NPTEL

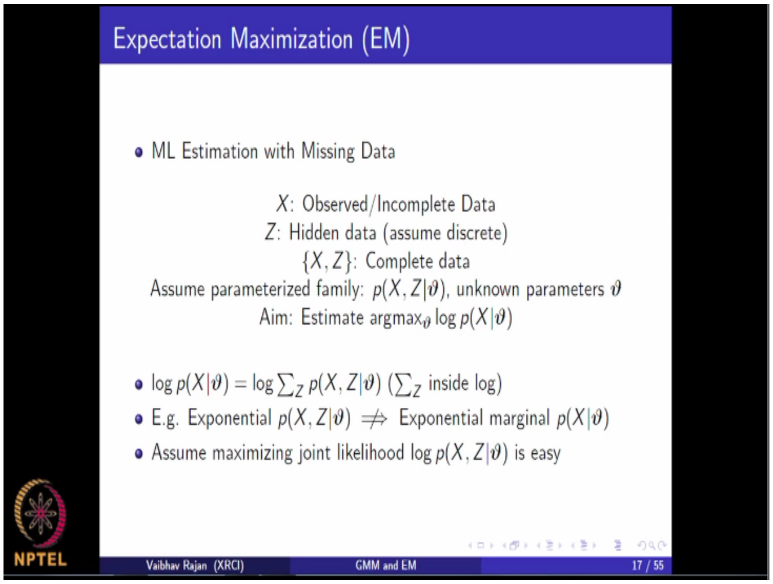
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-77
Expectation Maximization

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

(Refer Slide Time: 00:15)



Expectation Maximization (EM)

- ML Estimation with Missing Data

X: Observed/Incomplete Data
Z: Hidden data (assume discrete)
$\{X, Z\}$: Complete data
Assume parameterized family: $p(X, Z|\vartheta)$, unknown parameters ϑ
Aim: Estimate $\text{argmax}_\vartheta \log p(X|\vartheta)$

- $\log p(X|\vartheta) = \log \sum_Z p(X, Z|\vartheta)$ ($\sum_Z$ inside log)
- E.g. Exponential $p(X, Z|\vartheta) \not\Rightarrow$ Exponential marginal $p(X|\vartheta)$
- Assume maximizing joint likelihood $\log p(X, Z|\vartheta)$ is easy

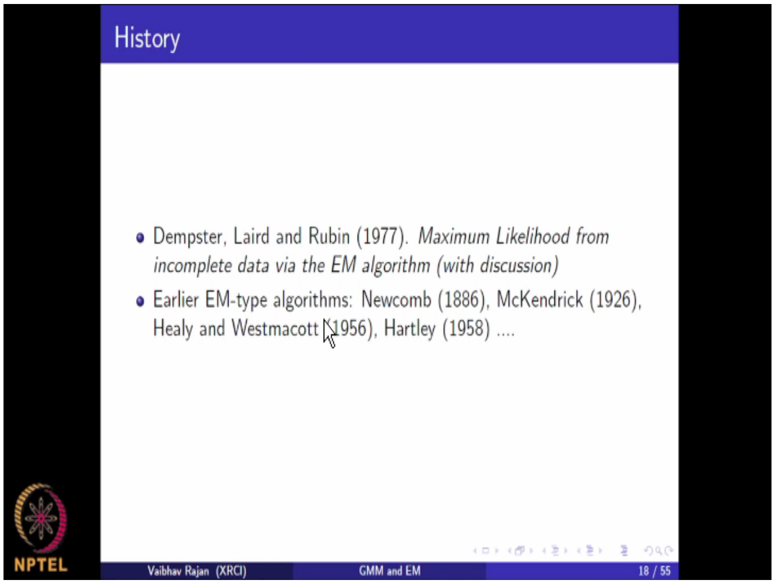
Vaibhav Rajan (XRCI) GMM and EM 17 / 55

This is expectation maximization the so it is a way to do maximum likelihood estimation initially it was proposed as a way to do maximum likelihood estimation when you have missing data so suppose you are given data X which is what you observe and is incomplete is known to be an incomplete but there are some values that are missing and you want to you want to estimate the maximum likelihood you want to get the maximum likelihood estimates of the parameters which are unknown.

But here you are assuming two things you are assuming that there is some parameterized family doing some parameter is fitting which now the h is for which the joint likelihood is easy to compute so we denote by X and Z the complete data the observed data plus the hidden data and we assume some parameterized family from which this data is generated like a Gaussian or exponential or something like that and we do not know the unknown parameters which you want to estimate.

So we see we start seeing connections with what we have seen in the case of Gaussian mixture because if you take the marginal probability here you again see a summation coming inside the log right and this again poses problems and they need not be of the same family so if they have a joint exponential does not mean that you will have imagined was coming from the same family okay.

(Refer Slide Time: 02:04)

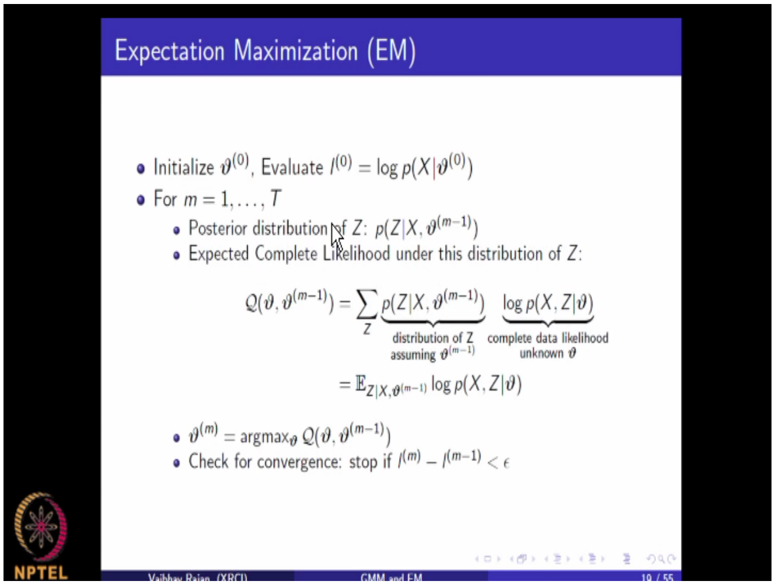


History

- Dempster, Laird and Rubin (1977). *Maximum Likelihood from incomplete data via the EM algorithm (with discussion)*
- Earlier EM-type algorithms: Newcomb (1886), McKendrick (1926), Healy and Westmacott (1956), Hartley (1958) ....

So EM as it is most commonly used now was proposed by Dempster Laird and Rubin in their seminal paper maximum likelihood from incomplete data by the EM algorithm this was in 1977 and even before that a lot of statisticians mostly they have developed EM like algorithms but usually when we cite EM we usually site the 1977 paper so slide 17 so this is the problem okay.

(Refer Slide Time: 02:32)



Expectation Maximization (EM)

- Initialize $\vartheta^{(0)}$, Evaluate $l^{(0)} = \log p(X|\vartheta^{(0)})$
- For $m = 1, \ldots, T$
  - Posterior distribution of Z: $p(Z|X, \vartheta^{(m-1)})$
  - Expected Complete Likelihood under this distribution of Z:

$$Q(\vartheta, \vartheta^{(m-1)}) = \sum_{Z} \underbrace{p(Z|X, \vartheta^{(m-1)})}_{\substack{\text{distribution of Z} \\ \text{assuming } \vartheta^{(m-1)}}} \underbrace{\log p(X, Z|\vartheta)}_{\substack{\text{complete data likelihood} \\ \text{unknown } \vartheta}}$$

$$= \mathbb{E}_{Z|X, \vartheta^{(m-1)}} \log p(X, Z|\vartheta)$$

  - $\vartheta^{(m)} = \text{argmax}_{\vartheta}\, Q(\vartheta, \vartheta^{(m-1)})$
  - Check for convergence: stop if $l^{(m)} - l^{(m-1)} < \epsilon$

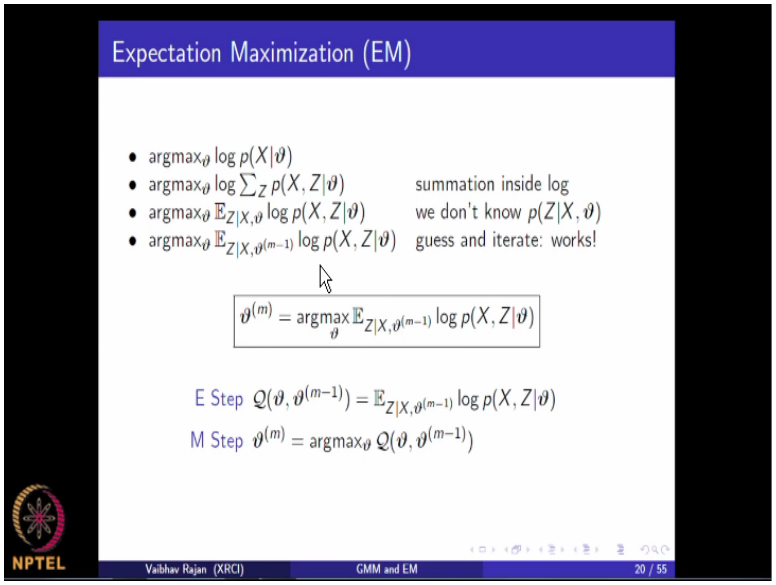
Vaibhav Rajan (XRCI) GMM and EM 19 / 55

You have incomplete data you have hit hidden data yeah I forgot to mention something for, for the rest of the discussion will assume that this hidden data is discreet but all the derivations will work if you assume this to be continuous as well you just have to replace the summations with the intent with integrals so you have observed or incomplete data and he did not deter the complete data is the combination of these two you are assuming some parameterized family EM solves.

And now we will see what EM is so EM is an iterative algorithm just like what we saw for just like what we designed for Gaussian mixture you again started the guess, guess of the parameters that you have you evaluate your first guest likelihood and then you iteratively to two steps first compute the posterior distribution of Z given the current estimate of the parameters okay.

After that you compute the expected complete log likelihood under this distribution so we call this will call this Q function okay now notice that this, this expectation takes the complete data likelihood here the parameters are unknown right and this expectation assumes the distribution of Z given the parameters that you have guests in the previous room okay and then you again get a new guess and that new guess is got by maximizing taking maximizing this Q function.

(Refer Slide Time: 05:08)



Expectation Maximization (EM)

- $\text{argmax}_\vartheta \log p(X|\vartheta)$
- $\text{argmax}_\vartheta \log \sum_Z p(X,Z|\vartheta)$ summation inside log
- $\text{argmax}_\vartheta \mathbb{E}_{Z|X,\vartheta} \log p(X,Z|\vartheta)$ we don't know $p(Z|X,\vartheta)$
- $\text{argmax}_\vartheta \mathbb{E}_{Z|X,\vartheta^{(m-1)}} \log p(X,Z|\vartheta)$ guess and iterate: works!

$$\boxed{\vartheta^{(m)} = \underset{\vartheta}{\text{argmax}}\, \mathbb{E}_{Z|X,\vartheta^{(m-1)}} \log p(X,Z|\vartheta)}$$

E Step $Q(\vartheta, \vartheta^{(m-1)}) = \mathbb{E}_{Z|X,\vartheta^{(m-1)}} \log p(X,Z|\vartheta)$

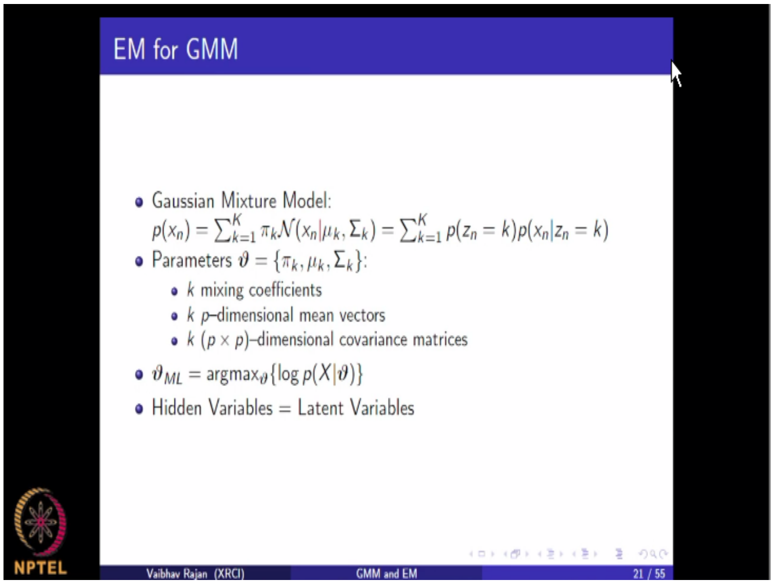
M Step $\vartheta^{(m)} = \text{argmax}_\vartheta\, Q(\vartheta, \vartheta^{(m-1)})$

Vaibhav Rajan (XRCI) GMM and EM 20 / 55

And taking the that argument theta curly theta which maximizes this Q function so I think I have it there what we wanted was to get the maximum likelihood estimate of theta right and we, we see X but X is not complete X has some missing data Z so and so we express this as a marginal distribution of the complete data right and we get into the same problem of the summation being inside the logarithm.

So we decide to not compute the maximum likelihood in this way but instead compute the max compute the maximum taking the expectation of the log likelihood of the complete data but again we under the distribution of Z but we do not know the real distribution of Z because we do not know the parameters so we take the guess that we had in the last the previous round right so we compute the expectation of the complete data likelihood under the distribution of Z given the current guess of the parameters.

And this works you will see why it works but this works so the entire EM algorithm can actually be represented by just this one line you start with some guess and then for the next guess you calculate the expectation of the complete data log likelihood under the distribution of the missing data using the previous guess this can actually be broke this is broken down into two steps.
(Refer Slide Time: 07:39)



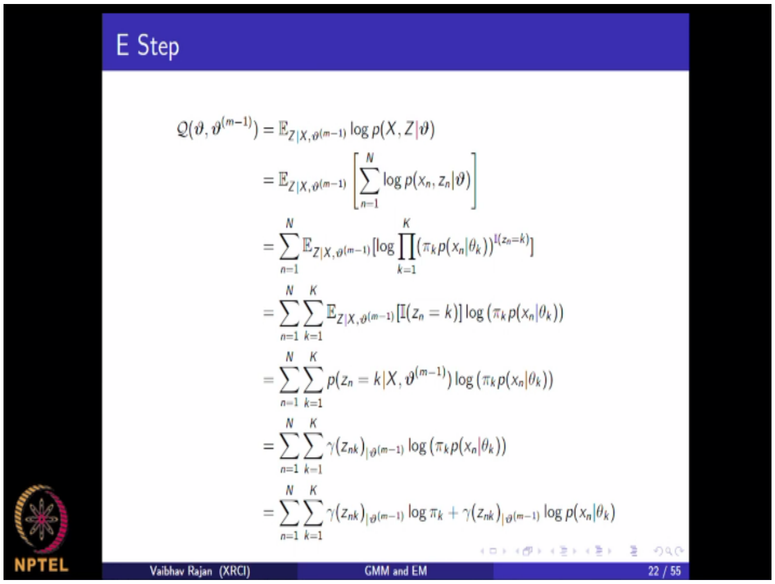
EM for GMM

- Gaussian Mixture Model: $p(x_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) = \sum_{k=1}^{K} p(z_n = k)p(x_n|z_n = k)$
- Parameters $\vartheta = \{\pi_k, \mu_k, \Sigma_k\}$:
  - k mixing coefficients
  - k p–dimensional mean vectors
  - k $(p \times p)$–dimensional covariance matrices
- $\vartheta_{ML} = \text{argmax}_{\vartheta}\{\log p(X|\vartheta)\}$
- Hidden Variables = Latent Variables

Vaibhav Rajan (XRCI) GMM and EM 21 / 55

We are in the Estep you compute this expectation in the m-step you maximize and get the parameters maximize this Q and get the next set of parameters so let us see how, how we can get the EM algorithm for Gaussian mixtures the key thing here is we did not assume anything we did not say anything about hidden variables but the trick here is to use these latent variables as hidden variables and this is how EM is used in a lot of different models not just Gaussian mixture lot of latent variable models.

(Refer Slide Time: 08:31)

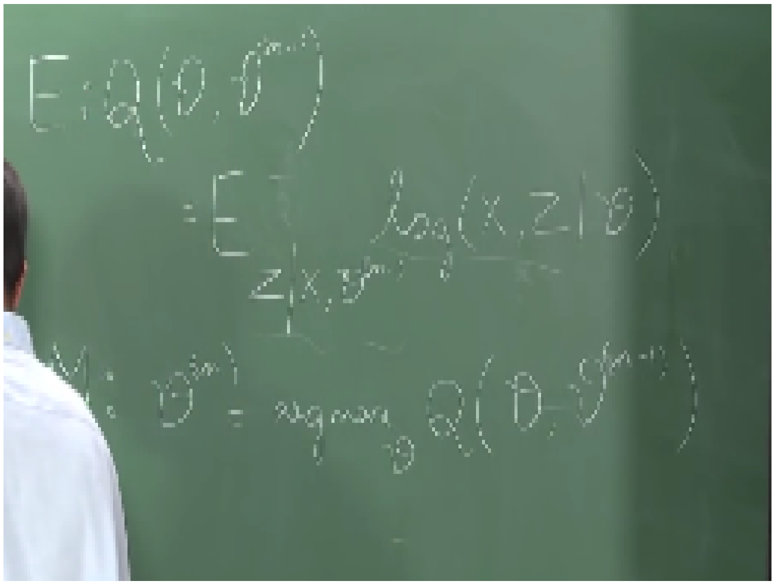


E Step

$$
\begin{aligned}
\mathcal{Q}(\vartheta, \vartheta^{(m-1)}) &= \mathbb{E}_{Z|X,\vartheta^{(m-1)}} \log p(X, Z|\vartheta) \\
&= \mathbb{E}_{Z|X,\vartheta^{(m-1)}} \left[\sum_{n=1}^{N} \log p(x_n, z_n|\vartheta) \right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{Z|X,\vartheta^{(m-1)}} [\log \prod_{k=1}^{K} (\pi_k p(x_n|\theta_k))^{\mathbb{I}(z_n=k)}] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{Z|X,\vartheta^{(m-1)}} [\mathbb{I}(z_n = k)] \log (\pi_k p(x_n|\theta_k)) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_n = k|X, \vartheta^{(m-1)}) \log (\pi_k p(x_n|\theta_k)) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})_{|\vartheta^{(m-1)}} \log (\pi_k p(x_n|\theta_k)) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})_{|\vartheta^{(m-1)}} \log \pi_k + \gamma(z_{nk})_{|\vartheta^{(m-1)}} \log p(x_n|\theta_k)
\end{aligned}
$$

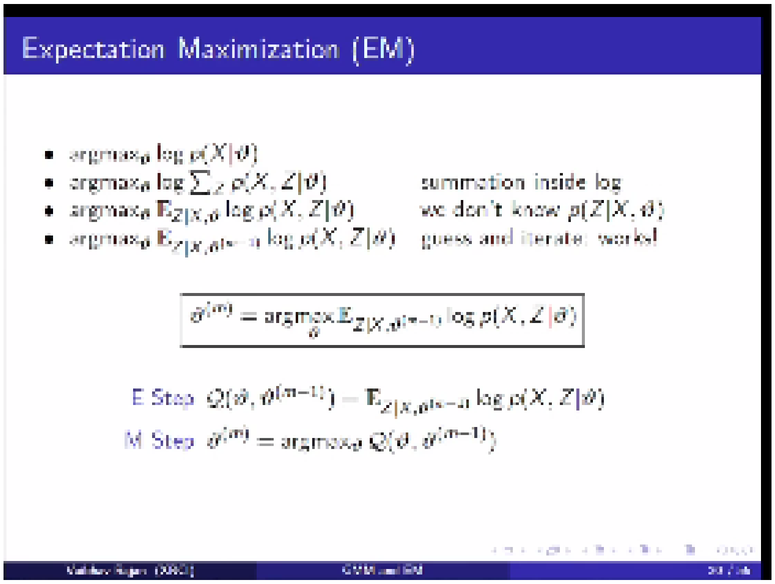
You assume that the latent variables are hidden you do not know them and run the whole EM machinery okay so this is just a reminder we have the Gaussian mixture model where we want to estimate all the parameters represented by curl theta you have K components pi k mu k sigma k for each for each component right. And what we want to find out is the maximum likelihood estimate all right we have x so.

(Refer Slide Time: 09:20)


$E: Q(\theta, \theta^{(n-1)})$
$= E_{Z|X,\theta^{(n-1)}} \log(X,Z|\theta)$
$M: \theta^{(n)} = \mathrm{argmax}_{\theta}\, Q(\theta, \theta^{(n-1)})$

So let us let us write this down here yeah the most important thing to remember is that this distribution is taken over the previous guess where as the expectation is for the complete log-likelihood over the unknown parameters right and the m-step just gets the next guess alright so if you want to get the maximum likelihood parameters we first need to compute this Q function and then usually this step is easier it is just the expectation that requires some work once you get Q then the maximization step is just computing the derivatives.

(Refer Slide Time: 10:45)



Expectation Maximization (EM)

- $\text{argmax}_\theta \log p(X|\theta)$
- $\text{argmax}_\theta \log \sum_Z p(X, Z|\theta)$ summation inside log
- $\text{argmax}_\theta \mathbb{E}_{Z|X,\theta} \log p(X, Z|\theta)$ we don't know $p(Z|X, \theta)$
- $\text{argmax}_\theta \mathbb{E}_{Z|X,\theta^{(m-1)}} \log p(X, Z|\theta)$ guess and iterate: works!

$$\boxed{\theta^{(m)} = \underset{\theta}{\text{argmax}}\, \mathbb{E}_{Z|X,\theta^{(m-1)}} \log p(X, Z|\theta)}$$

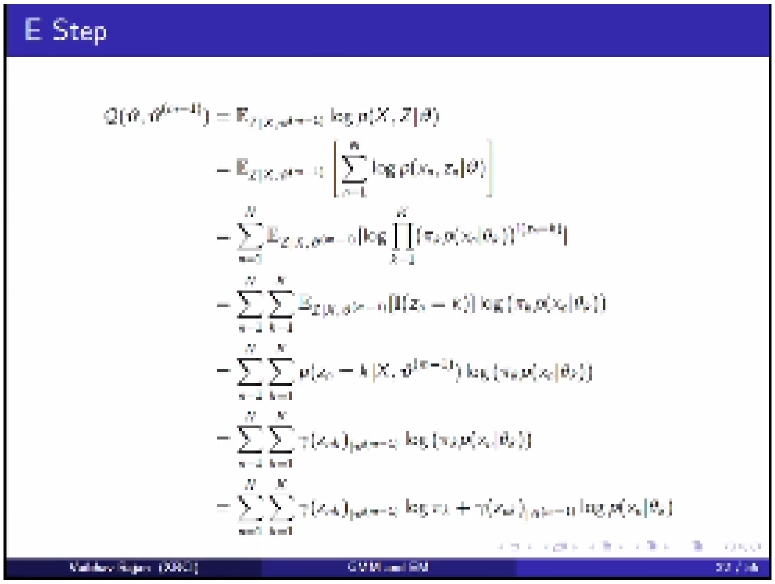
E Step $Q(\theta, \theta^{(m-1)}) = \mathbb{E}_{Z|X,\theta^{(m-1)}} \log p(X, Z|\theta)$

M Step $\theta^{(m)} = \text{argmax}_\theta\, Q(\theta, \theta^{(m-1)})$

Vaibhav Rajan (XRCI) GMM and EM [illegible]

And as you can see this works only if the e-step gives you something which you can easily maximize and in the case of Gaussian mixture we will see that by using the complete data likelihood we will be able to get something that we can easily maximize.

(Refer Slide Time: 11:06)



E Step

$$\begin{aligned} Q(\theta, \theta^{(t-1)}) &= \mathbb{E}_{Z|X,\theta^{(t-1)}} \log p(X, Z|\theta) \\ &= \mathbb{E}_{Z|X,\theta^{(t-1)}} \left[\sum_{n=1}^{N} \log p(x_n, z_n|\theta) \right] \\ &= \sum_{n=1}^{N} \mathbb{E}_{Z|X,\theta^{(t-1)}} \left[\log \prod_{k=1}^{K} (\pi_k p(x_n|\theta_k))^{\mathbb{I}(z_n=k)} \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{Z|X,\theta^{(t-1)}} [\mathbb{I}(z_n = k)] \log (\pi_k p(x_n|\theta_k)) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_n = k|X, \theta^{(t-1)}) \log (\pi_k p(x_n|\theta_k)) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})|_{\theta^{(t-1)}} \log (\pi_k p(x_n|\theta_k)) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})|_{\theta^{(t-1)}} \log \pi_k + \gamma(z_{nk})|_{\theta^{(t-1)}} \log p(x_n|\theta_k) \end{aligned}$$

Vaibhav Rajan (IISC) GMM and EM

So let us see this one by one the Q function is this is just by definition it is the expectation of the complete data likelihood using under the distribution of Z given the previous the guest parameters right, so this log likelihood is just the log this is the likelihood of X_n Z_n given the unknown parameters over all the data points now I can take this summation outside by linearity of expectations so the summation goes in expectations here and now this complete log likelihood can be written in this form right.

So we can derive this formally but intuitively it is very clear you see this is this is just an indicator function which gets a value of one when Z_n is equal to k gets a value of zero and zero is not equal to K so this prod in this product here will be only one term which will have all the terms except one will get an exponent of zero and so you will have one and only one term out of the K will remain for each for this right and so the log-likelihood is just it comes straight for straight away from this formula after using the indicator function here.

And this gives us the complete data likelihood all right so then after that the product becomes a sum when you take it outside the log and the exponent comes down and again the expectation can be brought inside by linearity, so you get both the summations out and with respect to the distribution of z using the previously guest parameters this is just a constant, so the expectation is only over this indicator function right and this indicator function expectation of an indicator function just gives us the probability and you have the probability of Zn = k again given x and the previously guest parameter values log just remains.

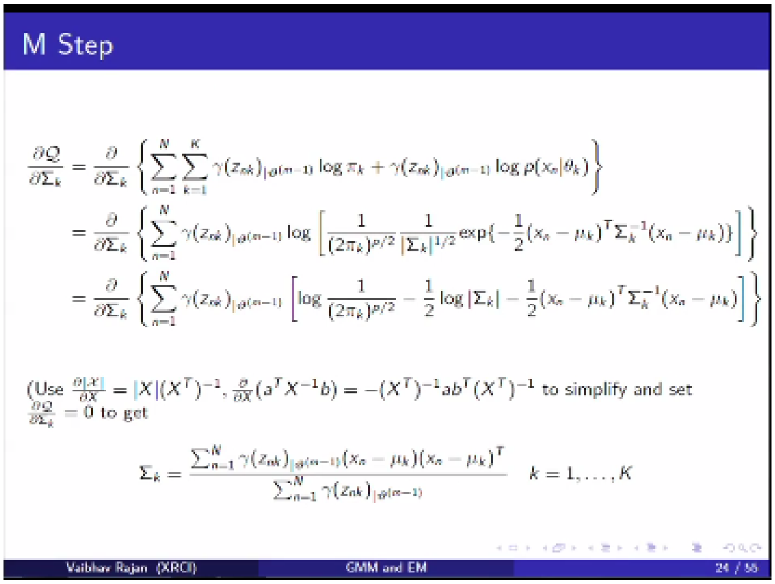
So this is again the responsibility is the posterior probability of Zn = k but in this case this responsibility is not with respect to the exact the original parameters this posterior probabilities with respect to the guest parameters right, so I am indicating that by this subscript here so it is the responsibility times the log that remains, so what have we achieved we the reason so we have we have got a expression for Q in terms of again the responsibility but this time the responsibility is with respect to the guest parameter.

And if you just look at this function you can see that this is easier to differentiate because the summations are all outside and the normal distribution will come here and the differentiation will be just like what you do for the in the case of fitting a single Gaussian okay and how did how did that happen it happened mainly because we were taking the complete data likelihood the complete data likelihood gave us a nice mathematical form here which due to the expectation all the summations got pushed out and we got this nice form for Q any questions here.

Yeah but EM is used so that that comes in many contexts not just in the case of Gaussian mixture in a lot of those cases EM is useful yeah but it is me if you could if you could get the maximum likelihood easily you would not need to use EM for Gaussian mix of course now the m-step is just this which is we differentiate with respect to each of the parameters so you have the Q function right and this is the same Q function here now one thing to remember here is that this parameter the guest parameter at the previous iteration we know these parameters.

So we know the responsibilities so the responsibilities are just constants in this case right and so differentiating this becomes very easy so this is with respect to μk this whole term is not necessary we just we focus only on this term and we again get so this is for each of the different components you get you get the entire normal distribution here and no summations log of this is exactly the same division as for a single Gaussian it and you use the same you use matrix derivatives to get very simple forms here and you can get you will see that you again get back μk the same form that you get the same form that we saw earlier for our guests for our adhoc iterative algorithm except that these responsibilities are with respect to the previously guest parameters okay right.

(Refer Slide Time: 17:43)



M Step

$$\frac{\partial Q}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k}\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})_{|\theta^{(n-1)}}\log \pi_k + \gamma(z_{nk})_{|\theta^{(n-1)}}\log p(x_n|\theta_k)\right\}$$

$$= \frac{\partial}{\partial \Sigma_k}\left\{\sum_{n=1}^{N}\gamma(z_{nk})_{|\theta^{(n-1)}}\log\left[\frac{1}{(2\pi_k)^{p/2}}\frac{1}{|\Sigma_k|^{1/2}}\exp\{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\}\right]\right\}$$

$$= \frac{\partial}{\partial \Sigma_k}\left\{\sum_{n=1}^{N}\gamma(z_{nk})_{|\theta^{(n-1)}}\left[\log\frac{1}{(2\pi_k)^{p/2}} - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right]\right\}$$

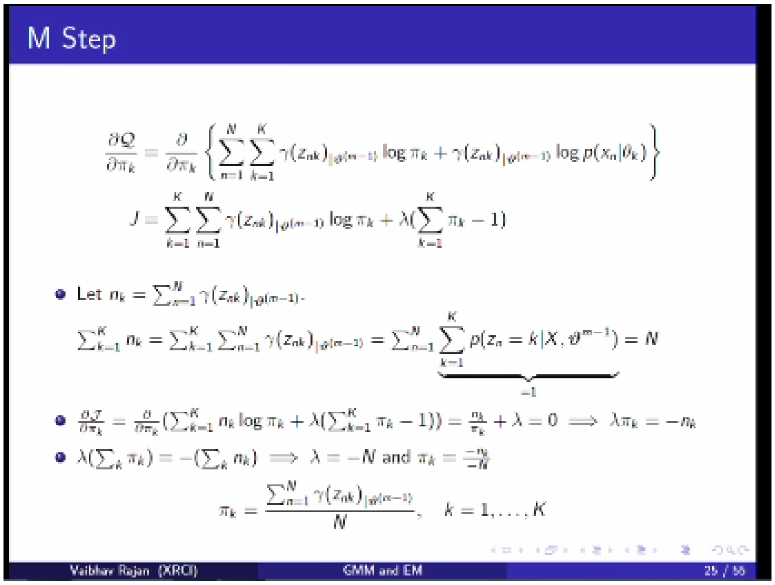
(Use $\frac{\partial |X|}{\partial X} = |X|(X^T)^{-1}$, $\frac{\partial}{\partial X}(a^T X^{-1} b) = -(X^T)^{-1}ab^T(X^T)^{-1}$ to simplify and set $\frac{\partial Q}{\partial \Sigma_k} = 0$ to get

$$\Sigma_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})_{|\theta^{(n-1)}}(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{N}\gamma(z_{nk})_{|\theta^{(n-1)}}} \quad k = 1,\ldots,K$$

Vaibhav Rajan (XRCI) GMM and EM 24 / 55

So when we do this for Σ k again we do not need to worry about this part we only differentiate for this part this is also very simple you can you can simplify this further by my first applying the logarithm here for each of these parts and then the logarithm for the determinant is given by a simple formula you can apply the logarithm here I mean to a derivative here and again you get back the same form for Σ k which you found earlier.

(Refer Slide Time: 18:22)



M Step

$$\frac{\partial Q}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})_{|\vartheta^{(m-1)}}\log \pi_k + \gamma(z_{nk})_{|\vartheta^{(m-1)}}\log p(x_n|\theta_k)\right\}$$

$$J = \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk})_{|\vartheta^{(m-1)}}\log \pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

- Let $n_k = \sum_{n=1}^{N}\gamma(z_{nk})_{|\vartheta^{(m-1)}}$.

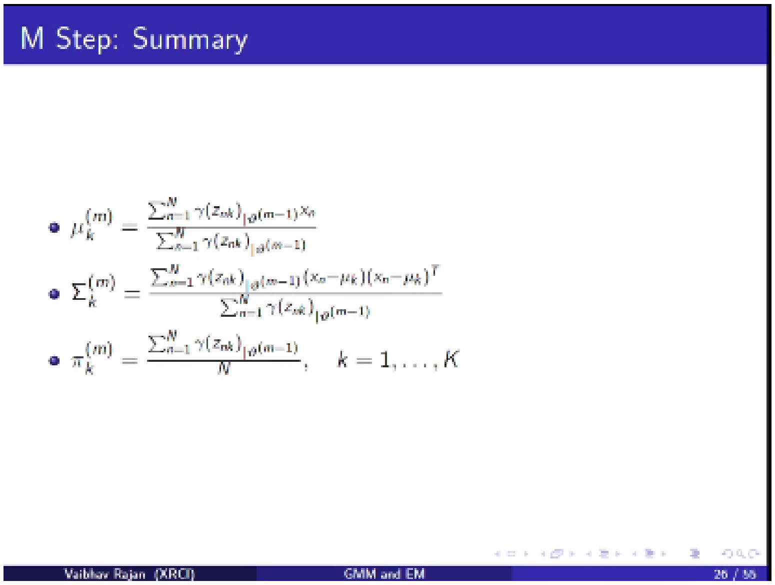
  $\sum_{k=1}^{K} n_k = \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk})_{|\vartheta^{(m-1)}} = \sum_{n=1}^{N}\underbrace{\sum_{k=1}^{K} p(z_n = k|X, \vartheta^{m-1})}_{=1} = N$
- $\frac{\partial J}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}(\sum_{k=1}^{K} n_k \log \pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1)) = \frac{n_k}{\pi_k} + \lambda = 0 \implies \lambda\pi_k = -n_k$
- $\lambda(\sum_k \pi_k) = -(\sum_k n_k) \implies \lambda = -N$ and $\pi_k = \frac{-n_k}{-N}$

$$\pi_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})_{|\vartheta^{(m-1)}}}{N}, \quad k = 1, \ldots, K$$

Vaibhav Rajan (XRCI) GMM and EM 25 / 58

So what and similarly for the M-step or Π k right this time this term is not this term goes away you use Lagrange multipliers to get to add this term and differentiate this function to again get so this is quite straightforward you can check it yourself right.

(Refer Slide Time: 18:50)



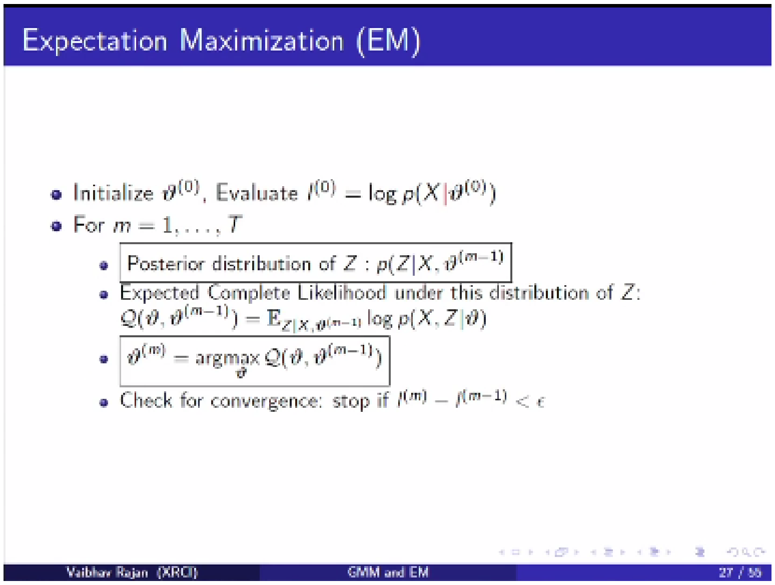
M Step: Summary

- $\mu_k^{(m)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})_{|\theta^{(m-1)}} x_n}{\sum_{n=1}^{N} \gamma(z_{nk})_{|\theta^{(m-1)}}}$
- $\Sigma_k^{(m)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})_{|\theta^{(m-1)}} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})_{|\theta^{(m-1)}}}$
- $\pi_k^{(m)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})_{|\theta^{(m-1)}}}{N}, \quad k = 1, \ldots, K$

Vaibhav Rajan (XRCI) GMM and EM 26 / 56

So what have we done we first we first found a we first found these same formulas by assuming that we know the responsibility and then we used the EM steps to find that these values are exactly the same as what we found earlier right.

(Refer Slide Time: 19:13)



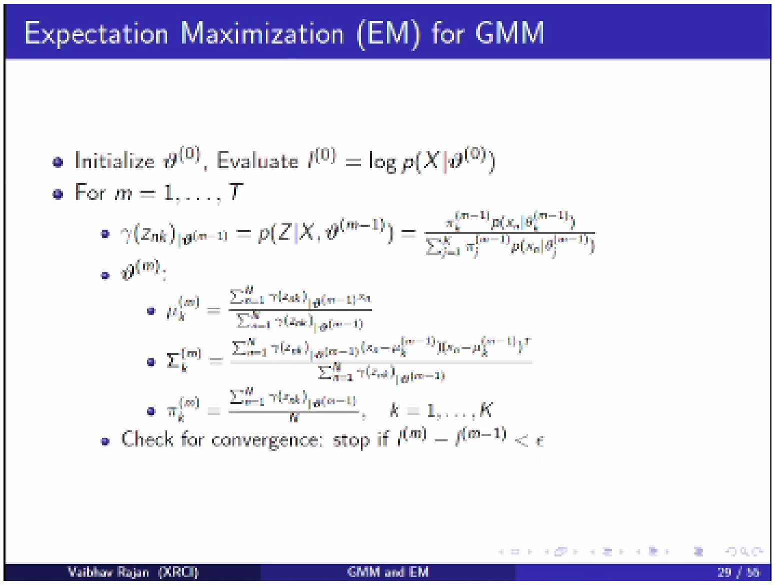
Expectation Maximization (EM)

- Initialize $\vartheta^{(0)}$, Evaluate $l^{(0)} = \log p(X|\vartheta^{(0)})$
- For $m = 1, \ldots, T$
  - $\boxed{\text{Posterior distribution of } Z : p(Z|X, \vartheta^{(m-1)})}$
  - Expected Complete Likelihood under this distribution of Z: $\mathcal{Q}(\vartheta, \vartheta^{(m-1)}) = \mathbb{E}_{Z|X,\vartheta^{(m-1)}} \log p(X, Z|\vartheta)$
  - $\boxed{\vartheta^{(m)} = \underset{\vartheta}{\operatorname{argmax}}\, \mathcal{Q}(\vartheta, \vartheta^{(m-1)})}$
  - Check for convergence: stop if $l^{(m)} - l^{(m-1)} < \epsilon$

Vaibhav Rajan (XRCI) GMM and EM 27 / 55

And if we plug this into the EM framework this is the EM framework you start to the gas you iterate by first finding the posterior distribution of Z and then you find the expected complete likelihood under this distribution of Z and finally maximize to get the new guesses right, so the posterior distribution of Z just gives us the responsibilities and then this derivation we have seen.

(Refer Slide Time: 19:43)



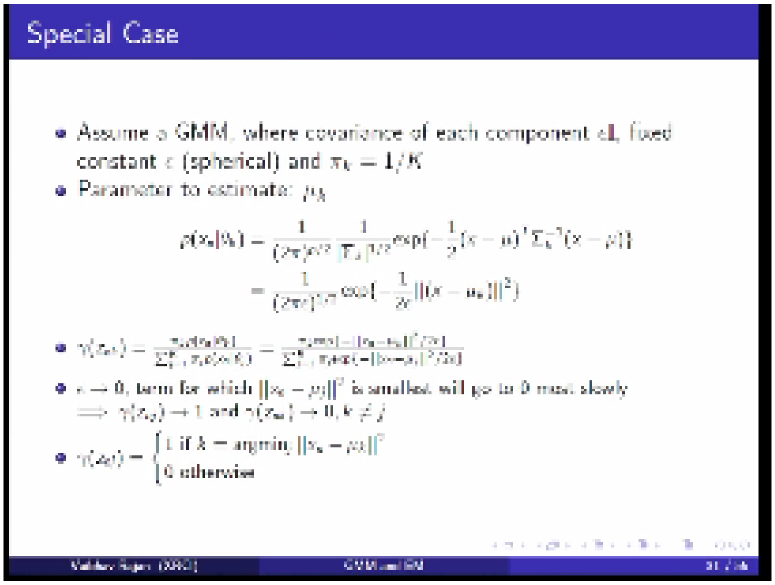
Expectation Maximization (EM) for GMM

- Initialize $\vartheta^{(0)}$, Evaluate $l^{(0)} = \log p(X|\vartheta^{(0)})$
- For $m = 1, \ldots, T$
  - $\gamma(z_{nk})_{|\vartheta^{(m-1)}} = p(Z|X, \vartheta^{(m-1)}) = \frac{\pi_k^{(m-1)} p(x_n|\theta_k^{(m-1)})}{\sum_{j=1}^{K} \pi_j^{(m-1)} p(x_n|\theta_j^{(m-1)})}$
  - $\vartheta^{(m)}$:
    - $\mu_k^{(m)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})_{|\vartheta^{(m-1)}} x_n}{\sum_{n=1}^{N} \gamma(z_{nk})_{|\vartheta^{(m-1)}}}$
    - $\Sigma_k^{(m)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})_{|\vartheta^{(m-1)}} (x_n - \mu_k^{(m-1)})(x_n - \mu_k^{(m-1)})^T}{\sum_{n=1}^{N} \gamma(z_{nk})_{|\vartheta^{(m-1)}}}$
    - $\pi_k^{(m)} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})_{|\vartheta^{(m-1)}}}{N}, \quad k = 1, \ldots, K$
  - Check for convergence: stop if $l^{(m)} - l^{(m-1)} < \epsilon$

Vaibhav Rajan (XRCI) GMM and EM 29 / 56

This gives us the next set of guess and this iteratively we can we can check for convergence when the likelihood does not change much we stop that is the EM algorithm for GMM I still not told you why this works but we will see that we will see the theoretical properties of why this is why this works well yeah I just wanted to show you that what we have got through by doing all the math for EM is exactly the same as what we found during the height relative algorithm that we guessed okay.

(Refer Slide Time: 20:31)



Special Case

- Assume a GMM, where covariance of each component $\epsilon\mathbf{I}$, fixed constant ϵ (spherical) and $\pi_k = 1/K$
- Parameter to estimate: μ_k

$$p(x_n|\theta_k) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma_k|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma_k^{-1}(x-\mu)\}$$
$$= \frac{1}{(2\pi\epsilon)^{D/2}}\exp\{-\frac{1}{2\epsilon}||(x-\mu_k||^2\}$$

- $\gamma(z_{nk}) = \frac{\pi_k p(x_n|\theta_k)}{\sum_{j=1}^{K}\pi_j p(x_n|\theta_j)} = \frac{\pi_k \exp\{-||x_n-\mu_k||^2/2\epsilon\}}{\sum_{j=1}^{K}\pi_j \exp\{-||x_n-\mu_j||^2/2\epsilon\}}$
- $\epsilon \to 0$, term for which $||x_n-\mu_j||^2$ is smallest will go to 0 most slowly $\implies \gamma(z_{nj}) \to 1$ and $\gamma(z_{nk}) \to 0, k \neq j$
- $\gamma(z_{nk}) = \begin{cases} 1 \text{ if } k = \arg\min_j ||x_n-\mu_j||^2 \\ 0 \text{ otherwise} \end{cases}$

Vaibhav Rajan ([illegible]) GMM and EM 31 / [illegible]

So let us do one more thing for today let us look at a special case let us assume Gaussian mixture model where the covariance of each component is fixed to be epsilon times the identity matrix so epsilon is a fixed constant you have epsilon times identity will give you a spherical Gaussian we also fixed pi k Picasa so each component is gives exactly the same contribution towards the Gaussian mixture now the only parameter to estimate is Mookie right.

So when you when you since **π** k is known sorry since Sigma K is just absolute times identity the formula for the normal distribution simplifies to just this okay so this is identity this goes away inside you just have epsilon from here the formula forth responsibility also simplifies you just plug this probability of x n given theta K here and you get ratio over the so you get just the exponential here and the sum over the different Exponential.

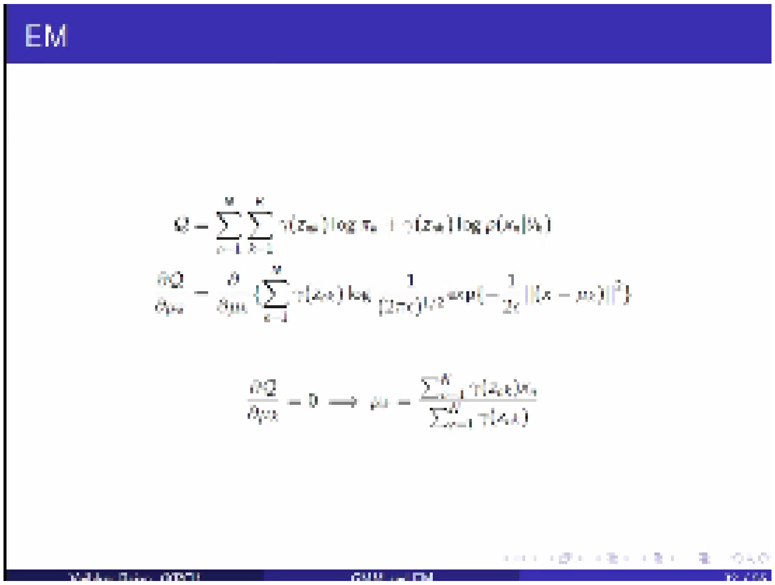
Here now if you look at this expression and see what happens to the denominators epsilon tends to 0 as epsilon tends to 0 the term for which this difference the smallest will go to 0 most slowly and so the responsibility will for that particular that jet component here because this these the numerator will be equal to the denominator in the in the limits and for all others the responsibility will go to 0.

Right so this is the special case of hard clustering that I was mentioning earlier right and what it turns out to be is just setting the responsibility to one for that component where the me is for this for that component where the data point is closest to the mean otherwise the responsibility is zero so we so what are we trying to say the responsibility is just the posterior probability of Inbeing equal to K for the net data point.

we want to know which component it has come from right and you are saying the responsibility set to one for that component where this is with respect to a particular data point the net data point for the n theta point the responsibility is one for that component whose mean the data point is closest to and it is 0.

For all others which means if you look at the data and look at x n and want to know the posterior probability of which component it came from it is that component whose mean is that data point is closest to and let us do the e/m for this so II m the first step is to calculate Q so you plug in this formula for this formula it's the same as before because we are doing Gaussian mixture is this the special Gaussian mixture.

(Refer Slide Time: 24:23)



EM

$$Q = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k + \gamma(z_{nk}) \log p(x_n|\theta_k)$$

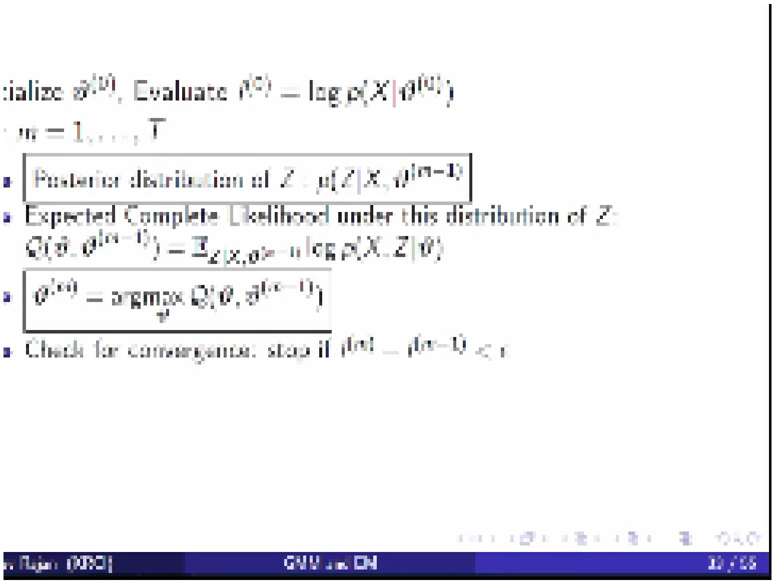
$$\frac{\partial Q}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \{ \sum_{n=1}^{N} \gamma(z_{nk}) \log \frac{1}{(2\pi\epsilon)^{1/2}} \exp(-\frac{1}{2\epsilon} ||(x - \mu_k)||^2 \}$$

$$\frac{\partial Q}{\partial \mu_k} = 0 \Longrightarrow \mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

GMM and EM

And when you do the differentiation herewith the simplified normal distribution you again get the same formula but the only difference is that this responsibility is defined in the way it was earlier but what is it what is it basically saying it is saying the something that you set the mean as so for the K component you take all these so all the other responsibilities will be 0so you take only the k ate component for whatever is assigned to the case component take all those data points and take the mean of those data points.

(Refer Slide Time: 25:13)



ialize $\theta^{(0)}$, Evaluate $l^{(0)} = \log p(X|\theta^{(0)})$

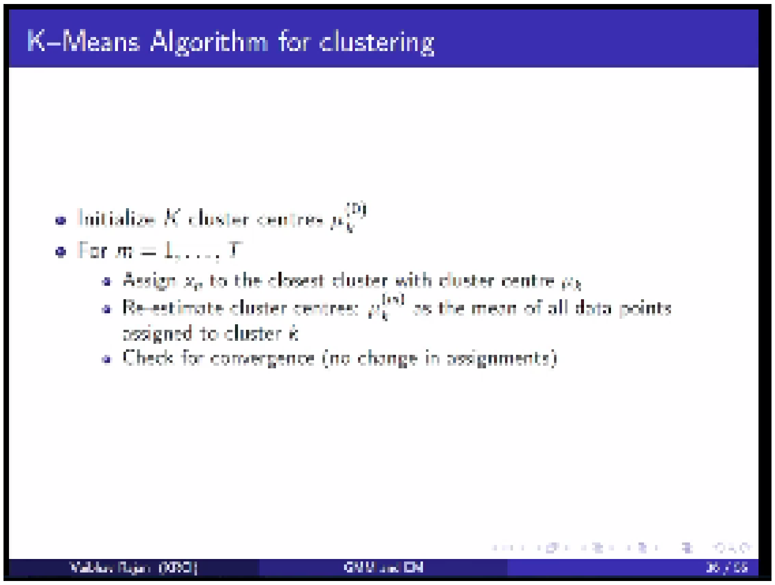
$n = 1, \ldots, T$

- $\boxed{\text{Posterior distribution of } Z : p(Z|X, \theta^{(n-1)})}$
- Expected Complete Likelihood under this distribution of Z: $Q(\theta, \theta^{(n-1)}) = \mathbb{E}_{Z|X,\theta^{(n-1)}} \log p(X, Z|\theta)$
- $\boxed{\theta^{(n)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{(n-1)})}$
- Check for convergence: stop if $l^{(n)} - l^{(n-1)} < \epsilon$

So this is the general am we first calculate the posterior distribution of Z which we saw is exactly this you assign the latent variable of X n to the closest mean and then set the new mean as the mean of all data points with the same latent variable which is just the k-means algorithm so you are assigning x n to the closest cluster with the cluster Center meek and then r as timatein the cluster centers as the mean of all data points.

That is assigned to that cluster so k-means is just a special case of Gaussian mixture where the covariance matrix is an is epsilon times the identity matrix that is why you just you just have to compute the means you do not worry about the covariance and it follows the same framework of en that we saw okay so I think we will stop here.

(Refer Slide Time: 26:31)



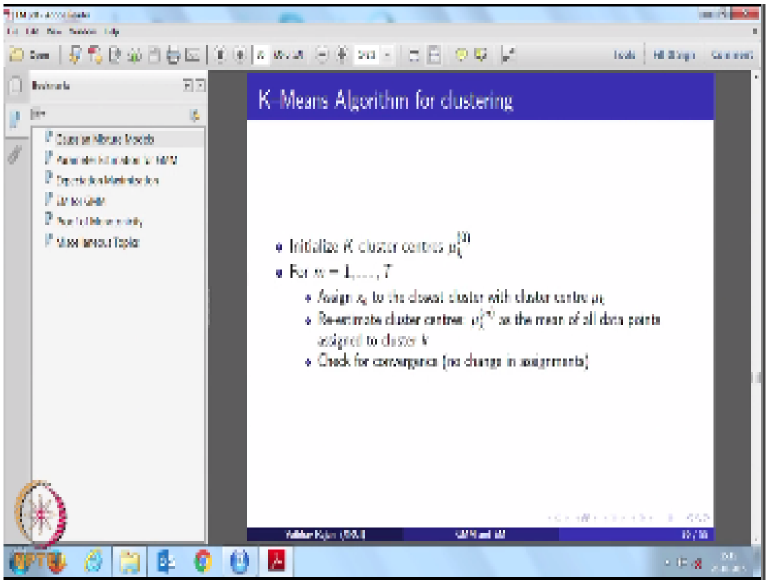
K–Means Algorithm for clustering

- Initialize K cluster centres $\mu_k^{(0)}$
- For $m = 1, \ldots, T$
    - Assign x_p to the closest cluster with cluster centre μ_k
    - Re-estimate cluster centres: $\mu_k^{(m)}$ as the mean of all data points assigned to cluster k
    - Check for convergence (no change in assignments)

Vaibhav Rajan (XRCI) GMM and EM 36 / 55

Because after this we will talk about all the theoretical properties of e m and why oh I wanted to show you one more thing.

(Refer Slide Time: 26: 44)



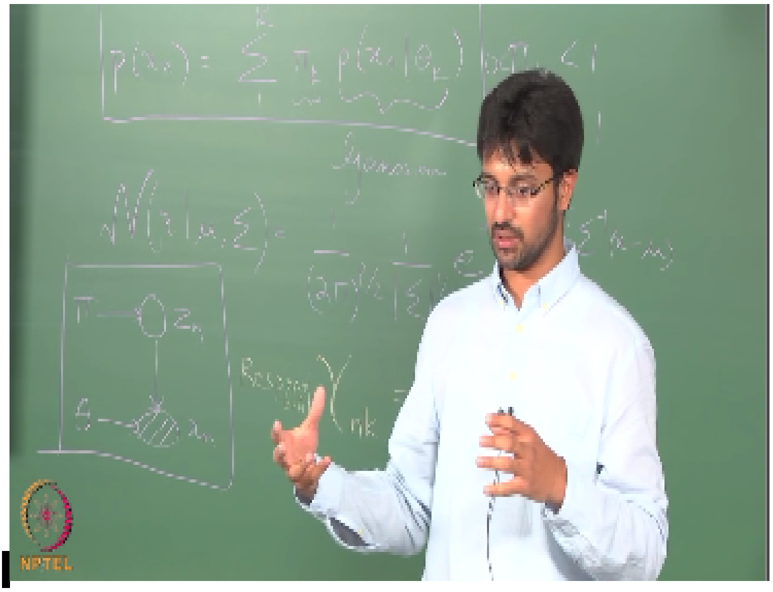
K–Means Algorithm for clustering

- Initialize K cluster centres $\mu_k^{(0)}$
- For $n = 1, \ldots, T$
  - Assign x_n to the closest cluster with cluster centre μ_k
  - Re-estimate cluster centres $\mu_k^{(t)}$ as the mean of all data points assigned to cluster k
  - Check for convergence (no change in assignments)

So let us let us take this data three it is generated from three Gaussians these are the three Gaussians and their cluster centers is what the data looks like this is the density and if we run e-m1 sonnet and plot it you see it recovers but you have to believe me in this case but it actually recovers exactly the class the means and covariance’s exactly how the way it was generated and now let's run it ten times.

And plot it so I ran e m 10 times and what I am plotting here is on the x axis we have iterations and on the y axis we have the likelihood in each iteration the likelihood keeps increasing until it reaches a point and remains steady there so this is a property of the e/algorithm which we will prove on Friday that the likelihood always increases during the iterations and now let us just look at these ten likelihoods so these are the 10 values of the when we stopped the iteration.

(Refer Slide Time: 28:54)



So usually if you if you see the slides you will see that I have always put t here which is like a hard bound on the number of iterations because sometimes the likelihood may not converge right and so we usually give a heart we give an upper bound on the number of iterations as well and stop it there so these are when I ran am ten times those are the 10 final likelihood values that I got and the minimum is four.

So in this case the likelihood values this is a very easy case so this demo is not going to work no wok so I made a mistake the reason is I gave k equal to 3 here so it is ok so you see what it has done I gave k equal to k equal to 5 so it has estimated five different components seethes are the means of the components so it has try to fit five Gaussians instead of three scared of more difficult case here so this is a more difficult case because the three components.

Are not very well separated and now if we run the e/m algorithm ten times here so you see in this case this was the data a.m. ran the likelihood always increased but what it estimated was this so when the data is very well separated like we saw earlier en will usually give very good results but when the data is not very well separated like this it starts giving very weird results but no matter what it does the likelihoods will always increase thank you.

IIT Madras production

Funded by
Department of the higher education
Ministry of the human resource department
Government of India