NPTEL

NPTEL ONLINE CERTIFICATION COURSE

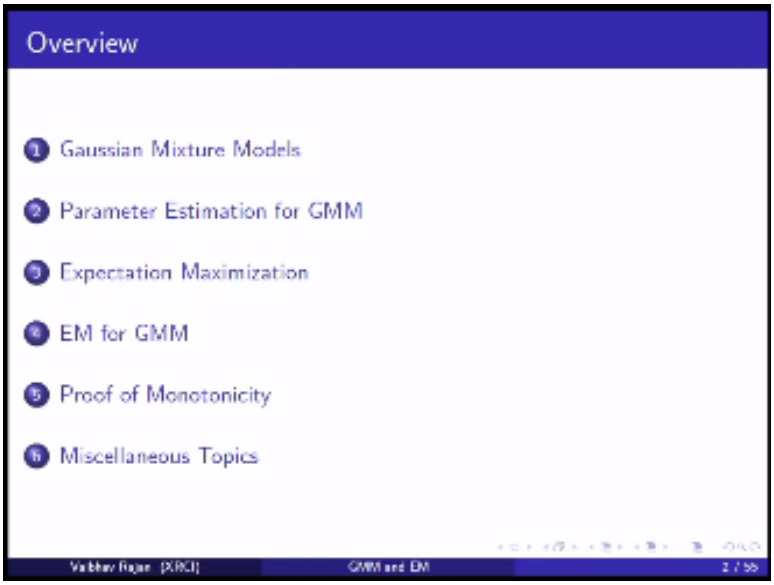
Introduction to Machine Learning

Lecture-76
Gaussian Mixture Models

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

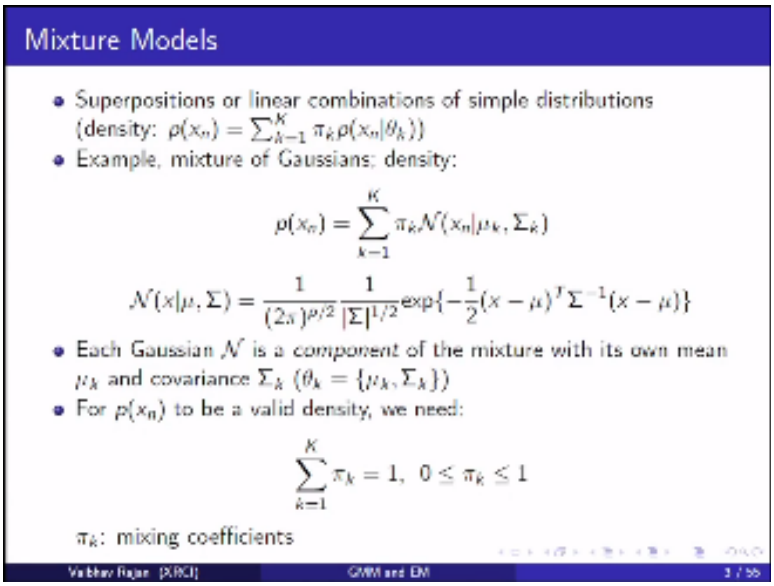
So I will be talking about Gaussian mixture models and the expectation-maximization algorithm.

(Refer Slide Time: 00:21)



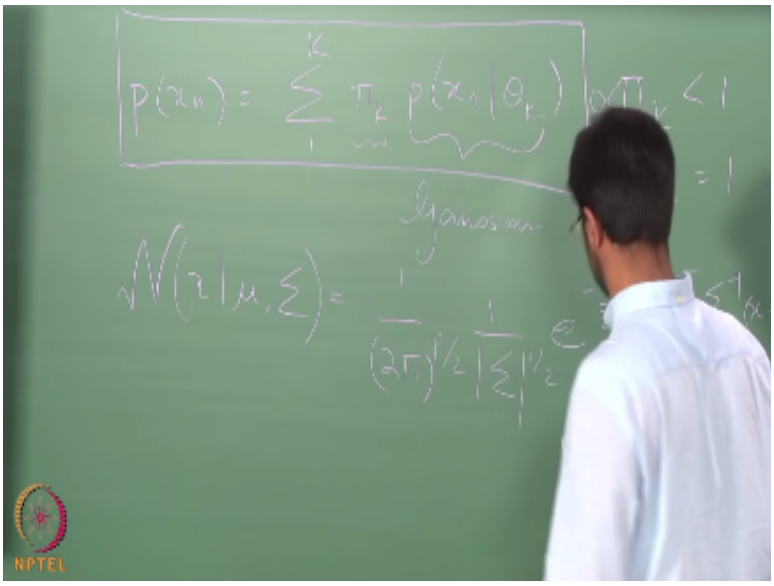
So the plan is to start with introducing Gaussian mixture models and then talk about mainly how we estimate parameters for a Gaussian mixture model and then through that introduced what expectation-maximization is because that is the iterative algorithmic framework that will be using for parameter estimation that is in general and then we will come back to Gaussian mixture models and see how EM can be used for Gaussian mixture models and then talk a little bit about why EM is a nice way to, what talk a little bit about theoretical properties of EM and why it is interesting.

(Refer Slide Time: 01:03)



So mixture models as the name suggests they are a mixture of different they are a mixture of models like formally they are linear, combinations of simpler not necessary not always simpler but linear combinations of distributions. So they typically have a form like this these so the density of a mixture model is a linear combination of other densities p and different mixture models will have different forms for the probability distribution here, right. π_k so let me write this down, so because we will see this quite often.

(Refer Slide Time: 01:50)

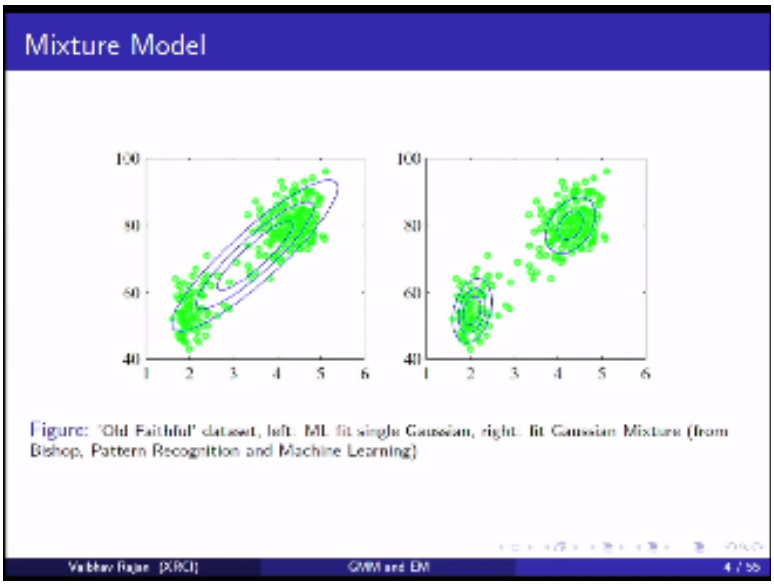


So the density is given by a linear combination of different probability densities and here we have k components and each of these components have a mixer wait so this π_k is called a mixture weight or a mixing coefficient, so this probability here we can it can assume different parametric forms the most common is the Gaussian, so when this takes when this follows a Gaussian this is called a Gaussian mixture model, right.

And this is one of the most commonly used mixture model in a lot of different domains in bioinformatics in speech processing you will see this everywhere. One of the reasons why it is used is because it is mathematically tractable but there are other nice properties too. So you know I guess you all know what a Gaussian is but let me write it down anyway so that because we will use this very often in the coming few slides, right so this is the form of a Gaussian I am assuming you all know this but let us keep it.

So each Gaussian each component here is a Gaussian and each of these Gaussians have its own parameters, the mean parameter and the covariance parameter right, and for this to be a valid density we need the π case to be between 0 and 1 and also the sum of all the π case to be exactly equal to 1 we can show this mathematically, okay.

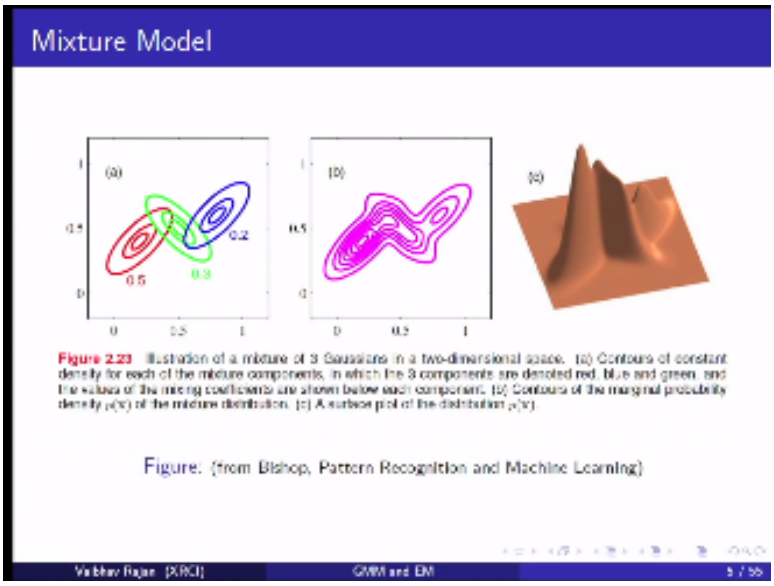
(Refer Slide Time: 04:20)



So why do we need these mixtures, why do we need these super positions of densities, so here is an example from Bishop's book so very, very well known data set old faithful data set on the left so we have plotted the data set in green points it is a two dimensional data set and when we try to fit a Gaussian this is what you get right, when you try to fit a Gaussian here. Now visually it clearly looks not okay, because the Gaussian he is most dense around the mean but when you see the data it does not look like the data is most dense around the mean, right.

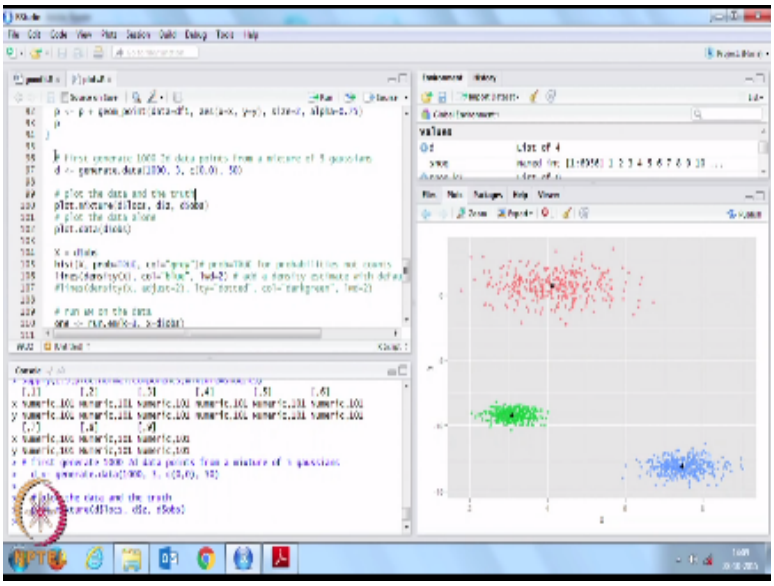
But if we use instead of a single Gaussian two different Gaussians and try to fit to a mixture of two Gaussians to this data it looks somewhat okay, the data is dense here and data has dense here and it looks like to, it looks like this a mixture of two different two Gaussians would be a good fit for this data. So let me let me show you some more examples.

(Refer Slide Time: 05:36)



Before I go into this let me show you some more examples right,

(Refer Slide Time: 05:49)



So this is just some our code to generate the sample data from a Gaussian mixture and what I am going to do is just sample and then plot the data right. As you can see every time I sample from a Gaussian mixture it is basically here I am setting the number of components to 3 so every time it is sampling from three different Gaussians and whenever you see this what you typically see is this clustered kind of data, so you see class three clusters.

If I change this to 5 or 6 you will start seeing more clusters here not necessarily well separated like this they may be overlapping as well, you see so whenever you see such clustered kind of structure in the data the first thing that comes to mind is that first thing that comes to mind when you try to model such data is to try to use Gaussian mixtures, because Gaussian mixtures can nicely fit such clustered data.

So this is another figure from the same book illustration of three different Gaussians you see as I was saying these Gaussians need not always be well separated in this case these Gaussians because the mean and variance that are chosen they are overlapping right, and here you know that when somebody is telling you that there are three different Gaussians, but when you look at the data and try to plot the density like the fitted density it looks something like this right, and this is the surface plot of the distribution.

Another thing that you should observe here is that this was generated from a three-component Gaussian mixture with weights 0.5, 0.3 and 0.2 right, which means that this the red Gaussian here is contributing most of the mass and that is again observed here when you plot the density

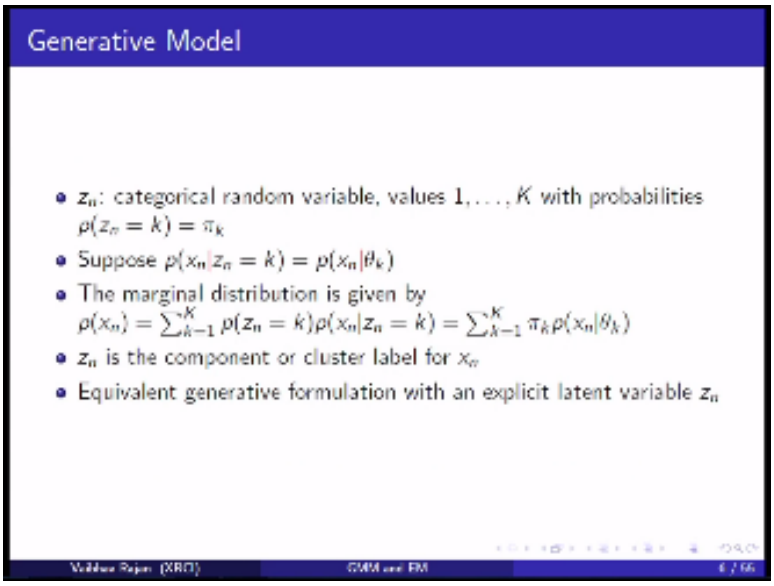
so this mean here is, the high the probability here is the highest and then lower for a bit lower for the green Gaussian and the lowest for the blue Gaussian, right.

Let me show you some more examples of these densities, so again I generated from six different components let me reduce the number of components here, 3 there you see this is an example that I wanted, so here there are three components but the two components are highly overlapping, so if you see how it was generated it was generated by one Gaussian here a second Gaussian here and a third Gaussian here.

But when you look at the data you do not know how it was generated it looks like this, so sometimes it may not be apparent there are just three, there are exactly three different components and when you plot the fitted density you typically see a density curve like this which has multiple modes, right. Now this is for this is a fitted density for a three component Gaussian you see it does not it does not necessarily have exactly three modes so it depends on the samples that you have.

So let us see a few more density plots, you see this is the data that I generated and this is the fitted density which has four modes let us run it again, this is exactly three modes for this data this is very well separated three clusters okay.

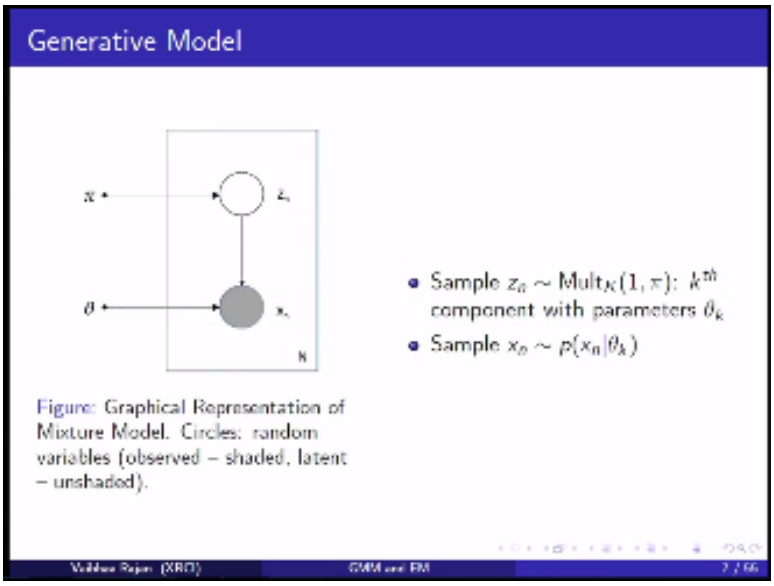
(Refer Slide Time: 10:28)



So you might have already understood this by now at least intuitively we can formulate this as a generative model as well right, and the generative model would be to select a component right. Once you select a component you know you have selected the Gaussian corresponding to that component and you know the parameters of that Gaussian and then you sample data from that Gaussian so that would be a generative model for a Gaussian mixture right, to make it more formal so let us take z_n to be a categorical random variable right, with the probability of z_n=k being exactly equal to this π_k right.

So it is a categorical random variable which takes values from 1 to k, and now suppose that the probability of the data x_n given z_n=k so what is the probability of which is just this the probability of x_n given that you know the parameters for that particular component I express it as probability of x_n given $\theta_{k,}$ θ_k represents the parameters. So the marginal distribution you can express it as a probability of z_n=k you select the component and then the probability of x_n given z_n=k the probability of x_n coming from exactly that component. By what we have assumed the first probability is just k and the second probability is probability of x_n given θ_k so this is an equivalent generative formula with an explicit latent variable z_n.

(Refer Slide Time: 12:24)



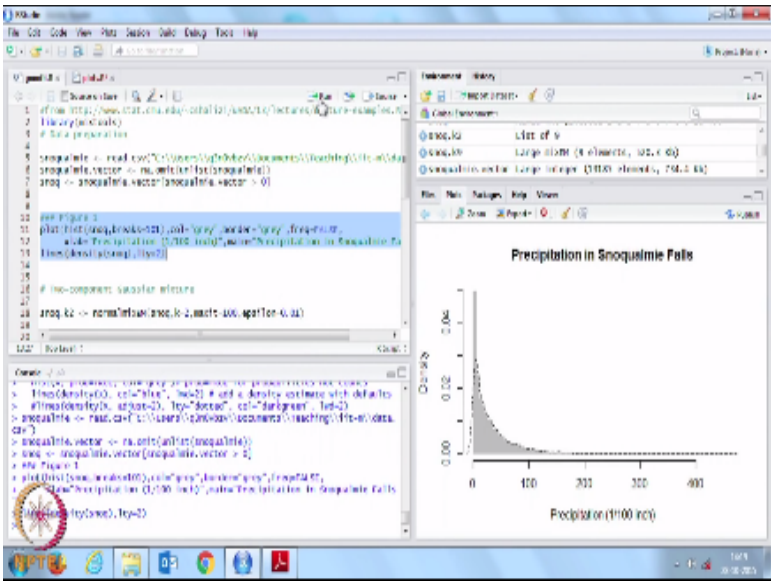
And we can represent this graphically the shaded circle here is the observed data this is the latent variable the latent variable is governed by the parameter π and x_n is generated once you know the latent variable it is generated by known parameters corresponding to that latent variable. So if you have to generate data from such a model you first sample z_n from the categorical distribution which is just so the categorical distribution is a special form of multinomial distribution with parameter 1.

And so that once you samples z_n you get the k^{th} component with parameters θ_k and then you sample x_n from that probability distribution, so in our case this is Gaussian but it need not be Gaussian it could be exponential it could be any complicated probability distribution right, so is it any questions please ask me if there any questions. Could you go to the pervious slide again, could you explain is that?

So the probability of x_n so we are just getting the so we are assuming that there is a joint distribution and the marginal distribution can be written down like this you for each of the k components you take the probability that probability of z^n being equal to k and then the probability of x_n being generated from that component so this is exactly what the graphical model represents right, if you write down the probability distribution represented by this graphical model it will be exactly this except that there is a sigma n term that takes care of all the different data points.

But for a single data point it is exactly this you are choosing a component and then choose choosing the you are sampling the data point given the distribution parameters for that component right, so before I go into that this is just so remember this is another distribution one use of it as I just showed is to model clustered data but you can also model other kinds of data with it. For example, so let us look at another data set I got this from the web.

(Refer Slide Time: 15:25)



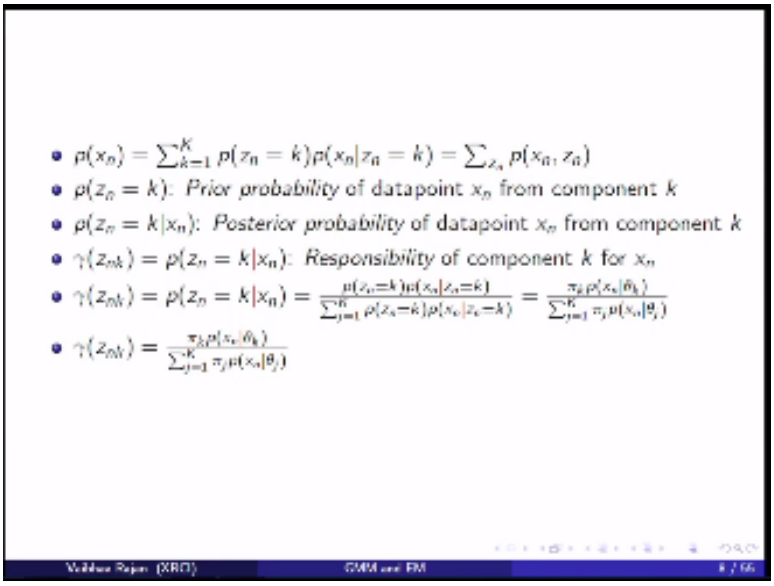
So here is the density curve from the fitted density curve for some data set which is records precipitation in some false I cannot pronounce it also snoqualmie falls okay, and so suppose we want to model this density right, we can we can model this with a Gaussian mixture and let us see how it looks when we do it with the Gaussian mixture with two components right, so we are trying to model this with a Gaussian mixture this is one Gaussian and this is another Gaussian okay, does not look like it is modeling this well, but we can increase the number of components and when we use nine components take some time and there you see it is getting closer and closer to this distribution right. So Gaussian mixture with nine components because each of these Gaussians have different means and covariance parameters is modeling this reasonably well.

So Gaussian mixture by that way is very versatile it can model a lot of different distributions by just choosing the right number of components and choosing the parameters appropriately, so it models not just clustered structure yeah, yeah you do not know that is, so that is the parameter estimation task right, when you want to estimate when you are fitting the model you want to

estimate what the right number of components is and you want to estimate what these parameters are right.

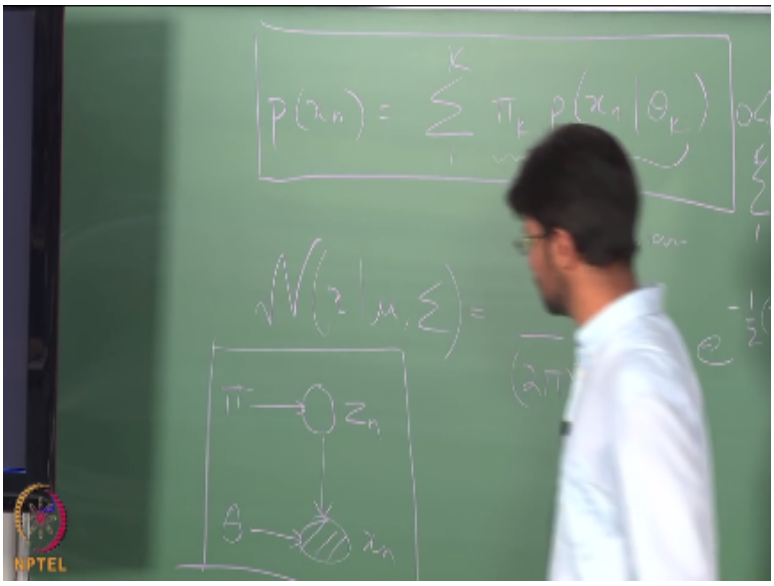
So if you are just given the data you do not know what those parameters are and that is in fact going to be the bulk of the lecture, how are we going to estimate the parameters.

(Refer Slide Time: 17:47)



Alright, so whenever you see this actually I think I am going to keep this model handy here.

(Refer Slide Time: 18:03)

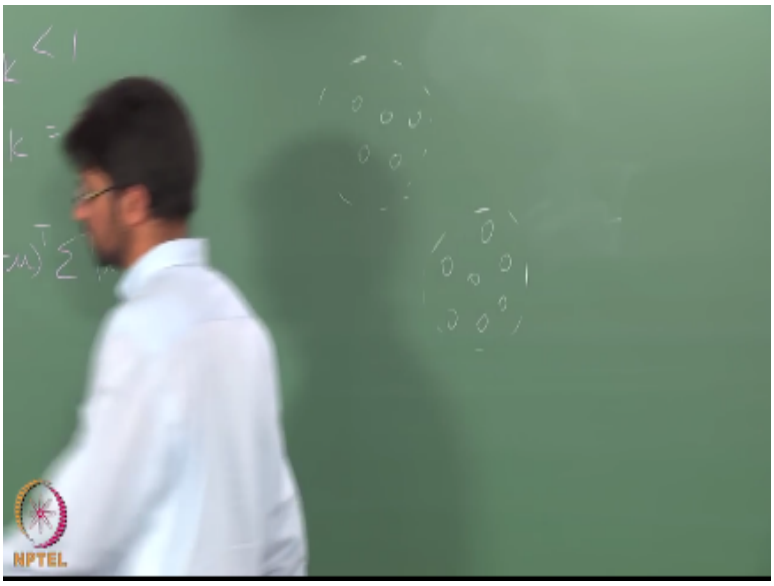


So it would be very good a few sort of keep this model in your mind as we go through the lecture, because all the different all the math that we will see will sort of start making sense when you have this model in your mind right, so when I talk about some formula like this the probability of x_n being equal to σ_k this probability times this probability you can see that it is just the probability of x_n being generated and the generative model is usually more easy to think with, so the generative model would be that I choose the component and then once I choose the component I choose the corresponding parameters and generate this x_n.

So these components if you are doing a clustering task they are just the cluster labels right, so suppose you have three clusters and you want the cluster labels and if you fit a Gaussian mixture model there these components the component values 1,2,3 would just become the cluster labels in that case okay, so this would be a probabilistic way of doing clustering. Alright so the probability of z_n=k equal to k probability of z_n the latent variable or the cluster label taking a particular value is the prior probability of the data point x_n coming from the component k okay.

Now suppose you are given some data set like the data set we saw from snoqualmie falls or something, suppose you are given that data set and then you are asked to find what is the label z_n for the corresponding data point right, so what I mean is suppose you have data points here.

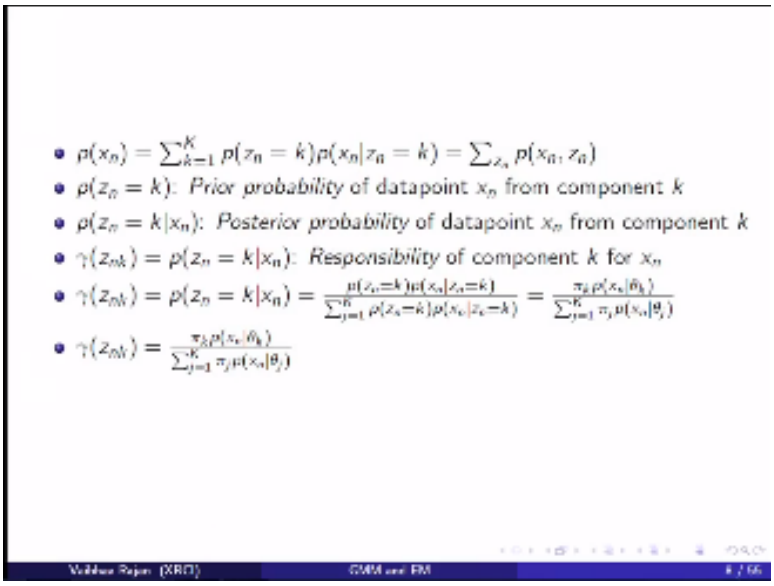
(Refer Slide Time: 20:34)



You have two clusters and if you knew how it was generated then the cluster label for all these points would be say 1 and the cluster labels for all these points would be 2, right but you do not know this you do not know how the data was generated okay, so you once you, once you have given this data then you have to infer what these parameters are and you have to infer given that this given that you are using this model to fit you have to infer what is z_n value is for each of the data points.

So the z_n values for all the data points in one component will have the same value 1 and the z_n values for all the other data points in the other component will have the same the other values and equal to 2 of course in the for clustering it can be interchanged the exact value does not matter, right.

(Refer Slide Time: 21:47)



So this probability the posterior probability of the data point x_n coming from component k is so important that it is given a name of its own is called the responsibility I am going to write that down as well, because we will you use it again and again and again. So important that it is given a name of its own is called the responsibility I am going to write that down as well because will you use it again and again and again oh, he asked me not to use this the posterior.

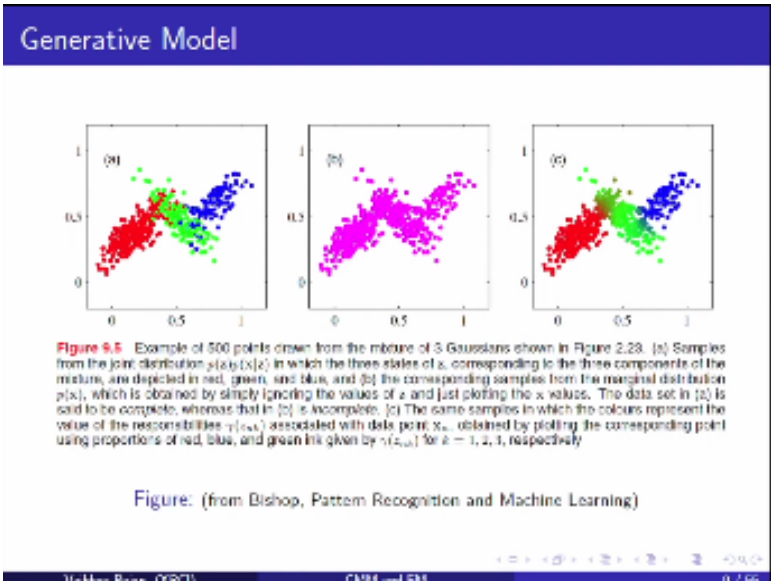
(Refer Slide Time: 22:14)

Probability of zn = k given the data right and it is also called the responsibility it is the responsibility of component k for xn right, so till now the way I have described it is that there are these different components and only one of these components is responsible for giving rise to that data but that is the generative point of view right, but if you look at it probabilistically each of the components is contributing something to that towards the probability of that data point and that contribution the weight of that contribution is given by πk.

So your clustering need not always be a hard clustering of this component versus that component it can be a soft, clustering where each the cluster label can be probabilistic so it can be cluster it can be cluster label 1 with probability 0.5 cluster label to with probability 0.3 and so on right so when you express it as a probability then you get the option of doing both hard clustering as well as soft clustering so now you can use Bayes rule to get this formula for the responsibility.

So this is straightforward and with so and when you substitute for probability of zn = k which is the prior you get πk here and the probability of xn given that you know the component again keep this in mind given that you know the component is the probability of xn given θ k the parameters for that component it so let us write this down okay, so observe that you do not know the responsibility values until you know all the parameters you need to know all the π case and all the θ case, which in the case of Gaussian is all the μ and all the σ case for all the k components and of course you need to know k to know the responsibilities, so let us see another very interesting picture.

(Refer Slide Time: 25:26)

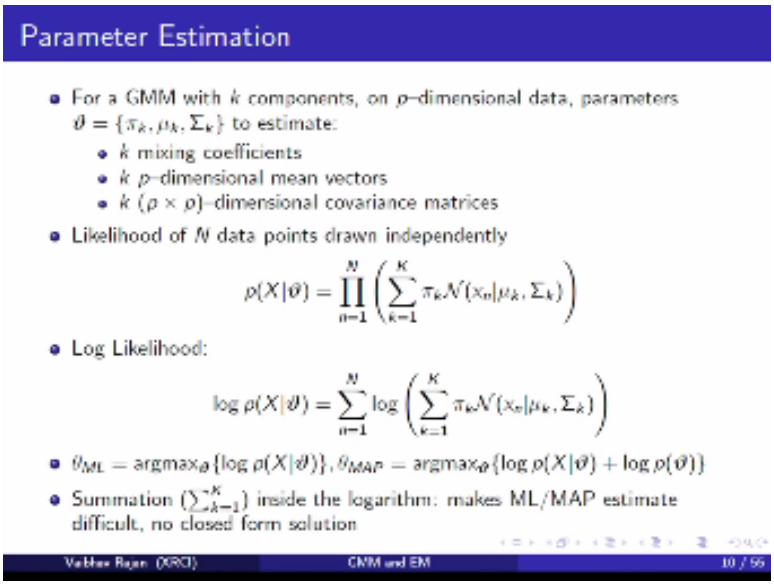


When you are generating the data here he is again from the same book is generating three different Gaussians the red, green and blue Gaussians and of course you do not know how the data was generated when you see the data you see something like this okay and then you try to fit three Gaussians into 3 a mixture model with 3 Gaussians 3 component mixture and what he has done is he has colored the responsibilities right so all these guys have been given the responsibility corresponding to component one.

Or competent red here these so the blue comp blue component is responsible for all these data points green component for this and here you see in the border it is a mixture of green and blue depending on what probability values area sign, right so this is one example of soft clustering these values these data points are not assigned completely to either green class to the green cluster or to the blue cluster and you can see there are mistakes in influence because of course if you do something like a maximum.

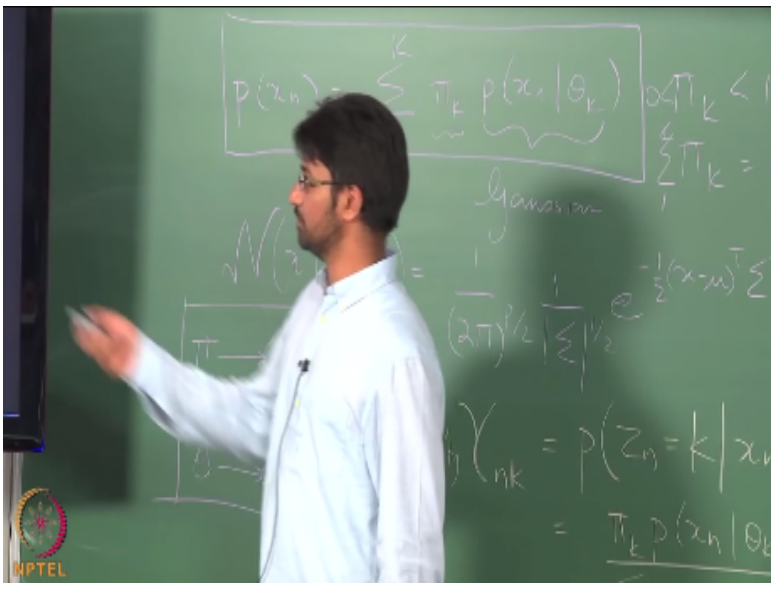
Likelihood estimates these will be most likely from this from a single Gaussian right and these blue points and these red points will not be identified correctly alright.

(Refer Slide Time: 27:10)



So now suppose you are given data you want to fit a Gaussian mixture model to it to either infer the clusters or to just fit the density for either case you need to estimate the parameters, and what are the parameters if you have P dimensional data k components for each of the Gaussians we need to find out the k mixing coefficients.

(Refer Slide Time: 27:42)

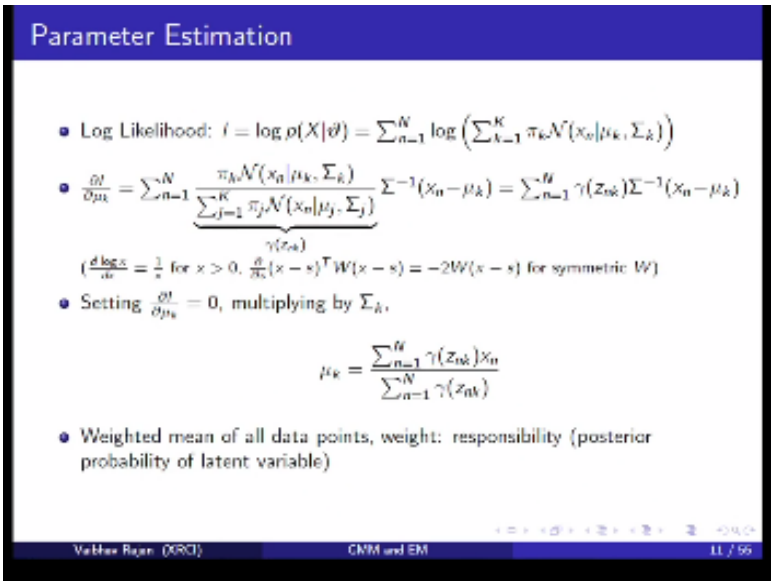


We need to find out the mean parameters the mean vectors which are P dimensional and we need to find out infer the covariance matrices, right so for now we will assume that mean okay and I will come back to how k is estimated later yeah for now we will just assume for example we see data like this and we decide that k is going to be 3, it need not be the case but yeah and from this one from here itself you can see that this is going to get difficult on high dimensional data because the covariance matrices are going to scale quadratically with dimensions.

So as the dimensionality of the data increases your influence task becomes more and more difficult right, we will keep we will not worry about that right now so how do we estimate these parameters will take the standard route of maximum likelihood the likelihood is just given by this for the case of Gaussian mixture model with k components this is for all the n data points and we take the logarithm which will convert this product to a sum here the problem is it will only convert this to a sum the inside summation remains.

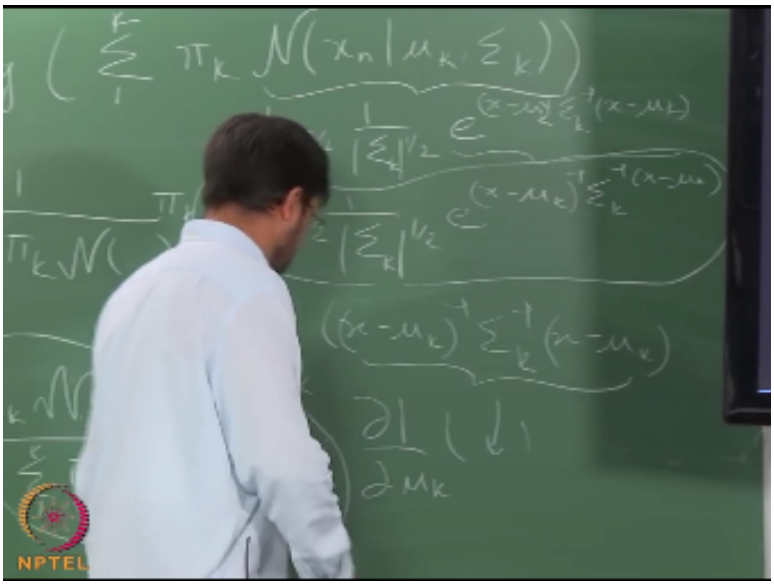
So what you get when you look at the log likelihood would be would be the ∑ of n different log logarithm terms we are inside each logarithm there is a summation and this turns out to be this gives us a lot of problems, so if you go by just differentiating this and getting you are equating it to 0 the standard route what you probably did for normal distribution this summation is going to cause problems and you are not going to get a closed-form solution because you cannot easily differentiate this and that is one of the main problems for estimating the parameters of a Gaussian mixture.

(Refer Slide Time: 30:13)



All right but just to play with the math let us still do it let us take the derivatives but we will make one crucial assumption, one assumption which is that we know the responsibility terms let us see how it works out so we have the we have the log likelihood and then let us say we want to find out each of these parameters the standard maximum-likelihood ways you take one of the each one of the parameters start with μk and take the derivative equate it to 0 and then see what you get by doing some algebra so let me just do this.

(Refer Slide Time: 31:04)



Right this is the log-likelihood and what we want to do is find the partial derivatives with respect to μk first, so this ∑ just remains outside and log of this would just give us the whole thing below whatever is inside there then we differentiate this part, so we also this let us write it down it is these are constants σk is a constant here with respect to μk and then we need the derivative of this derivative of an exponential would just be an exponential right, and this times so what you get you get back the entire formula for Gaussian here.

Multiplied by the derivative of this complicated term here let us change the I am just changing the this is, so what I really want to say here is when you do this derivation what you will find here if that this term comes up and this is the same as the responsibility term this is what I have written here and then the rest of it is just algebra there is a very nice book called Matrix cookbook where you will find all these derivatives complex matrix derivatives and you can see that the derivative of this term.

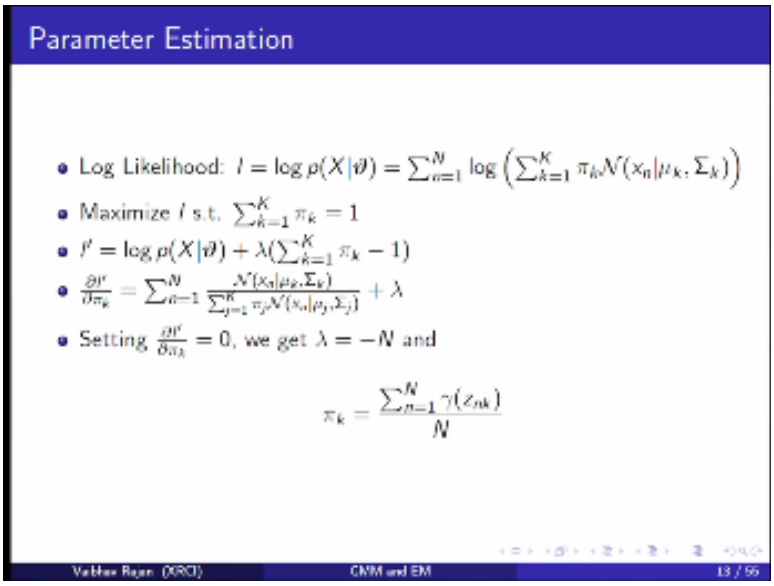
X-μk inverse x σ inverse times x - μk is nothing but - 2 times the – 2 σ inverse times x - μk when and the covariance matrix is symmetric, so you can just substitute that here in continue and what you get is σ and k how did it x so what is it – 2 times σ inverse so this ∂l / ∂μk just gives us this and when you equate this to 0 you can x σ k on both sides and finally turns out that you will get a very nice form μk = σ γ z and k times X n so what does this mean the mean that you have that you have inferred is the weighted mean.

Of all the data points where the weight is the responsibility right so remember we do not know the responsibilities we have assumed that we know this but the responsibility has all the unknowns inside it has π k μk and σk all of it is unknown we just assumed that we know this and we substituted it and did the math right and we got a nice, form for μk we can do the same math for σk again if it differentiate this with respect to σk we will we have to just use two different matrix derivative formulas.

These it is very simple the derivation is very simple do you want me to do it so I thing okay and again you get a reasonably nice form for σk provided that you assume that you know the responsibilities or the k^{th} cluster is given by the weighted mean of all the data points where the may the weight is nothing but the responsibility, responsibility of that cluster towards this data point responsibility is the posterior probability of the latent variable given that you have this data so essentially it is trying to capture what generated that data point.

Which of the components generated that data point right, so all this is just so right now we do not really have anything we do not have the me the main parameters right, now we just know that if we knew the responsibilities which you do not know we could get a nice form for this and without this assumption this derivation does not hold okay and the same thing with respect to with respect to the σk values when we take the derivatives with respect to πk we have to be careful.

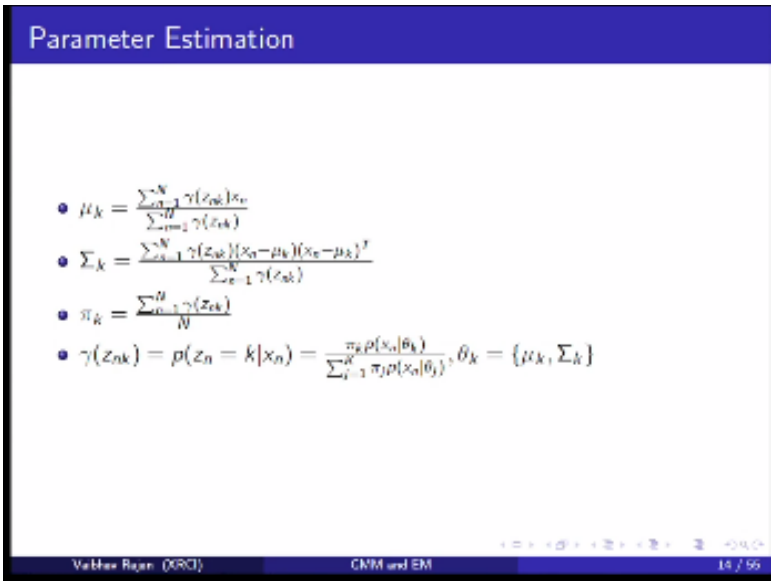
(Refer Slide Time: 39:17)



We have to make sure that we use the constraint the remember σk the sum of all the σ case must be equal to 1 so we cannot directly differentiate this right you have to use Lagrange multipliers is everybody familiar with the Lagrange multipliers requires yes no okay, so then you just take a Lagrange multiplier here take this constraint and do the differentiation you again get you again get back the responsibilities so every time you do this derivative you see the responsibility bring up and if you set this to 0.

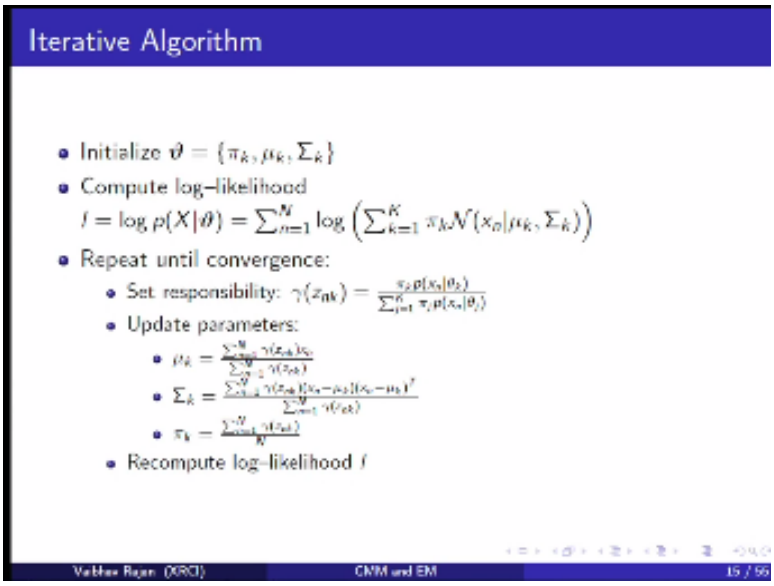
We can get the parameter the γ parameter as - n and we get the value of πk equal to again this proportion this proportion again is intuitively very clear what it is saying is πk the mixture weight is the proportion so if you take the sum of all the responsibilities overall data points over all components it is equal to n right and, so the k^{th} mixture weight is nothing but the proportion of the responsibilities that are coming from the k^{th} component towards all the data points and you are taking the proportion that is given by that by those responsibilities with respect to all of it some of it.

(Refer Slide Time: 40:53)



So let us summarize what we found is that we have very nice forms from you k σ k and πk a given that we know the responsibilities which is the posterior probability of the latent variable coming from that company coming from the k component for the nth data point, so can so can you think of can you think of how you can use this to create an algorithm for estimating the Gaussian mixture parameters in your exactly, yeah so this is our first case we start with some we initialize all.

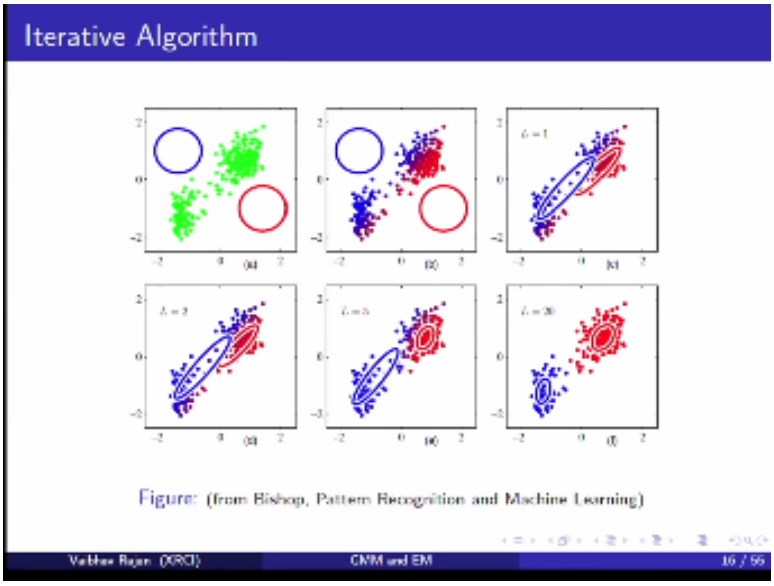
(Refer Slide Time: 41:53)



So I will denote by curly θ all the parameters and we start with some guess some gas of the parameters and then we will compute the log likelihood we can compute it because we know we have guessed the parameters and then we will set the responsibilities because we know again all these all these parameters are guess and since given the responsibility we can compute all the parameters again we will use this to get a new guess and that way we will iteratively keep refining our guess.

Now this looks very ad hoc but is actually is actually theoretically quite sound and this is we will see that why the reason why this is a good thing to do is we will show that this is guaranteed to increase the likelihood at every iteration and we will see that when we understand why when you understand how EM works so this turns out to be actually an instance of the am algorithm it is quite is very intuitive.

(Refer Slide Time: 43:11)



So if you so this is an example of exactly that algorithm we start with some data the same data set we start with some guesses for the Gaussians and iteratively we see that the Gaussians converge to nicely fitting the data the parameters that you get here will fit the data very well the only problem is that it will take it usually takes a long time to come to the right parameters yeah we know the k till now we know the K we are so in this whole iterative algorithm we are assuming that we know the k.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in