

**Indian Institute of Technology Madras  
Presents**

**NPTEL  
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING**

**Pattern Recognition**

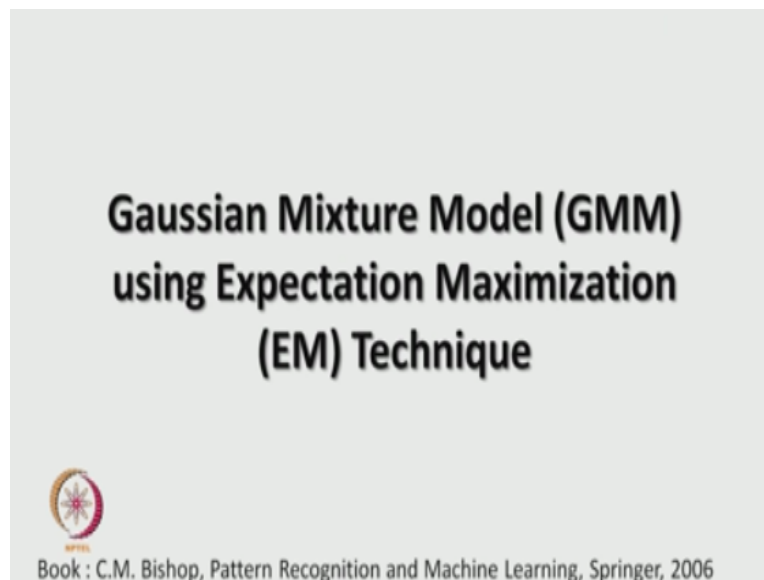
**Module 02**

**Lecture 15**

**Gaussian Mixture Model (GMM)**

**Prof. Sukhendu Das  
Department of CSSE, IIT Madras**

(Refer Slide Time: 00:15)



Gaussian mixture model using EM method where do you use and why do you need a mixture Gaussians we have seen in many of our analysis specifically if you recollect back some of the discussions which we had on modern business criteria the covariance matrix under the base bayes paradigm when we convert a distance short of multivariate Gaussian function into a distance criteria we made an assumption in most of those analysis that the distribution of scatter of the data in whatever dimension it may be 1, 2, 3 or even higher is a Gaussian distribution okay.

This sort of assumption may not happen in particular but of course most scientists engineers still use a Gaussian distribution for modeling many analysis in many different applications including single process communication theory violations whatever may be the advantage is that the Gaussian seems to be the one which is more closer to the natural distribution okay number the other main reason is it is easy to do mathematical manipulations if you have Gaussian functions.

Specifically differentiation exists up to as much of an order as you need infinite order it is suppose it as various other types of advantages which this function provides over other distribution functions but there may be situations where a distribution may not be sticky Gaussian in nature or nature is sickly not Gaussian in such cases there are methods which deal with multiple Gaussian it is like as if I want to cluster the data into several components or several parts.

And I assume or we assume that each of those clusters forms a Gaussian distribution so this leaves leads us to analysis which is based on GMM which is casually called or the Gaussian mixture model let us look at the expression of the Gaussian mixture.

(Refer Slide Time: 02:45)

## The Gaussian Distribution

### Univariate Gaussian Distribution

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

mean      variance

### Multi-Variate Gaussian Distribution

$$\mathcal{N}(x | \mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

mean      covariance

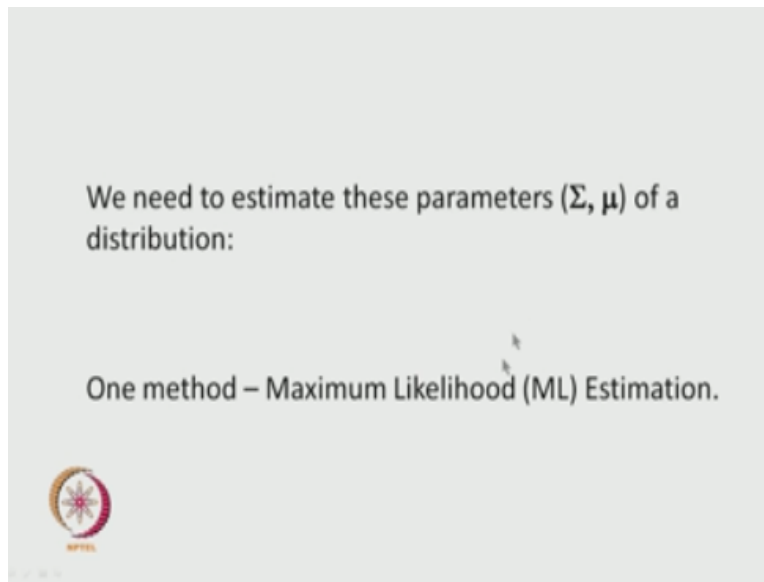
This is the univariate Gaussian distribution in 1D where the  $\mu$  is indicated by the mean okay and of course  $\sigma$  is a standard deviation  $\sigma^2$  is the variance which is here okay this is actually the variance remember the  $\sigma$  is called the standard deviation variance is  $\sigma^2$  which is here as well okay so this is the normalizing part of the function and this is the exponential function we had seen this and we had to also extend this to a multivariate Gaussian distribution case.

Where this  $\mu$  becomes a vector of the dimension of the data sample or instance  $x$  okay I have changed this  $G$  to  $\mathcal{N}$  here indicating this is a univariate Gaussian distribution in 1D in higher dimension it is an  $\mathcal{N}$  okay where this  $\sigma^2$  represent replaced by the covariance term root over this and this is modern distance function within the exponential term which you have inverse of the covariance matrix this is nothing new.

We had this discussion earlier under multi variant Gaussian distribution we put this under the bayes paradigm and we formulate distance functions and we know under what properties of the covariance matrix we are going to have between class linear decision boundaries or DP is or non linear boundaries actually the covariance matrix and it is inverse of the covariance matrix dictates the corresponding property of the decision boundary okay.

But now we what we will do is we will extend this to a case where we will have not only a multivariate Gaussian distribution one of them in higher dimensions but multiples of these speared over the data.

(Refer Slide Time: 04:33)



And to do this we basically need to estimate the covariance matrix and  $\mu$  for a particular distribution and one such method is actually called the maximum likelihood estimation when we need to estimate this for a particular data.

(Refer Slide Time: 04:47)

## ML Method for estimating parameters

□ Consider log of Gaussian Distribution

$$\ln p(x | \mu, \Sigma) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)$$



Data samples so do this we will look at this ML method which is a simpler one to visualize if  $y_i$  take the log of this probability function for the pervious expression let us go back.

(Refer Slide Time: 04:58)

## The Gaussian Distribution

### □ Univariate Gaussian Distribution


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

mean                  variance

### □ Multi-Variate Gaussian Distribution

$$\mathcal{N}(x | \mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{d/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

mean                  covariance



To the expression here this what we are talking the log about an we did this when we actually form the discriminate function for a particular class when we derived a distance criteria.

(Refer Slide Time: 05:08)

## ML Method for estimating parameters

- Consider log of Gaussian Distribution

$$\ln p(x | \mu, \Sigma) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)$$

- Take the derivative and equate it to zero

$$\frac{\partial \ln p(x | \mu, \Sigma)}{\partial \mu} = 0$$

$\downarrow$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\frac{\partial \ln p(x | \mu, \Sigma)}{\partial \Sigma} = 0$$

$\downarrow$

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})(x_n - \mu_{ML})^T$$



Where, N is the number of samples or data points

So if you do that these terms so this is all expression is also not nothing new to you have a nonlinear term here and you have a certain constant terms in this expressions okay we need to tell the derivative of this because you need to actually maximize this so take the derivative with respect to the mean and the covariance term because this is are the 2 parameters which you need to estimate for that function.

This will give you an expression based on to estimate the mean so the mean estimated by the maximum likelihood or ML method where N is the number of sample points is given by this which a trivial expression which you can get it from here and the covariance matrix is actually given by the overall scatter matrix is given here.

So the ML method for estimation of the parameters mean and the  $\sigma$  is giving you the same expressions which we have seen earlier this the covariance matrix this is the way you estimate the covariance matrix and the corresponding mean for the data what happens if you have multiple Gaussians or what is called as a mixture of Gaussians.

(Refer Slide Time: 06:12)

## Gaussian Mixtures

- Linear super-position of Gaussians
 
$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)$$

Number of Gaussians

←

→

Mixing coefficient: weightage for each Gaussian dist.
- Normalization and positivity require:
 
$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^K \pi_k = 1$$
- Consider log-likelihood:
 
$$\ln p(X | \mu, \Sigma, \pi) = \sum_{n=1}^N \ln p(x_n) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right\}$$

Or Gaussian mixture and this is sometimes called as a linear super position of a set of k number of Gaussians K is the total number of Gaussians typically K is more than 1 but of course in a very special case K can be equal to 1 where are just have one Gaussian and the overall probability is now a summation of all K number of Gaussians where this  $\pi_k$  is called the mixing coefficient hence we will call this as the mixing coefficient for the  $K^{\text{th}}$  Gaussian the subscript k indicate the corresponding Gaussian.

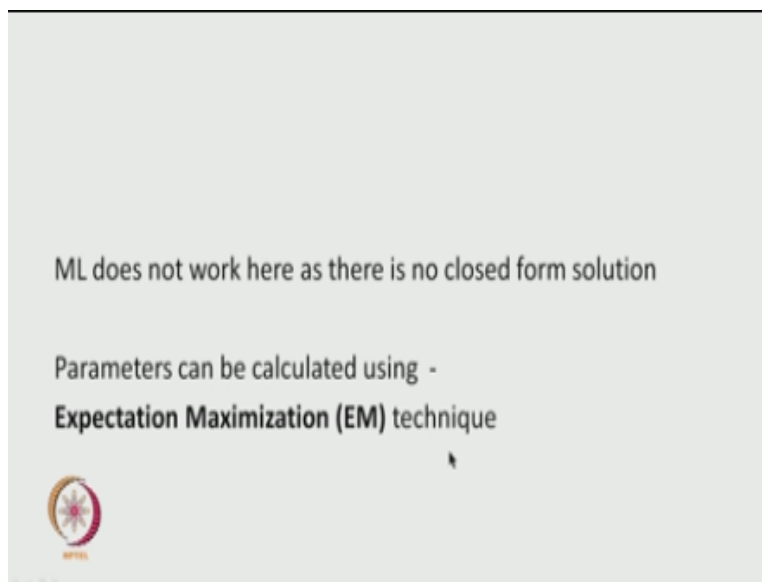
And this expression is the normal distribution normal multi variant Gaussian distribution normal multi variant Gaussian distribution for the class K  $\mu$  vector of the corresponding dimension as the sample K and this is the covariance matrix for the  $K^{\text{th}}$  Gaussian again repeat this is the mixing coefficient for the  $K^{\text{th}}$  Gaussian this is the normal multi variant Gaussian distribution for the  $K^{\text{th}}$  Gaussian okay the only constraint which we put with respect the mixing coefficients is that each of them live in 0 to 1 and the  $\sum$  of all this is equal to 1.

$\sum$  all are mixing so these are basically considered as weights they are also called the weights for the corresponding Gaussian function and if you take the log likelihood, if you take the log likely would of this overall function which is a function of the mean the covariance and the weight coefficients which is also given as a function overall the data samples N is the total number of samples and the p(x) is given here this is the p up to the probability for the distribution for a sample x as given as this.



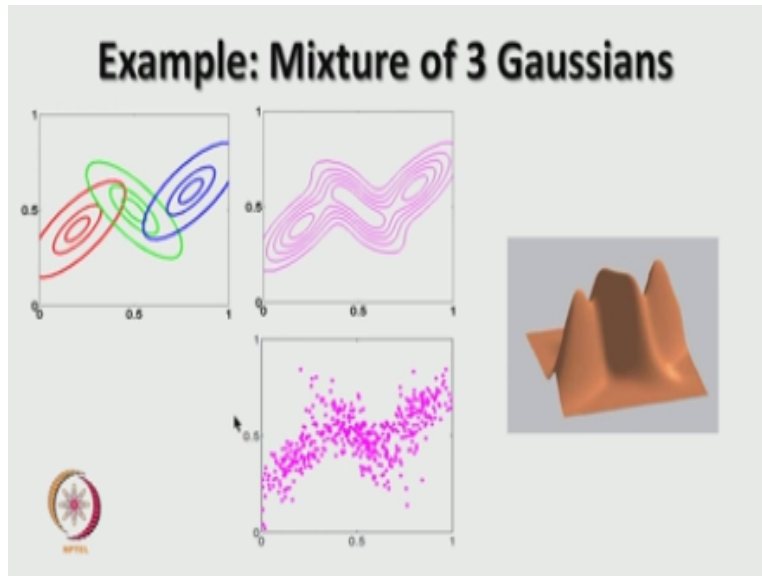
So this what you will get you can take the  $\sigma$  so this is the same so replace this expression  $p(x)$  sand by this here and this is what you will get okay remembered inside the logarithm you have the  $\sum$  over  $K$  Gaussians and then you have it for as many number of samples okay  $K$  is the index indicating  $K$  is the index indicating the  $K^{\text{th}}$  Gaussian and  $N$  is the indexes front to a particular instance or the sample total number of samples is  $N$  total number of Gaussians is  $K$  this is the expression you have for the lack likelihood for a mixture of Gaussians multi variate Gaussian distribution.

(Refer Slide Time: 08:53)



Is what we are considering in this case the maximum likelihood may not work to yield a close form solution and you need the method of optimization which is a attractive and that is what EM or expectation maximizations will give you.

(Refer Slide Time: 09:07)



So let us take this example to show what we are planning to intent if this is a short of a scatter of the data set of data samples which is obtained by the mixture of 3 Gaussian you can see here that this the overall trained of the does not follow a Gaussian distribution by itself, so we can cluster the data into three different components and say each of this individual components is a cluster following a Gaussian distribution in fact this particular data has been obtained or synthetically generated by.

Three different Gaussian distributions as given by this three different Iso contours these are asymmetric Gaussian distributions in 2D three different class means indicated by three different colors and they are Iso contour lines and actually if you look at this particular plot this is showing a surface plot in two dimensions where the height of the surface at each individual points reflects the probability density of a corresponding cluster or a Gaussian I repeat again if you look back into the slide.

This surface plot can be visualize to be an extension of this plot here where this plots indicated Iso contour lines or curves of equal distance which respect to the class mean but at each point if you compute the probability say you compute the probability at a point here and translate that to a height this is the plot which you will get so you will get surface plot where the height here on the right hand side is indicating the probability density of that function.

So what I mean is this is synthetic data obtained by 3 Gaussian functions and overall so this is cluster density may look like this which respect to the clusters so it is difficult actually model this

under single Gaussian distribution even in 2D and these are the three sample points corresponding to three clusters, if you remove the cluster level or the color of this data this is the data which you will have.

So cannot model this perfectly using a single Gaussian and this data shows the example why do you need multiple Gaussian's or a mixture of Gaussian's to model this data. Of course you could ask me a question how do you know a priori how many Gaussian functions you need, okay. So that is something which is not under the scope of the discussion today and it is a matter left for individual researchers to find out for a given data set what is the optimal number  $k$ .

There are methods to find out the optimal number  $k$ , if you have a data set a priori given to you often you can find out some methods by with the best  $k$  number of Gaussian's you can fit on it. In this case of course since I know the data before and I will say that the number of Gaussian's is three, but if you give you an arbitrary distribution which is non Gaussian in nature where  $k=2,3,4$  or 10 or even more is very difficult to visualize.

In practice in general but there are methods in which people adopt to find out what is the ideal value of  $k$  for the expression.

(Refer Slide Time: 12:31)

## Latent variable: posterior prob.

- ❑ We can think of the mixing coefficients as prior probabilities for the components
- ❑ For a given value of 'x', we can evaluate the corresponding posterior probabilities, called responsibilities



So we can think of mixing coefficients as prior probabilities for these individual components and so for given value of  $k$  we can evaluate the corresponding posterior probabilities called responsibilities which can be visualized as some latent variables in our expression which we saw couple of slides back.

(Refer Slide Time: 12:50)


□ From Bayes rule

$$z_k(x) = p(k|x) = \frac{p(k)p(x|k)}{p(x)}$$

Latent Variable

$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \quad \text{where, } \pi_k = \frac{N_k}{N}$$

Interpret  $N_k$  as the effective no. of points assigned to cluster  $k$ .



Let us get back in the expression of the likelihood but before that we will look at this base rule in which we define an expression of  $z_k$  as a latent variable given by under the base paradigm but this is nothing new few this is the posterior probability the class priors and so on and so forth which we are discussed earlier.

So what we are doing here is taking the class conditional probability to be our normal distribution a mixture of Gaussians so we have replace this term here by the numerator has given by the mixture coefficient and the corresponding unconditional prior to be the  $\Sigma$  of all the Gaussian's in the bottom.

What is the mixing coefficient  $\pi_k$  here, the number of samples for a particular class divided by the total number of samples here, okay. So interpret that the number of samples for a particular class as the number of points assign for a particular curve which is not assigned beforehand we do not know how many samples belong to a particular cluster  $k$ , so that has to be obtain and found out which in turn will actually give you the mixing coefficient  $\pi_k$ .

(Refer Slide Time: 14:01)

## Expectation Maximization

- ❑ EM algorithm is an iterative optimization technique which is operated locally
- ❑ Estimation step: for given parameter values we can compute the expected values of the latent variable.
- ❑ Maximization step: updates the parameters of our model based on the latent variable calculated using ML method.



So what does EM algorithm do, so the EM algorithm is an iterative optimization technique which is operated locally to find out the set of values of the parameters what are the parameters now we need to estimate here in a Gaussian mixture of Gauss, a mixing coefficients the set of class means for individual clusters so if there are  $k$  class, let us say somebody decides that I want to fit  $k$  mixture of Gaussian's or KGMM or  $k$  Gaussian mixtures to be very precise on the data.

So  $k$  is known, so if there are  $k$  Gaussian clusters which you want to fit so you have  $k$  different means, each of this mean has a dimension which is a same as the dimensional of the data. But there are  $k$  means,  $k$  covariance matrices and  $k$  mixture coefficients, so  $3 \times k$  seems to be the number of what not in terms of the number. But  $k$  sets parameters which you need to estimate okay, well of course the mixing coefficients each of them is a scalar quantity that is alright.

For  $\mu_k$  there are  $k$  means each of dimension  $D$ ,  $k$  covariance matrices how many elements  $D^2$  elements within each covariance matrix, so these are the set of parameters one needs to estimate. Let us see how the EM does it, so there are two steps in EM one is call the expectation another is called the maximization and this is an iteratively one after the another.

You have an expectation step or an estimation step as it is called followed by a maximization step, so an iteratively you follow this pair of steps one after another in a sequence unless you have a condition of conversions which is satisfied at the end. So estimation is for the given parameter values we can compute the expected values of the latent variable and hence it is called an expectation step as well.

And then you have a maximization step which updates the parameters of the given model based on the latent variable calculated using the ML method, okay.

(Refer Slide Time: 15:55)



So let us look at EM algorithm now, so given a Gaussian mixture model our goal is to maximize the likelihood function which we have seen a few slides back with respect to the parameters  $\pi$ ,  $\sigma$  and  $\mu$  comprising the means which is the  $\mu$ . The co variance matrix  $\sigma$  of the components and the mixing  $\pi_k$  okay, so the 1<sup>st</sup> step is initialize the means just an index from 1 to k the co variance terms and mixing coefficients and evaluate the initial value of the log likelihood.

You can start with the arbitrary set of random values here as well but instead of being absolutely random what we could also do is take since you do not have something like a class information you can take the overall of the entire data set and take the individual mean of the clusters or the Gaussian to be the data mean okay.

The co variance matrix could also start in fact there is lot of some degree of research which is happened to find out what should be the good starting point for any method the more closer you are to the final solution the faster you will convert the better solution is what you will have, so instead of starting with the absolute random values in this case it is possible to that you can have better estimates but I am not touching those aspects in this particular talk.

So initialize let us say with some initialize value which could be random values we go to the E step of the EM which is the expectation step and that you compute with the little variable as given as the expression here. So this is for the kth Gaussian which and the corresponding expression is given here kth. So the denominator you sum up all the Gaussian distributions for that value for the corresponding x which you have estimated using the random number so sum up all Gaussian in the denominator and the corresponding Gaussian after you have estimated this.

(Refer Slide Time: 18:13)

**EM Algorithm for GMM**

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

$$\mu_j = \frac{\sum_{n=1}^N \gamma_j(x_n) x_n}{\sum_{n=1}^N \gamma_j(x_n)}$$

$$\Sigma_j = \frac{\sum_{n=1}^N \gamma_j(x_n) (x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^N \gamma_j(x_n)}$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^N \gamma_j(x_n)$$

4. Evaluate log likelihood

$$\ln p(X | \mu, \Sigma, \pi) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k N(x_n | \mu_k, \Sigma_k) \right\}$$

If there is no convergence, return to step 2.

Let's go to the 3<sup>rd</sup> step which is the EM step and the 2<sup>nd</sup> step mind you have estimated this  $\gamma_j$  or the  $\gamma_k$  the index as changed but it is the same variable and that goes inside the expression here to compute the corresponding mean and the co variance from and this is the same as the EM step done earlier except that the written variable is sort of a weight here with comes here and more accurate value of the mean and co variance matrix.

The mixing coefficient must also be calculate using the variable computed in the e step in the step 2 earlier as given here, correspondingly so one the  $\mu_j$  are available here you can see the 3 expression here the mixing coefficient, co variance and the mean they all are computed using the data sample points and the variable this is the normal distribution expression as given here, so then what you do obtain the Gaussian mixtures using the parameters estimated in step number 3.

Now what you need to do here is find out if the corresponding likelihood estimated here truly represents the data samples if this is not which will not typically happened you back go back to

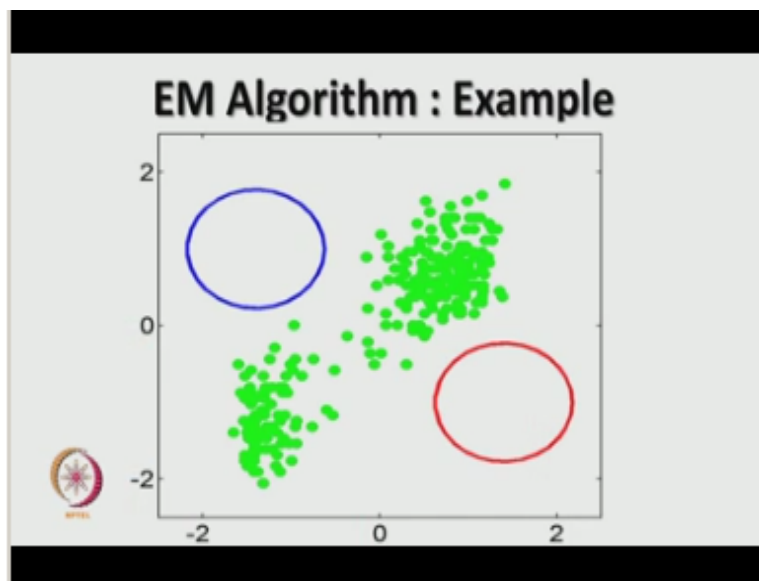


step number 2, so you keep repeating step number 2 , 3 and 4 what essentially which is the E and the M part that will help you to actually converge in better solution we will have an illustration of this very soon in next few slides.

And you put an converge and say that I will keep on repeating this process till my log likelihood of the distribution satisfies some criteria of converge. So the convergence criteria could be such that in successive iterations or over set few iterations the parameters do not change, what are the parameters the mixing co efficient  $\pi$  the mean  $\mu$  and the co variance scatter  $\sigma$  they do not change over successive iteration over a last few iteration.

The other some similar in which you can use is given in step number 4 is when you estimate this likelihood if this itself does not change over a set of iteration, so you compute and restore the log likelihood value which we have computed in the previous iteration compare that with the current one and the change is negligible below a threshold you say that you have met a criteria for conversions and you say that you have an estimated the corresponding Gaussians over the distribution which we have

(Refer Slide Time: 21:12)



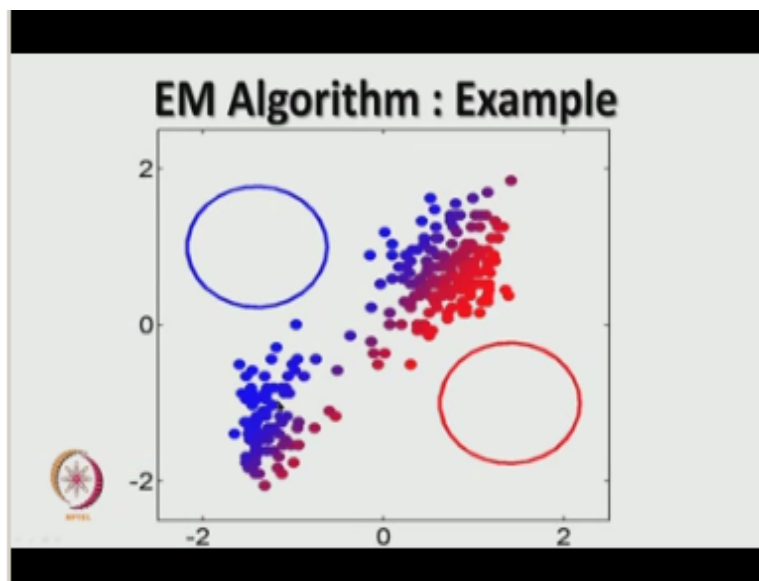
So to wind up let us take an example now in 2d and this images have been obtained also from the book by bishop so we have taken it from the book fashion recognition and machine learning the reference of the book has been given at the beginning of this particular lecture so what does it

show here this is a scatter you can almost see it is visible that there are two different clusters of data so it is possible to fit them with two different Gaussians distributions.

So almost blindly I am selecting the  $k$  to be 2 but you can select  $k$  to be 3 or 4 also there will be some amount of convergences whether good or bad another into be seen but in this case let us take the example of and let say the Gaussians have been initialized at these two places with the corresponding mean and scatter this two ellipse is show that this is the Gaussians here  $k=1$ ,  $k=2$  Gaussians is here that these are the corresponding means at centre of the ellipses and the distributions this scatter is in 2 d.

So as you keep proceeding remember what did we do in the EMI algorithm there were two typical steps one was the E step in which you are estimating after the initial set of ransom values have been has been used to start this cycle for the corresponding set of variable or the parameter of the Gaussians mixture model you use a E step to estimate the written variables then using the third step you go to the M step where you actually estimate the parameters again using the lateral variable and continue these process off course you keep in watch in the clock as you proceed.

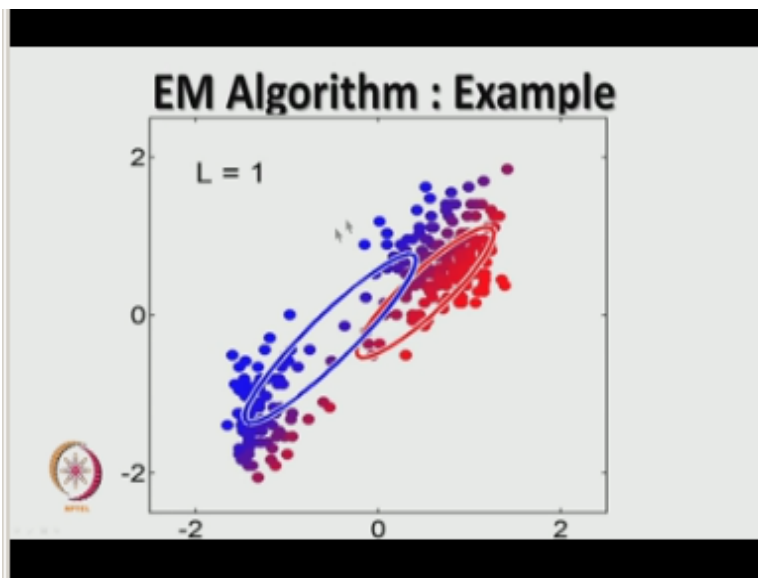
(Refer Slide Time: 22:56)



That these are initial set of points which are marked in blue will be assigned to this particular Gaussians and what will happen to this set of points because they are the once which are going to be closed closer to the corresponding Gaussian remember this was the initial stage okay and you start assigning the distribution because this will the points lying close to the blue cluster.

Here or the blue the Gaussians market with these blue ellipse will be closer to this set of points and hence they will be assign to this corresponding Gaussians and the points which are labeled in red now we will assign to this so what you do now after this assignment you recomputed the parameters the mixing coefficient the mean and the scatter let's see how the recomputed means after the first literature look like so once you recomputed they will appear like this.

(Refer Slide Time: 23:52)



From the data this is distribution for the points which have been assigned to the Gaussian this is step number one and the corresponding which is in two will be marked here this is after the second tip so this is how this start converting you reassign the points one second and this is how the literature at number step 5 we can see that the one of the Gaussians convert to this close cluster of points marked in blue.

The points here are red flowing another Gaussians distribution after 20 you have almost convert to this point and this is the final stage of interaction where if you still you will not have a change in either the parameter values or a log like layout criteria after you compute with this set of parameters okay let us have look at this animation once second.

So this is the starting point you not know where is the clusters where should put the you can start almost anywhere and then this is how the conversion takes place second interaction 5<sup>th</sup> interaction and the 20<sup>th</sup> interaction okay so this is the method by which you can fit a set of Gaussians to certain scatter of data points which do not point a invariant single Gaussians.

And this is the method which is actually used not only to form clusters but usually model data points after because after this you can actually apply all your methods of classification or the you want to transform clusters group them under certain criteria and so on so this is one method which is used to model as well as form clustering thank you very much.

### **Online Video Editing /Post Production**

K.R.Mahendra Babu

Soju Francis

S.Pradeepa

S.Subash

### **Camera**

Selvam

Robert Joseph

Karthikeyan

Ram Kumar

Ramganes

Sathiaraj

### **Studio Assistants**

Krishankumar

Linuselvan

Saranraj

### **Animations**

Anushree Santhosh

Pradeep Valan .S.L

### **NPTEL Web &Faculty Assistance Team**

Allen Jacob Dinesh

Bharathi Balaji

Deepa Venkatraman

Dianis Bertin

Gayathri

Gurumoorthi

Jason Prasad

Jayanthi  
Kamala Ramakrishnan  
Lakshmi Priya  
Malarvizhi  
Manikandasivam  
Mohana Sundari  
Muthu Kumaran  
Naveen Kumar  
Palani  
Salomi  
Senthil  
Sridharan  
Suriyakumari

**Administrative Assistant**

Janakiraman.K.S

**Video Producers**

K.R. Ravindranath  
Kannan Krishnamurthy