APPLIED MACHINE LEARNING

Probability Density Functions
Gaussian Mixture Models
Discrete Probabilities

Consider two variables $x$ and $y$ taking discrete values over the intervals $[1.....N_x]$ and $[1.....N_y]$ respectively.

\[ P(x = i) : \text{the probability that the variable } x \text{ takes value } i. \]
\[ 0 \leq P(x = i) \leq 1, \forall i = 1,..., N_x, \]
\[ \sum_{i=1}^{N_x} P(x = i) = 1. \]

Idem for $P(y = j), j = 1,...N_y$
Discrete Probabilities

The *joint probability* is written \( p(x,y) \).
The joint probability that variable \( x \) takes value \( i \) and variable \( y \) takes value \( j \) is:

\[
P\left( (x = i), (y = j) \right) \quad \text{or} \quad P\left( (x = i) \cap (y = j) \right)
\]

\( P(x \mid y) \) is the *conditional probability* of observing a value for \( x \) given a value for \( y \).

\[
P(x \mid y) = \frac{P(x, y)}{P(y)}
\]

Bayes' theorem:

When \( x \) and \( y \) are statistically independent:

\[
P(x \mid y) = P(x), \quad P(y \mid x) = P(y) \quad \text{and} \quad P(x, y) = P(x)P(y).
\]

Matlab Exercise I
Discrete Probabilities

The *marginal probability* that variable $x$ takes value $x_i$ is given by:

$$P_x(x = x_i) := \sum_{j=1}^{N_y} P_{x,y}(x = i, y = j)$$

- To compute the marginal, one needs the joint distribution $p(x,y)$.
- Often, one does not know it and one can only estimate it.
- If $x$ is a multidimensional variable $\rightarrow$ the marginal is a joint distribution!
Joint Distribution and Curse of Dimensionality

The joint distribution is far richer than the marginals.

The marginals of N variables taking K values corresponds to N(K-1) probabilities.

The joint distribution corresponds to ~N^K probabilities.

**Pros** of computing the joint distribution:

- Provides statistical dependencies across all variables and the marginal distributions

**Cons:**

- Computational costs grow exponentially with number of dimensions
  (statistical power: 10 samples to estimate each parameter of a model)

→ Compute solely the conditional if you care only about dependencies across variables (this will be relevant for lecture on non-linear regression methods)
Probability Distributions, Density Functions

$p(x)$ a **continuous function** is the **probability density function** or **probability distribution function (PDF)** (sometimes also called **probability distribution** or simply **density**) of variable $x$.

\[
p(x) \geq 0, \quad \forall x \in \mathbb{R}
\]

\[
\int_{-\infty}^{\infty} p(x) dx = 1
\]
Probability Distributions, Density Functions

The pdf is not bounded by 1.

It can grow unbounded, depending on the value taken by $x$. 

$p(x)$
PDF equivalency with Discrete Probability

The cumulative distribution function (or simply *distribution function*) of \( X \) is:

\[
D_x (x^*) = P \left( x \leq x^* \right)
\]

\[
D_x (x^*) = \int_{-\infty}^{x^*} p(x) \, dx, \quad x \in \mathbb{R}
\]

\( p(x) \, dx \sim \) probability of \( x \) to fall within an infinitesimal interval \([x, x + dx]\)
PDF equivalency with Discrete Probability

Uniform distribution on $x$

Probability that $x$ takes a value in the subinterval $[a,b]$ is given by:

$$P(x \leq b) := D_x(x \leq b) = \int_{-\infty}^{b} p(x)dx$$

$$P(a \leq x \leq b) = D_x(x \leq b) - D_x(x \leq a) = D_x\left(x^*\right)$$

$$P(a \leq x \leq b) = \int_{a}^{b} p(x)dx = 1$$
The *expectation* of the random variable $x$ with probability $P(x)$ (in the discrete case) and pdf $p(x)$ (in the continuous case), also called the *expected value* or *mean*, is the mean of the observed value of $x$ weighted by $p(x)$. If $X$ is the set of observations of $x$, then:

When $x$ takes discrete values: \[ \mu = E \{x\} = \sum_{x \in X} xP(x) \]

For continuous distributions: \[ \mu = E \{x\} = \int_{X} x \cdot p(x) \cdot dx \]
Variance

\( \sigma^2 \), the variance of a distribution measures the amount of spread of the distribution around its mean:

\[
\sigma^2 = \text{Var}(x) = E\left\{ (x - \mu)^2 \right\} = E\left\{ x^2 \right\} - \left[ E\left\{ x \right\} \right]^2
\]

\( \sigma \) is the standard deviation of \( x \).
Parametric PDF

The uni-dimensional Gaussian or Normal distribution is a distribution with pdf given by:

\[ p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \mu:\text{mean}, \quad \sigma^2:\text{variance} \]

The Gaussian function is entirely determined by its mean and variance. For this reason, it is referred to as a parametric distribution.
~68% of the data are comprised between +/- 1 sigma
~96% of the data are comprised between +/- 2 sigma-s
~99% of the data are comprised between +/- 3 sigma-s

This is no longer true for arbitrary pdf-s!
For other pdf than the Gaussian distribution, the variance represents a notion of dispersion around the expected value.
Multi-dimensional Gaussian Function

The uni-dimensional Gaussian or Normal distribution is a distribution with pdf given by:

\[ p(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \mu:\text{mean}, \quad \sigma:\text{variance} \]

The multi-dimensional Gaussian or Normal distribution has a pdf given by:

\[ p(x; \mu, \Sigma) = \frac{1}{N} \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \]

if \( x \) is N-dimensional, then
\( \mu \) is a N – dimensional mean vector
\( \Sigma \) is a \( N \times N \) covariance matrix
2-dimensional Gaussian Pdf

\[ p(x_1, x_2) \]

\[ p(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|^{1/2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \]

Isolines: \( p(x) = \text{cst} \)

if \( x \) is N-dimensional, then
\( \mu \) is a \( N \) – dimensional mean vector
\( \Sigma \) is a \( N \times N \) covariance matrix
Modeling Data with a Gaussian Function

Construct covariance matrix from (centered) set of datapoints \( X = \{ x^i \}_{i=1}^M \):

\[
\Sigma = \frac{1}{M} XX^T
\]

\[
p(x; \mu, \Sigma) = \frac{1}{\left(2\pi\right)^\frac{N}{2} |\Sigma|^{-\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}
\]

if \( x \) is \( N \)-dimensional, then
\( \mu \) is a \( N \) – dimensional mean vector
\( \Sigma \) is a \( N \times N \) covariance matrix
Modeling Data with a Gaussian Function

Construct covariance matrix from (centered) set of datapoints $X = \{x^i\}_{i=1}^M$:

$$\Sigma = \frac{1}{M} XX^T$$

$\Sigma$ is square and symmetric. It can be decomposed using the eigenvalue decomposition.

$$\Sigma = V \Lambda V^T,$$

$V$: matrix of eigenvectors, $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ \vdots & \ddots \\ 0 & \lambda_N \end{pmatrix}$: diagonal matrix composed of eigenvalues

For the 1-std ellipse, the axes' lengths are equal to:

$$\sqrt{\lambda_1} \text{ and } \sqrt{\lambda_2}, \text{ with } \Sigma = V \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} V^T.$$  

Each isoline corresponds to a scaling of the 1std ellipse.
Fitting a single Gauss function and PCA

PCA Identifies a suitable representation of a multivariate data set by **decorrelating** the dataset.

When projected onto $e^1$ and $e^2$, the set of datapoints appears to follow two uncorrelated Normal distributions.

$$p \left( (e^2)^T X \right) \sim N \left( X; \mu_2, \left( \lambda_2 \right)^{\frac{1}{2}} \right)$$

$$p \left( (e^1)^T X \right) \sim N \left( X; \mu_1, \left( \lambda_1 \right)^{\frac{1}{2}} \right)$$
Marginal, Conditional in Pdf

Consider two random variables $x_1$ and $x_2$ with joint distribution $p(x_1, x_2)$, then the *marginal probability* of $x_1$ given $x_1$ is:

$$p(x_1) = \int p(x_1, x_2) dx_2$$

The conditional probability is given by:

$$p(x_2 \mid x_1) = \frac{p(x_1, x_2)}{p(x_1)} \iff p(x_2 \mid x_1) = \frac{p(x_1 \mid x_2)p(x_2)}{p(x_1)}$$
Marginal, Conditional Pdf of Gauss Functions

The conditional and marginal pdf of a multi-dimensional Gauss function are all Gauss functions!

Illustrations from Wikipedia

Matlab Exercise II
Conditional Pdf and Statistical Independence

$x_1$ and $x_2$ are statistically independent if:
\[ p(x_2 \mid x_2) = p(x_1) \quad \text{and} \quad p(x_2 \mid x_1) = p(x_2) \]

\[ \implies p(x_1, x_2) = p(x_2) p(x_2) \]

$x_1$ and $x_2$ are uncorrelated if $\text{cov}(x_1, x_2) = 0$.
\[ \text{cov}(x_1, x_2) = E\{x_1, x_2\} - E\{x_1\} E\{x_2\} \]
\[ \implies E\{x_1, x_2\} = E\{x_1\} E\{x_2\} \]

The expectation of the joint distribution is equal to the product of the expectation of each distribution separately.
Statistical independence and uncorrelatedness

\[ p(x_1, x_2) = p(x_1) p(x_2) \implies E\{x_1, x_2\} = E\{x_1\} E\{x_2\} \]

\[ p(x_1, x_2) = p(x_1) p(x_2) \nLeftrightarrow E\{x_1, x_2\} = E\{x_1\} E\{x_2\} \]

Statistical independence ensures uncorrelatedness.
The converse is not true.
Statistical independence and uncorrelatedness

Are $x_1$ and $x_2$ uncorrelated?
Are $x_1$ and $x_2$ statistically independent?

<table>
<thead>
<tr>
<th></th>
<th>$x_2=-1$</th>
<th>$x_2=0$</th>
<th>$x_2=1$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1=-1$</td>
<td>3/12</td>
<td>0</td>
<td>3/12</td>
<td>1/2</td>
</tr>
<tr>
<td>$x_1=1$</td>
<td>1/12</td>
<td>4/12</td>
<td>1/12</td>
<td>1/2</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
</tr>
</tbody>
</table>
Statistical independence and uncorrelatedness

\[
\begin{array}{|c|c|c|c|c|}
\hline
& x_2=-1 & x_2=0 & x_2=1 & \text{Total} \\
\hline
x_1=-1 & 3/12 & 0 & 3/12 & 1/2 \\
x_1=1 & 1/12 & 4/12 & 1/12 & 1/2 \\
\hline
\text{Total} & 1/3 & 1/3 & 1/3 & \\
\hline
\end{array}
\]

\(x_1\) and \(x_2\) are uncorrelated as but not statistically independent.

\[
E\{x_1, x_2\} = E\{x_1\}E\{x_2\} = 0
\]

but \(p(x_1 = -1, x_2 = 1) = 3/12 = 0.25\ne

\[
p(x_1 = -1)p(x_2 = 1) = 1/2 \times 1/3 = 0.1667
\]
Part I - Exercises
Determining how well a model fits the data

Data are noisy → no model will fit perfectly the data (unless you fit the noise = overfitting)

→ need a mean to determine how much the model fits the underlying distribution.

Which of the two models fit best the data?
Likelihood Function

The Likelihood function or *Likelihood (for short)* determines the probability density of observing the set $X$ of $M$ datapoints, if each datapoint has been generated by the pdf $p(x)$.

$$L(X) = p(x^1, x^2, ..., x^M) \quad X = \{x^i\}_{i=1}^{M},$$

If the data are *independent and identically* distribution by $p$, then:

$$L(X) = \prod_{i=1}^{M} p(x^i)$$

The likelihood can be used to determine how well a particular model of $p(x)$ models the data at hand.
Likelihood of Gaussian Pdf Paramatrization

Consider that the pdf of the dataset $X$ is parametrized with parameters $\mu, \Sigma$.

One writes: $p(X; \mu, \Sigma)$ or $p(X | \mu, \Sigma)$

The likelihood function (short – likelihood) of the model parameters is given by:

$$L(\mu, \Sigma | X) := p(X; \mu, \Sigma)$$

Measures probability of observing $X$ if the distribution of $X$ is parametrized with $\mu, \Sigma$

If all datapoints are identically and independently distributed (i.i.d.)

$$L(\mu, \Sigma | X) = \prod_{i=1}^{M} p(x_i; \mu, \Sigma)$$

To determine the best fit, search for parameters that maximize the likelihood.
Values taken by the likelihood for two different fits using 1-D Gauss functions with different means.
Maximum Likelihood Optimization

The principle of *maximum likelihood* consists of finding the optimal parameters of a given distribution by maximizing the likelihood function of these parameters, equivalently by maximizing the probability of the data given the model and its parameters.

For a multi-variate Gaussian pdf, one can determine the mean and covariance matrix by solving:

\[
\max_{\mu, \Sigma} L(\mu, \Sigma | X) = \max_{\mu, \Sigma} p(X | \mu, \Sigma) \\
\frac{\partial}{\partial \mu} p(X | \mu, \Sigma) = 0 \quad \text{and} \quad \frac{\partial}{\partial \Sigma} p(X | \mu, \Sigma) = 0
\]

If \( p \) is the Gaussian pdf, then the above has an analytical solution (assuming that one has enough observations of \( x \) to draw the parameters from).
Expectation-Maximization (E-M)

EM used when no closed form solution exists for the maximum likelihood estimate.

Example: Fit Mixture of Gaussian Functions

\[ p(x; \Theta) = \sum_{k=1}^{K} \alpha_k p_k(x; \mu^k, \Sigma^k) \]

with \( \Theta = \{ \mu^1, \Sigma^1, \ldots, \mu^K, \Sigma^K \} \), \( \alpha_k \in [0, 1] \).

No closed form solution to:

\[ \max_{\Theta} L(\Theta | X) = \max_{\Theta} p(X | \Theta) \]

\[ \max_{\Theta} L(\Theta | X) = \max_{\Theta} \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k p_k(x^i; \mu^k, \Sigma^k) \]

E-M is an iterative procedure to estimate the best set of parameters

Converge to a local optimum

→ Sensitive to initialization!
Expectation-Maximization (E-M)

EM is an iterative procedure:

0) Make a guess, pick a set of $\Theta$ (initialization)
1) Compute likelihood $L(\hat{\Theta} | X, Z)$ (E-Step)
2) Update $\Theta$ by gradient ascent on $L(\Theta | X, Z)$
3) Iterate between steps 1 and 2 until reach plateau (no improvement on likelihood)

Ensured to converge to a local optimum only! (see more details next slides)
From K-means Clustering to Density Modeling with Mixture of Gaussians

The algorithm of K-means is a simple version of Expectation-Maximization applied to a model composed of isotropic Gauss functions.
K-means Clustering (probabilistic interpretation)

Computing the distance to the k-th centroid is equivalent to computing the probability that the data point has been generated by the k-th model.

\[ d(x^i, \mu^k) \sim p(x^i; \mu^k) = e^{-(x^i - \mu^k)^2} \]

The likelihood of the k-th model is:

\[ L(X; \mu^k) = \prod_i e^{-(x^i - \mu^k)^2} \]

Assignment Step (E-step):  
~ Compute expectation of the equivalent Gaussian model with unity variance and centered on the centroid
K-means Clustering (probabilistic interpretation)

Assignment Step (E-step):
~ Compute expectation of the equivalent Gaussian model with unity variance and centered on the centroid
• Assign the responsibility of each data point to its “closest” centroid.
~ Bayes’ rule: If \( p(x^i; \mu^2) > p(x^i; \mu^1) \), then \( x^i \) belongs to cluster 2.
K-means Clustering (probabilistic interpretation)

The new centroid is closer to the datapoints after the update step,

$$\mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k}$$

$$\Rightarrow d(x^i - \mu^k)$$

→ the likelihood of the k-th model increases.

$$L(X; \mu^k) = \prod_i e^{-(x^i - \mu^k)^2}$$

Update Step (M-step):
~ Maximize expectation of the equivalent Gaussian model with unity variance and centered on the centroid
Soft-K-means (probabilistic interpretation)

Soft K-means can be seen as fitting the data distribution with a mixture of isotropic (spherical) Gaussian pdf-s.

E-M updates the parameters of each Gaussian to optimize the likelihood that the Gaussians represent the distribution of the datapoints.

Likelihood of overall distribution (with uniform prior for each Gaussian):

\[ L(X; \mu^1, \sigma^1, \mu^2, \sigma^2) = L(X; \mu^1, \sigma^1) + L(X; \mu^2, \sigma^2) \]

![Poor fit](image1.png)

\[ L(X; \mu^1, \sigma^1) = 0.6 \]
\[ L(X; \mu^2, \sigma^2) = 0.4 \]

![Better fit](image2.png)

\[ L(X; \mu^1, \sigma^1) = 0.8 \]
\[ L(X; \mu^2, \sigma^2) = 0.9 \]
Assignment Step (E-step):

The responsibility factor gives a measure of the likelihood that cluster $k$ generated the dataset.

$$r_i^k : \text{responsibility of cluster } k \text{ for point } x^i$$

$$r_i^k = \frac{\alpha_k p(x^i; \mu^k, \sigma^k)}{\sum_k \alpha_k p(x^i; \mu^k, \sigma^k)}, \quad \alpha_k \in [0,1]$$

$p(x^i; \mu^k, \sigma^k) \in [0,1]$: Gauss pdf evaluated at $x^i$

Normalized over clusters: $\sum_k r_i^k = 1$

Relative importance of each of the $K$ clusters (measure of number of datapoints in each cluster)

→ In GMM, we will see that this is a measure of the likelihood that the Gaussian $k$ (or cluster $k$) generated the whole dataset.
Soft-K-means (probabilistic interpretation)

Update Step (M-Step):

- \( r_i^k \): responsibility of cluster \( k \) for point \( x^i \)
  \[
  r_i^k = \frac{\alpha_k p(x^i; \mu^k, \sigma^k)}{\sum_k \alpha_k p(x^i; \mu^k, \sigma^k)}, \quad \alpha_k \in [0,1]
  \]

- \( p(x^i; \mu^k, \sigma^k) \in [0,1] \): Gauss pdf evaluated at \( x^i \)

Normalized over clusters: \( \sum_k r_i^k = 1 \)

- \( \mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k} \)

- \( (\sigma^k)^2 = \frac{\sum_i r_i^k (x^i - \mu^k)^2}{M \cdot \sum_i r_i^k} \)

- \( \alpha_k = \frac{\sum_i r_i^k}{\sum_k \sum_i r_i^k} \)

Relative importance of each of the \( K \) clusters (measure of number of datapoints in each cluster)

\( \rightarrow \) In GMM, we will see that this is a measure of the likelihood that the Gaussian \( k \) (or cluster \( k \)) generated the whole dataset.

This fits only a mixture of \textit{spherical} Gaussians!
From spherical to diagonal Gaussian pdf-s.

Update Step (M-Step):

\[ r_i^k : \text{responsibility of cluster } k \text{ for point } x^i \]
\[ r_i^k = \frac{\alpha_k p \left( x^i ; \mu_k , \sigma_k^j \right)}{\sum_k \alpha_k p \left( x^i ; \mu_k , \sigma_k^j \right)} \]

\[ j = 1, \ldots, N : \text{dimension of dataset} \]
\[ p \left( x^i ; \mu_k , \sigma_k^j \right) \in [0,1] : \text{Gauss pdf evaluated at } x^i \]

Normalized over clusters: \[ \sum_k r_i^k = 1 \]

\[ \mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k} \]

\[ \left( \sigma_j^k \right)^2 = \frac{\sum_i r_i^k \left( x_j^i - \mu_j^k \right)^2}{M \cdot \sum_i r_i^k} \]

One covariance element per dimension, but still aligned with the axes of the original frame of reference.
Fitting data with Diagonal Mixture of Gaussians

Mixture of diagonal Gaussians can only fit Gaussians whose axes are aligned with the data axes, i.e. the covariance matrices of the Gaussians are diagonal.
Gaussian Mixture Models (GMM) can learn mixtures of Gaussians with arbitrary (full) covariance matrices.

→ Gaussian Mixture Model can exploit local correlations and adapt the covariance matrix of each Gaussian so that it aligns with the local direction of correlation.

→ Each Gaussian performs a local linear PCA.
Tradeoff between computation costs and better fit

- In addition to better fit the local non-linearities, GMM may also reduce the number of Gaussians required to fit the data.
- But this comes at the cost of an increase in the number of parameters: *Full covariance matrices require $N*(N+1)/2$ parameters against $N$ for diagonal matrices.*

How to derive an algorithm for fitting data with complex mixtures of Gaussians?
**Clustering with Mixture of Gaussians**

Likelihood of the mixture of Gaussians: 

\[ L\left(X; \{\mu^k, \Sigma^k\}_{k=1}^K\right) = \sum_{k=1}^K \alpha_k \cdot p\left(X; \mu^k, \Sigma^k\right) \]

with 

\[ p\left(X; \mu^k, \Sigma^k\right) \sim \prod_{i=1}^M e^{-\frac{(x^i - \mu_k)^T (\Sigma_k)^{-1} (x^i - \mu_k)}} \] 

(UNORMALIZED likelihood of Gaussian k)

\( \mu^k, \Sigma^k \): mean and covariance matrix of Gaussian k

**Mixing Coefficients are normalized**

\[ \sum_{k=1}^K \alpha_k = 1 \]

\[ \alpha_k \sim \frac{1}{M} \sum_{i=1}^M \sum_{l=1}^M \frac{p\left(x^i; \mu^k, \Sigma^k\right)}{p\left(x^i; \mu^l, \Sigma^l\right)} \]
Gaussian Mixture Modeling with Expectation-Maximization

The parameters of a GMM are the means, covariance matrices and priors:

$$\Theta = \{ \mu^1, \ldots, \mu^K, \sum^1, \ldots, \sum^K, \alpha_1, \ldots, \alpha_K \}$$

Estimation of all the parameters can be done through *Expectation-Maximization* (E-M). E-M tries to find the optimum of the likelihood of the model given the data, i.e.:

$$\max_{\Theta} L(\Theta | X) = \max_{\Theta} p(X | \Theta)$$
Expectation-Maximization

One usually can safely assume that the datapoints are i.i.d. (identically and independently distributed).

$$\max_{\Theta} \ p(X \mid \Theta) = \max_{\Theta} \ \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p(x_i; \mu^k, \Sigma^k)$$

Computing the log of the likelihood yields the same optimum:

$$\max_{\Theta} \ \log p(X \mid \Theta) = \max_{\Theta} \ \log \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p(x_i; \mu^k, \Sigma^k)$$

$$\max_{\Theta} \ \log \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p(x_i; \mu^k, \Sigma^k) = \max_{\Theta} \ \sum_{i=1}^{M} \log \left( \sum_{k=1}^{K} \alpha_k \cdot p(x_i; \mu^k, \Sigma^k) \right)$$

No close-form solution unlike the case for one Gaussian. See derivation of E-M for GMM in the annexes posted on the website.
E-M Steps for GMM

**Initialization:**

The priors $\alpha_1, \ldots, \alpha_k$ can be uniform for starters.
The means $\mu^1, \ldots, \mu^k$ can be initialized with K-means.

**Estimation Step (E-step):**

At each step $t$, estimate, for each Gaussian $k$, the probability that this Gaussian generated the dataset by computing:

$$p\left(k \mid \Theta^{(t)} \right) = \frac{1}{M} \sum_{i=1}^{M} \frac{p_k\left(x^i; \mu^{k(t)}, \Sigma^{k(t)}\right) \cdot \alpha_k^{(t)}}{\sum_j p_k\left(x^i; \mu^{j(t)}, \Sigma^{j(t)}\right) \cdot \alpha_j^{(t)}}$$
E-M Estimate for Gaussian Mixture Models

Update Step (M-step):

Recompute the means, covariances matrices and prior probabilities so as to maximize the log – likelihood of the current estimate: $\log \left( L(\Theta^{(t)} | X) \right)$

and using current estimate of the probabilities: $p(k | \Theta^{(t)})$

$$
\mu^{k(t+1)} = \frac{\sum_j p\left(k | x^j, \Theta^{(t)}\right) \cdot x^j}{\sum_j p\left(k | x^j, \Theta^{(t)}\right)}
$$

$$
\alpha_k^{(t+1)} = \frac{1}{M} \sum_j p\left(k | x^j, \Theta^{(t)}\right)
$$

$$
\sum_{k(t+1)} = \frac{\sum_j p\left(k | x^j, \Theta^{(t)}\right) \left( x^j - \mu^{k(t+1)} \right) \left( x^j - \mu^{k(t+1)} \right)^T}{\sum_i p\left(k | x^i, \Theta^{(t)}\right)}
$$
Clustering with Mixtures of Gaussians using spherical Gaussians (left) and non-spherical Gaussians (i.e. with full covariance matrix) (right).

Notice how the clusters become elongated along the direction of the clusters (the grey circles represent the first and second variances of the distributions).
Estimating a Pdf through Maximum Likelihood

Matlab and mldemos examples
Gaussian Mixture Model
Gaussian Mixture Model

GMM using 4 Gaussians with random initialization
Gaussian Mixture Model

Expectation Maximization is very sensitive to initial conditions:

GMM using 4 Gaussians with new random initialization
Gaussian Mixture Model

Very sensitive to choice of number of Gaussians. Number of Gaussians can be optimized iteratively using AIC or BIC (see later slides):

Here, GMM using 8 Gaussians
Summary of this lecture

This class revisited some basic notions of statistics, with standard definitions of pdf, cdf, marginal and conditional distributions.

It emphasized the notion of statistical independence and how one can recognize it numerically and geometrically from looking at the distributions.

It exemplified these concepts with multi-dimensional Gaussian functions.

Finally, it introduced the notion of maximum likelihood fit first with one Gaussian and then with a mixture of Gaussians.