Machine Learning

Session 4: Generative Models II. Mixture of Gaussians

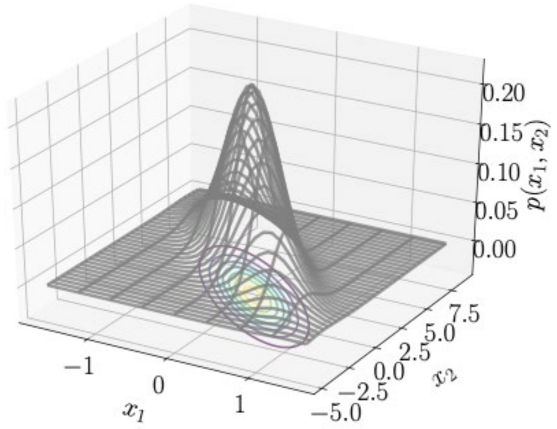
1. Limitations of the Gaussian Model
2. Mixture of Gaussians and hidden variables
   1. Mixture Sampling
   2. Mixture interpretation
   3. Learning parameters
3. The Expectation Maximization Algorithm
4. Examples

Bibliography:

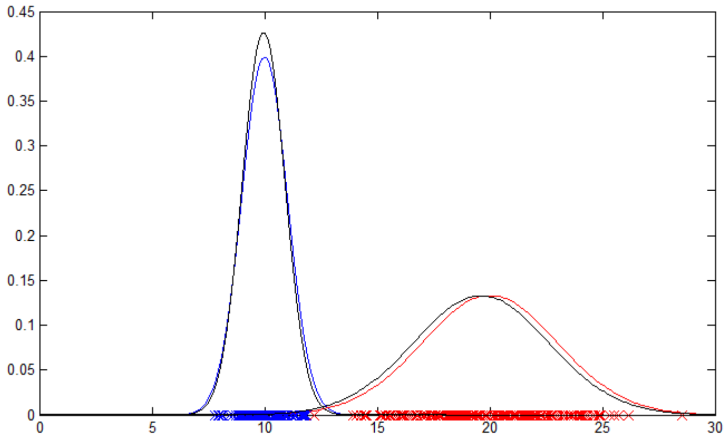
- Deisenroth et al. Mathematics for Machine Learning CUP 2020 Ch 11.
- C. Bishop: 2.3.9, 9.2, 9.2.1, 9.2.2, 9.3, 9.3.1, 9.3.2
- Video of EM 1D: https://www.youtube.com/watch?v=XLKoTqGao7U

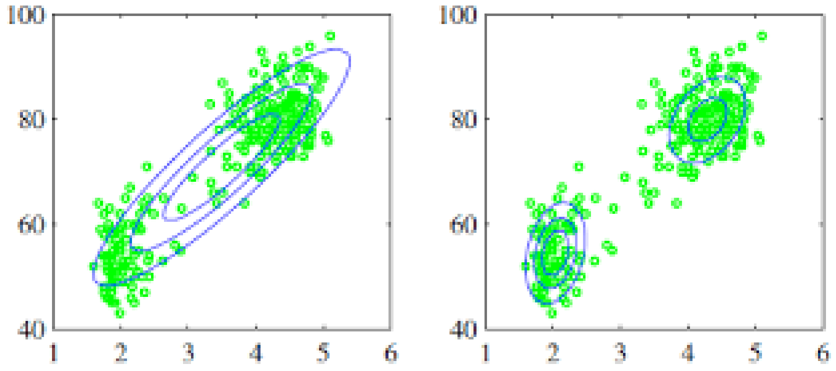
Limits of the previous models of Clustering

- **K-means**: for circular clusters, converges to a local optimum. Assignments of data points to clusters is hard
- **Gaussian Model**: good model but it is only unimodal



- We need a multimodal model: the points come from different populations



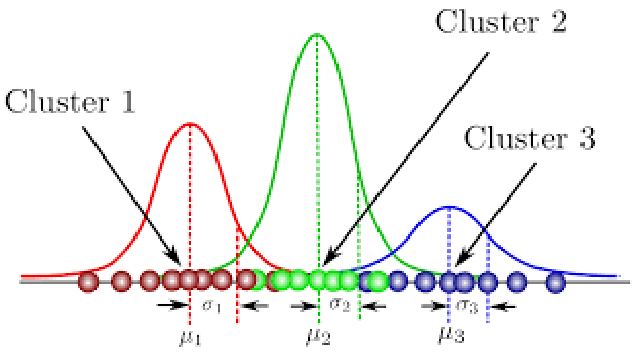


- We need a more powerful model: "linear superposition" of two o more Gaussians formulated as a probability model to assure its predictive and generative abilities.

Going to Mixtures of Gaussians

Imagine if we knew the assignment: the component of each observation:

- The calculus of each Gaussian cluster is easy:



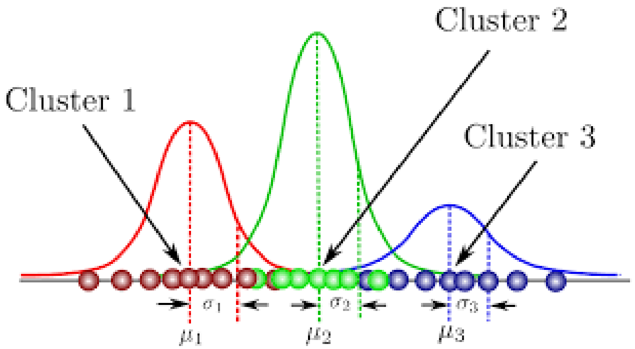
$$\mathcal{N}_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu_{ML} &= E(x) = \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \\ \sigma^2_{ML} &= E\{(x-E(x))^2\} = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)}-\mu_{ML})^2 \end{aligned}$$

Going to Mixtures of Gaussians

Imagine if we knew the assignment: the component of each observation:

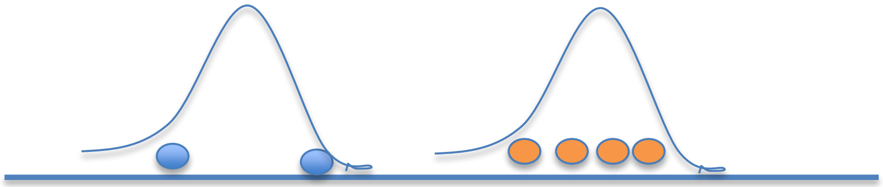
- The calculus of each Gaussian cluster is easy:



$$\mathcal{N}_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu_{ML} &= E(x) = \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \\ \sigma^2_{ML} &= E\{(x-E(x))^2\} = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)}-\mu_{ML})^2 \end{aligned}$$

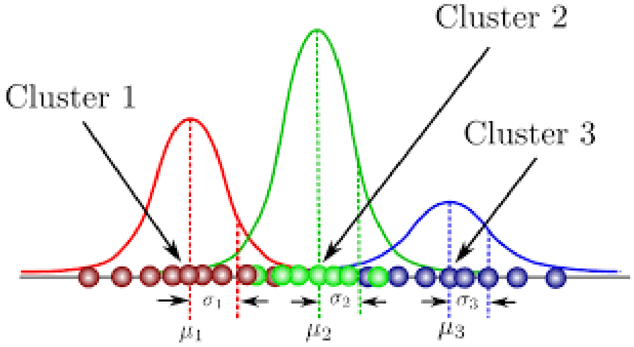
-We need also to reflex the weight of each cluster, otherwise two clusters with very different number of points could be summarized as two Gaussian displaced



Going to Mixtures of Gaussians

Imagine if we knew the assignment: the component of each observation:

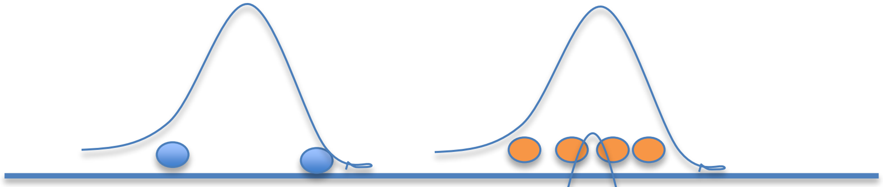
- The calculus of each Gaussian cluster is easy:



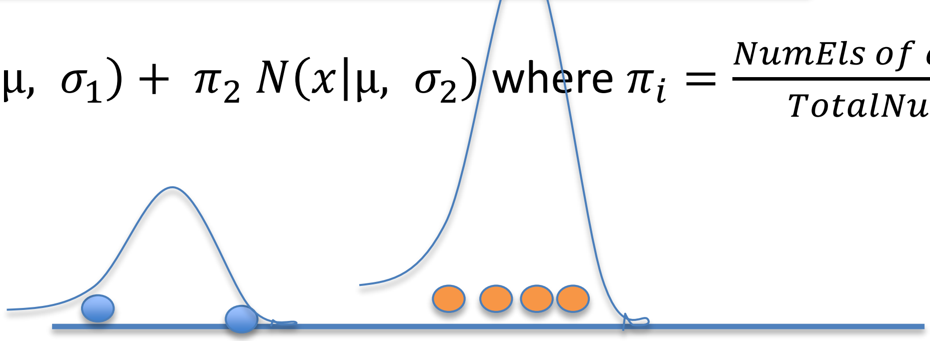
$$\mathcal{N}_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu_{ML} &= E(x) = \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \\ \sigma^2_{ML} &= E\{(x-E(x))^2\} = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)}-\mu_{ML})^2 \end{aligned}$$

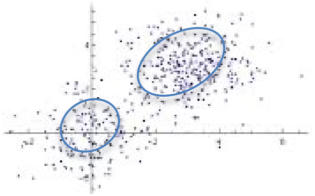
-We need also to reflex the weight of each cluster, otherwise two clusters with very different number of points could be summarized as two Gaussian displaced



-We want $\pi_1\, N(x|\mu,\ \sigma_1) + \pi_2\, N(x|\mu,\ \sigma_2)$ where $\pi_i = \frac{NumEls\ of\ cluster\ i}{TotalNumEls}$



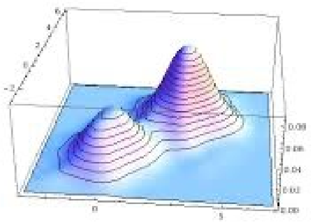
Going to Mixtures of Gaussians



What is the problem? The responsibility of each Component is unknown

Imagine if we knew the component corresponding to each observation:

- Each Gaussian cluster model is:



$$\mathcal{N}_k(\mathbf{x}|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$$

1. Calculate the Mean $\boldsymbol{\mu}_k$ and $\mathbf{Z}_k = \mathbf{Y} - \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{\boldsymbol{k}} = \frac{\mathbf{1}}{\boldsymbol{N}}\mathbf{Z}_{\mathrm{k}}\boldsymbol{Z}_k^\top$ is the covariance matrix

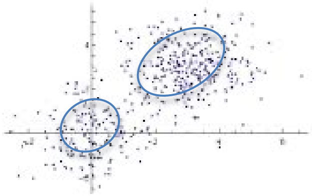
2. The SVD of $\boldsymbol{\Sigma}_{\boldsymbol{k}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ The columns of $\mathbf{V}$ are the new basis

3. The values of $\mathbf{D}$ are the variance along each new vector basis

4. The Mahalanobis distance $\boxed{\Delta^2 = (\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu) = \frac{y_1^2}{\lambda_1} + \cdots + \frac{y_D^2}{\lambda_D}}$

5. The ellipses will be the points: $\mathrm{V}\sqrt{D}\begin{pmatrix} \cos\left(\frac{2\pi}{M}\right) & \ldots & \cos\left(M\,\frac{2\pi}{M}\right) \\ \sin\left(\frac{2\pi}{M}\right) & \ldots & \sin\left(M\,\frac{2\pi}{M}\right) \end{pmatrix} + \boldsymbol{\mu}_k$

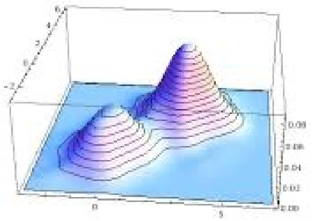
Going to Mixtures of Gaussians



What is the problem? The responsibility of each Component is unknown

Imagine if we knew the component corresponding to each observation:

- Each Gaussian cluster model is:

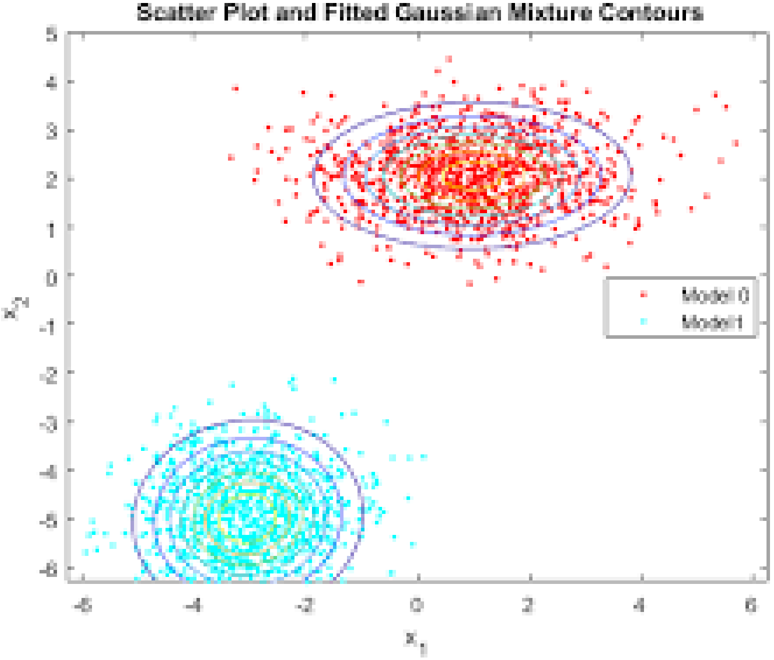


$$\mathcal{N}_k(\mathbf{x}|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\top}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$$

- We want (for two clusters):

$$\pi_1 N_1(\boldsymbol{x}|\,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+\ \pi_2 N_2(\boldsymbol{x}|\,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$$

where $\pi_i = \frac{NumEls\ of\ cluster\ i}{TotalNumEls}$



Mixtures of Gaussians

Mixtures of Gaussians:

- A generalization of the Gaussian model to multiple modes

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

Mixtures of Gaussians

Mixtures of Gaussians:

- A generalization of the Gaussian model to multiple modes

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

- Each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a ***component***

Mixtures of Gaussians

Mixtures of Gaussians:

- A generalization of the Gaussian model to multiple modes

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

- Each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a ***component***
- Parameters π_k are called ***mixing coefficients***

Mixtures of Gaussians

Mixtures of Gaussians:

- A generalization of the Gaussian model to multiple modes

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

- Each Gaussian density $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a ***component***
- Parameters π_k are called ***mixing coefficients***
- $\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a ***constituent*** of the Mixture

– They satisfy the requirements to be probabilities

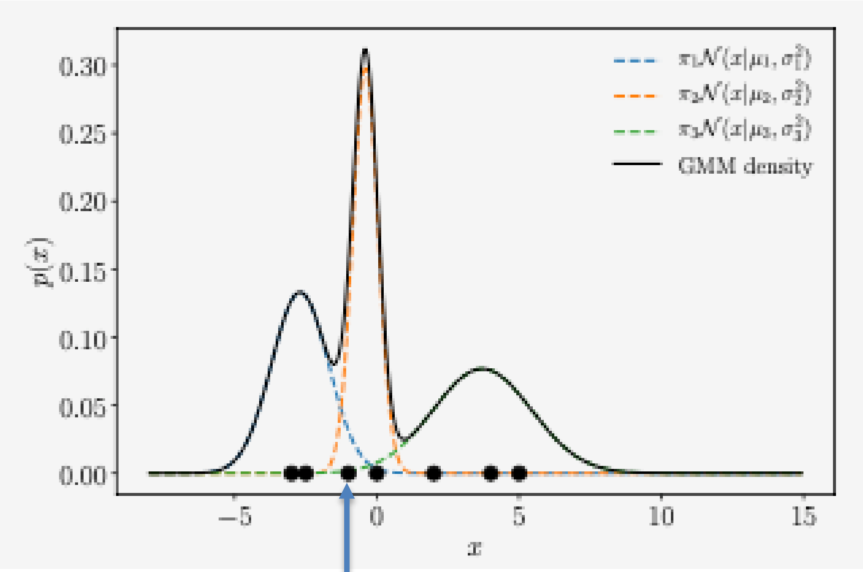
$$\sum_{k=1}^{K} \pi_k = 1, \qquad 0 \leq \pi_k \leq 1$$

Mixtures of Gaussians

- **The 1D case (example)**:
  - Simple superpositions of Gaussians can capture complex input densities
  - $p(x) = \pi_1 \cdot \mathcal{N}(x \mid \mu_1, \sigma_1^2) + \pi_2 \cdot \mathcal{N}(x \mid \mu_2, \sigma_2^2) + \pi_3 \cdot \mathcal{N}(x \mid \mu_3, \sigma_3^2)$

    where:

$$\mathcal{N}_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Gaussian Mixture in one dimension
  - three Gaussians (each scaled by a coefficient π_k).
  - The components in blue, green and orange.
  - Their sum in black.
- Note the importance of the scaling coefficient (mixing coefficient)

This point could be generated for all three constituents

Mixtures of Gaussians

- **The 2D case (example):**

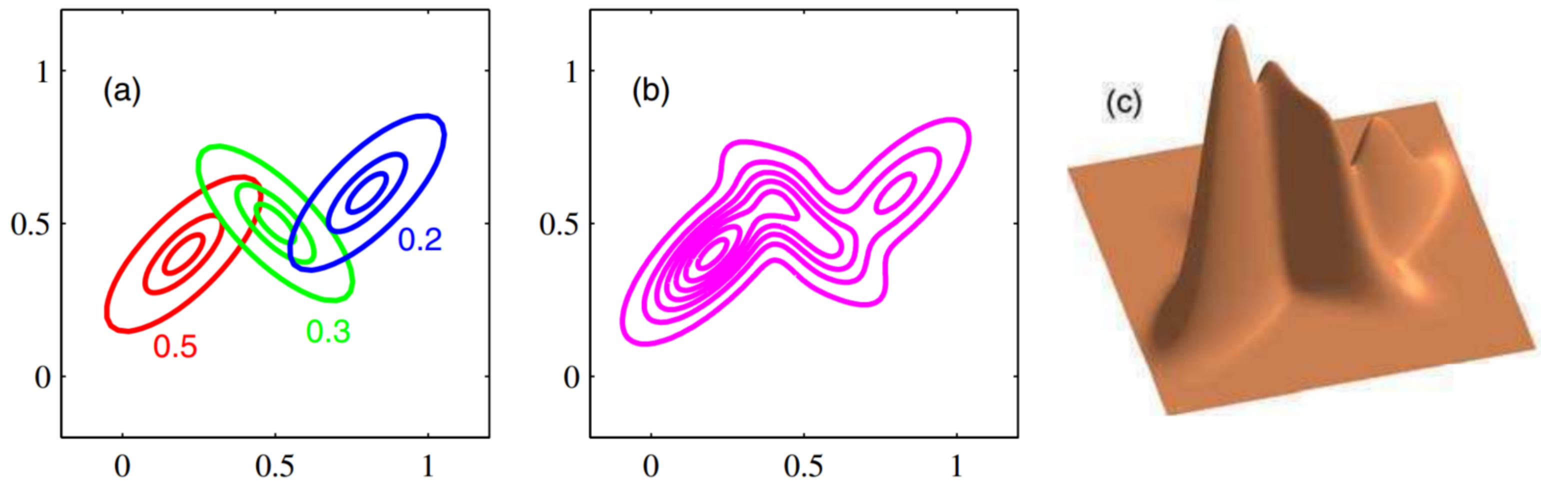


Figure 2.23 Illustration of a mixture of 3 Gaussians in a two-dimensional space. (a) Contours of constant density for each of the mixture components, in which the 3 components are denoted red, blue and green, and the values of the mixing coefficients are shown below each component. (b) Contours of the marginal probability density $p(\mathbf{x})$ of the mixture distribution. (c) A surface plot of the distribution $p(\mathbf{x})$.

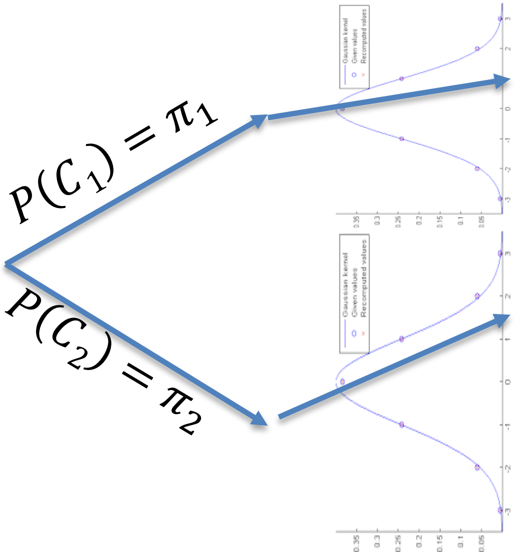
It can be shown that Mixture of Gaussian modelling can approximate any continuous density function using a sufficient number of constituents

Mixtures of Gaussians

- **Mixture Sampling**: A mixture of Gaussians is a generative model. **How can we generate points from the model**?

$$p(\mathbf{x}) = \pi_1 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

- Two steps:
  - Sample (choose) a component k (with probability π_k)
  - Generate a sample point from the component $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Note that π_k will represent the proportion of points that we will be generated from the component k

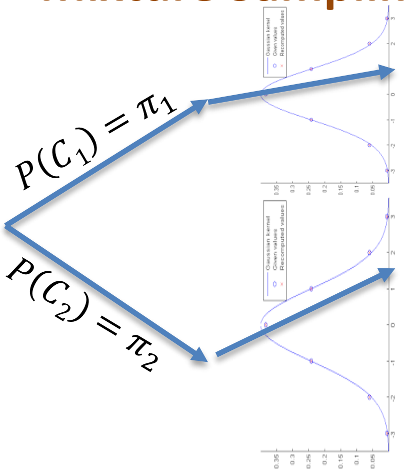


$\mathbf{x}^{(n)}$ $\quad p(\mathbf{x}^{(n)}, \mathrm{C}_1) = p(\mathrm{C}_1) p(\mathbf{x}^{(n)}|\mathrm{C}_1) = \pi_1 \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$

$\mathbf{x}^{(n)}$ $\quad p(\mathbf{x}^{(n)}, \mathrm{C}_2) = p(\mathrm{C}_2) p(\mathbf{x}^{(n)}|\mathrm{C}_2) = \pi_2 \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$

Mixtures of Gaussians

Mixture Sampling: Bayesian interpretation



$\boldsymbol{x}^{(\boldsymbol{n})}$ $$p\left(\mathbf{x}^{(n)}, \mathrm{C}_1\right) = p(\mathrm{C}_1)p\left(\mathbf{x}^{(n)}\middle|\mathrm{C}_1\right) = \pi_1\mathcal{N}\left(\mathbf{x}^{(n)}\middle|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$$

$\boldsymbol{x}^{(\boldsymbol{n})}$ $$p\left(\mathbf{x}^{(n)}, \mathrm{C}_2\right) = p(\mathrm{C}_2)p\left(\mathbf{x}^{(n)}\middle|\mathrm{C}_2\right) = \pi_2\mathcal{N}\left(\mathbf{x}^{(n)}\middle|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$$

- What is the total probability (or marginal prob)to obtain the point $\mathbf{x}^{(n)}$?

p($\boldsymbol{x}^{(\boldsymbol{n})}$)= $p\left(\mathbf{x}^{(n)}, \mathrm{C}_1\right)$+ $p\left(\mathbf{x}^{(n)}, \mathrm{C}_2\right)$= $\pi_1\mathcal{N}\left(\mathbf{x}^{(n)}\middle|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) + \pi_2\mathcal{N}\left(\mathbf{x}^{(n)}\middle|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$

- Posterior probability: What is the probability that a data point $\mathbf{x}^{(n)}$was generated by constituent k?

$$r_k^{(n)} = p\left(k|\mathbf{x}^{(\boldsymbol{n})}\right) = \frac{p\left(\mathbf{x}^{(n)}, \mathrm{C_k}\right)}{p(x)} = \frac{p(k)p\left(\mathbf{x}^{(n)}\middle|k\right)}{\sum_{l=1}^{K} p(l)p\left(\mathbf{x}^{(n)}\middle|l\right)} = \frac{\boxed{\pi_k\mathcal{N}\left(\mathbf{x}^{(n)}\middle|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}}{\boxed{\sum_{l=1}^{K} \pi_l\mathcal{N}\left(\mathbf{x}^{(n)}\middle|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)}}$$

Constituent k

Sum K Const.

This posterior $r_k^{(n)}$= $p\left(k|\mathbf{x}^{(n)}\right)$ is known as the responsibility that component k takes for ‘explaining’ the observation $\mathbf{x}^{(\mathrm{n})}$

Mixtures of Gaussians

Mixture Interpretation as a marginalization of constituents:

A constituent view:

$$p(\mathbf{x}) = \pi_1 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \pi_3 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

$$p(\mathbf{x}^{(n)}, \mathrm{C}_1) = p(\mathrm{C}_1) p(\mathbf{x}^{(n)}|\mathrm{C}_1) = \\ = \pi_1 \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

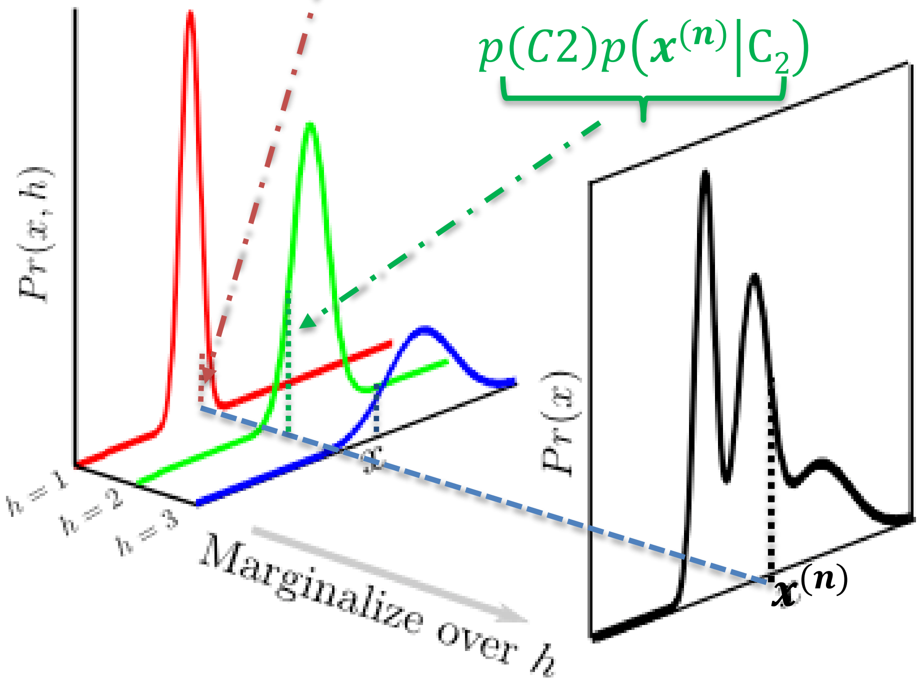
$$p(\mathbf{x}^{(n)}, \mathrm{C}_2) = p(\mathrm{C}_2) p(\mathbf{x}^{(n)}|\mathrm{C}_2) = \\ = \pi_2 \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

$$p(\mathbf{x}^{(n)}, \mathrm{C}_3) = p(\mathrm{C}_3) p(\mathbf{x}^{(n)}|\mathrm{C}_3) = \\ = \pi_3 \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

Constituent 1 of the point $\boldsymbol{x}^{(\boldsymbol{n})}$

$p(C1)p(\boldsymbol{x}^{(\boldsymbol{n})}|\mathrm{C}_1)$

$\underbrace{p(C2)p(\boldsymbol{x}^{(\boldsymbol{n})}|\mathrm{C}_2)}$



$p(x) = \sum_h p(x,h)$ where

- h are the Hidden variables: variables in the model not present in the data
- For each value of h, $p(x,h)$ is a constituent of the mixture
- $p(x)$ Is the marginal of the joint probability $p(x,h)$

Mixtures of Gaussians

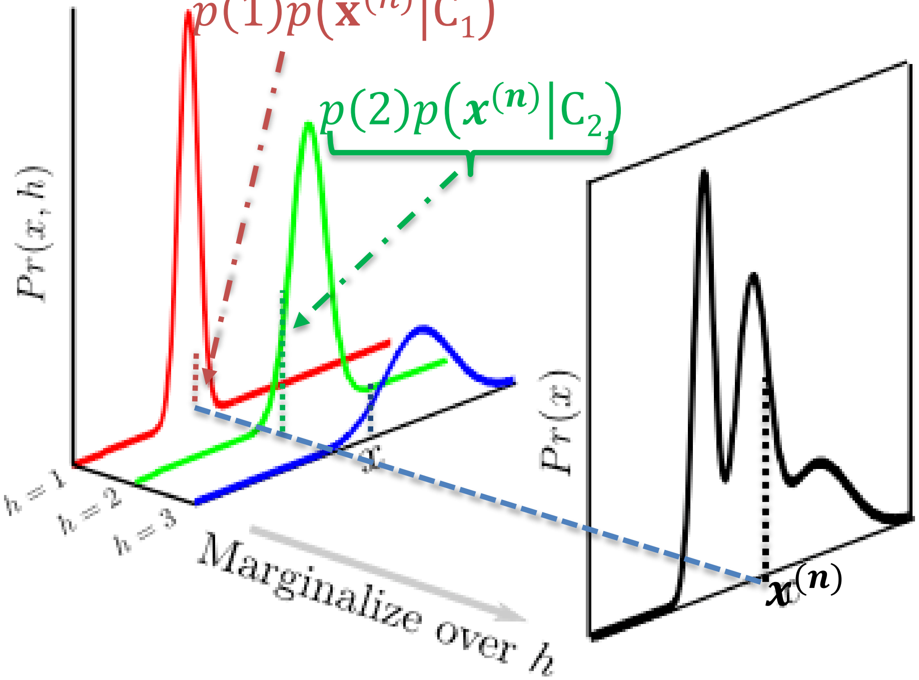
Mixture Interpretation as a marginalization of constituents:

A constituent view:

$$p(\mathbf{x}) = \pi_1 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \pi_3 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

Reponsibilities

$$r_k^{(n)} = \frac{\mathit{Constituent\ k\ of\ } x^{(n)}}{\mathit{Sum\ of\ the\ K\ Constituents}}$$



Constituent 1 of the point $\mathbf{x}^{(n)}$

$p(1)p(\mathbf{x}^{(n)}|\mathrm{C}_1)$

$\underbrace{p(2)p(\boldsymbol{x}^{(n)}|\mathrm{C}_2)}$

- Mixture of Gaussians as a marginalization. The mixture of Gaussians can also be thought of in terms of a joint distribution p($\boldsymbol{x}^{(\boldsymbol{n})}$,h) between the observed variable $\boldsymbol{x}^{(\boldsymbol{n})}$ and a discrete hidden variable h.
- To create the mixture density we marginalize over h.
- The hidden variable has a straightforward interpretation: it is the index of the **constituent** normal distribution.

Mixtures of Gaussians

From data points to model parameters

- The log-likelihood function is given by

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \ln\{p(\boldsymbol{x}^{(1)}) \ldots p(\boldsymbol{x}^{(N)})\} = \sum_{n=1}^{N} \ln\{p(\boldsymbol{x}^{(n)})\}$$

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

where row n of $\mathbf{X}$ is$\boldsymbol{x}^{(\boldsymbol{n})\boldsymbol{T}}$and $\{\pi, \mu, \Sigma\}$ denote the set of all parameters

Objective find the parameters that maximizes log-likelihood:

$$\{\widehat{\pi, \mu, \Sigma}\} = \mathrm{argmax}_{\{\pi,\mu,\Sigma\}} \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}^{(\boldsymbol{n})}\middle|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\}$$

Mixtures of Gaussians

From data points to model parameters

- We need to find:

$$\{\widehat{\pi, \mu, \Sigma}\} = \text{argmax}_{\{\pi,\mu,\Sigma\}} \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(\boldsymbol{x}^{(\boldsymbol{n})} \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\}$$

- Setting the derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t. μ_k to zero

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \boldsymbol{x}^{(\boldsymbol{n})}, \qquad N_k = \sum_{n=1}^{N} r_k^{(n)}$$

- Setting the derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t. $\boldsymbol{\Sigma}_k$ to zero

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \left(\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{\mu}_k\right)\left(\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{\mu}_k\right)^{\mathrm{T}}, \qquad N_k = \sum_{n=1}^{N} r_k^{(n)}$$

- Setting the derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t. π_k to zero

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

Mixtures of Gaussians

- From data points to model parameters
  - But these equations depend on the responsibilities $r_k^{(n)}$!! (and these depend on $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$)

    $$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \boldsymbol{x}^{(\boldsymbol{n})}$$

    $$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \left(\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{\mu}_k\right)\left(\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{\mu}_k\right)^{\mathrm{T}}$$

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

  - A simple iterative algorithm
    - Guess some initial parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$
    - Compute responsibilities $r_k^{(n)}$ for those parameters
    - Update the parameters according to the top equations and repeat
  - **Does it sound familiar?**

EM Algorithm for Mixtures of Gaussians

Expectation-Maximization

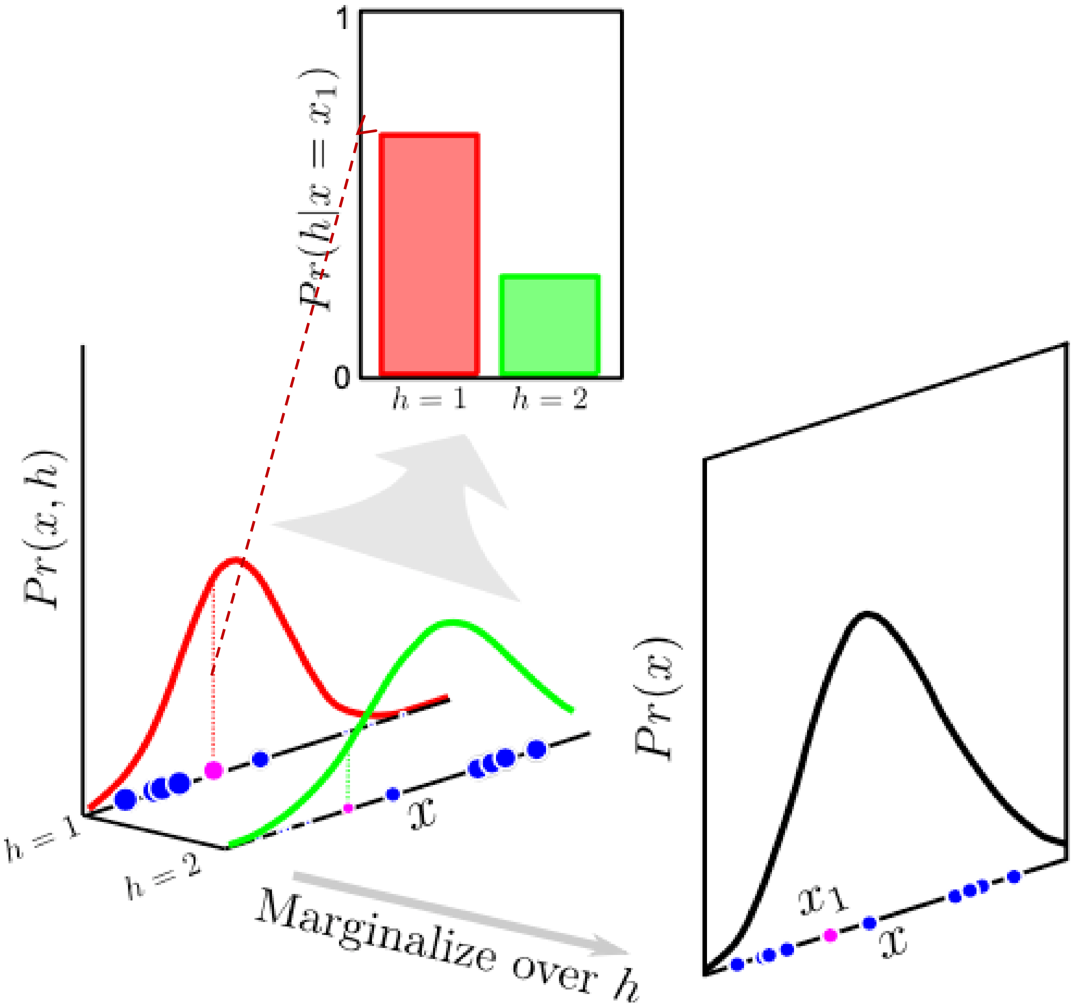
- The Expectation-Maximization algorithm (EM) is an algorithm for fitting the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ in models with hidden variables
- The basic idea of EM in this context is to:

1. **Pretend** that we know the parameters of the components and then infer the probability that each data point belongs to each component (**the responsibility**)
2. **Refit** the components to the data using the responsibilities. Each component is fitted to the entire dataset with each point weighted by the probability that it belongs to that component

- EM alternates between these two steps until convergence.
- Note the similarity with the “assignment step” and ”update step” of k-means

EM Algorithm for Mixtures of Gaussians

E-Step



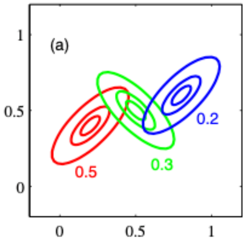
For each data point $x^{(n)}$

- We calculate the posterior $r_1^{(n)}$ and $r_2^{(n)}$ probabilities (responsibilities).
- For the data point x_1 (magenta circle), constituent 1 (red curve) is more than twice as likely to be responsible than constituent 2 (green curve).
- Note that in the joint distribution (left), the size of the projected data point indicates the responsibility

EM Algorithm for Mixtures of Gaussians

E-Step

example 2D



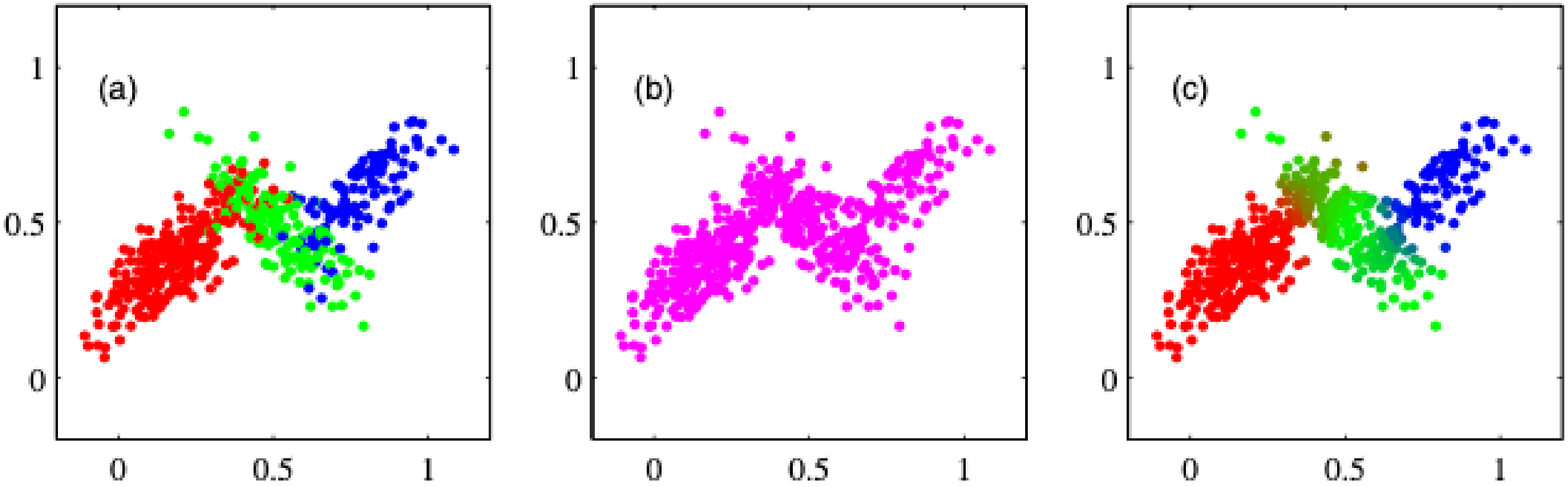
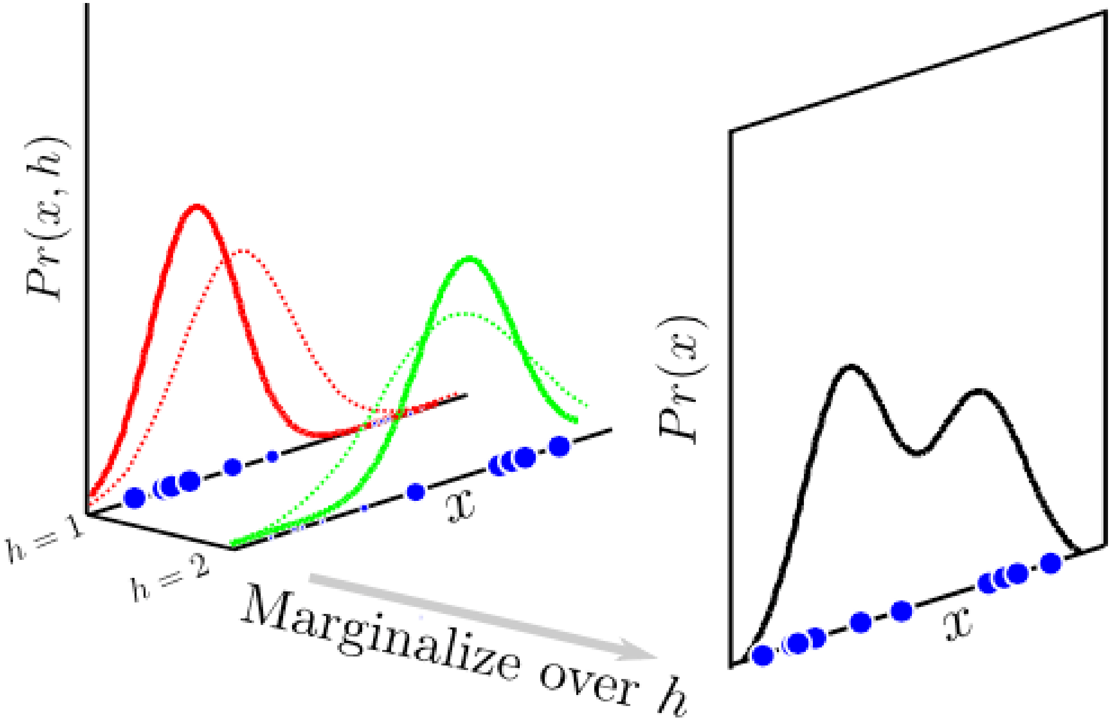


Figure 9.5 Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. **(a)** Samples from the joint distribution p(x,h) in which the three states of h, corresponding to the three constituents of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution p(x), which is obtained by simply ignoring the values of h and just plotting the x values. The data set in (a) is said to be complete, whereas that in (b) is incomplete. (c) The same samples in which the colours represent the value of the responsibilities $r_k^{(n)}$ associated with data point $\boldsymbol{x}^{(\boldsymbol{n})}$, obtained by plotting the corresponding point using proportions of red, blue, and green ink given by $r_k^{(n)}$ for *k = 1,2,3*, respectively

EM Algorithm for Mixtures of Gaussians

M-Step



- Dashed and solid lines represent fit before and after update, respectively.
- The ith data point x$^{(i)}$ contributes to update the components according to the responsibility $r_k^{(i)}$ (indicated by size of point) assigned in the E-step;
- Data points that are more associated with the kth constituent have more effect on the parameters.

EM Algorithm for Mixtures of Gaussians

Algorithm:

Take initial parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ k=1..K, alternate until convergence:

0. Evaluate the log-likelihood:

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

1. **E-step**: Compute responsibilities keeping $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ k=1..K fixed.

$$r_k^{(n)} = \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}^{(\boldsymbol{n})} \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{l=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}^{(\boldsymbol{n})} \middle| \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} \qquad N_k = \sum_{n=1}^{N} r_k^{(n)}$$

2. **M-step**: update component parameters keeping the $r_k^{(n)}$ fixed

$$\boldsymbol{\mu}_{\boldsymbol{k}}^{\boldsymbol{new}} = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \mathbf{x}^{(n)} \qquad \boldsymbol{\pi}_{\boldsymbol{k}}^{\boldsymbol{new}} = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{k}}^{\boldsymbol{new}} = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \left(\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{\mu}_{\boldsymbol{k}}^{\boldsymbol{new}}\right)\left(\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{\mu}_{\boldsymbol{k}}^{\boldsymbol{new}}\right)^{\mathsf{T}}$$

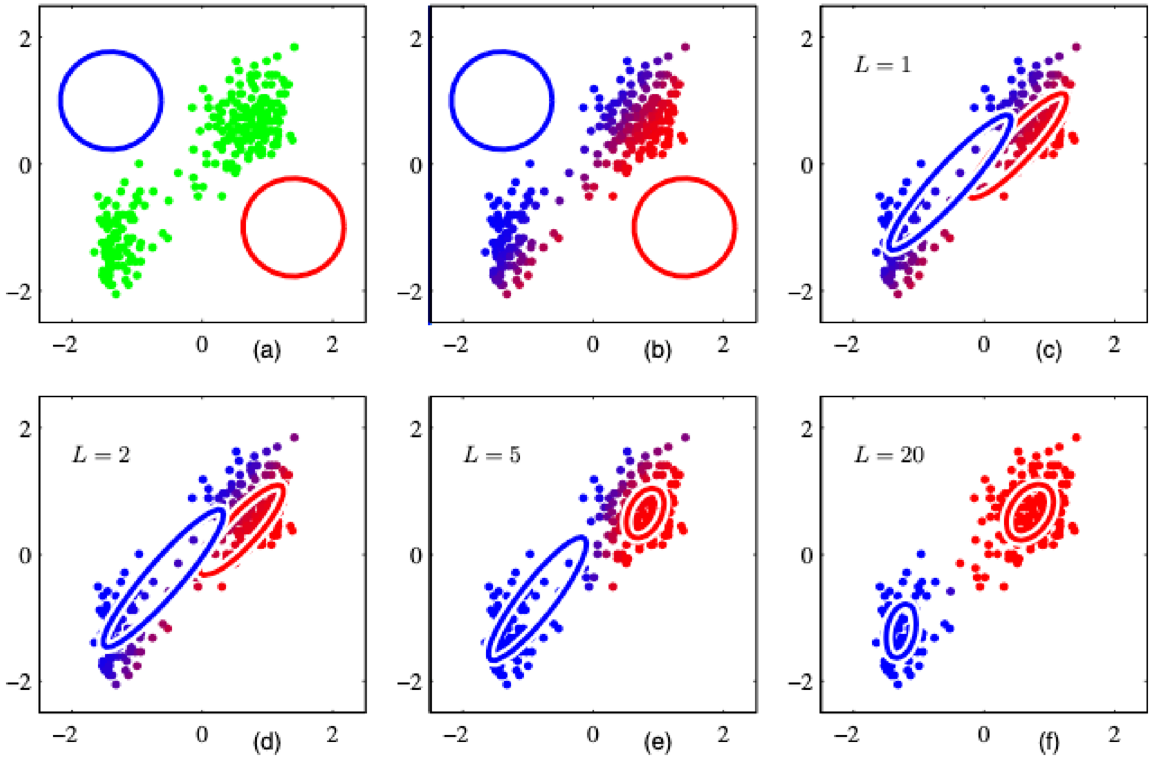
EM Algorithm for Mixtures of Gaussians

Initialization

- The performance of the EM depends strongly on the initialization of the parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, k = 1, \ldots K$, and the number of components
- It is common to run k-means to initialize EM: covariances can be initialized to the sample covariance of the clusters found by k-means, the mixing coefficients can be set to the fractions of cluster points
- This makes sense because EM is typically much slower to converge and more expensive to compute
- As with k-means, K can be determined by running EM with different values of K
- K-means can be derived from EM for the case of spherical covariances of equal constant size ε for all components.

Example

EM algorithm for the Old Faithful set



a) Initialization of 2 Gaussians $\Sigma_i = \sigma I, \pi_i = 0.5$

b) After 1st E-step: each data point according to the posterior probability $r_k^{(n)}$ of having been generated from each constituent. Thus, points with similar posteriors appear purple.

c) After 1st M-step: the mean of the blue Gaussian has moved to the center of mass of the blue ink. Analogous results hold for the red component

(d), (e), and (f) show results after 2, 5, and 20 complete iterations, respectively. The algorithm converges after 20 iterations

Example

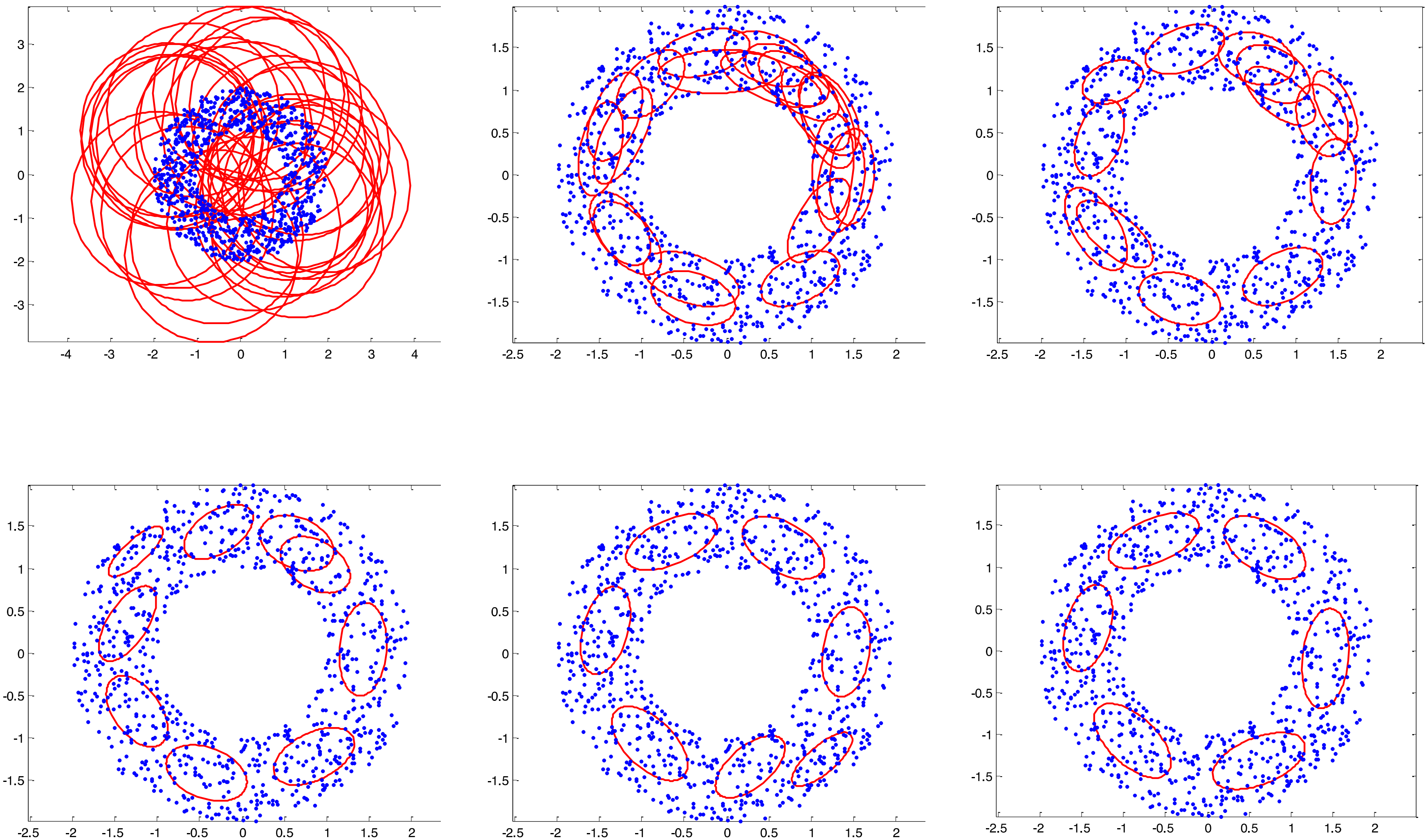
The annulus problem

- A training set of $N = 900$ examples was generated using a uniform pdf inside an annulus with inner and outer radii of 1 and 2 units, respectively
- A mixture model with $C = 30$ Gaussians was used to model the distribution of the training set

Training procedure

- The centers of the Gaussians were initialized by choosing 30 arbitrary points from the training set
- The covariance matrices were initialized to be diagonal, with a large variance compared to that of the training data
  - To avoid singularities, at every iteration the covariance matrices computed with EM were regularized with a small multiple of the identity matrix
- Components whose mixing coefficients fell below a threshold were trimmed
  - This allowed the algorithm to produce a compact model with only a few of the initial C=30 Gaussian components

Illustrative results are provided in the next page


3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
1.5
1
0.5
0
-0.5
-1
-1.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Summary

- Mixture of Gaussians allow to model multimodality in unsupervised clustering
    - Probabilistic model: allows prediction and sampling
- EM is a two steps iterative algorithm to estimate the mixture parameters
    - E-step: estimate the responsibilities (cluster memberships)
    - M-step: estimate the parameters using all the points weighted by the responsibility
    - The EM algorithm converges, at least to a **local** maximum and stops when the log-likehood of data improves less than fixed (and small) quantity