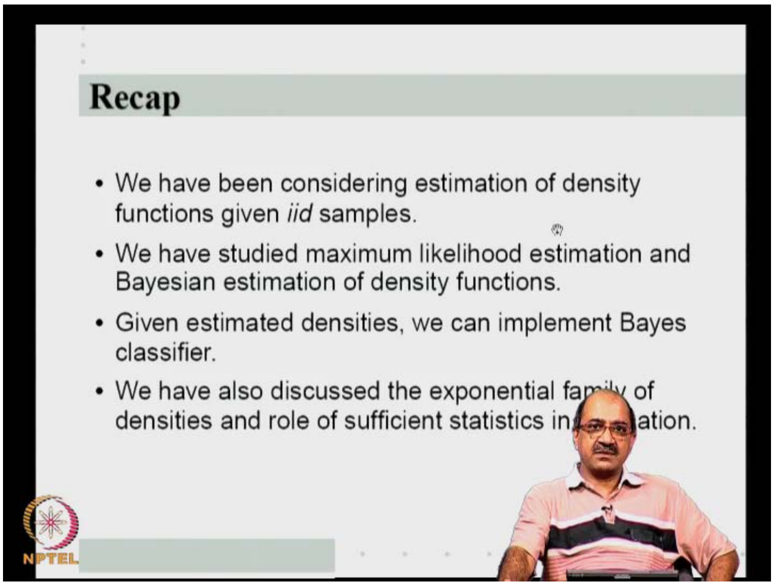
Pattern Recognition

Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 10
Mixture Densities, ML estimation and EM algorithm

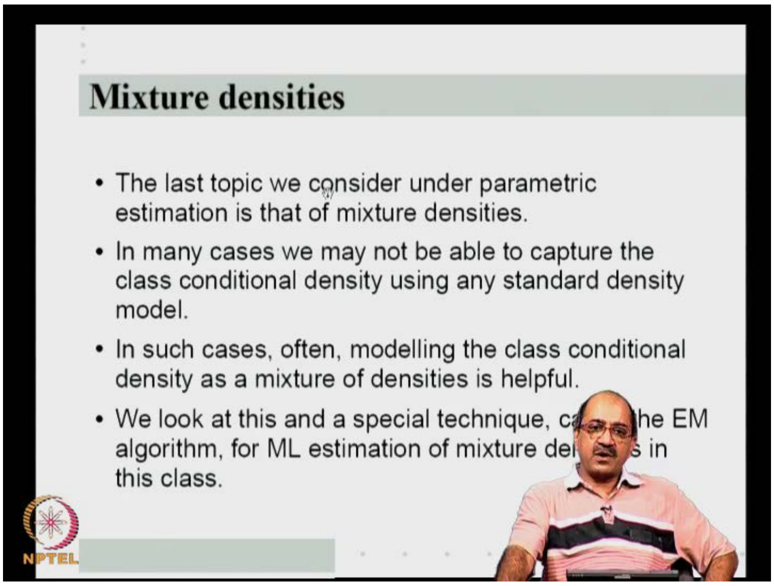
Hello and welcome to the next talk in Pattern Recognition. This would be a last topic we look at parametric estimation and then, we will move on to non-parametric estimation. So, to briefly recall what we have been doing in the last few classes, we will considering estimation of density functions, given IID samples from the density. We studied maximum likelihood estimation, Bayesian estimation, density functions.

(Refer Slide Time: 00:42)



For all the standard density functions, we seen how to compute the ML estimates as well as the Bayesian estimates. The idea is that, given the estimate densities, we can implement the Bayes classifier with the estimated densities; while considering ML estimation and so on. We had also looked at the issue of exponential family of densities, as we have seen that is a very good generic density model, all the standard densities are captured there. We looked at exponential family of densities and we also studied the role of sufficient statistics in estimation, in the last class.

(Refer Slide Time: 01:20)

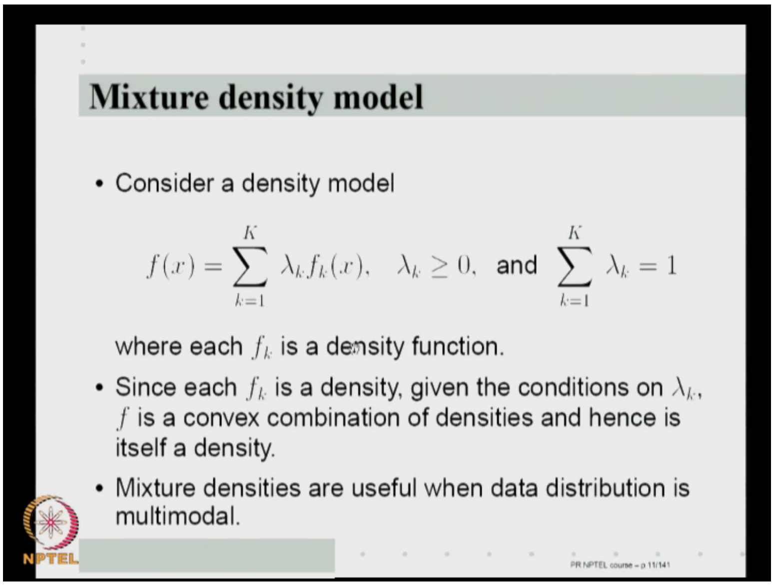


So, as I said, we are going to look at the last topic namely mixture densities, in the parametric estimation. So, today we will consider estimation of mixture of densities, the basic idea is that, in many cases, the standard density model such as exponential or normal or gamma whatever, do not necessarily capture the the underlying data distribution.

One reason could be for example, most densities are unimodal and you know, data distribution may not be unimodal model. In such cases, it is often very helpful to consider a mixture of densities as the class conditional density models. That is, instead of thinking of a class conditional density a single normal, we may think of it as a mixture of normal's or mixture of exponentials and so on.

So, that is where, mixture densities come and in this class, we are going to look at, how one estimates mixture densities. Our main reason for looking at mixture densities is that, through this problem, we will introduce a very important algorithm in estimation, which is called the EM algorithm. EM stands for Expectation and Maximization, we will see the algorithm later so, this EM algorithm is a is a very important algorithm estimation, useful in many probabilistic models including things like HMM models, hidden Markov models and graphical models and so on. While many of those things we may not consider in this class, this since this algorithm is important, we look at it from the point of view of estimating mixture of densities.

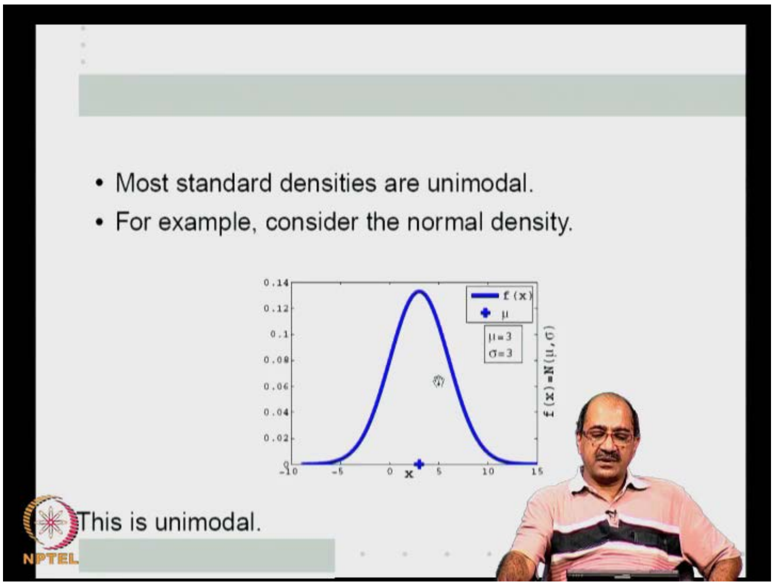
(Refer Slide Time: 03:05)



So, this is the mixture density model, consider a density model where, the density of x is given by summation over k, k going from 1 to capital K, lambda k f k x where, lambda k is greater than or equal to 0 and lambda k sum to 1. And each of the f k's is a density of course, may be same kind of density or different kind of densities, that does not matter but, each of f k's is a density.

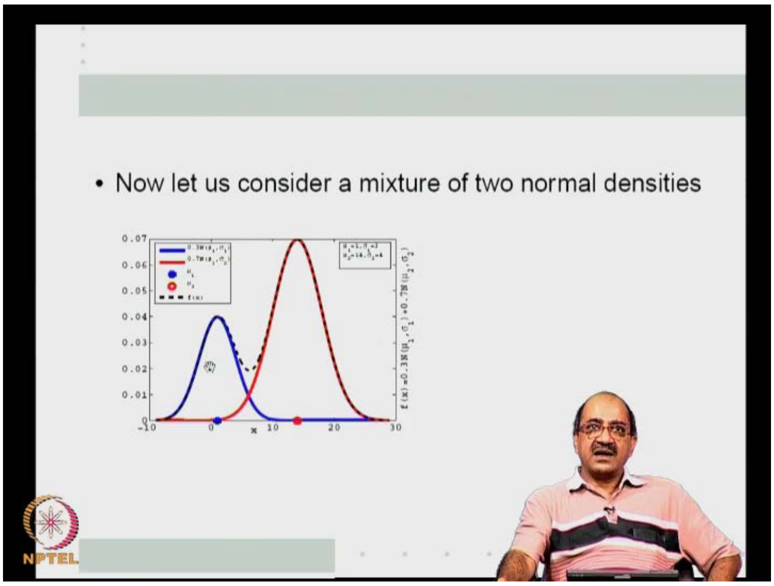
Since each of k is a density and all the lambda k's are positive, the summation lambda k of k will be positive, also if you integrate this over to over x, each of k will integrate to 1. And since summation lambda k is equal to 1, f x will itself integrate to 1; so, essentially because, f is a convex combination of densities, it will itself be a density. So, this is a well formed density model, if lambda satisfy these conditions for convex combination, and f k's are densities. We often call lambdas as the mixing coefficients and f as the component densities. Mixture density are very useful especially, when data distribution is multimodal, that instead of being having only one maximum, it may have many many maxima.

(Refer Slide Time: 04:23)



For example, if you consider most standard densities are unimodal for example, normal, there is a normal density curve, which has exactly one maximum right. Everything falls off monotonically on either side of the maximum, exponential density is a similar thing, more standard density is Poisson binomial, all of them are unimodal.

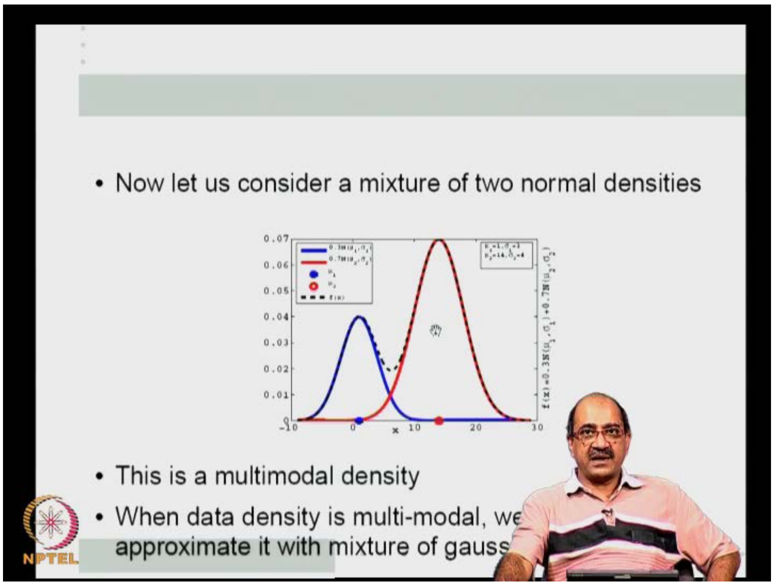
(Refer Slide Time: 04:45)



But, mixture densities allow us to capture densities, that are multimodal so, here is a mixture of two normal densities. The blue is one normal density, the red is another normal density and the black one, the the that is going like like that is a mixture of the two, here, I just taken 0.3 and 0.7 as the mixing coefficients. Essentially, if I take the two means of the normal, sufficiently far off, on one side the sum is completely dominated by one density, and the other side the sum is completely dominated by other density.

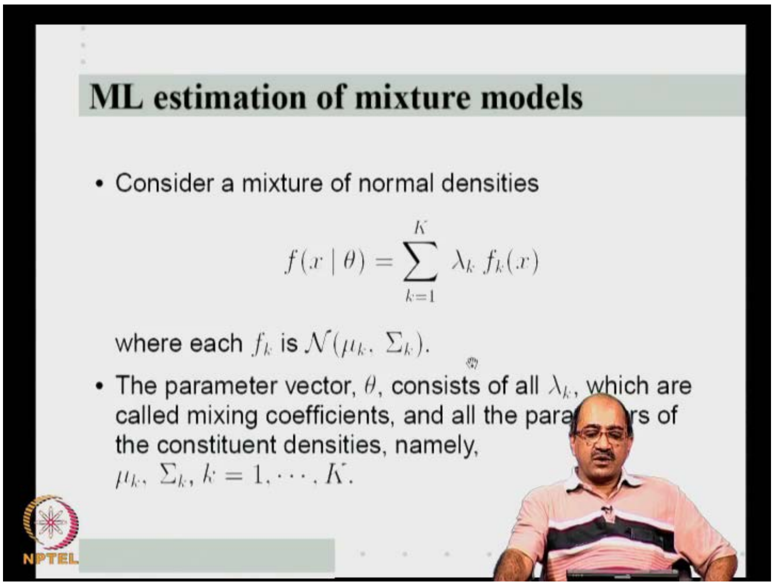
In between, I get a smooth transition from one maximum to another. So, many data distributions, which have such multimodal characteristic right can be easily captured using mixture of normal.

(Refer Slide Time: 05:34)



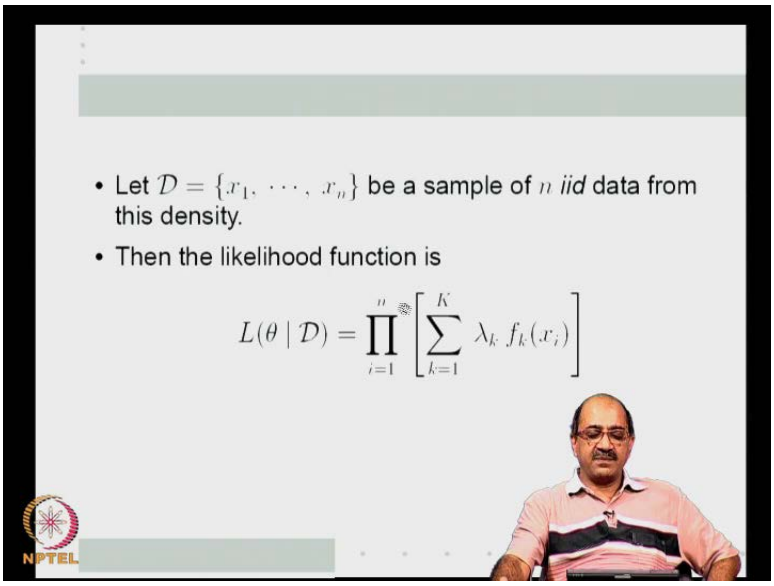
So, these are multimodal densities, the mixture is a multimodal density and when data density is multimodal, we can often approximate it with sum of normal's like this. So, as a matter of fact, what I call Gaussian mixtures, mixture density modal is a very often used density modal, for class conditional densities in pattern recognition problems.

(Refer Slide Time: 05:58)

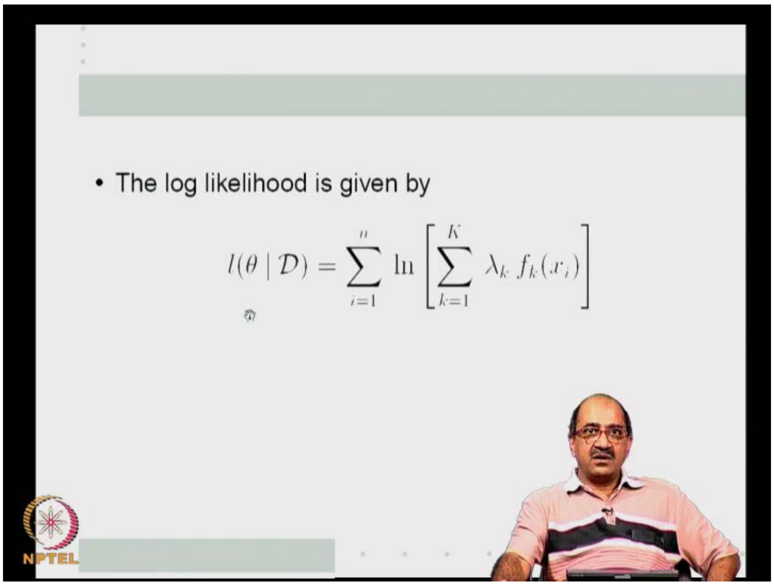


So, let us say, we have a mixture of k normal densities, f x given theta, k is equal to 1 to k lambda k f k x where, each f k is normal with mean vector mu k and covariance matrix sigma k right. So, the parameter vector theta consists of all the lambda case, which are called mixing coefficients and the parameters of all the constituent densities namely, all mu k's and all sigma k's. The f k's are sometimes called component densities or constituent densities, let us say, lambda k are called the mixing coefficients.

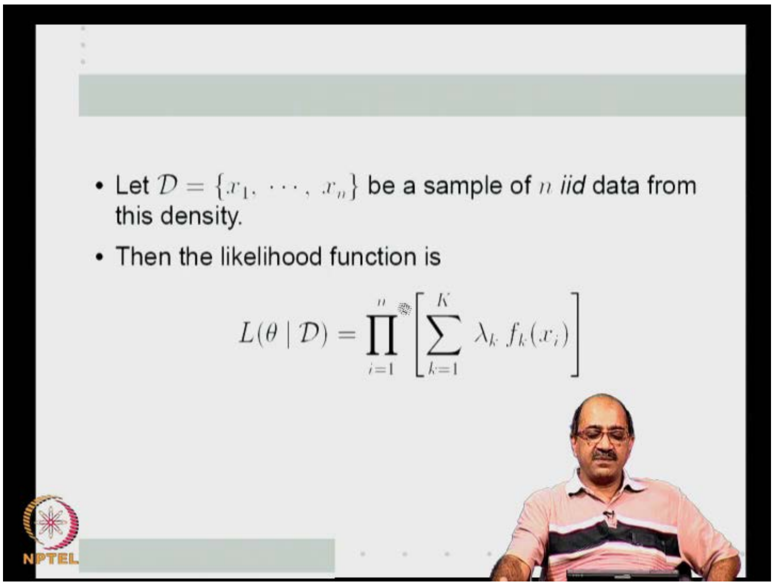
(Refer Slide Time: 06:37)



(Refer Slide Time: 07:18)



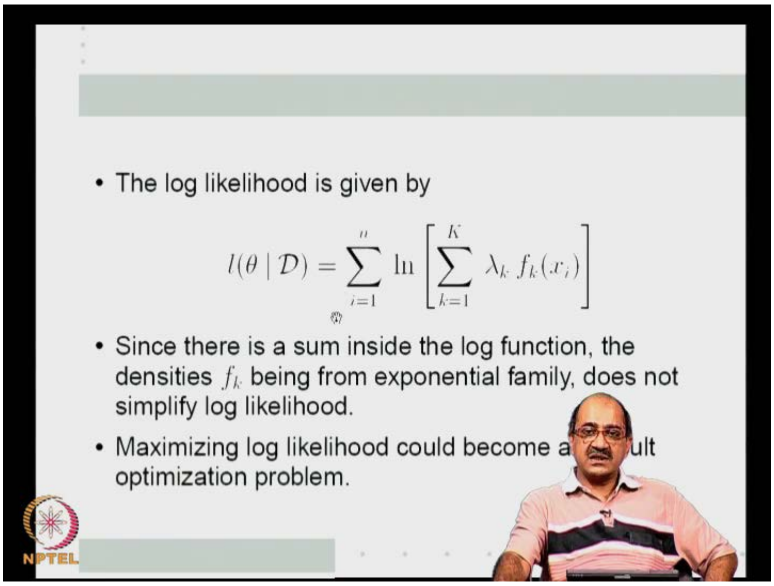
(Refer Slide Time: 07:18)



Now, let us say, we we have this density model and let us say, we have a sample of n iid data from this density model that is, the standard scenario in any maximum likelihood estimation. We have a density model and we have data and then, we form the likelihood and maximize the likelihood to estimate the parameters. So, the likelihood would be what l theta given D, is simply product of this density model right.

So, if I do product of this density model, this is what I get, product i is equal to 1 to n of the density model, which k is equal to 1 to k lambda k of k . As you have seen, we often take log likelihood, in all our earlier examples taking log simplified the problem. So, let us take the log likelihood, if I take log likelihood, this product of course, will certainly become sum when I take a log.

(Refer Slide Time: 07:24)



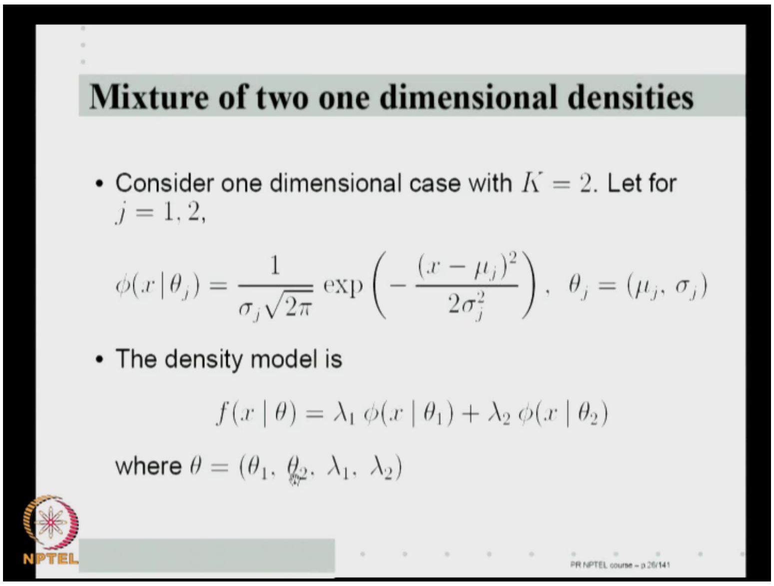
So, I get sum of l n of this now, this kind of tells you, why mixture density model estimation is difficult, see what we have here is the sum inside the log function. Earlier, we got sum log of directly the f k, and f k being from the exponential family for example, Gaussian. Because, f itself will be exponential something, if I take log, I get a very nice expression and that is, how we have solved analytically for the ML estimation, of all the standard densities.

But here, because, there is a sum inside the log and we cannot simplify it, the fact that f k is from an exponential family, does not give us any analytical simplification. So, this means, that maximizing log likelihood can become a difficult optimization problem earlier because, as directly got l n suppose, capital k is 1 then, you have to simply get l n lambda times l n f.

And if f is exponential that, what is inside the exponent comes out and as we have seen through sufficient statistics, that immediately gives rise to very simplified ML analytical expressions for ML estimates. But here, we have a sum inside a log and hence, those simplifications do not occur then, we may have to solve it numerically and it can become a difficult optimization problem.

What we will see in this class is, we will take a specific example of this and try to solve it by explicitly calculating the partial derivatives and out of that, we will try and formulate a general procedure for all such densities.

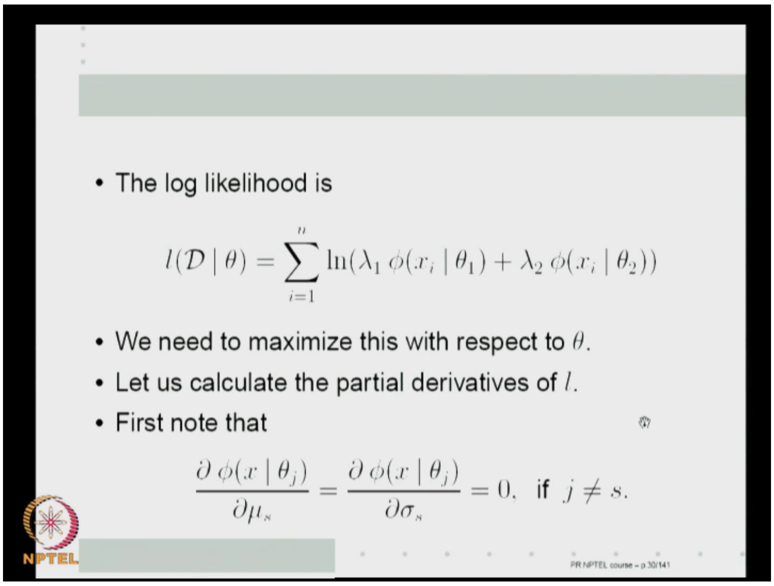
(Refer Slide Time: 09:08)



So, for our simplification, we will consider one dimensional problem with only two component densities, we want both component densities to be normal. So, let us put a special symbol for it, let us say, phi of x given theta j is normal with parameters mu j and sigma j. So, theta j consists of mu j sigma j so, phi of x given theta j is normal with mean mu j and variance sigma j square right, this is the phi x given theta j.

And let us say, our density model is mixture of two such normals so, f x given theta is lambda 1 times phi x given theta 1 plus lambda 2 times phi x given theta 2. So, this is phi x given theta 1 is normal with parameters theta 1 namely, mu 1 and sigma 1, phi x given theta 2 is normal with parameter theta 2 namely, mu 2 and sigma 2. So, f itself will have parameters given by theta, as theta 1 parameters of this density, the theta 2 parameters of this density and the two mixing coefficients, that will be the parameter vector.

(Refer Slide Time: 10:12)



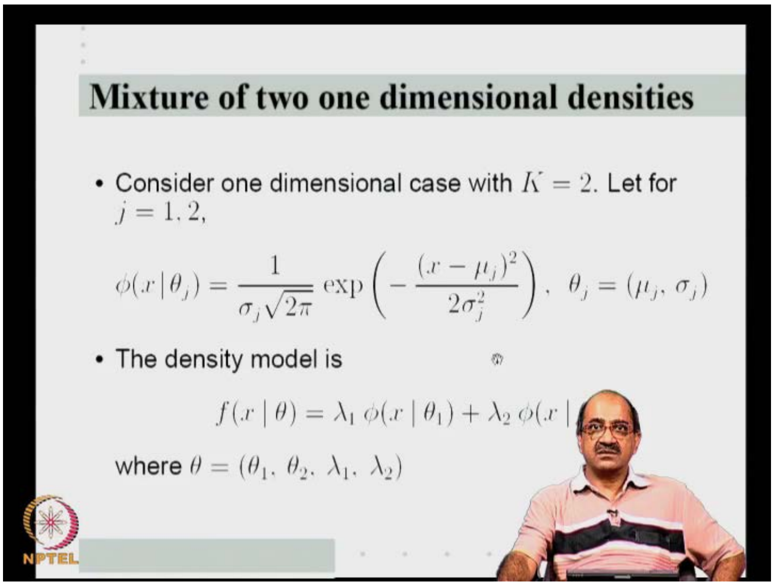
Now, what is the log likelihood, as we have just now seen is, summation i is equal to 1 to n, l n of lambda 1 phi x i, given theta 1 plus lambda 2 phi x i, given theta 2. So, this is what, I have to differentiate with respect to various components of theta, what are the components of theta mu 1, mu 2, sigma 1, sigma 2, lambda 1, lambda 2. So, I have to differentiate this, with respect to the components to maximize so, to obtain the ML estimate of theta, given this log likelihood, we have to maximize this with respect to theta.

To maximize this with respect to theta, we have to calculate partial derivatives of l and equate it to 0 and see, if we can solve for it. So, if I want partial derivatives let us say, with respect to mu or sigma of this, essentially this l n will give me 1 by that lambda 1 phi, given theta 1 plus lambda 2 phi, given theta 2, into derivative of phi, given phi x i given theta 1, with respect to mu 1 or sigma 1 or whatever similarly, derivative of phi x i, given theta 2 with respect to mu 1 sigma 1.

So, first, let us calculate this partial of these these derivates of phi, with respect to mu and sigma. The way we put this theta 1 and theta 2, let us first notice that, partial derivative of phi x, given theta j with respect to mu s or phi x, given theta j with respect to sigma s is 0, if j is not equal to s. So, if I want phi x, given derivative of phi x, given theta 1 with respect to mu 2 because, phi x given theta 1 depends only on mu 1 and sigma 1, is derivative with respect to mu 2 would be 0.

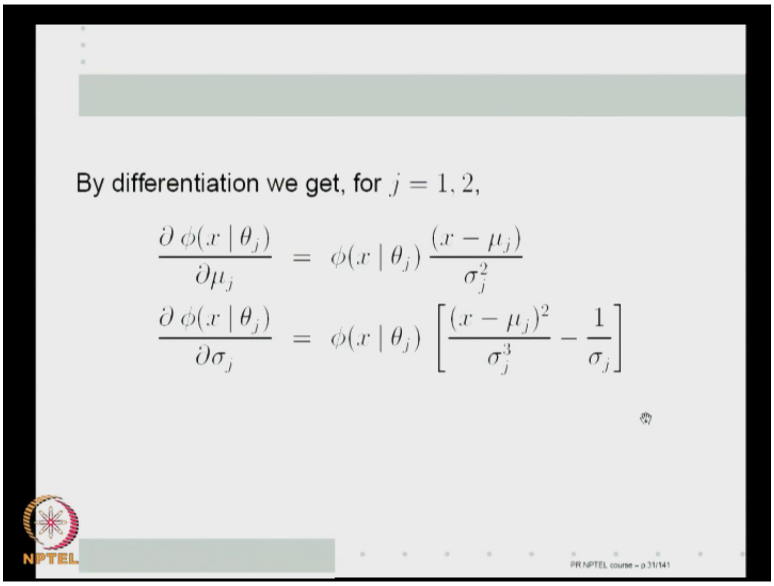
Similarly, the derivative with respect to sigma 2 will also be 0 right so, only phi x given theta 1 will have non-zero derivative with respect to mu 1 and sigma 1 only. And phi x given theta 2 will have non-zero derivatives with with respect to mu 2 and sigma 2 only. So, we do not have to calculate the crossed derivatives so, we have to only calculate derivatives of phi x, given theta j with respect to mu j and sigma j.

(Refer Slide Time: 12:11)



That is easily done simple differentiation of this, if I want derivative of this let us say, with respect to mu j, I get this factor as it is, into exponential of this as it is, multiplied by, the 2 will come out, we will cancel with this 2, x minus mu j by sigma j square, and another minus will cancel this minus. So, essentially what I get is the same phi x, given theta j multiplied by x minus mu j by sigma j square right.

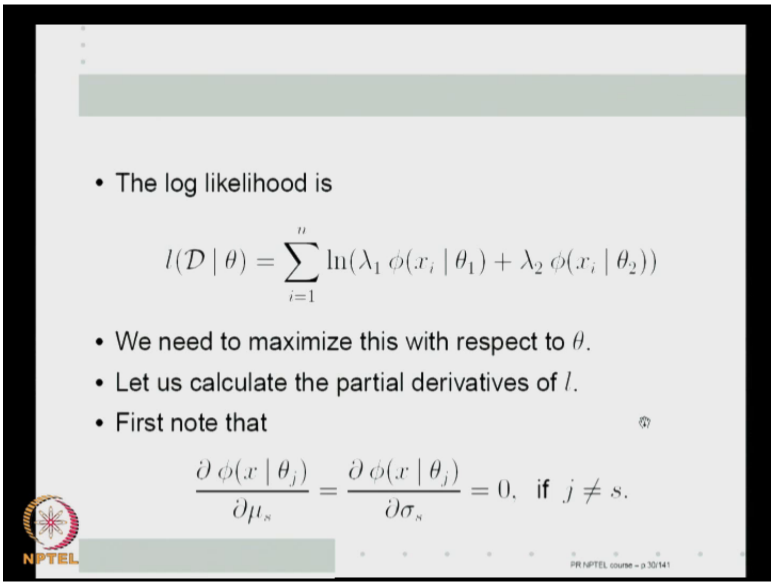
(Refer Slide Time: 12:45)



So, that is what I will get, del phi x given theta del by del mu j is phi of x given theta j into x minus mu j by sigma j square. Similarly, differentiating the normal function with respect to sigma j, we get minus 1 by sigma j square into exponential of this in one term and 1 by sigma j root 2 pi exponential of this into x minus mu j whole square, into minus 2 by sigma j cube in this term. So, in both of them the the phi x given theta there will be a common factor.

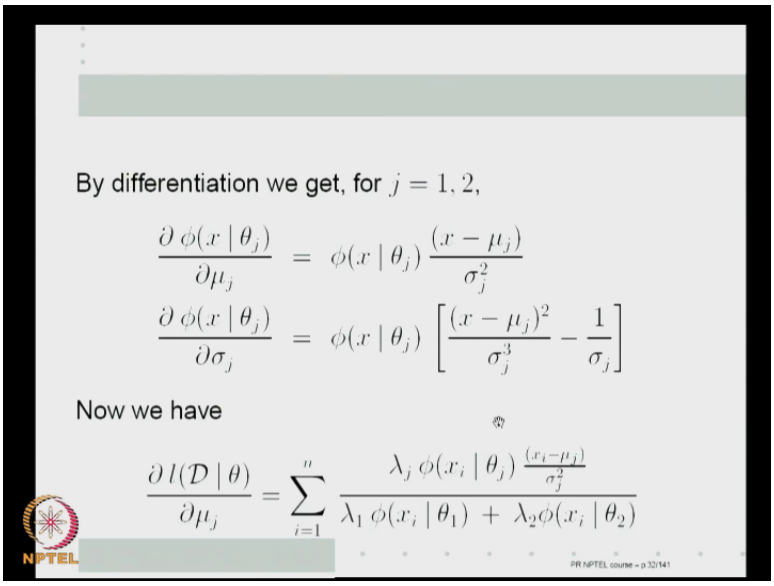
So, with that algebra, one can show that, phi x given theta j, del phi x given theta j by del sigma j will once again be phi x given theta j multiplied by this x, minus mu j whole squared by sigma j cube minus 1 by sigma j.

(Refer Slide Time: 13:40)



Given this now, we know how to calculate derivative with respect to this, if I want derivative of this with respect to mu 1, that will be because, this l n, this derivative will go inside the sum. l n will give me 1 by lambda 1 phi x i, given theta 1 plus lambda 2 phi x i, given theta 2 into derivative of this with respect to mu 1. Derivative of this second term with respect to mu 1 will be 0 and thereby, the first term with respect to mu 1, we already know.

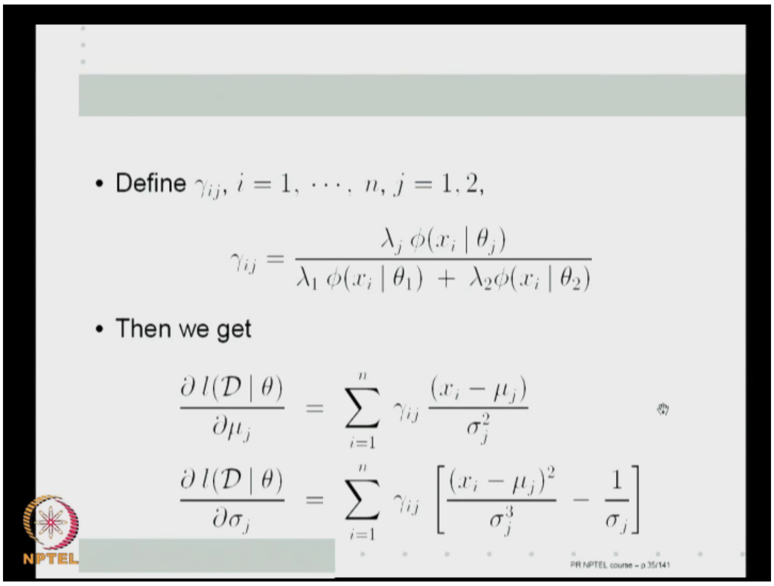
(Refer Slide Time: 14:04)



So, doing all that, we get derivative of l with respect mu j, will be lambda j phi x i given theta j that comes from here into x i minus mu j by sigma j square, this is the del phi del phi x theta j by del mu j, this comes because of the log. So, this is the partial derivative with respect to mu j similarly, we will get this has to be sigma j, if you want with respect to sigma j once again this 1 by the lambda 1 plus lambda 2 term will be there.

And with respect to sigma j, I once again get that lambda j phi x, given theta that term and instead of this x minus mu j by sigma j square, I will get this term right. So, in all these terms, as you can see, we have this factor lambda phi x, given theta j by lambda j of x i, given theta j by lambda 1 phi x i, given theta 1 plus lambda 2 phi x i, given theta 2 plus given name to that term.

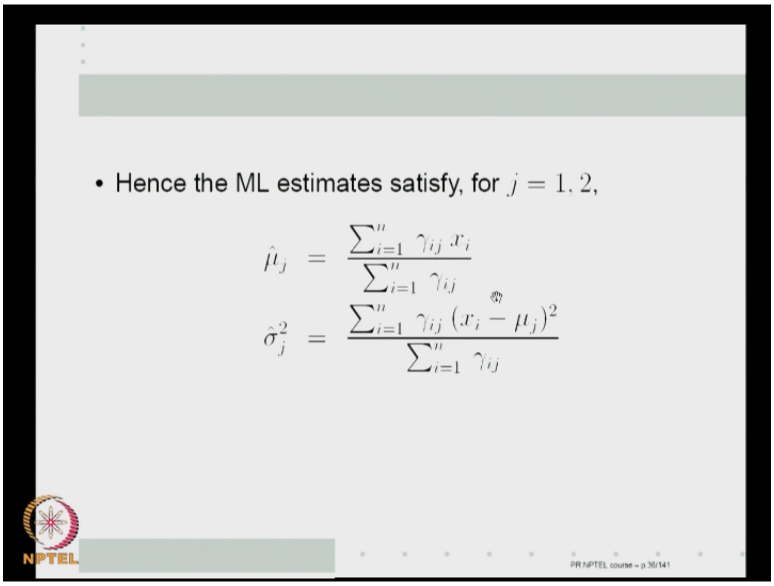
(Refer Slide Time: 15:03)



Let us call it gamma i j, gamma i j is lambda j phi x i, given theta j by the sum of the two densities then, the earlier equation can be written as del l by del mu j, is summation i is equal to 1 to n, gamma i j into x i minus mu j by sigma j square right. Just by taking that factor out, we got a very simple equation, as I said derivative with respect to sigma j will also be same so, let us write that also.

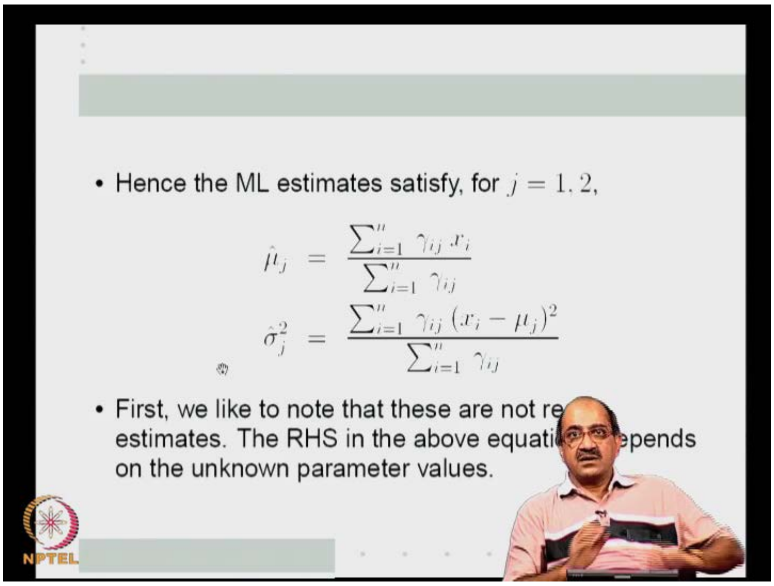
So, once again, it will be summation over i gamma i j instead of this term, the derivative the sigma j derivative term will come there. So, if I want to solve for these, I have to just equate them to 0 so, for example, if I equate them to 0, what do I get, summation gamma i j, x i minus mu j equal to 0. By multiplying by sigma j square right, that will give me summation gamma i j, x i is equal to mu j times summation gamma i j.

(Refer Slide Time: 16:05)



So, I can easily solve from mu j that is, what I get so, I know the ML ML estimates satisfy mu j, if summation i is equal to 1 to n gamma i j, x i by summation gamma i j. similarly, if I equate this to 0, take sigma j cube this side, it becomes sigma j square so, I get sigma j square times summation gamma i j is equal to gamma i j times x i minus mu j whole square. So, sigma j square will be summation gamma i j, x i minus mu j whole square by summation gamma i j right.

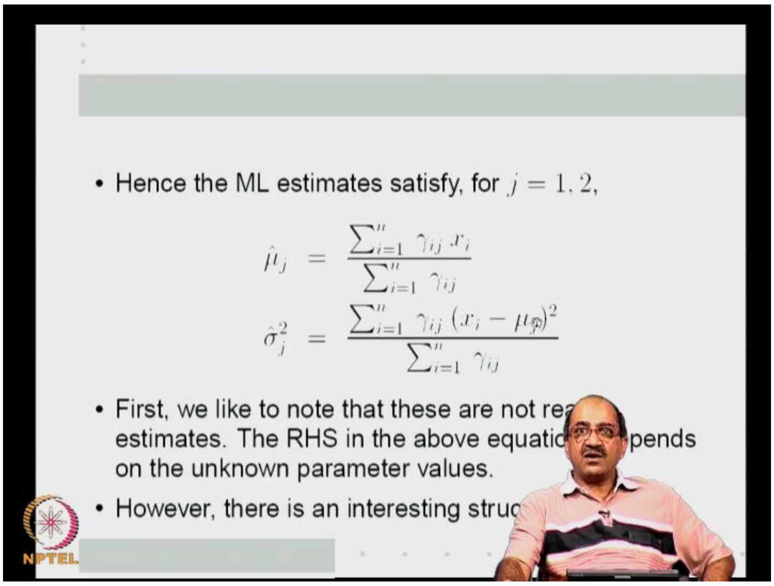
(Refer Slide Time: 16:46)



That is the term, first let us understand that, we have not solved the problem at all, the gamma i j right depend on theta j. That means, they they need to to calculate gamma i j, I need to know mu j sigma j for j is equal to 1 to 2. So, even though this looks like an estimate, it is not an estimate because, the RHS depends on the parameters we want to estimate.

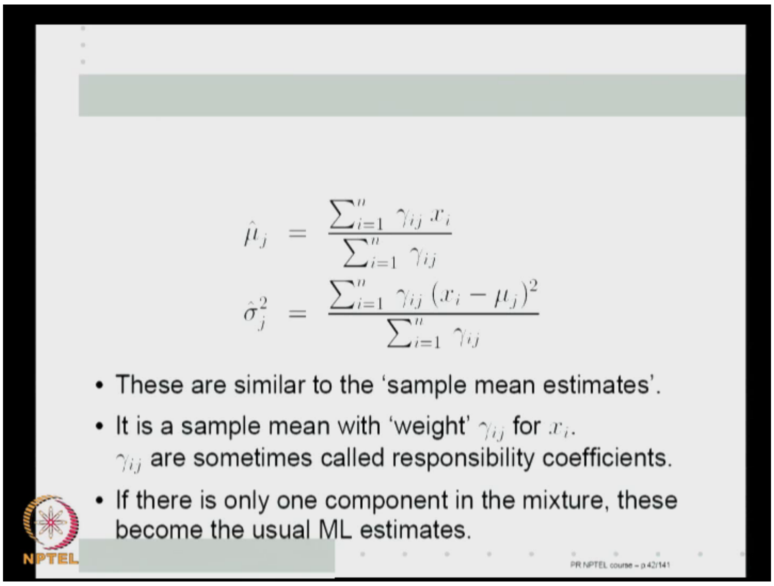
But, if I put the same mu j hat sigma j hat in the RHS, this is some equations that, the mu j hat sigma hat j hat have to satisfy that is why, I I I wrote here ML estimate satisfy this equation. The inside, these are the ML estimates given by the the equation right because, both sides I will have the mu's and sigma's right. But, on the other hand, this is very interesting structure here right say, for example, if gamma i j is equal to 1 for all i, it will be simply summation x i by n, which is the old sample mean estimate.

(Refer Slide Time: 17:42)



So, there is a lot of structure here so, let us look at the structure more closely.

(Refer Slide Time: 17:46)



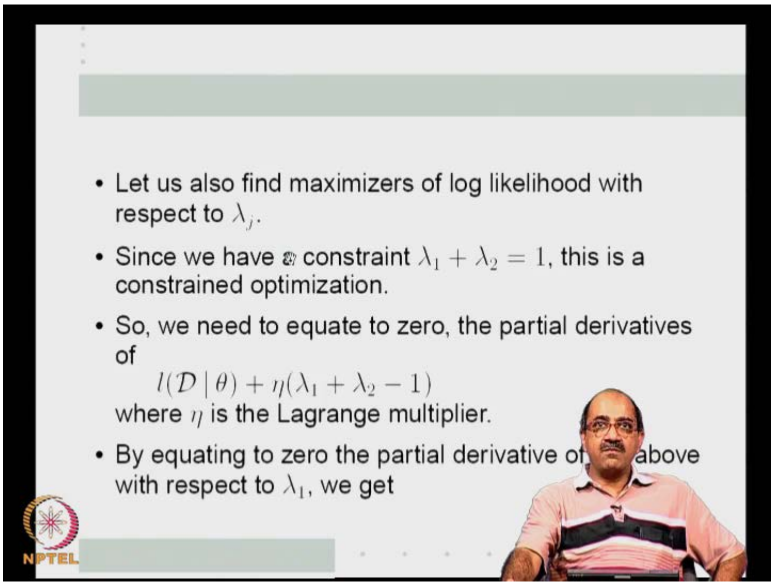
These are my estimates so, first thing is, if they are they are like sample mean estimates, if I look at the mu j estimate, it is summation something into x i plus summation the same thing right. The sample mean would have taken this gamma i j's to be 1 instead of getting 1, I am getting some other numbers so, I can same thing about sigma j i i j square estimate. So, I can think of this, as it is a sample mean where, all samples are not weighted equally, it is a sample mean but, the sample x i is weighted with gamma i j right.

So, if I think of if i think of giving separate weight to each sample so, when I am estimating mu j, I I give the weight gamma i j to sample x i. So, you can think of gamma i j, as how much responsibility x i has, for estimation of the j-th component density. So, if I think of gamma i j as weights then, this is essentially a weighted sample mean. See somehow, if gamma i j's are 1 and 0 so that, I know this exercise are from the mu j density then, this should have an exactly the sample mean.

But, instead of that, this it has become a weighted sample mean with weights gamma i j and the gamma i j sometimes called the responsibility coefficients. Of course, if if there is only one component then, j will instead of being one, to do j is 1 then, gamma i j's are all 1 then, this has the same sample mean estimates, we got earlier for ML estimation of mean and variance, of a single normal density.

So, they are kind of, in the limiting case of only one component the in the density, they go back to the single density estimates. They still retain the nice sample mean structure, with gamma i j is being the weights associated with the i th sample and if I think of these these as weights then, these are essentially weighted sample mean estimates. Of course, they are not really estimates, as they are said because, the gamma is a dependent mu i mu j. So, let us come back to this but, let us remember that, mu j and sigma j are not the only parameters, lambda 1 lambda 2, the mixing coefficients are also to be estimated.

(Refer Slide Time: 20:19)

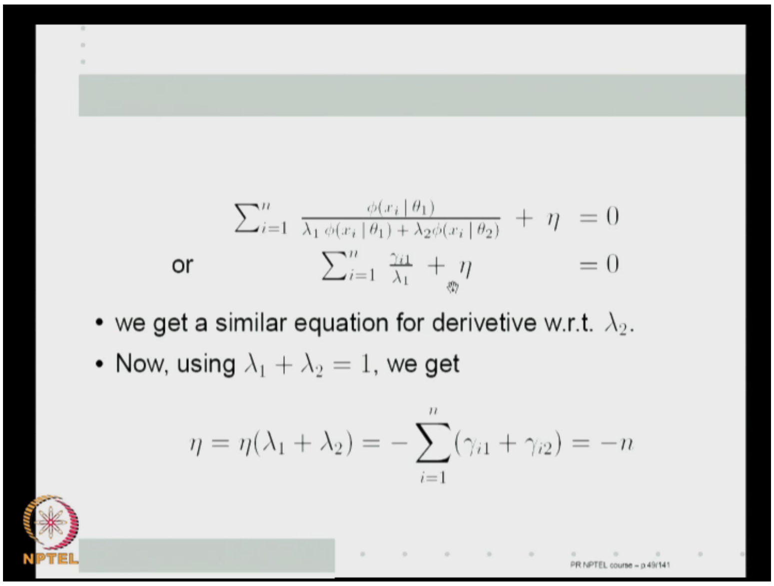


So, let us complete our estimation so, we have to maximize the log likelihood with respect to the lambda j's also but, maximization with respect to lambda j s is not straight forward. Because, lambda j's have a constraint lambda 1 plus lambda 2 is equal to 1 of course, they also have a constraint that, lambda 1 lambda 2 greater than 0. But, we have to certainly, satisfy the constraint lambda 1 plus lambda 2 is equal to 1 and hence, it is a constraint optimization problem.

No problem, we already seen this when we estimated Bernoulli multinomial density. We had similar constraint and we used constraint optimization for it. So, what do we do, we form the Lagrangian so, which means, instead of equating to 0, the partial derivatives of the log likelihood, we take the log likelihood plus a Lagrange multiplier times the constraint lambda 1 plus lambda 2 minus 1.

Take the derivative of this and equate that to 0 so, if you do that let us say, we take derivative of this with respect to lambda 1. So, the second term only gives me eta, the first term I wanted to differentiate with respect to lambda 1 that is, fairly straight forward, we will get this. Let us look at l, if I differentiate with respect to lambda 1, I get 1 by this into d by d lambda 1 of this, will give me phi x i given theta 1 right.

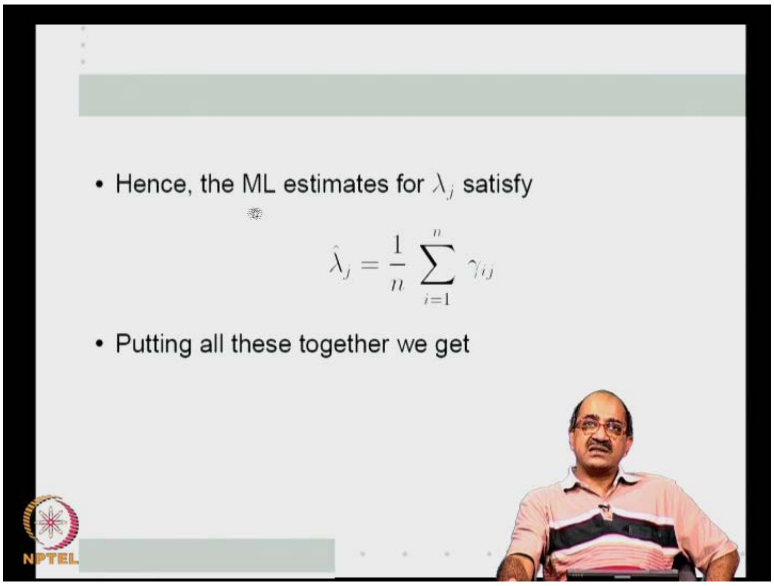
(Refer Slide Time: 21:57)



So, this is phi x i, given theta j by lambda 1 phi x i given theta 1 plus lambda 2 phi x i, given theta 2 plus eta equal to 0. Since we already defined the gamma's, this term is nothing but, gamma i 1 because, this is phi x i given theta 1. So, if there is a lambda 1, here this term would have been gamma because, there is no lambda 1, we will write it as gamma i 1 by lambda 1 plus eta equal to 0.

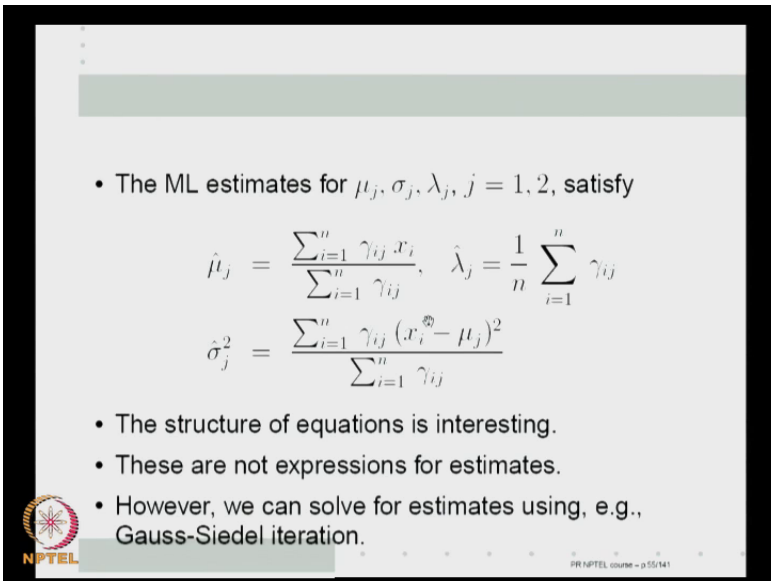
That means, summation gamma i 1 is equal to minus eta lambda 1 so, we get a similar equation for lambda 2 meaning, summation gamma i 2 will be equal to minus eta lambda 2. Now, using lambda 1 plus lambda 2 is equal to 1, we get eta is equal to eta times lambda 1 plus lambda 2. Now, eta lambda 1 is summation gamma i 1, eta lambda 2 is summation gamma i 2 so, that is what, we get and by definition, gamma i 1 plus gamma i 2 is equal to 1 so, this gives me n so, if I substitute that in this equation, I get my lambda 1 estimate.

(Refer Slide Time: 23:04)



My lambda 1 estimate is 1 by n summation gamma i j but, both for lambda 1 lambda 2 I written one equation for lambda j right so, we now obtained estimates for mu's mu j's sigma j's and lambda j's.

(Refer Slide Time: 23:23)



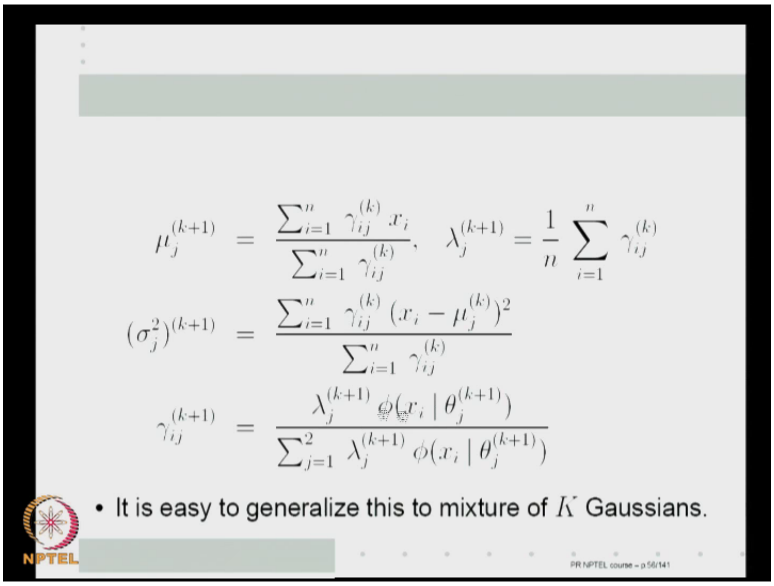
So, we can put all of them together into like this, mu j hat is summation gamma i j x i by summation gamma i j, sigma j square hat is summation gamma i j x i minus mu j whole square by summation gamma i j and lambda j is 1 by n summation gamma i j. As I said of course, these are not estimates because, both sides involve the unknown mu's and sigma's. But, we know that, the final estimates for mu sigma and lambda's have to satisfy this set of simultaneous non-linear equations.

This first note that, as I said this structure is very interesting see, lambda j you are asking what what is the chance that, a given x i, a random x i comes from the j th component density, for for each i, if x i is responsibility to j is gamma i j then, gamma i is summation gamma i j by n, should be the chance of a random x i coming from the j th component density right.

Similarly, we have already seen that, these are simple sample mean estimators somehow, this gamma i j, the responsible coefficients seem to give us some weight of, how much of x i should I assign, while estimating things for j th component density. So, this structure is interesting even though, they are not expressions for estimates because, both sides involve.

But then, we know, that the estimates mu mu j sigma j lambda j satisfy these simultaneous equations, given any set of simultaneous non-linear equations, we can solve them iteratively using let us say, Gauss-Siedel iteration. What is Gauss-Siedel iteration do, at each iteration I use the values of the previous iteration, put them in the RHS of the equations and then, the LHS will give me my new equations.

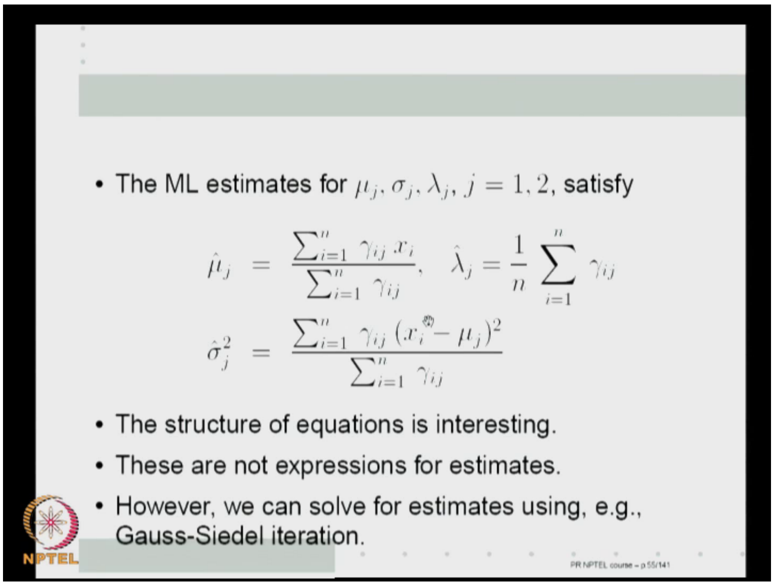
(Refer Slide Time: 25:22)



So, if I did that, I get these iterates so, if I want the new mu j that is, the value of the k plus first iteration, for my estimate for mu j, that will be gamma i j k, x i by summation gamma i j k where, I am using the previous iteration values of gamma i j right. Similarly, for lambda j similarly, for sigma square j and gamma i j k plus 1 is always given the current values of theta j's and lambda j's, I can calculate gamma as a k plus 1.

So, this is an iterative procedure, I start with some initial guesses then, given the initial theta 1 0, theta 2 0 and lambda 1 0 and lambda 2 0, I can calculate gamma i j 1. Once I know gamma i j 1, I can calculate mu j 1, sigma j 1, lambda j 1 using that, I will calculate gamma i j 2 and so on. So, this is a nice iterative method, which is essentially the Gauss-Siedel iteration for solving the previous set of simultaneous equations right.

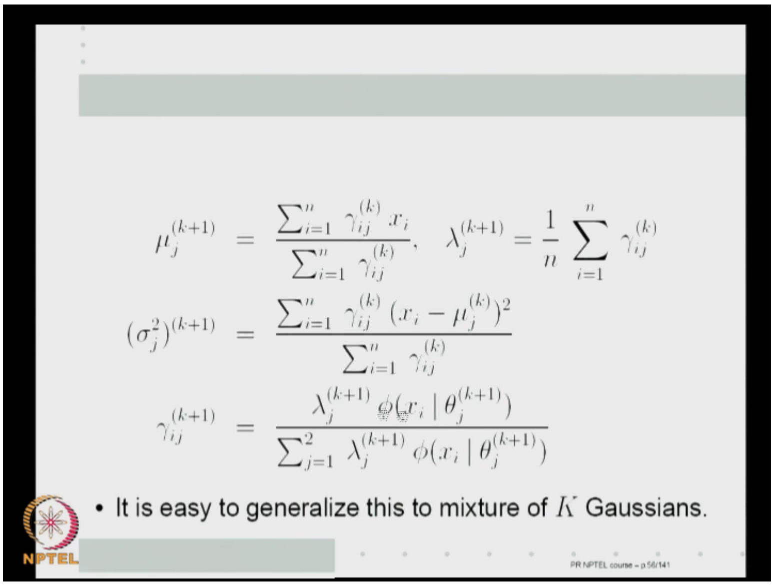
(Refer Slide Time: 23:32)



We know, that the ML estimates satisfy these simultaneous equations, these are only equations because, as I said the gamma i j's on the right hand side themselves involves mu j's sigma j's and lambda j's. So, this is the set of simultaneous equations satisfied by mu j sigma j lambda j. So, we are iteratively solving it, each iteration we use the previous values in the RHS and that value, we assign as the new value for the corresponding variable in the on the LHS, that is the standard Gauss-Siedel iteration.

So, if I did that, get these iterative equations of course, these simultaneous equations have to be well behaved for the Gauss-Siedel iteration to properly converge, what rate it converges, depends on the nature of these equations and so on. These are of course, non-linear equation, they are not linear equations but, generally for most well behaved non-linear equations, the Gauss-Siedel iteration converges.

(Refer Slide Time: 27:33)

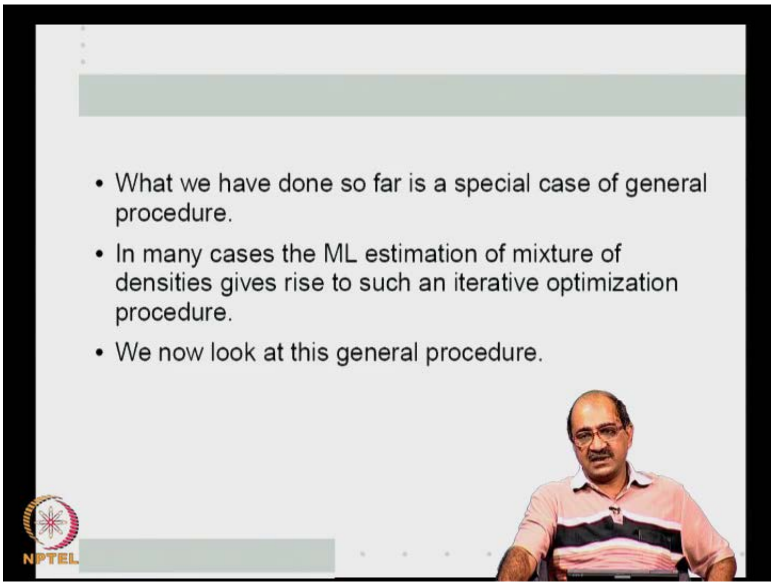


So, we can say that, this iterative procedure is the is the optimization procedure to optimize log likelihood and hence, find the ML estimates of a mixture density, which in this specific example of mixture of two Gaussians. By the way, my component index is j we have taken care to always write sum over j is equal to 1 to 2 like here so that, whatever we have done easily extends to a k component mixture too. Of course, because all our these derivatives have come from Gaussian, these specific set of equations are particular to Gaussian mixtures.

But, even that, they are not specific to only two components, is easily generalises to k components, we use it two only so that, my equations do not overflow out of the slide otherwise, even if you have k component mixtures, we can still do this. So, this we obtained by simply differentiating the log likelihood function and equating to 0. And since we cannot analytically solve it, we used a Gauss-Siedel iteration to solve the resulting simultaneous equation, that the estimates have to satisfy, that is what gave us this iterative procedure.

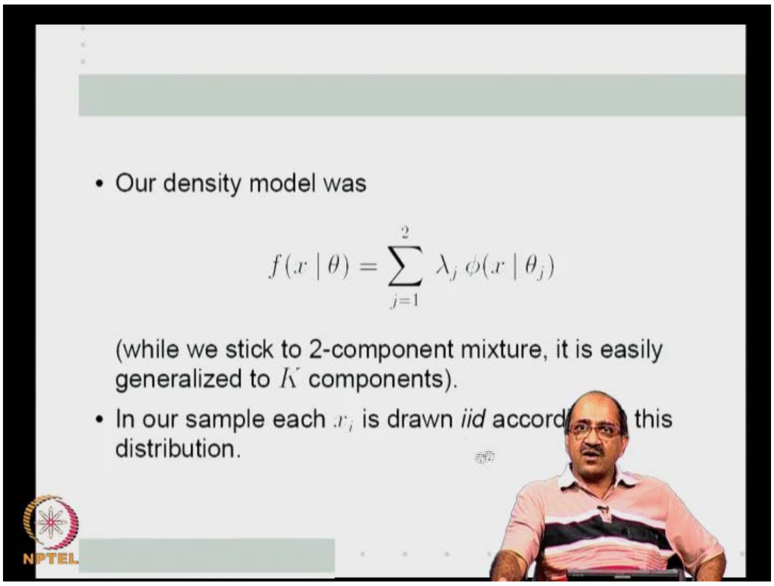
Now, we will of course, there must be something nice about it because, as I said, this is an essentially sample mean estimates right. So, there is some structure here, something interesting happening here so, let us try to take a look at that.

(Refer Slide Time: 29:04)



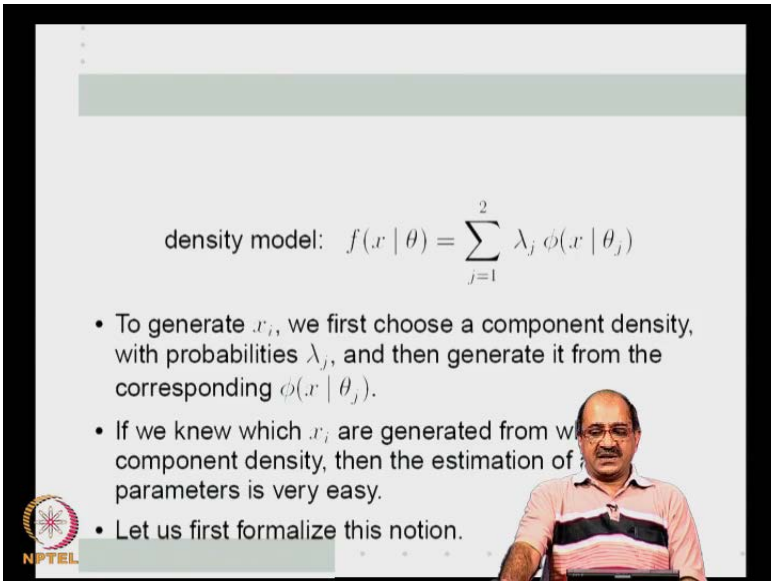
So, as it turns out, what we have done is a special case of a very general procedure, that allows you to do the ML estimation of any k component mixture, not necessarily Gaussian. In many case of ML estimation of mixture on densities, it gives rise to the same kind of iterative optimization procedures and we will now look at this general procedure. As I said in the beginning, this general procedure, is what will lead us to the so called EM algorithm, expectation maximization algorithm.

(Refer Slide Time: 29:40)



So, let us go back, our original density model is f x given theta j is equal to 1 to 2, lambda j phi x given theta j, I still put j is equal to 1 to 2 so, I am sticking to my example of two components. But, as I said, is only because, it is easier to write the expressions as you will see wherever, I put the summation up to 2, if you put summation up to k it is easily generalizes to k component mixtures so, when we have our samples, each excised on iid, according to this density.

(Refer Slide Time: 30:17)



So, I can ask, how would I draw such samples if this is my density model, how would I draw samples from it because, they have different densities, from which you draw. We are assuming that, if phi's are normals, we can always generate random numbers with respect to a particular normal density. So, given any density, there are ways to generate random numbers, how do I generate with respect to mixture density, one way of thinking about it is, what I will do is, to generate each x i, I will first choose a component density.

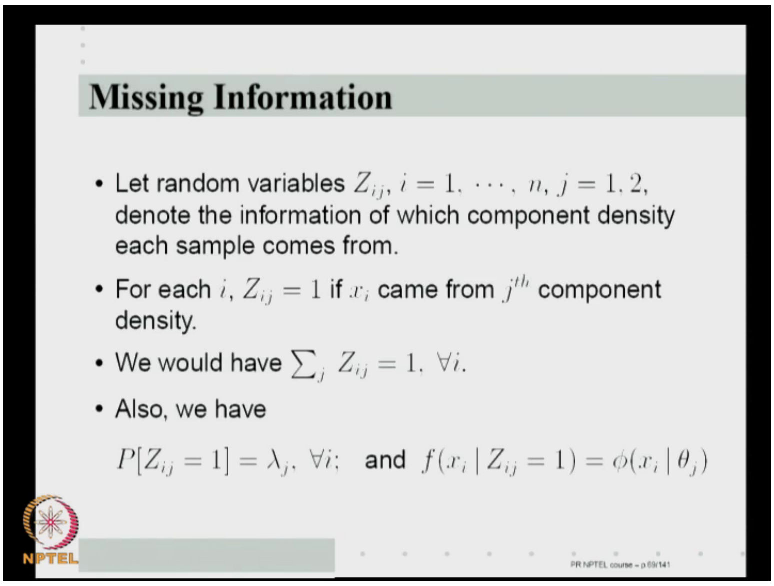
I have to decide, whether the next x i say, x 1, x 2 whatever, the next x comes from phi x given theta 1 or phi x given theta 2, this decision I have to make with probability lambda 1 or lambda 2. That is, with probability of lambda 1, I choose the first density, probability of lambda 2, I choose the second density that is why, lambdas are positive and sum to 1.So, using lambda 1 lambda 2 so on, as the probabilities for choice, I choose a component density and then, I generate it from the component density.

So, each x i, is the finally what I get, is generated either from one of this phi x given theta 1 or phi x given theta 2. When you think, as I do not know, which x i is generated from which density right because, that I am not told probabilistically some x i are generated from phi x given theta 1, some x i are generated phi x given theta 2. And this is this probabilistic or stochastic choice is controlled by lambda 1 and lambda 2 that is how, I am going to generate my sample that is how, iid samples come from such a mixture density estimation.

However, I am only given x 1, x 2, x n so, I do not know, whether x 1 is come from the first component density or second component density. But, just for a moment pretend that, somebody tells us this, if somebody tells us this then, the estimation is absolutely trivial. Then, given my full data, I can separate it out into data of x i, that have come from a phi x given theta 1 now, that is simply estimating a single normal density.

So, I know how to do it so, if I can separate data, as data that come from the first component density, phi x given theta 1 and data that comes from the second component density, phi x given theta 2 then, estimation of the parameters is very simple. Now, this is a very interesting characteristic so, we have a situation here where, essentially I have a mixture density and to actually estimate it, I have to go through a complicated iterative, complicate otherwise, I have go through an iterative procedure. Whereas, if somebody tells me, which x i are come from, which component density, the estimation is trivial so, let us try to formalize this.

(Refer Slide Time: 33:09)



Let us say, we define random variables Z i j, i going from 1 to n and j going from 1 to 2 recall that, n is the number of data samples we have. So, you define random variable Z i j, the idea is Z i j gives me information about, which component density i has come from. So, if I look at Z i 1 and Z i 2 I can tell, whether x i has come from first density or second density.

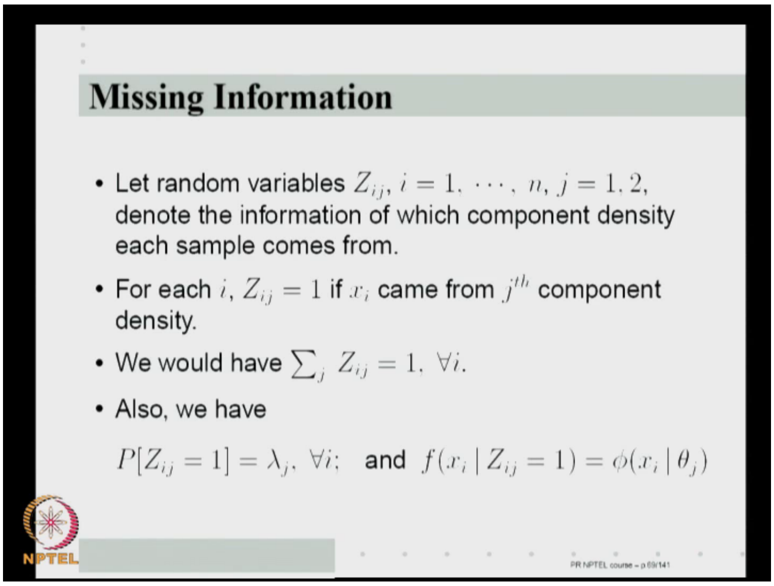
How am I doing this, for each i, I will make Z i j 1 if x i came from j th component that is, if x i came from the first component then, I will make Z i 1 equal to 1 and Z i 2 is equal to 0. And if x i came from the second component density then, I will make Z i 1 equal to 0 and Z i 2 is equal to 1 right. So, each Z i j is binary like that and we have summation j, z i j is equal to 1 of course, it looks needlessly complicated, I have only two numbers Z i 1 and z i 2.

And from way, the way I have defined it because, each x i, if x i has come from the first component is, it is not coming from the second component. So, I might as well use only one number Z i right the reason why I have used two numbers is that, this naturally generalizes to k components, if I had k components then, I would have Z i 1, Z i 2, Z i k, only one of them is 1.

So, this is this is a representation of a binary vector, much like what we saw in the multinomial density estimation. So, even though the two, only there are only two components and I know Z i 1 plus Z i 2 is equal to 1, let us still stick with the notation of

Z i j for our indicators. So, Z i j is 1 if x i came from the j th component, for j is equal to 1 to 2 and Z i 1 plus Z i 2 is equal to 1.

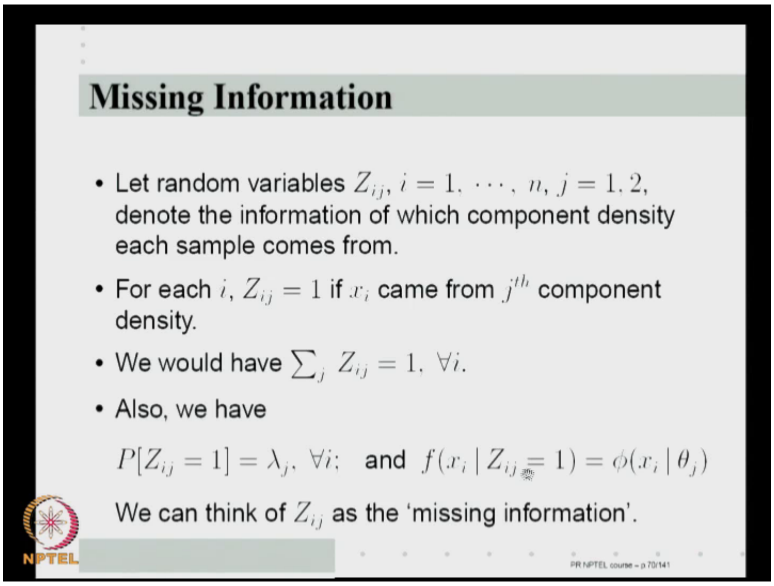
(Refer Slide Time: 33:10)



Now, by definition, when probability Z i j is equal to 1 Z i j is equal to 1, the even z i j is equal to 1 is that, the x i comes from the first from the j th component, component density. Now, because each x i are iid, irrespective of what the remaining samples are doing, for that x i with probability lambda j, I choose the j th component density. So, probably Z i j is equal to 1 is lambda that, this probably Z i 1 is equal to 1 is lambda 1, probably Z i 2 is equal to 1 is lambda 2.

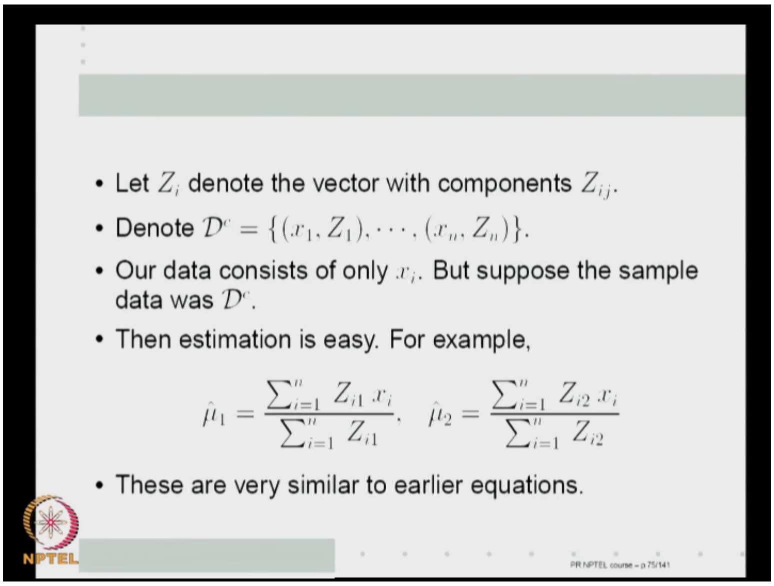
And the second thing, that is interesting here is the density of x i unconditional density of x i, I know what it is here is my density model. But, if you give me Z i j's, conditioned on Z i j is equal to 1, what is the density of x i, Z i is equal to 1 means, x i is coming from the j th component density. So, conditioned on z i, j is equal to 1, the density of x i is simply phi x i given theta j so, conditioned on Z i 1 is equal to 1, density of x i is phi x i given theta 1. And condition on Z i 2 is equal to 1, density of x i is phi x i given theta k that is that is how, this Z i j's can also be equivalently specified.

(Refer Slide Time: 36:12)



We can think of Z i j's as the missing information essentially, the density estimation problem for the mixture densities has become difficult because, I am not given the Z i j information, I am only given x i's. If I am also given the random variable Z i j, for i is equal to 1 to n and j is equal to 1 to 2 here or in general, 1 to k then, as we can see, the the estimation we see so, we can think of Z i j as the missing information.

(Refer Slide Time: 36:43)



Let Z i denote the vector of component z i j in our case, Z i will be Z i 1, Z i 2 and let us say, D superscript c denote the data written as x 1 Z 1, x n Z n, this c means complete.

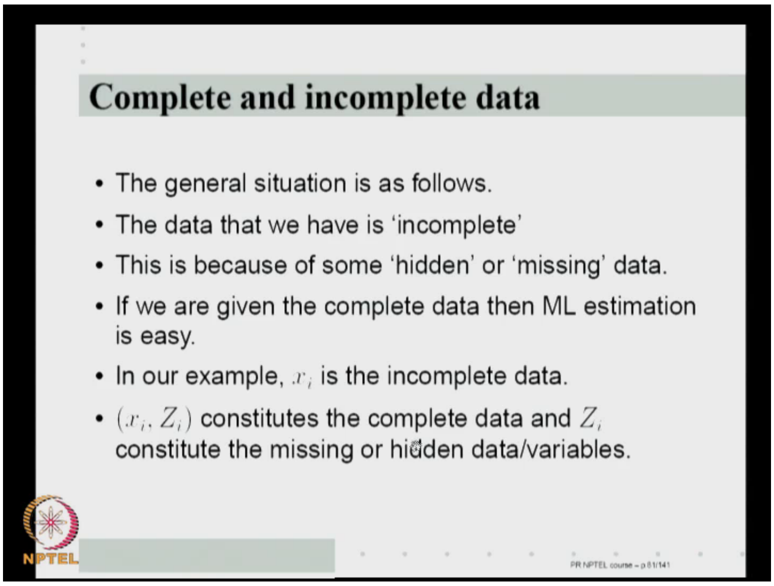
So, I am I am thinking of x 1 x n is the data I am given but, there are some missing components there, if you put in the missing components then, my new data is become D superscript c, c for complete.

We will see complete again later, our data consists of only x i but, suppose for a minute, that the sample data that we are given is actually x 1 Z 1, x n Z n. then obviously, estimation is trivial. What will be mu 1 hat, summation i is equal to 1 to n, Z i 1 x i by summation i equal to 1 to 1, Z i 1. Only if Z i 1 is 1, the x i has come from the first component so, I will pick all the all all my sample that have come from the first component and take their sample mean, that is the mean of the first component, that is the ML estimation for the mean of the first component.

Similarly, mean of the second component right, this is very straight forward by the definition of Z i's so, if the Z i's are given, the estimation is trivial. More importantly, as you can see, these equations are exactly like the equations we saw earlier, except that wherever, I am getting Z i j, there I am getting gamma i j. Because, I did not have Z i j, I am kind of using the ratio of likelihoods as so, to say the probability, that i th one comes from component one or component two that is what, gamma i j's are giving me.

So, these equations tell me that, if the missing information Z i is given to me then, I can trivially finish my estimation of all the component densities. Of course, I have written only for mu's but similarly, I can write for everything sigma's as well as lambda's, if Z i's are known.

(Refer Slide Time: 38:40)

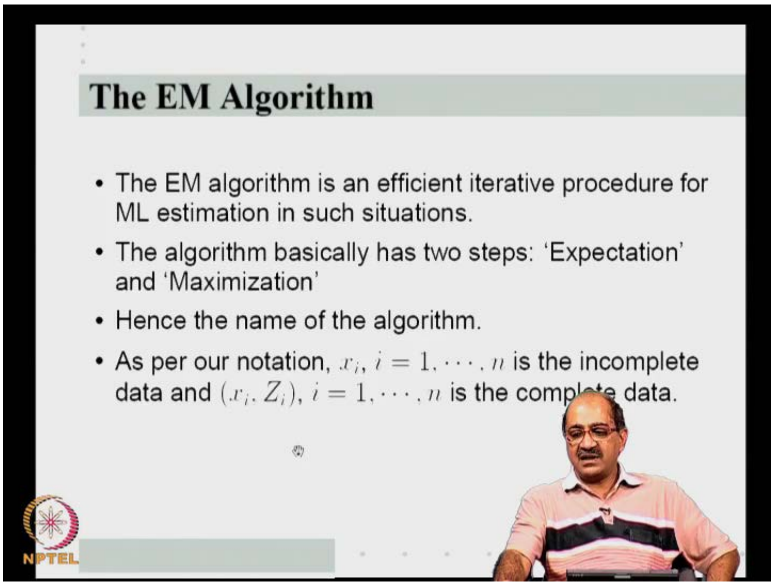


Now, let us put all this intuition together into characterize the general situation that we have, the data that we have is somehow, what we call incomplete right. That that does not have all the information, that we would ideally like to have to make our job of estimation easy. So, the data we have is incomplete, it is incomplete because, there are some hidden or missing data, or hidden or missing variables right.

If we are given the complete data, what we call D c in the previous slide namely, the original incomplete data plus the missing data then, ML estimation is very easy, this is the general situation. In our example, x i's are x i's all the x i's together come constitute the incomplete data and if I have given x i Z i, that constitutes the complete data so, Z i are the missing or hidden data or hidden variables so, that is the general situation.

So, we are given incomplete data, we can hypothesize some missing data, like the Z i j variables here such that, if I have given x i Z i as my actual complete data then, the estimation is simple right, this kind of thing is true for all mixture density estimation right. Now, we have, we are seeking a general method whereby, if I can hypothesize some variables Z i at the missing variables in such a way that, if I am given x i Z i together, my estimation is very simple. But, I am only given x i, what can I do about estimation, that is the question we would like to ask.

(Refer Slide Time: 40:21)

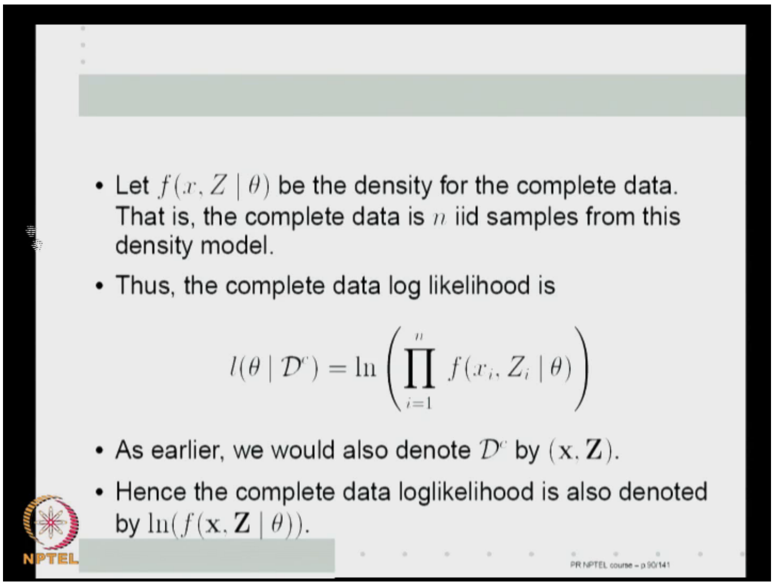


And the algorithm that specifically addresses this question is called the EM algorithm, is an efficient iterative procedure for maximum likelihood estimation, in such situations. The algorithm basically has two steps, so called expectation step and maximization step and that is the reason, for the name EM, expectation maximization, is called expectation maximization algorithm or EM algorithm.

The notation is that, x i, i is equal to 1 to n is the incomplete data, x i Z i, i is equal to 1 to n is the complete data so, the EM algorithm is like this. So, we have a notion of incomplete data, which is the actual observed data and we have notion of complete data by hypothesizing some variables, some extra random variables Z i, we define what is called complete data.

So, an EM algorithm starts with deciding what is my complete data so obviously, the incomplete data is whatever data I am given, my complete data consists of incomplete data plus the missing variables. So, I design the missing variables to in such a way that, if the complete data is given then, my estimation become simple.

(Refer Slide Time: 41:38)



Say, let f x Z given theta be the density for the complete data, what do I mean, that the complete data that is, x i Z i, i is equal to 1 to n or n iid samples from this density. So, the way we hypothesize z, we should because, we already know the density for a f x given theta and we are hypothesizing z so, we should be able calculate the density model for the complete data. Once we get the density model of a complete data, this is the density model for the, the the log likelihood for the complete data. i is equal to, product i is equal to 1 to n of x i Z i given theta, is the likelihood for the a complete data and putting a log, I get a log likelihood for the complete data.

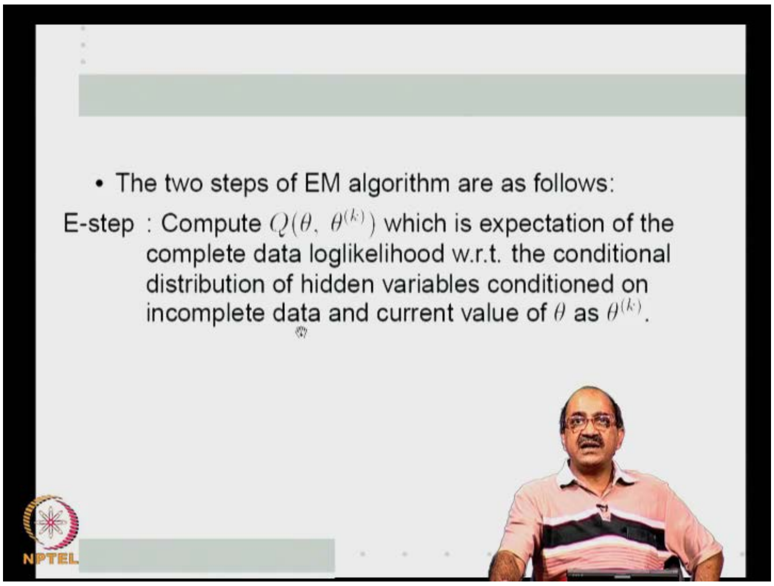
Recall that earlier, we have been, we said that the data x 1 to x n can also be represented by a bold X similarly, the data D c, which consists of X 1 Z 1, X 2 Z 2, X n Z n this is a complete data, will represented by bold X comma Z. So, in the in that sense, the complete data log likelihood can also be written as log of this product will now become f of bold X bold Z given theta.

So, we will use this notation also wherever, mostly we will use this notation for the complete data log likelihood and similarly, all the other log likelihoods. So, the the EM algorithm set up is, we have incomplete data x, x i, x 1, x 2, x n and we have the model f x given theta, that is the density model. We hypothesize some missing variables Z then, we compute we the complete data density model affects given Z f X Z given theta.

So, now, we have the complete data model, that the joint density of X and Z because, Z is what we hypothesize, f x given theta is the given density model. So, we have now, the density for X, the joint density for X and Z and of course, all the other conditional densities we can compute from there. For example, you can compute given this, we can compute the conditional density of Z x given theta where, we are essentially hypothesizing the missing variables Z.

So, EM algorithm says that, if I have the notion of the complete data and the incomplete data, how do I actually estimate given that, I can only use the incomplete data, I do not know the values for Z.

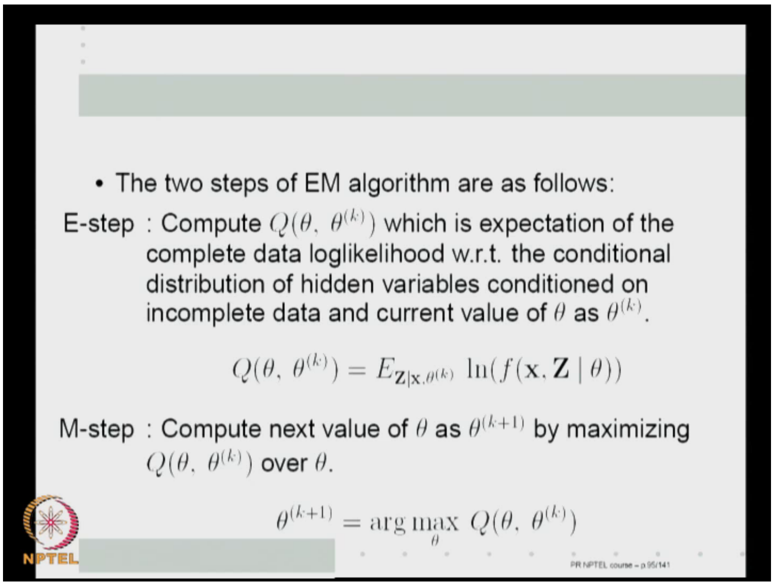
(Refer Slide Time: 44:15)



So, this is how the EM algorithm works, there are two steps of the EM algorithm, the E-step and the M-step, E for expectation M for so, it is a iterative algorithm. So, at a given step, I I am in the k th iteration let us say, my current values for the estimated parameters are represented by theta superscript k, this is the current estimated values. Given the current estimated values theta k, I compute a function of theta, which I call Q theta, theta k because, while it is a function of theta, it also depends on the current estimated value.

So, I call the function Q theta, theta k what is this function, it is the expectation of the complete data or log likelihood, as you take expectation of this quantity ln f X Z given theta right. It is the expectation of the complete data log likelihood with respect to the conditional distribution of hidden variables, conditioned on the incomplete data and current value of theta is theta k.

(Refer Slide Time: 45:16)



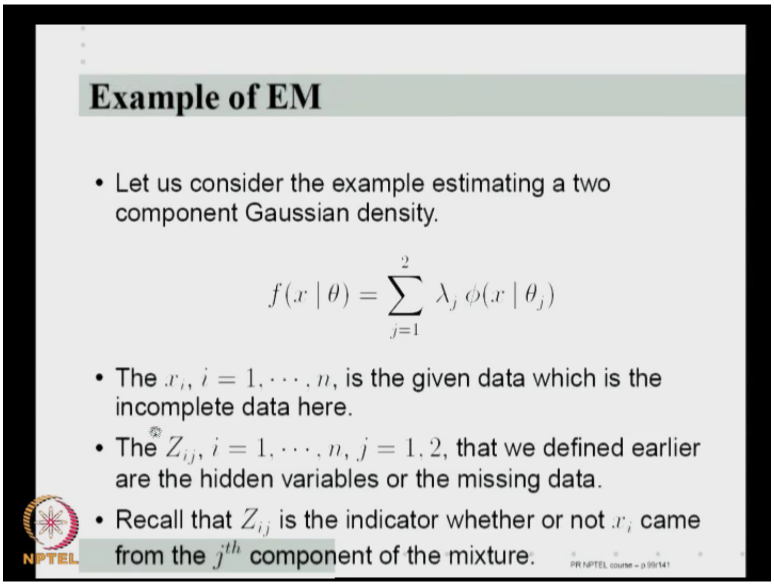
In symbols, what it means is, Q theta, theta k is the expectation with respect to only Z but, using the distribution of Z, conditioned on X and theta k of the complete data log likelihood. So, in this complete data log likelihood, all the terms involving Z are average wrote by the expectation and to average wrote, I use the conditional distribution of Z, given X and theta k. So, only in averaging over the Z terms here, I use this conditional distribution, the rest of the terms depend on this theta.

So, this this conditional expectation being ultimately by a function of both theta k and theta that is why, it is called Q theta, theta k. So, to say it again, Q theta, theta k is the conditional expectation of the complete data log likelihood, with respect to the conditional distribution of the missing variables or hidden variables, conditioned on the incomplete data and the this iteration values of the parameters. Once we got this, the M-step computes the next value of theta that is, theta k plus 1 by maximizing this function over theta.

So, theta k plus 1 is computed as maximizer of Q theta, theta k over theta so, this is a argument of the max over theta of Q theta, theta k. So, I take this function as a function of theta, find which value of theta maximizes this function and that is, given at the next iteration theta k plus 1. These are that two steps E-step and M-step, it is a little complicated because of, this funny conditional expectation.

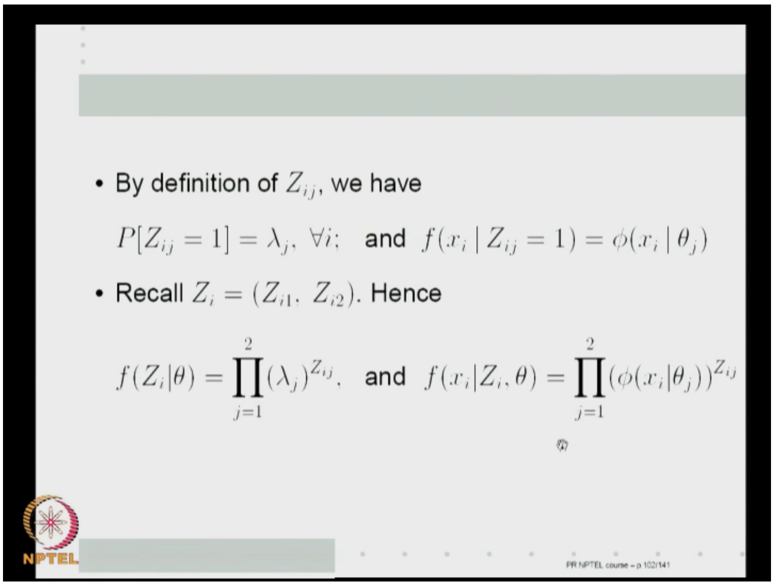
We will we will look at the example once again, we will go back to our two component Gaussian mixture example and compute these two steps so that, we understand how to compute these two steps.

(Refer Slide Time: 47:28)



So, let us consider the example of estimating two component Gaussian densities, this is our given data model this is our incomplete data generator. So, x i is the given data, which is the incomplete data, the Z i j's that we have defined earlier are the hidden variables or the missing variables, and x i Z i constitute the full data. Now, recall that, Z i j is an indicator of, which of these j component densities, that x i comes from.

(Refer Slide Time: 48:02)

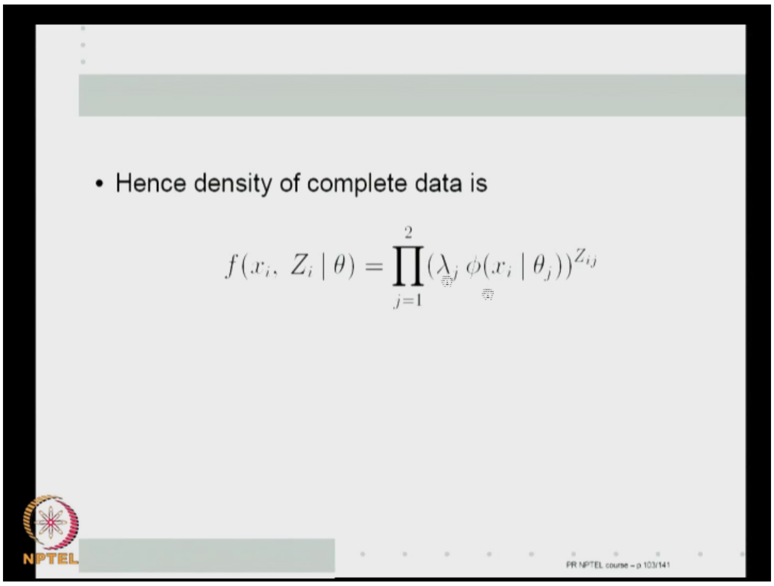


So, as remember, by definition Z i j, we have probability Z i j equal to 1, is lambda j that is, x i coming from j th component has probability lambda j and conditioned on Z i j is equal to 1, x i is the density, is the j th component density. Given these two, we can now write the marginal density of Z as follows, remember that, my complete data is x i Z i, Z i has two components Z i 1, Z i 2.

If you want to write marginal density then, f Z i given theta is pi, j is equal to 1 to 2 lambda j, Z i j this is, if I take away the product, it will be lambda 1 Z i 1 into lambda 2 Z i 2. Only one of Z i 1 and Z i 2 is 0, the other is 0 so, if Z i is 1 0, this becomes lambda 1, Z i is equal to 0 1, this becomes lambda 2 that is, exactly this. So, I can think of this just like in the Bernoulli model, I can take the marginal of Z like this and conditional of x, conditioned on Z and theta, once again x i conditioned at Z i theta is product phi x i given theta j to the power Z i j.

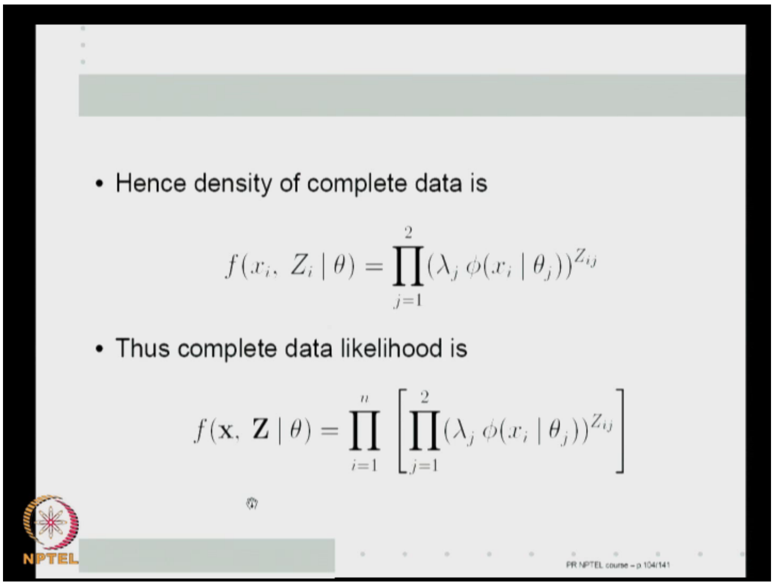
Once again Z i Z i j's are 0 on 1, for each i only one j is 1 so, whichever 1 is 1, I get that particular component density. So, I have got conditioned conditional density of x given Z's and has the marginal of Z's.

(Refer Slide Time: 49:34)



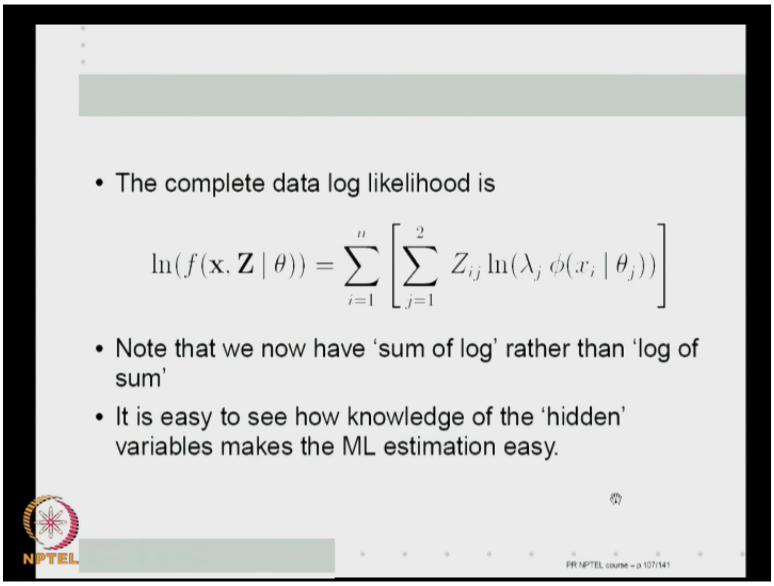
So, by multiplying, I get the joint right so, the joint density model is, I have to multiply these two, both the product. So, the lambda j also comes inside this so, that this is the density model for the complete data. So, this is how, I compute the density model complete data because, given the density model for the incomplete data. And because, I am I am hypothesizing, which are the missing variables that is how, I compute the density model for the complete data.

(Refer Slide Time: 50:05)



So, the the complete data likelihood will be, this this symbol we are using for the complete data likelihood f bold X bold Z, is product over i is equal to 1 to n of f X i Z i given theta, which is product over j is equal to 1 to 2 of this. So, this is my complete data likelihood.

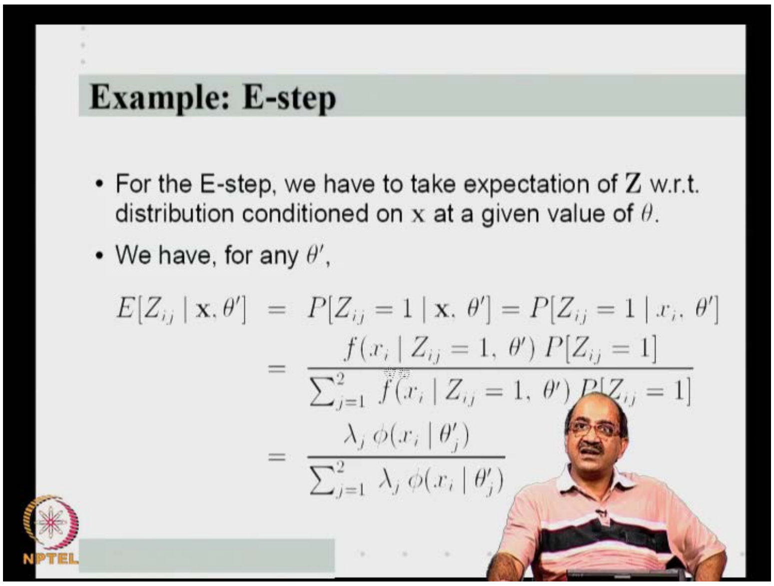
(Refer Slide Time: 50:26)



If I want complete data log likelihood, that is the complete data log likelihood now, before we will go to our E and M steps, let us look at this expression now, we have got rid of the sum of log. Instead of sum of log, we have got log of sum see now, log directly applies to phi's so because, phi is exponential I can now, simplify this. The moment I hypothesize Z i, my complete data log likelihood is the old structure of sum of log rather than, log of sum, that I was getting, if I used only the mixture density.

So, this is actually how, the knowledge of Z i j has made the made the estimation simple because, the complete data log likelihood is the usual form of sum of log, and the way we have done n of examples of, how to maximize such things all right.

(Refer Slide Time: 51:29)

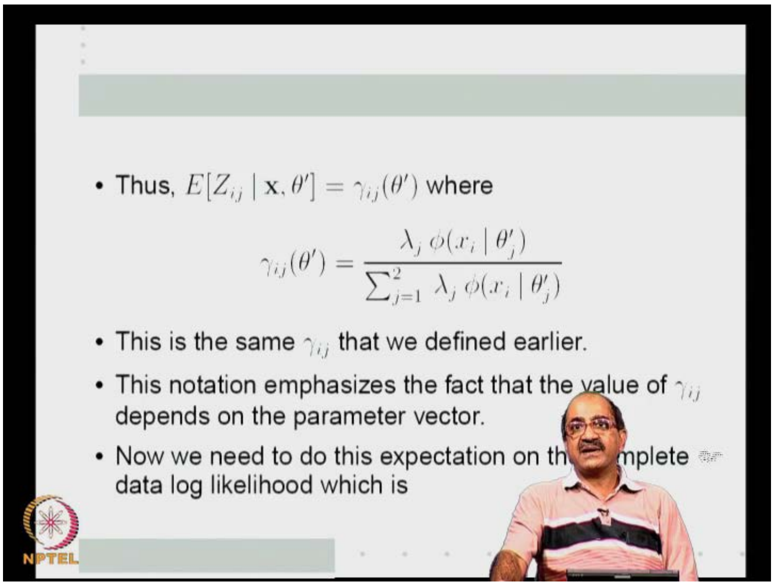


So, now, let us do the E-step, in the E-step we have to take expectation of Z with respect to the distribution conditioned on X, at a given value of theta. So, what is the expected value of Z i j conditioned on x and sum theta prime, Z i j is a binary valued random variable so, it is expectation will be probability it takes value 1. So, this is same as probability Z i is equal to 1, conditioned on x and theta now, X's are X 1, X 2, X n they are iid and hence, Z i j only depends on x i, it does not depend on others.

So, this is probably Z i j is equal to 1, conditioned on x i and theta prime now, this expression now, i can do what, using Bayes theorem. So, using Bayes theorem I can write this as, f of x i given Z i is equal to 1, probability Z i is equal to 1 and summation over j of the same thing. f of x i given Z i is equal to 1 and at a theta primes is nothing but, phi x i given because, Z i j is equal to 1, theta j prime and this is lambda j right.

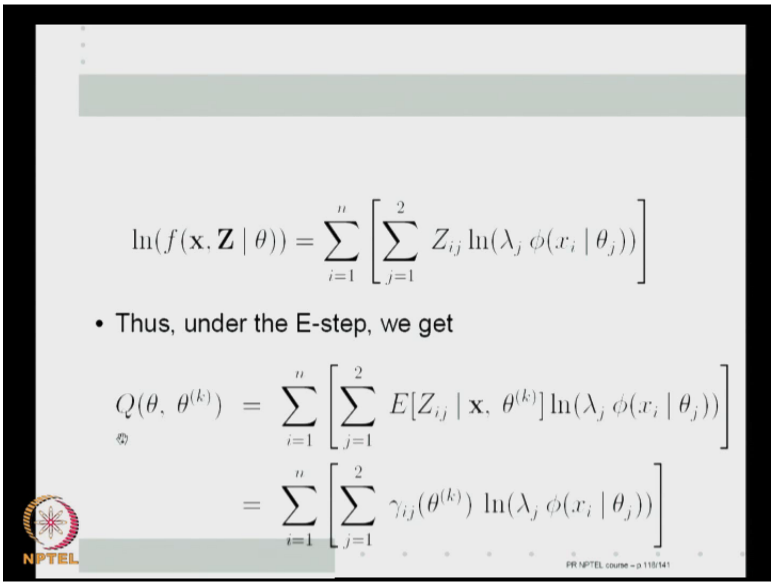
So, what does this gives me, lambda j from this term and this will give me phi x i given theta j prime and the denominator is sum of the same thing over j. This expression is very familiar right, this is exactly the gammas, that we defined earlier.

(Refer Slide Time: 52:53)



So, Z i j conditioned on X and theta prime is gamma i j of theta prime now, I will write it as a function of theta prime where, gamma i j is a theta prime, is lambda j phi x j given theta j prime by summation over j of the same thing right. First notice that, this is the same gamma i j's that, we defined earlier only thing is, earlier also a saying gamma i j's are functions of the parameters, this time this our notation makes it absolutely clear that, they are functions of the theta's right.

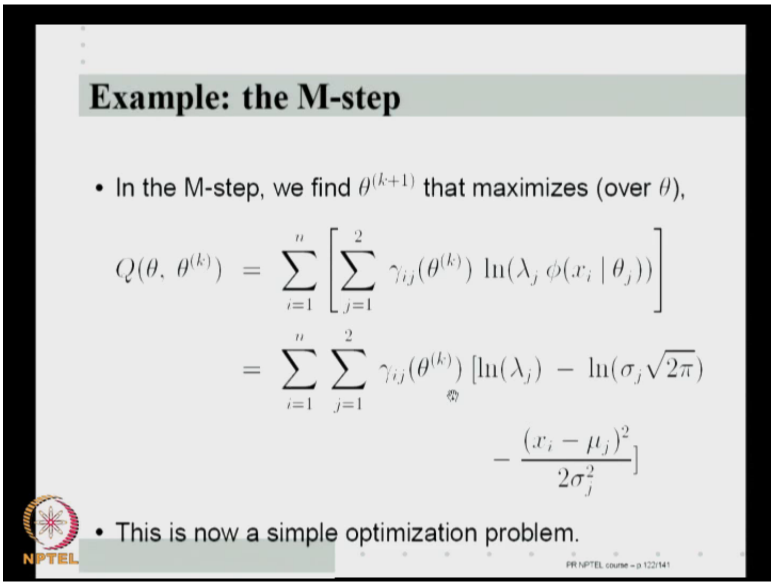
(Refer Slide Time: 53:33)



Now, we need to do this expectation and the complete data log likelihood right, which is the complete data log likelihood, this. So, to take expectation of this, expectation only of Z's so, only this term will get into the expectation. So, in the Q step what will i get, if I take expectation of log likelihood that is, Q theta, theta k is, i is equal to 1 to n expectation goes in, j is equal to 1 to 2 expectation goes in, expectation of Z i j given X gamma theta k, rest of it is not function of Z i j.

So, that comes out of the expectation all right this is the, as you can see this is the conditional expectation of the log likelihood of the complete data, conditioned on X and theta k. That is how see, this term is dependent on theta k, what comes out of that expectation, still depends on whole theta j's, the the the the variable theta j that is why, this is a function of Q theta, theta k all right. So now, this I have already calculated I can substitute for it, if I substitute for it, I get this so, my E-step is complete, I have computed my Q theta, theta k.

(Refer Slide Time: 54:39)

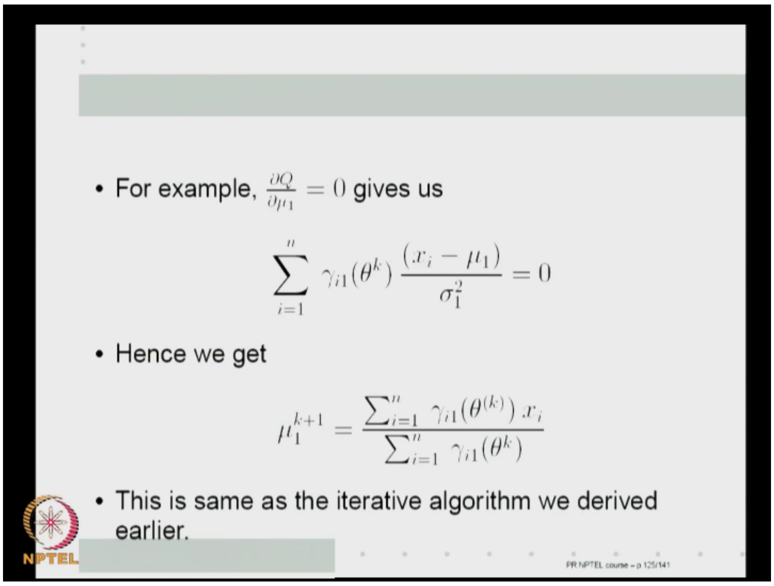


In the M-step, we have to find theta k plus 1, that maximizes Q theta, theta k over theta, this is my Q theta, theta k, as you can see this gamma i j's have now come in, at the expectation of those Z i j's conditioned on x and theta k. Now, if I substitute for my Gaussian, this will be ln lambda j minus, by the Gaussian this will be ln sigma j root 2 pi x i minus mu j by sigma j square.

So, this is a very straight forward maximization note that, gamma i j's are only a function of theta k, they are not function of theta. When I am giving maximization with respect to theta, these are constant right so, if I want to maximize this subset to mu, only this term will contribute a derivative with respect to mu. So, if I did that, this is a very straight forward optimization problem simply because, these are no longer a functions of theta, this is like doing a single Gaussian estimation and we get that right.

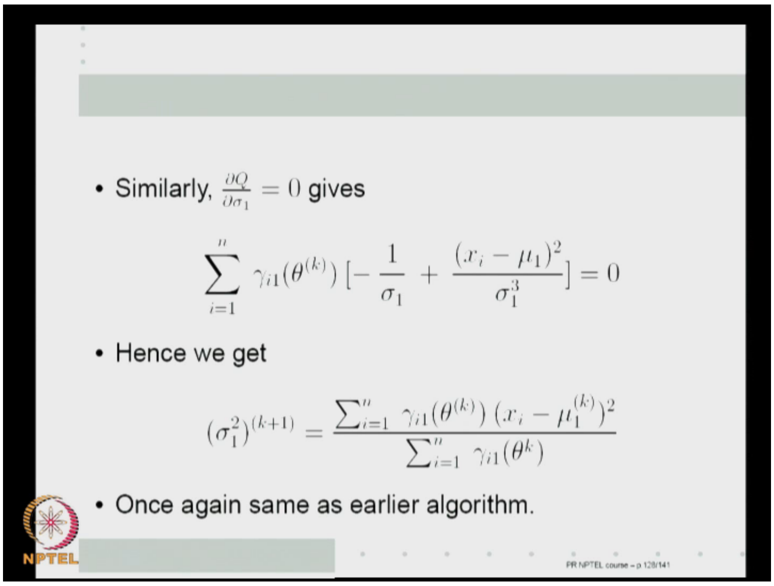
Because, if I differentiate with respect to theta right I have say, with respect to mu, this 2 will come, this 2 will cancel, x i minus mu j by sigma j square and this gamma i j theta k, as coefficient all right.

(Refer Slide Time: 55:54)



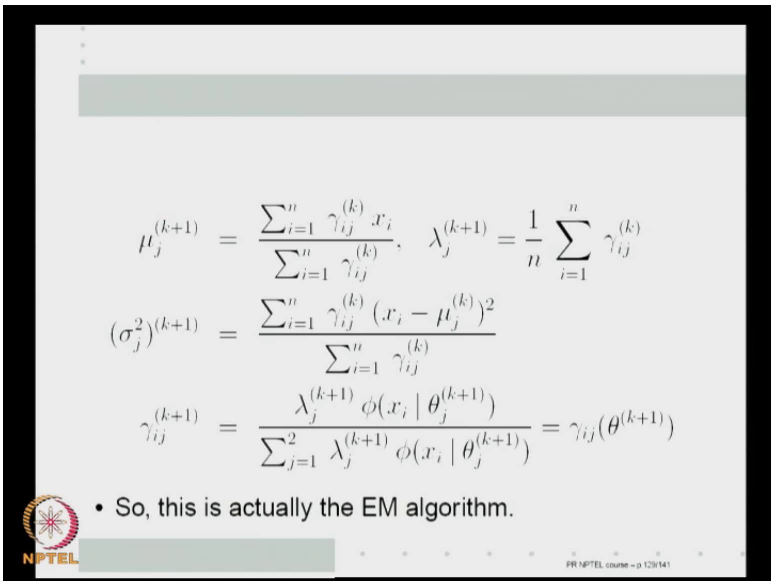
So, that will give me mu 1 k plus 1 is this, is exactly like the expression we have got earlier right so, the earlier expressions we got are, what the EM algorithm gives us right.

(Refer Slide Time: 56:10)



Similarly, if I had done with respect to sigma 1, I would have got the sigma 1 derivative and once again, I would have got the same expression as earlier.

(Refer Slide Time: 56:26)



So, now, I can go back and rewrite those equations right now, I will just wrote gamma i j k plus 1 as gamma i j theta k plus 1 right. This is the same slide, that we had earlier except the last line, this tells you, that the earlier iterative algorithm we got, is actually the EM algorithm for this particular example. So, I will stop here for this lecture, we completed this example so, next class, we will just briefly review this again and ask the in general, what the EM algorithm is about. And then, briefly see, why the EM algorithm should converge and why, it gives us very nice way of doing mixture density estimation.

Thank you.