption. Within a margin of error of ±10 minutes, Old Faithful will erupt either 65 minutes after an eruption lasting less than 2 $\frac{1}{2}$ minutes, or 91 minutes after an eruption lasting more than 2 $\frac{1}{2}$ minutes.



Single Gaussian

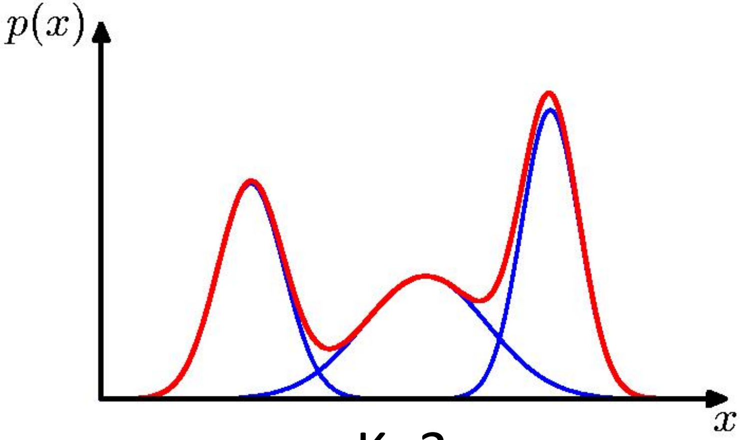


Mixture of two Gaussians

Mixtures of Gaussians (2)

Combine simple models
into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underset{\uparrow \atop \text{Mixing coefficient}}{\pi_k} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

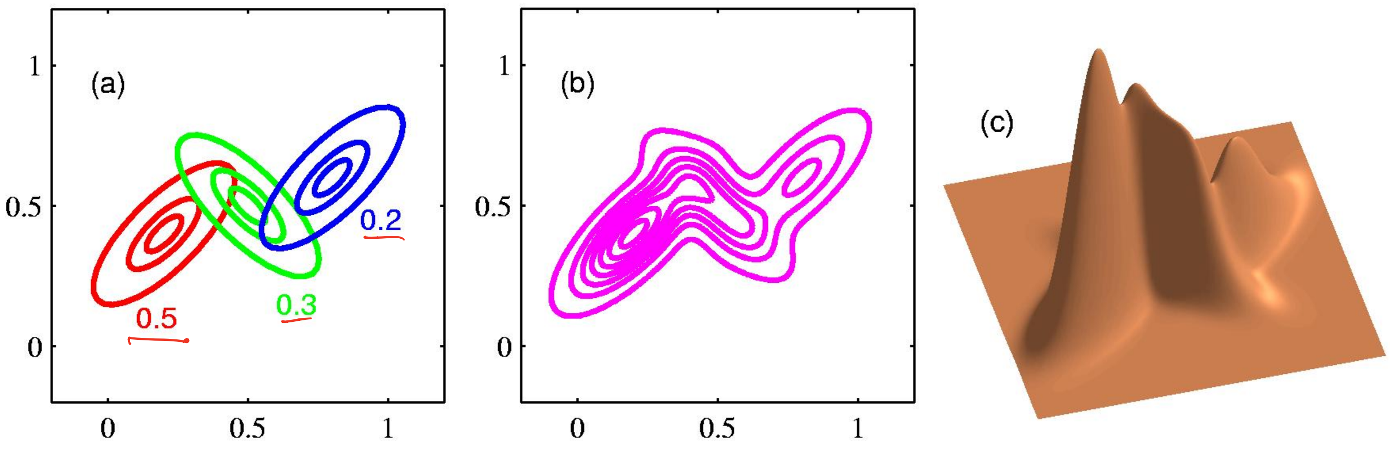


K=3

$$\forall k : \underline{\pi_k \geqslant 0} \qquad \sum_{k=1}^{K} \pi_k = 1$$

$\pi_k \leq 1$

Mixtures of Gaussians (3)

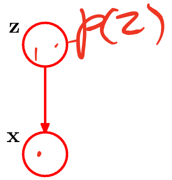


- Gaussian mixture
  - $p(x) = \sum_k^K \pi_k N(x;\mu_k,\Sigma_k)$

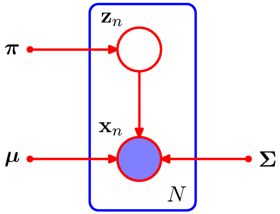
$z = \left[\quad\right]_K \qquad z_k \in \{0, 1\}$

- Latent variable:
  - Un-observed
  - Often hidden

- Here $p(z_k) = \pi_k$



p(z)p(x|z)



N iid {$\mathbf{x}_n$} with latent {$\mathbf{z}_n$}

$x \in R^D$ $\mu \in R^D$ $\Sigma_k \in R^{D \times D}$

$= \sigma^2 I$

$$p(\boldsymbol{x}|z_k = 1) = N(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$p(x|z_k) = N(x; \mu_k, \Sigma_k)^{z_k}$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod N(x; \mu_k, \Sigma_k)^{z_k}$$

$p(z_k = 1) = \pi_k$; $p(z) = \sum_k \pi_k^{z_k}$

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(x|z)\ p(z)$$

$$p(\boldsymbol{x}) = \sum_z p(x,z) = \sum_z p(x|z)\, p(z) = \sum \pi_k N(x; \mu_k, \Sigma_k)$$

Responsibilities Bayes

$$\gamma(z_k) = p(z_k = 1|\boldsymbol{x}) = \frac{p(x|z_k=1)\, p(z_k=1)}{\sum_j p(x|z_j=1)\, p(z_j=1)} = \frac{\pi_k N(x, \mu_k, \Sigma_k)}{\sum_j \pi_j N(x, \mu_j, \Sigma_j)}$$

Mixture of Gaussians

- Mixtures of Gaussians

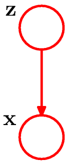
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Expressed with latent variable **z**

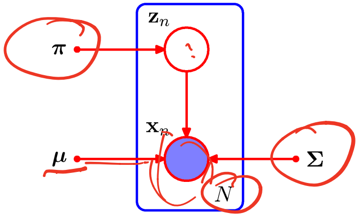
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Posterior probability: **responsibility**

$$\underline{\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})} = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$



p(z)p(x|z)



N iid {**x**$_n$} with latent {**z**$_n$}

Max Likelihood

$$X = \begin{array}{c} \overset{D}{\longleftrightarrow} \\ \begin{bmatrix} - x_1^T - \\ \\ x_n \end{bmatrix} \end{array} N$$

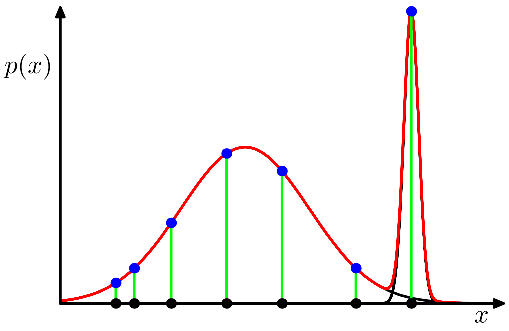
- $\underline{p(x)} = \sum_k^K \pi_k N(x; \mu_k, \Sigma_k)$
- N observations $\boldsymbol{X}$
- $\ln[p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)] = \prod_N \underline{\ln[\sum_k^K \pi_k N(x_n; \mu_k, \Sigma_k)]}$

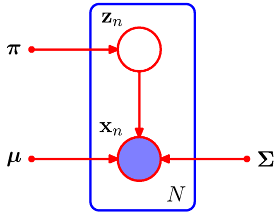
$\ell(\pi, \mu, \Sigma) =$ $\quad \frac{\partial L}{\partial \Sigma_k}$ $\quad \sum \pi_k = 1$

$\frac{\partial L}{\partial \mu_k}$

$g(x) = \ell + \lambda(\underline{\sum \pi_k - 1})$

$\frac{\partial g}{\partial \pi_k} = 0 = \sum \frac{N(x; \mu_k, \Sigma_k)}{\gamma} + \lambda$





N iid $\{\mathbf{x}_n\}$ with latent $\{\mathbf{z}_n\}$

EM Gauss Mix

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{9.27}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{9.28}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

General EM

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.

2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

   where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

   and return to step 2.

EM in general

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.69}$$

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{9.70}$$

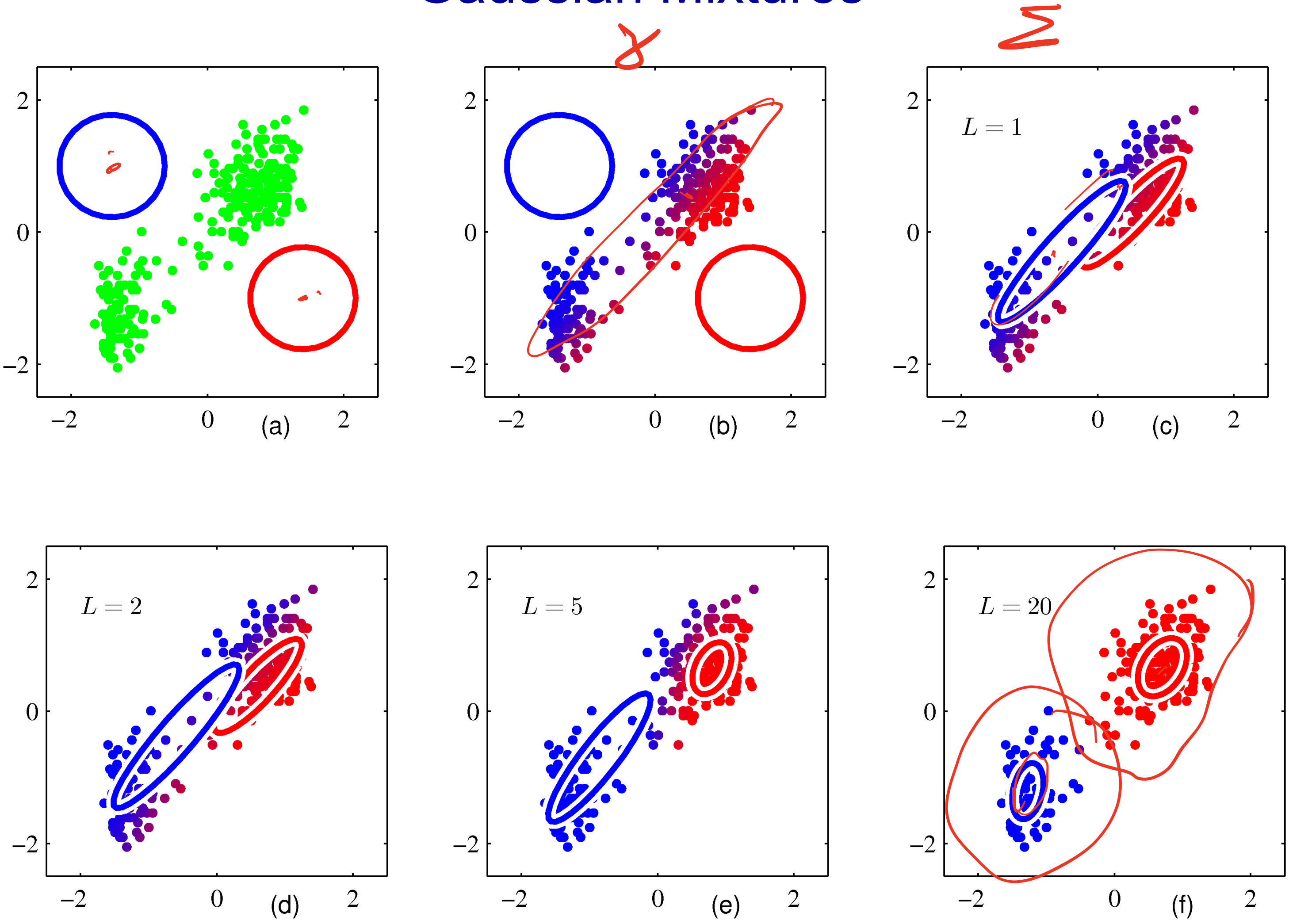
where we have defined

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} & (9.71) \\ \mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. & (9.72) \end{aligned}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta}) \tag{9.73}$$

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{const} & (9.74) \end{aligned}$$

Gaussian Mixtures



Kmeans and EM (9.3.2)

$$\boldsymbol{\Sigma}_k = \underline{\epsilon}\boldsymbol{I}$$

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \underline{\frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}}. \qquad (9.41)$$

Whereby the responsibilities

$$\gamma(z_{nk}) = \left. \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}. \right] = 1 \qquad (9.42)$$

Becomes delta functions. $\mu_k = x_n$

And the EM ~~means~~ approach the Kmeans

$$\underline{\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n} \qquad (9.17)$$

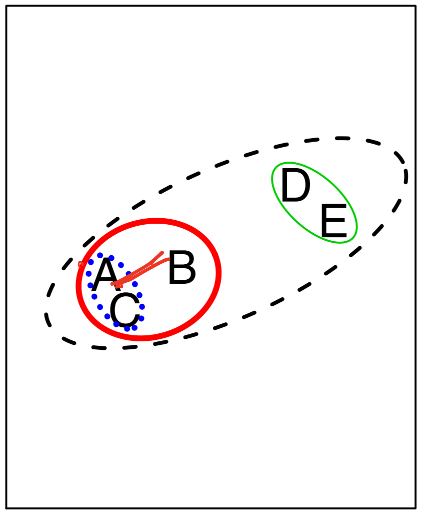
Hierarchical Clustering

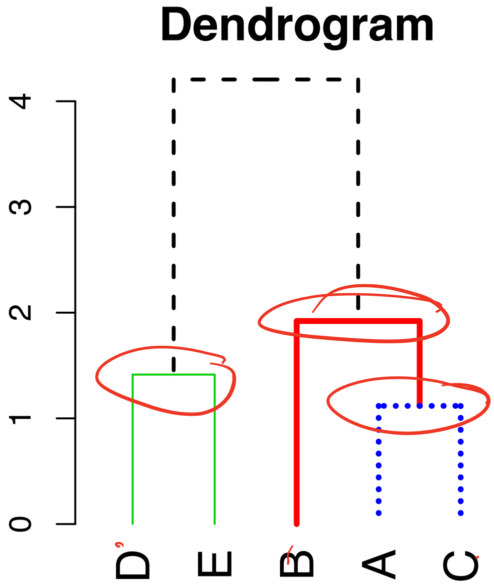
- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K)
- *Hierarchical clustering* is an alternative approach which does not require that we commit to a particular choice of K.
- In this section, we describe *bottom-up* or *agglomerative* clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

Hierarchical Clustering Algorithm

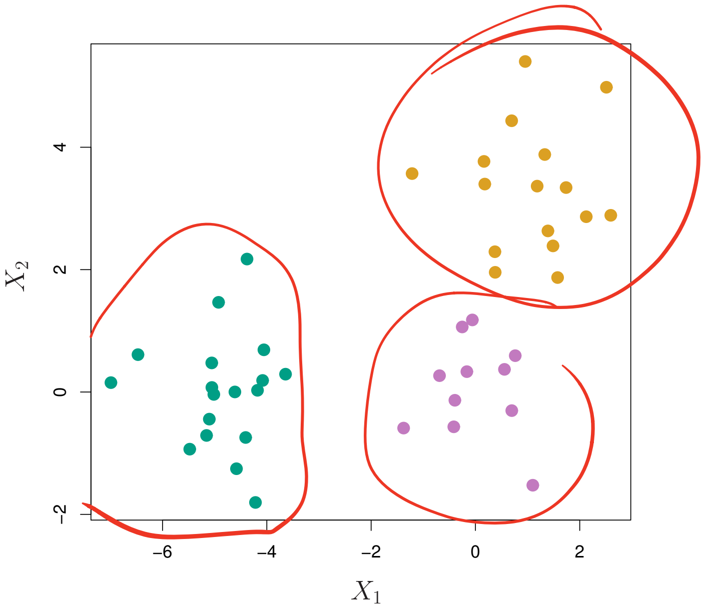
The approach in words:

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.



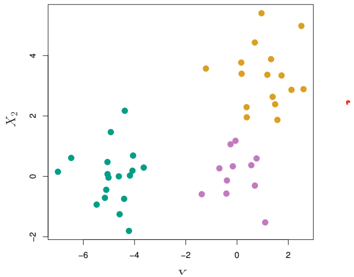


An Example

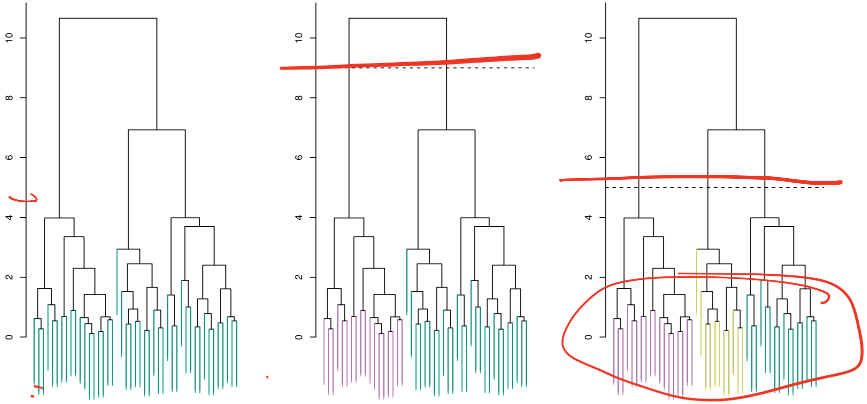


45 observations generated in 2-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

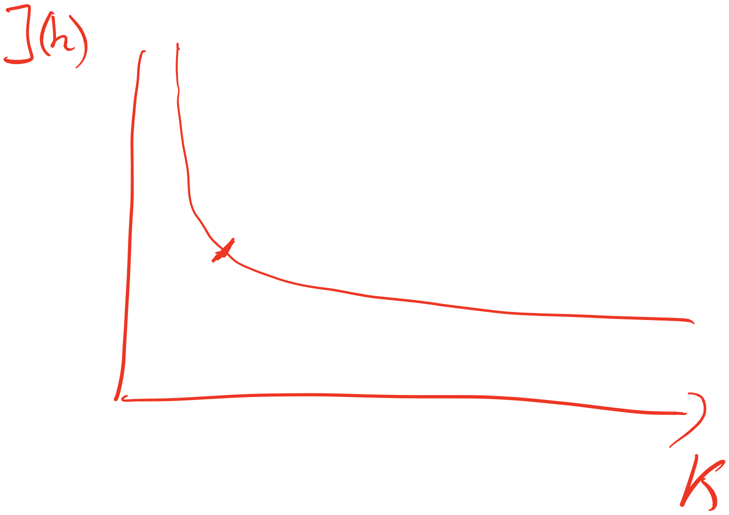
Example



45 observations generated in 2-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.



- *Left:* Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- *Center:* The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- *Right:* The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure


J(h)
K