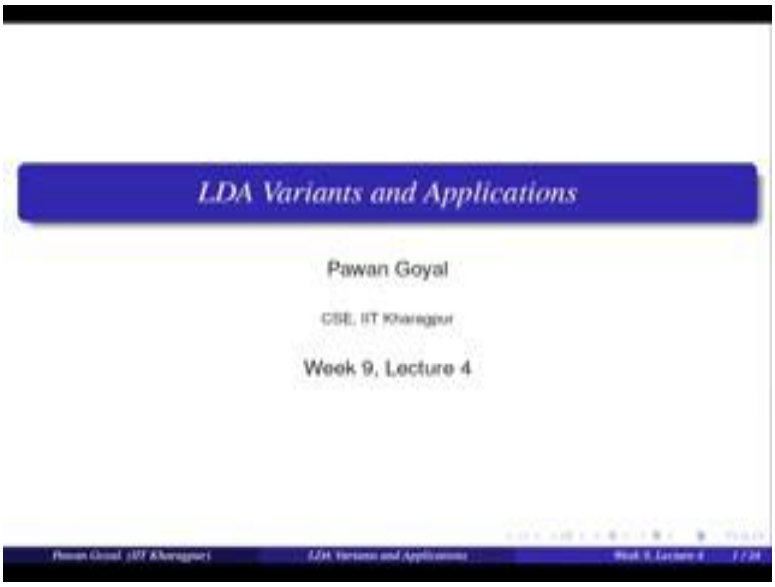
National Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 44
LDA Variants and Application – I

So welcome back for the fourth lecture of this week. So we have been talking about topic models.

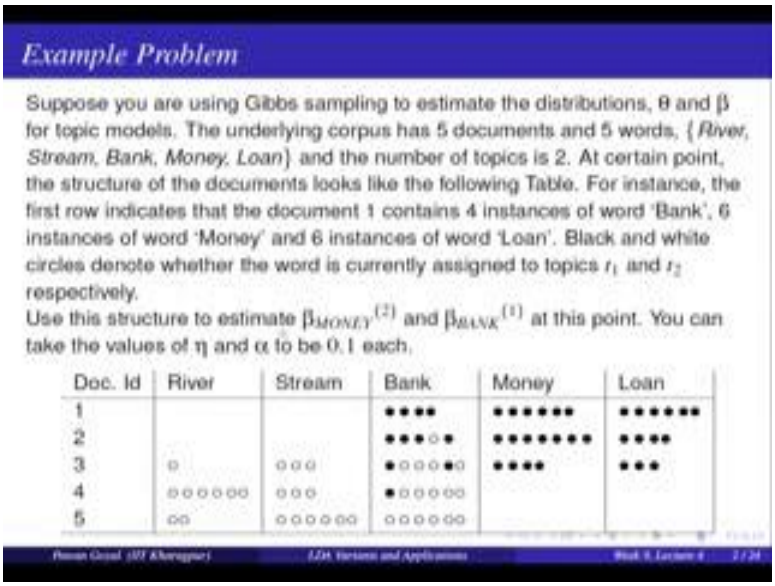
(Refer Slide Time: 00:24)



And in the last lecture we had covered what is generative model of LDA. And how do you Gibbs sampling to estimate the parameters of LDA by using the observations as a various documents and whatever works regarding the documents. So in this lecture in the next we will be talking about different variants of LDA and how do we use that for different applications. So you may not cover many of these topics in detail, but once you get the idea of what these variants are given application you can go and look more into these topics.

So this lecture I will be starting with some sort of classes size on Gibbs sampling that how do you estimate parameters from a given Gibbs sample and then we will go forward for the variants of LDA.

(Refer Slide Time: 01:10)



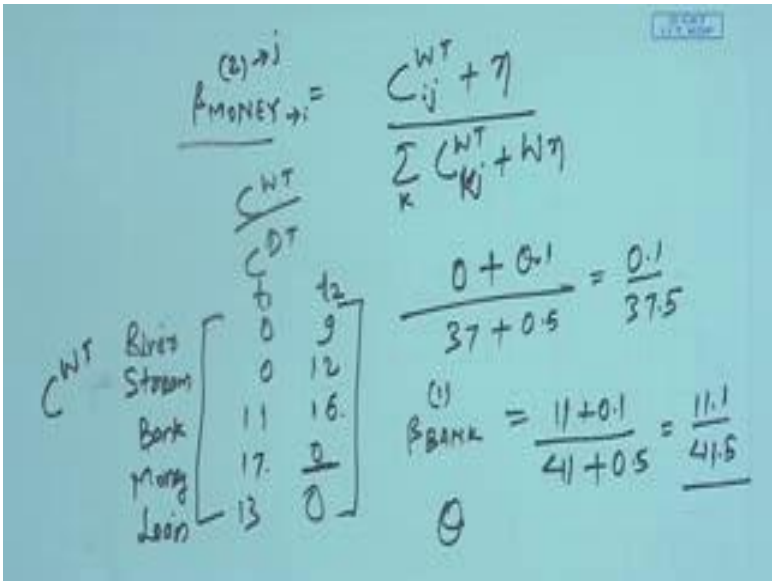
So let us take this example problem. So what we are given here. So you are saying that there is a corpus that has 5 documents, so 1 2 3 4 5 a document id is here. And 5 words river stream bank money and loan. And there are only 2 topics that you want to estimate. Now this is when you are doing Gibbs sampling at certain point of time you are given what are the different assignments of topics to different words in the document. So what do you see here document one has 4 words bank 1 2 3 4 5 6, 6 times money and 6 times loan and at that point of time all these 16 words have been assigned to topic 1.

So black is topic t 1 document 2 has most of the words assigned to topic t 1 and one word to topic t 2 and so on. You are given the topic of assignment at a given point of time. So you can see that the first show indicates that the document 1 contains 4 words for instances of the word bank 6 of word money and 6 of word loan and black and white circles are topic t 1 and t 2.

Now, your task is that you want to use the system share to estimate different parameters of your mountain. So remember what are the 2 main parameters, one was your theta another virtual beta. Beta is what is the probability of a word given a topic and theta is what is the probability of a topic given this document. So in this example we will try to estimate 2 different beta values. That is what is beta money to probability of money in topic 2 and beta bank 1 probability of bank in topic t 1 and you are given that eta and alpha are 0.1.

Now, if you remember the formula how do you compute beta money to for that you will need to use your matrices.

(Refer Slide Time: 03:10)



Remember you consider 2 matrices C wt and C dt. C wt this word assigned to what topic dt. This document what are topics that are sent. For this problem for beta money integrally what is that, what is the probability of the word money for the topic - t 2. So we will need only this matrix this word assigned to what topic in terms of this matrix. How do you write this for 2 parameters you say C wt? So if you have to write ij ith word jth topic. So let us say this is my i and this is my j. So this is ij plus you have used the hyper parameter eta divided by now you see all the different words that are assigned to this topic. Summation over k C wT, wj sorry it should be kj for all k plus w times eta. So this summation k for all w and that is why you have w eta here. That is why you estimate this parameter beta money too.

So let us see how do we estimate this parameter from this matrix. So one thing is that you will have to first construct this matrix, so let us see what does this matrix look like wt will have 5 words right. River, stream, bank, money and loan and you have to find out how many times this word has been assigned to topic t 1 and t 2 and not including this instance fine, so that we cannot do at this time. So we will take all the instances. So let us see. River is not assigned to only topic t 1 black. So river sorry river is not assigned to topic t 1 at all only the topic t 2. So I have 1 2 3 4, 1 2 3 4 5 6 7 8 9 rivers is send to topic

to 9 times. A string again 3 3 6 and 6 12 0 12 bank 1 2 3 4 5 6 7 8 9 10 11 11 times topic t 1 and 1 2 3 4 5 6 7 8 9 10 11 12 13 14 to 16 topic t 2, money only topic t1 1 2 3 4 5 6 1 2 3 4 5 6 7 13 and for 17 and loan 1 2 3 6 10 13 that is your matrix C wt.

Now, let us see how do we come to beta money, for 2 you have to compute C wt ij, i is money yes and j is t 2. So this is 0 it is not a second topic t to any number of times I write 0 plus eta h 0.1. Now divided by summation over k all the words, any all the words that when you have been assigned to topic to here. So I will just add this column 9 plus 12 plus 16. So that will give me 37 plus w number of words is 5 times eta this ones 0.5. So this comes out to be 0.1 divided by 37.5 similarly can I come to beta bank 1 for that I will find out how many times bank has been assigned to topic t 1 11 plus eta 0.1 divided by summation over k c kj wT. So that will be how many words that have been assigned to topic t 1. So it will be 11 plus 17 20 8 plus 13 41. So 41 plus 0.5, so this comes out be 1.1 divided by 41.5.

So like that you can compute all there your different betas at this given time point. So well you can compute here thetas. So this can this you can take as an exercise find out what is theta for document 1 or 2 for different topics. This is something that you can do.

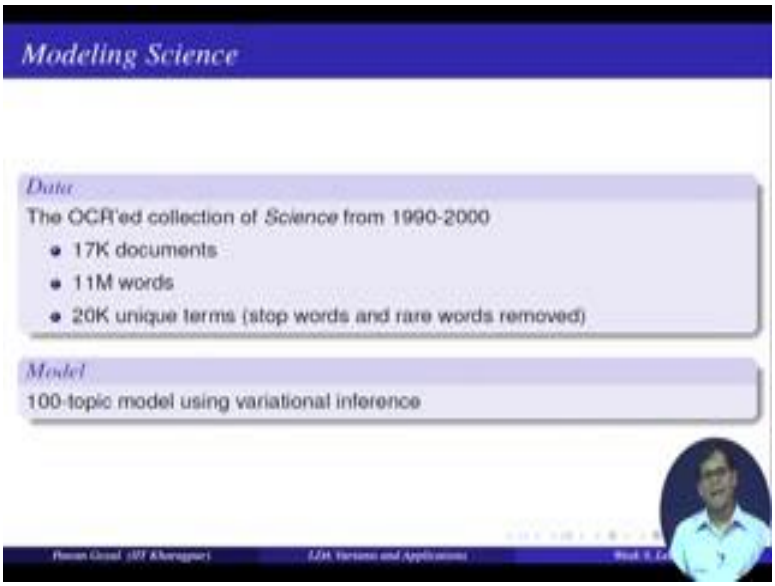
Now one more thing that that might be interesting, suppose I ask you question that find out in this iteration what is the topic that will be assigned or what will be the multinomial distribution from which you will sample a topic for a given work. Like the first bank in this document. First they are for instance is your bank for the first national bank you have to assign a new topic. That is what you doing iterations. So for that you will have to again compute different betas and thetas, but what you have to keep in mind you have to exclude the current instance. So when you are computing this you will remove the current instance.

So suppose this is 11. So you are removed then one from here and so on. You will compute each values and from by removing the conditions you will compute the betas and thetas and use your formula for find out what is the probability for topic t 1 quality to from this distribution you will sample a topic. There is something that you would keep in mind. So this was a simple example for how do you use Gibbs sampling to estimate your parameters.

Now, we talked about certain applications of LDA in the last lecture. So we saw that we can use it for computing similarity between words to complete similarity between documents that are one of some of them very promising applications, but what are some other different tasks where you can use these topic models for.

First let us see the simplest task. That is can we model the documents using the topics that is the straightforward thing that you can do using this LDA.

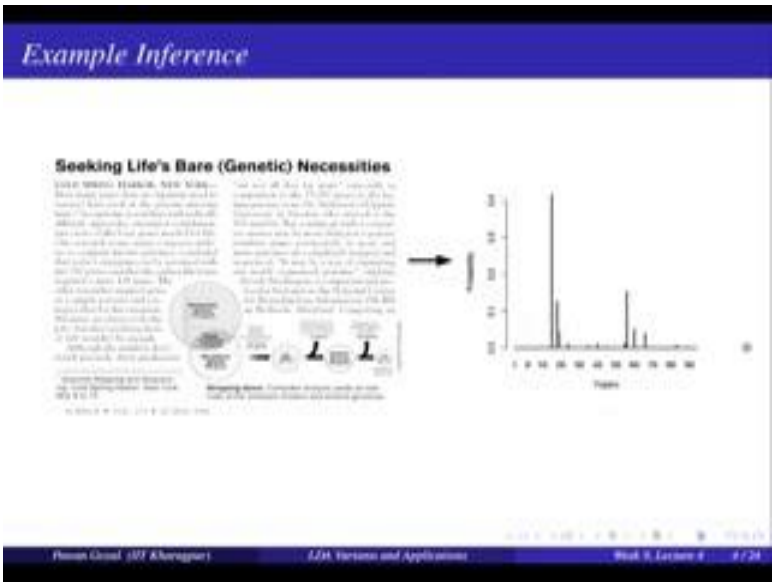
(Refer Slide Time: 09:39)



So here is something the collection that was also one of the motivation with which we started these topic models. So we are taking collection of science papers from 1990 to 2000. So there are 17000 documents and 11 million words there and there are 20k unit terms 20000 different terms after removing the stock words and real words. Now on this collection suppose you run your LDA model. So for running the LDA model you need to tell what is the number of topics.

Suppose you see it 100 topics. So once you have earned your model using your 100 topic models and you can use either Gibbs sampling or variational inference. So these are 2 different possibilities for estimating your parameters. Now once you have done that try to see what are what do your documents look like, what are the topic distributions there.

(Refer Slide Time: 10:30)



So when we do that, remember this was the article that we were looking at seeking lifes bare necessities and this for genetic in the parentheses. So we found 3 4 topics there right, compositional some data analysis some genetics evolutionary biology and so on.

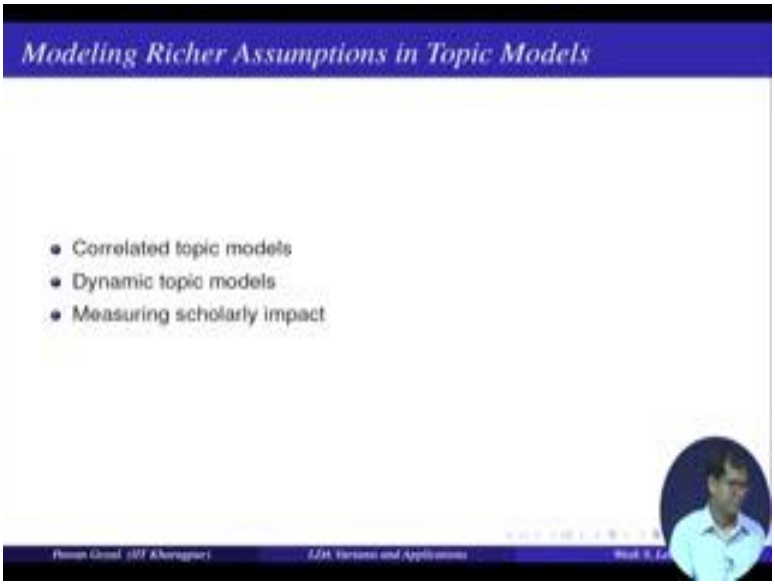
Now, suppose we run this topic model over this whole corpus we find out what happens to this document. So this document gets a probability assignment like that. So there are hundred topics and some topics get high probability. So they are few topics that are getting high probability. And then we go back and look at these 4 topics what are the most common words in these 4 topics. So we see something that we were looking for.

(Refer Slide Time: 11:08)

So we saw from first topic contains words like human genome dna genetic. So it is about genetics. Then the second topic is evolution in biology third about different disease and bacteria and forth about the data analysis. So these are the 4 topics that come on top.

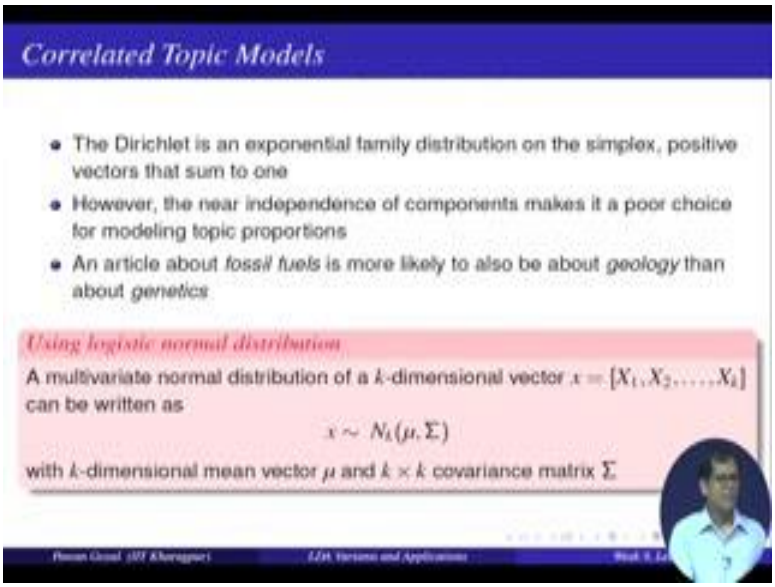
And this looks very interesting that from by you do not give any information to this model that this document contains these topics or which document contains these topics, is still by learning from a large corpus it was le able to learn different topics and the topic assignment for a given document. So this is very interesting aspect of LDA.

(Refer Slide Time: 11:46)



So, now apart from modeling simple topics that are there in the document what else can be modeled using these topic models. So we will see how do we model different other junctions in the data it. So till now what we are saying. So we have a static data. So we have a static data or whatever time it spends. So there is fix set of topics. And the topics are also kind of independent of each other. You do not say if in the document our t 1 occurs then t 2 should also occur we do not say that, but can we also model these assumptions. So for that we have different model models like correlated topic models dynamic topic models and measure measuring scholarly impact.

(Refer Slide Time: 12:33)



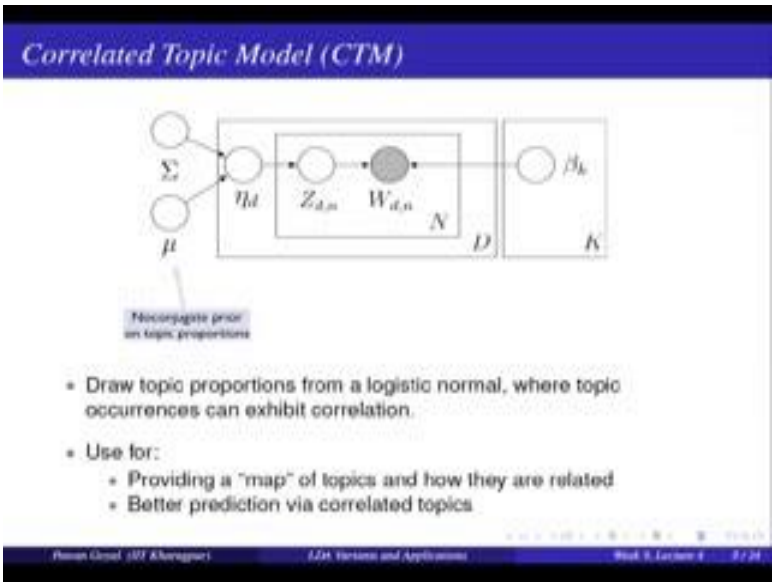
So we will see how do we go from LDA to any of this variance. So let us see the correlated topic models. So right now what we are doing? You are having additional distribution that helps me sample the probability distribution of topics for a given document. So this is what this is some simplex where there are positive vectors not in the probability that I add up to one; however, in this digital distribution the components of the probability distribution are quite independent of each other so; that means, they do not model various dependence between the topics.

So suppose I want to say that these are article about fossil fuels. And if I know the topic fossil fuels occurs in the article probably the topic about geology may also occur rather than genetics. There is something that I might know that these 2 topics are quite correlated and these 2 topics are not correlated. So can I use this intuition to battle on my

topics and distributions within the documents that certain topics are co related they will occur together certain topics are not correlated they will probably not occur together. So this cannot be modeled by using the distribution vision. So we use a different distribution to model the topics in in a document and that is where we use the multivariate normal distribution.

So something like this. So you are having k topics. So you will have a multivariate normal distribution, where you are having, your sampling this k dimensional distribution, but now from this normal distribution with a mean and covariance. So mean will be what is the private information that we have about different topics, what will be the mean of the different topics and sigma will be how are these different topics co related with each other. That is what you will try to give us in your model.

(Refer Slide Time: 14:25)

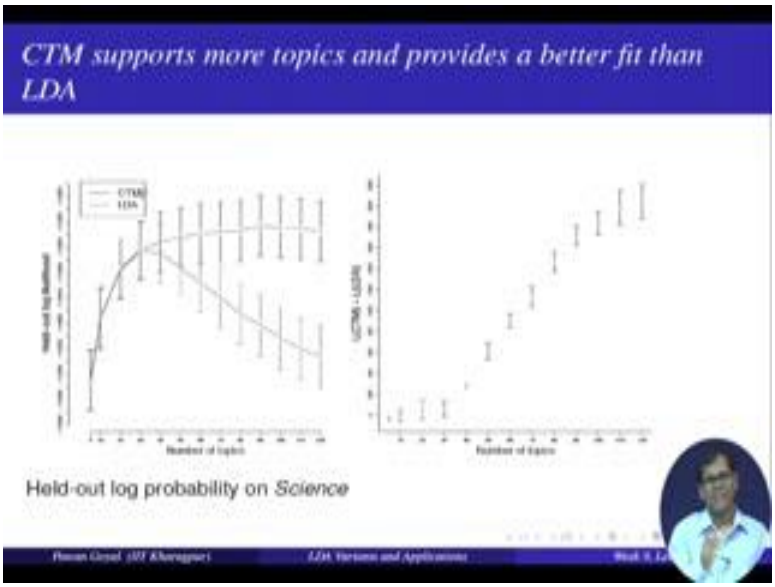


So how does your model change? So everything remains the same, except that instead of sampling from additional distribution you are now starting a sample from a multivariate logistic normal distribution with a mu and sigma. So eta these are samples from this distribution so that is where the topics can exhibit various correlations, that these 2 topics are correlate with each other while these 2 are not correlated.

So once you have done that finally, what you will get you will again get your k topics right like hundred topics you are doing in the case of science, plus you will also know which 2 topics or which pair of topics are correlated with each other. And this can be

very nicely used to give a map that these topics are make a make a single cluster a single group they are correlated to each other, these topics are against other no group that are correlated with each other.

(Refer Slide Time: 15:18)



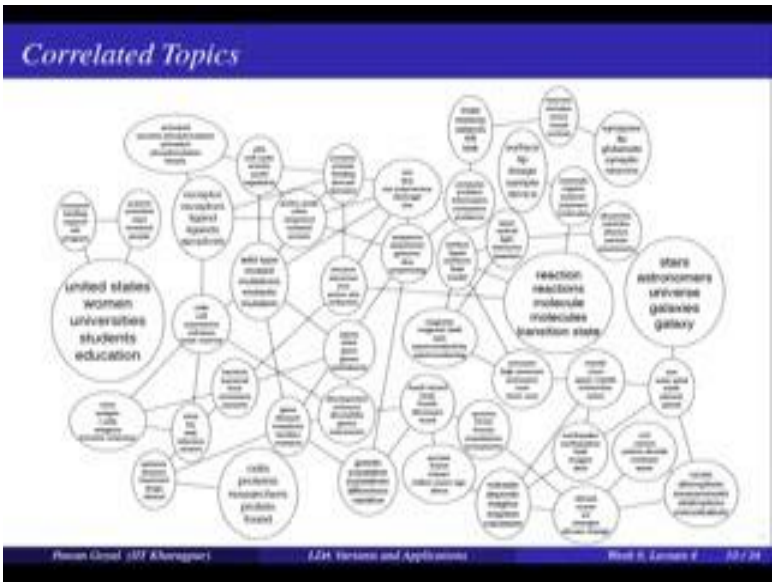
So for example, if it is and how do you know that this works better than LDA. So one good method of evaluation is that you try to find out what is the log likelihood that this providing to a held out data held out deta is some data that you did not use for training of your our topic model. So you do not give it as an input for Gibbs sampling and or variational inference. So once you have learnt the topics, try to see what probability it gets to a held out data some separate data again from a same domain. So whatever topic model gives you a better likelihood that is probably better that is a better model. This is similar to what we did in the case of language modeling.

We found out what is the perplexity that it assigns to a held out data. Similarly, here what is the log likelihood that is essentially different held out data. So we see if the colder top model gives a better likelihood then the LDA simple LDA and that is what you see here. So this is on the held out likelihood you want a number of topics. So interestingly if the number of public is small say 30 to 40 both models give the same log likelihood, but as you increase the number of topics the LDA model the likelihood given LDA model is starts decreasing, but this does not happen with CTM model; that means,

if you want to have more topics then CTM is a better choice than LDA. CTM can model reach your assumption in the data then LDA if you have more number of topics.

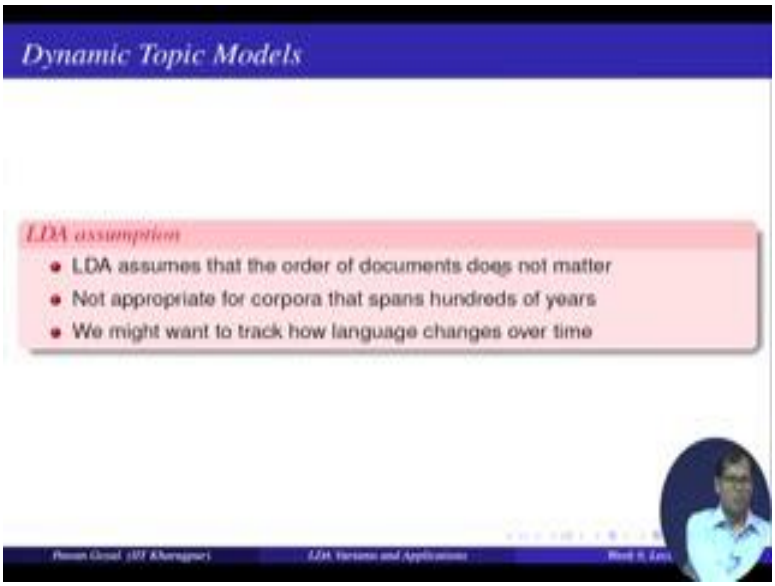
So that you can see here also likelihood difference between CTM and LDA it keeps on increasing as you increase number of topics.

(Refer Slide Time: 17:01)



So this is how the map will look like. So you will see here this topic talks about United States women universities students in education and it is about research funding support sciences scientists and research people. So these are correlated topics, but they are not so correlated to the topics like here a stars astronomers universe galaxies and galaxy. So which again make different sort of clusters for topics and this which model simply by using the covariance matrix. These topics are connected together now. So this was one assumption that we can model.

(Refer Slide Time: 17:33)



Now, suppose just take another assumption. That is which topics are correlated to each other and which topic, sorry how do topics change over thing. So right now what we are assuming, so you have a static corpus and in which there are the same topics over time the same set of topics over time working. And by topic I mean the distribution of words in the topic are also same over time, but this is not true in general. Suppose you have a collection that it spends on multiple decades or even centuries say 200 years of data.

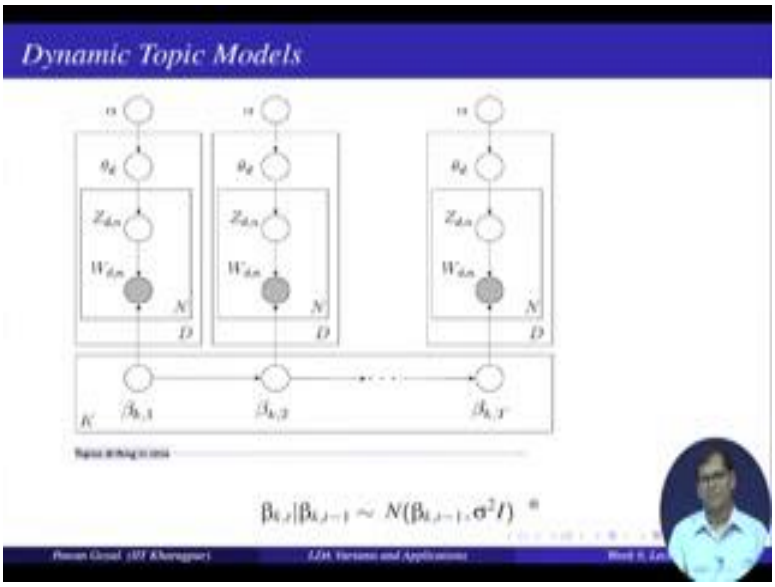
So what you will see? The same topic the number of work, the type of words that you are seeing over the time are changing initially you will see some different sort of words and later on you will see some different sort of words topic might be the same, but then the kind of words will keep on change changing. Also the probabilities of words will keep on changing. Now this you cannot model by a simple LDA model. So how do you actually specify this and that is where a dynamic topic model is used?

So what is the problem with LDA? So it assumes that the order of document does not matter and this is not appropriate for the corpus that are spending for hundreds of years. So we might want to track how the language with within the topics are changing over time and for that we use dynamic topic models. So it is very interesting this is just diet extension of LDA, but now when you model that how the topics are changing over time. So how do you do that? So when you have a large collection we will divide you into

multiple different time points. So you say this is your corpus 1 corpus 2 corpus 3 corpus 4 and so on.

Over time you are starting from one up to the last corpus. Now when you see, when you define your topic distributions you say that let us say the initial corpus had distribution beta k 1 beta k for time step 1. So what you will say as you go from time stamp 1 to time stamp 2 the next beta will not be the same as the beta 2 will not be the same as beta 1, but will be again a distribution is starting from with the mean of beta k 1 with some variance. So that is you are allowing to change the probabilities of words within the topic model. And that you can do over time. So that is the previous topic topics influence the next topics, but the next topics can also change with certain variance and this is how the model looks like.

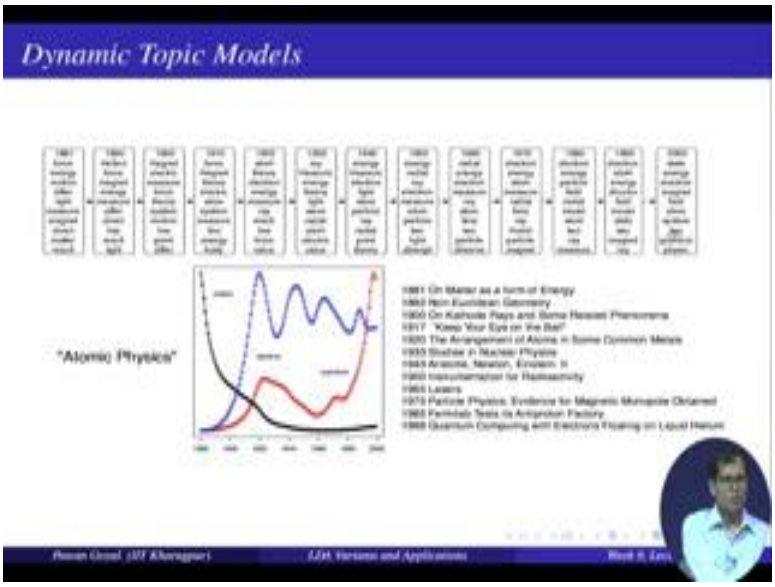
(Refer Slide Time: 20:05)



So we are having time stamps from 1 to t. So this is same as if you are having T, capital T different different copies and you are running topic model for each corpus. But now you are not doing it independently because your betas are connected. So you are saying beta k 1 is an input to beta k 2. So this is like a normal distribution beta k2 is like a normal distribution with a mean of beta k 1, but with some variance. So you are biased to take same words with same probability, but with certain variance. So that will allow changing the probability distribution of words and also having new words in the topic model.

So that is the only thing that happens. So you are having different betas over time, but they are connected it is starting from the first thing up to last time. So what you are seeing the previous time point topic models will influence the top model at the next time point. Now once you have this dynamic topic model how it can help.

(Refer Slide Time: 21:06)



So suppose you are modeling how in science a particular topic is changing over time.

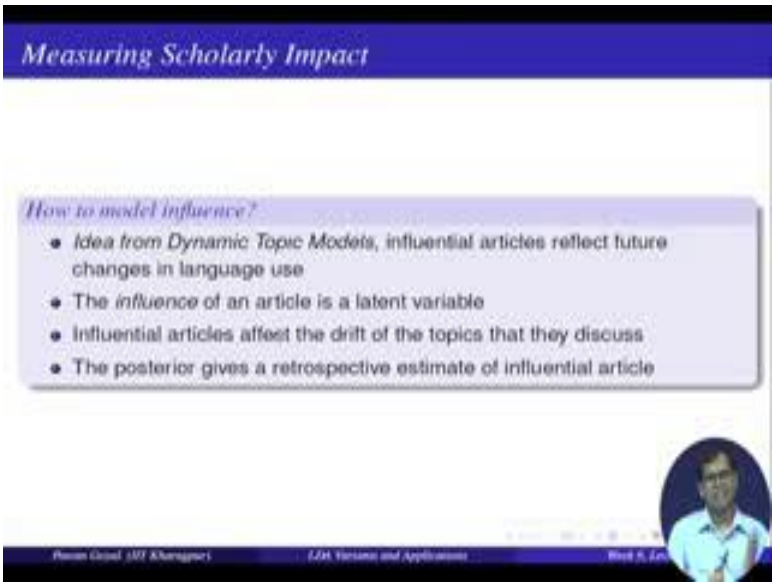
So let us say this is modeling of science is starting from at 1881 to 2000 and the topic is atomic physics. So what you see the kind of words that are there in the topic keeps on changing the over time. So here you have words like force, energy, motion, differ light, major magnet, direct matter and result, but as you go over time the words here are energy electron magnet field atom system to quantum physics. So you see the word quantum comes up. And the word electron comes up that are not there in the initial time point. So this you can see also over the time and this is a nice plot that shows how this 3 words vary over the decades in this in this topic.

So see, you initially start with matter having a very high probability, but then we start decreasing over the decades the word electrons come up at certain time point this is around 1900 with this cathode rays experiment and is start, this is like a stable like on each stable over time, but then the new topic on quantum also comes up, and that is having a very high probability. So this gives you a nice visualization that within this given topic how the words are evolving over time, how is topic evolving over time.

Similarly, if you see the topic of neuroscience, you can similarly observe that initially the word nerve was having a very high probability, but over the time you get the word like neuron coming into picture and then the ca2 that is like in area, where you will the particular area ca2. And this you can also correlate with what the difference sort of papers says seminal papers that are published, that might have given rise to these terms coming up into these topics. So this was interesting that in science in the same topic how the different words keep on coming over time.

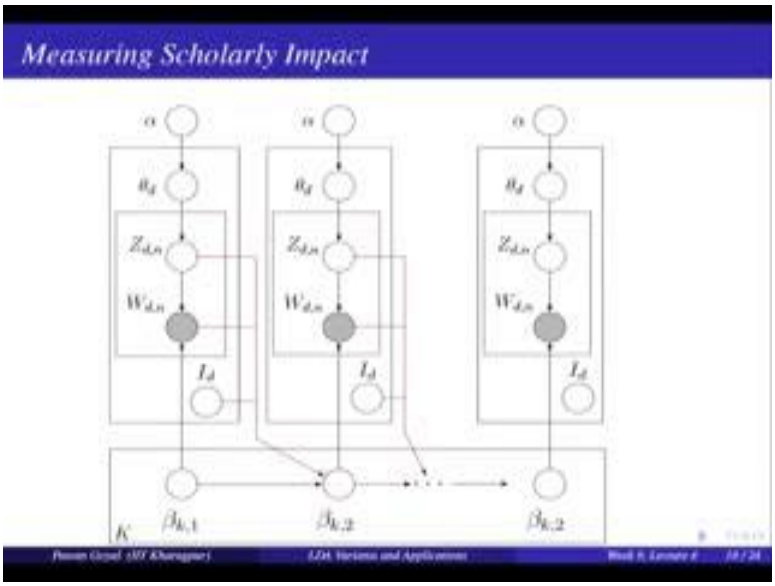
Now, this also gives rise to a nice application, that is can you model what are the most influential articles in science influential papers in science. So what will be the idea influential paper is the one that will affect the topic model. That will affect the change in the topic model. So with each document you might have an influence variable and the idea is that the change in topics are more affected by the influential vapors than the non-influential papers. And that way we can model which article is more influential in another. So how will this model look like?

(Refer Slide Time: 23:42)



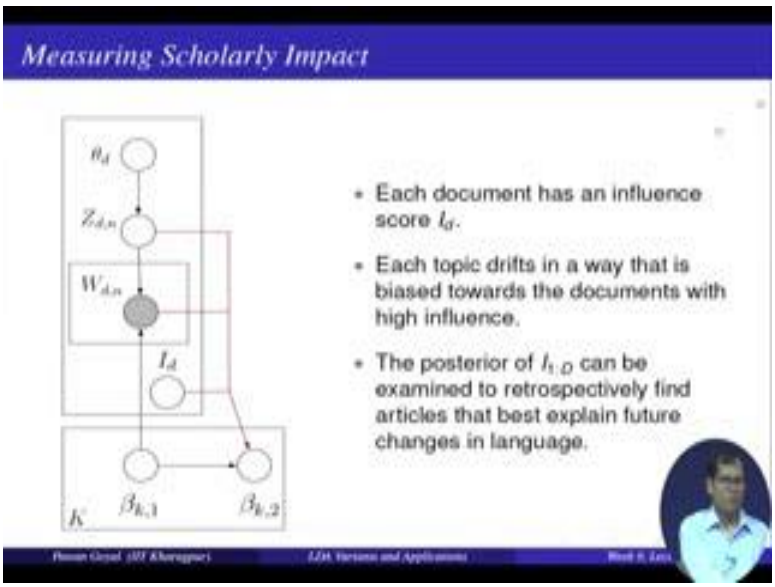
So that is influential articles reflect the future change in the language usage in the topic. And the influence of the article can be thought of as a latent variable and what we will do influential articles will affect the drift of the topics that they discuss. So that way we can model it as a variable and the posterior that will be read for this variable will tell me how influential this article was.

(Refer Slide Time: 24:12)



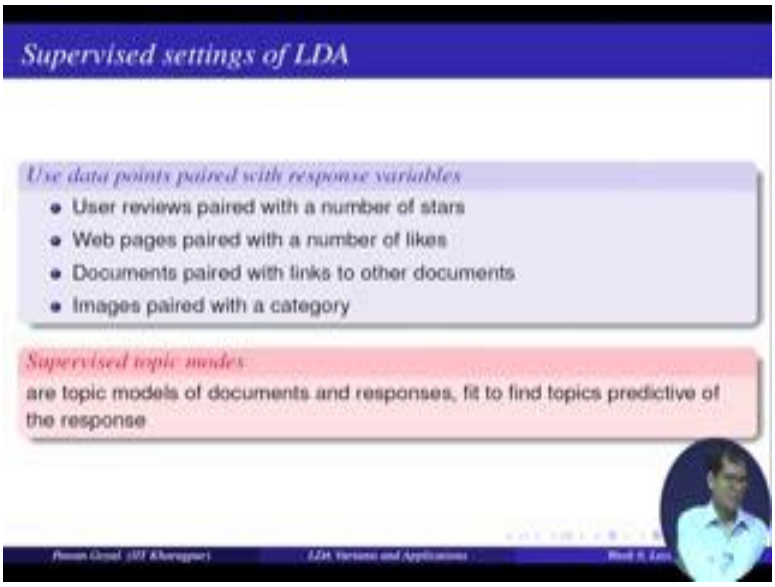
So again I make a very small change to the model. That is now beta k 2 instead of depending only on beta k 1 it also depends on this id in influence variable. So how influential this article was again depending on the topic of this article. So now, this id is there with each document and finally, while I compute the posterior I find out which documents get the highest influence. So which ever get document will get the high higher id will be the influential article. So this this will remain the same as the dynamical models only now it is also estimated by this id parameter.

(Refer Slide Time: 24:53)



So each document you see here as an influence score id, and each topic drifts in a way that it is biased towards the documents that are having a high influence. This is a posterior that I will estimate from the data. And you can explain the changes in the future.

(Refer Slide Time: 25:14)



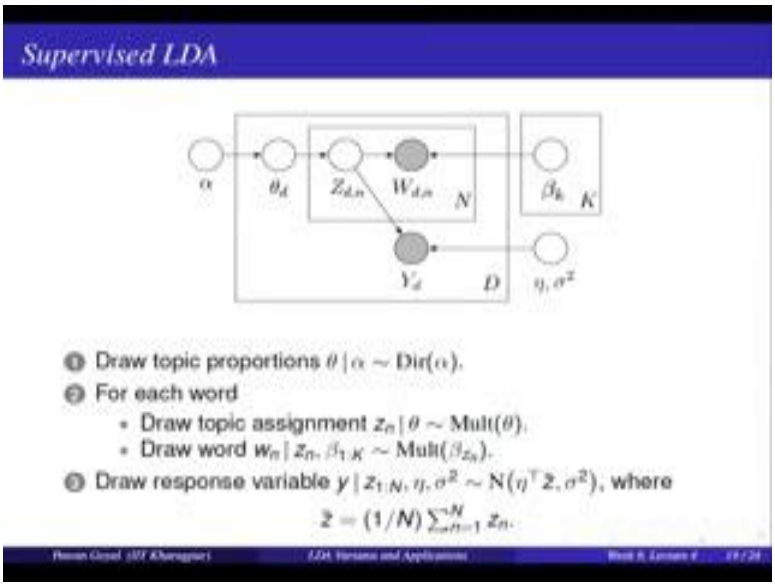
Now, let us see another very interesting variant of LDA, that is can be used in the supervised setting what do you mean by a supervised setting. So till now we are saying I have a set of documents in the document certain data occurs and I give it to my model. And by using Gibbs sampling evaluation influence I can find out what are different document which different topics occur in different documents that is what we can do and we can model certain assumptions like how topic change over timings, and which are correlated etcetera.

What you are saying now in the supervise settings, can we also use it to do certain prediction like think about the movie reviews with certain ratings. So can I use this model to say, this review will get that many ratings. Suppose in the text can I predict the ratings or web pages' link paired with a number of likes. How many likes this web page will get and the document pays with the link of other documents or individual with pair with the category. So lot of examples where the data points have some sort of class or a category, can you also model it using topic models. So this is where 2 different ways in

which it can be done we will talk about supervised topic models and see the, what is the other variation.

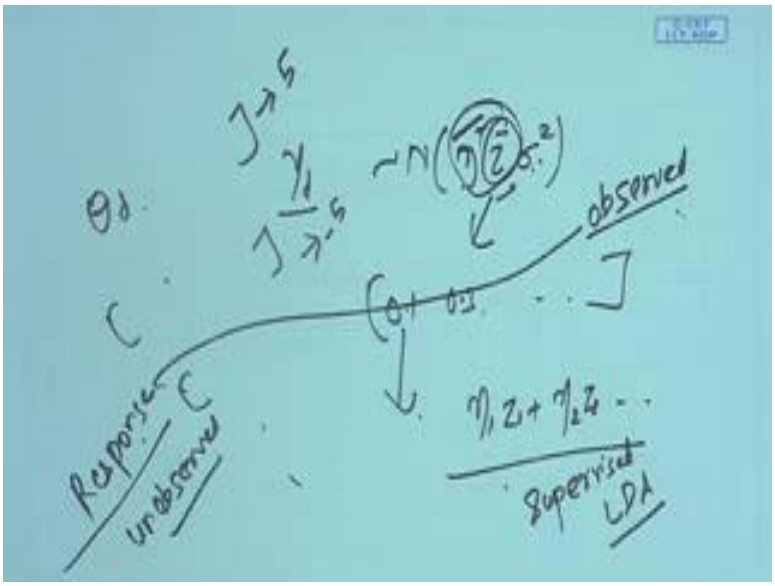
So what is the idea? So you are modeling the documents along with the responses or the categories. And the responses are those they are fitted they are fitted to find topics that are predictive of the response that is how do the topics tell about the response, so how do the topics correlate with the responses.

(Refer Slide Time: 26:56)



So how is it done? See you; this is the plate notation for the LDA model. So this is what we had seen earlier. So your beta k alpha and all that and what is additional here. So the draw topic proportional in each word what is the topic assignment plus for a given document that is your data point. So you are also finding out. So you are seeing what is the topic, distributions for the document, from there you are sampling what is your response. So from your Zdn you are sampling your response that is like a normal distribution over eta transpose z bar and a variance sigma square. So what we are seeing this response variable depends on the topic distributions with the variance. So let us just quickly see what it means is that.

(Refer Slide Time: 27:49)

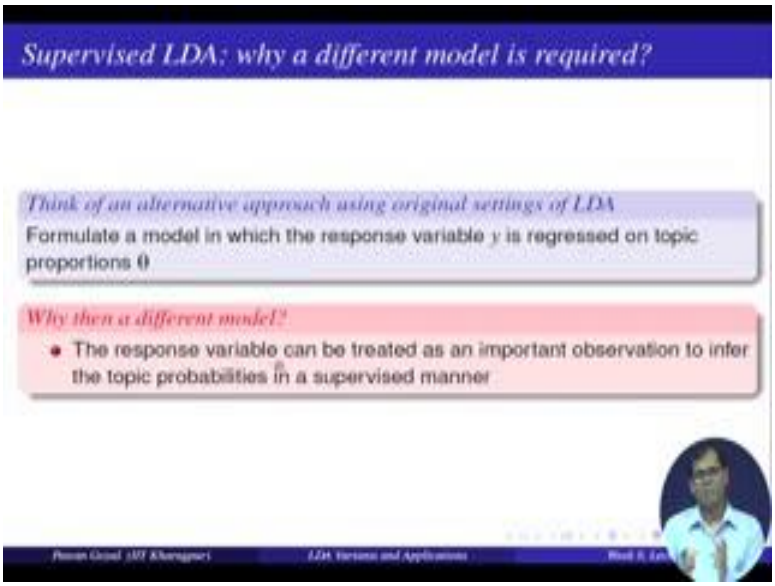


So Yd it is sampled from a normal distribution over eta transpose z bar and sigma square.

Now, what is z bar? Z bar as a topic proportions of this documents. So I know this topic t1 occurs 0.1 times t 2 occurs 0.3 times and so on. Eta like the weights given to different topics so; however, will be like if this topic is coded with the higher response and it can be negative also if there is a negative response eta can be negative. So these are like the weights given to different topics. So what whether this topic will go to a higher response or lower response. And sigma square is this will give you the mean and then you will sample the actual response with this p n certain variation.

So this will give you a scalar wait, eta transpose z plane will give you a scalar eta 1 z1 plus eta 2 z2 and so on. This will be a scalar. So this will be a link and with this way with this variance you will sample your response. So what you need to do? You need to find out what are your z prime and you need to estimate what are your etas. So you have to estimate both z prime and eta from your model.

(Refer Slide Time: 29:05)



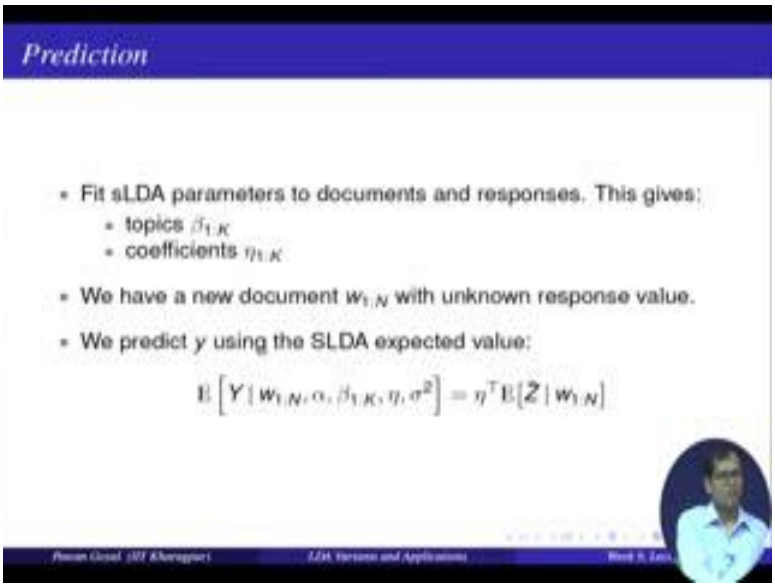
So now we were saying there are there is also an alternative to using the supervising LDA. So you can say why do we here, why are we doing taking this complicated model where we are having to estimate eta or using the model. Why cannot we run our topic model get my theta for each document theta d; that means, again we have a distribution over topics and then then there are something like a regression. So how does this theta give me certain is course - if liking of 5 another theta gives me minus 5 and so on that is another possibility. So I have different thetas for different documents and I run a regression from this theta to actually score. And this is called LDA plus regression model and what we are doing right now is a supervised LDA, where we are sampling this inside my model during the estimation time.

So what will happen here in this case, you are not using the response as an observed variable it is remains unobserved. In this case where wherever whereas, here response is also observed, so what is the intuition is it? If you take your response also as an observed variable then your topics can be much more aligned to the actual responses whereas, here the responses are not aligned to the topic topics are not aligned to the actual responses, and you have to later fit on later find out a mapping from the topics to the actual response. So this can be done here in the model itself that is why supervised LDA works better for taking rating the responses then in LDA model plus regression. So this is where you are taking the topic proportions theta and building your regression to find out

y another model. So here response variable can be taken as an important observation to infer the topics in this lowest manner.

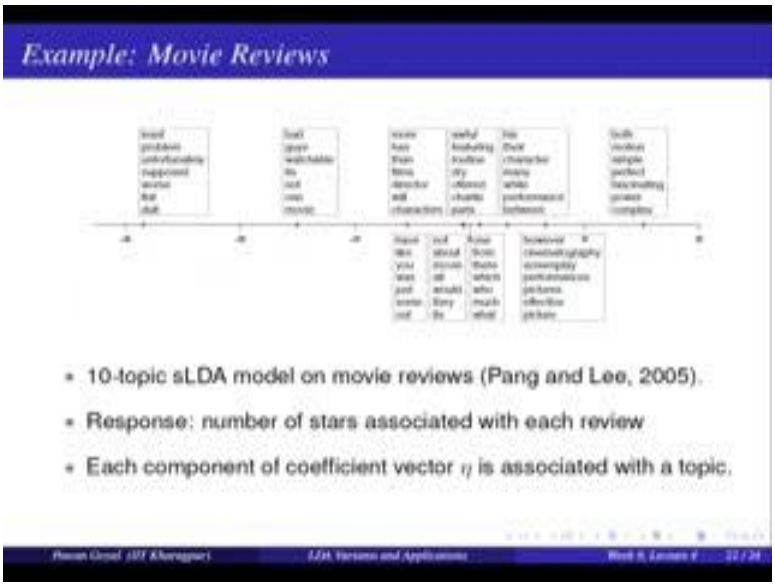
So that is the interesting thing here.

(Refer Slide Time: 31:19)



So what will happen? So we fit the LDA parameters to document responses and you will get the topics and the coefficients topics, from the coefficients. So right and using these together given a new document you can estimate what is the response eta transpose times expected value of z bar given the words in the document, this we can estimate from your a SLDA model the same way you are doing from your LDA. So what the topic distributions is for a new document and this is what you get.

(Refer Slide Time: 31:49)



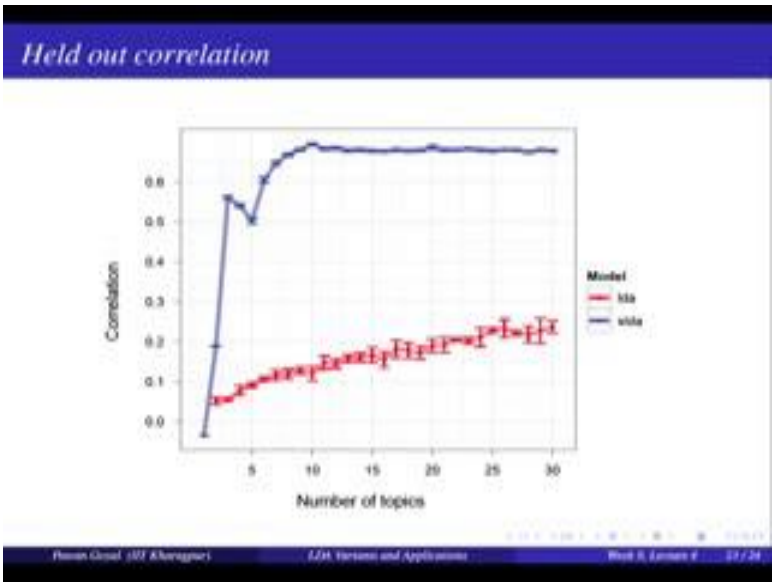
This is from the Pang and lee paper 2005.

So what there is they took 10 topic LDA model and put it on movie reviews. So you see here that 10 topics the most important words and they are plotted with their etas corresponding etas. So a high eta mean this topic corresponds to a high score or a high rating and in negative value of eta means this is called corresponding to a negative score negative rating. So what do you see here? On the higher side you have words like both motion simple perfect fascinating powering complex. So perfect in presenting a nice words that are coming with on the higher side and here you have least problem unfortunately supposed was flagged up and you will immediately. See they are like a negative image. So they are coming with a very low negative value, and this very nicely also puts your topics in a scale of minus 2 point plus that this was not available earlier.

Earlier you have different topics, but you did not know whether this topic is for a positive or negative rating.

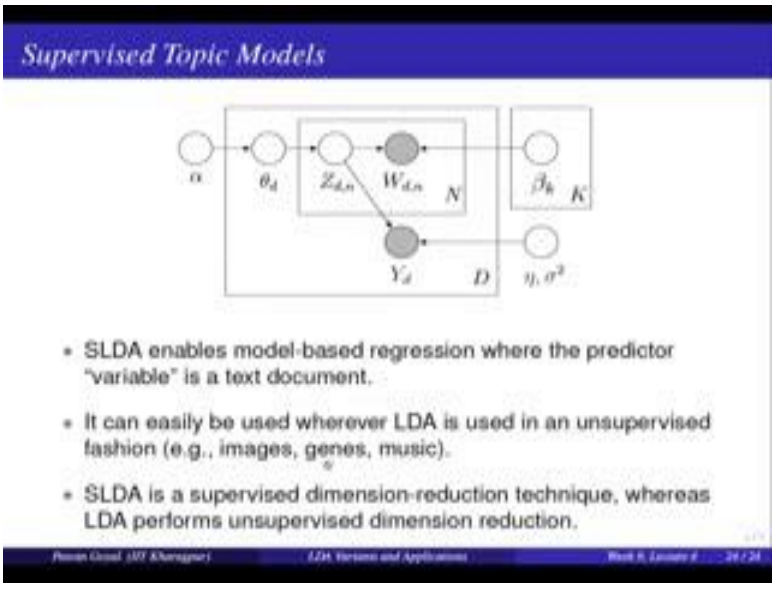
Now, you can also find out from your model itself.

(Refer Slide Time: 33:01)



And it gives a very good held out correlation also this is from number of topics; even if you increase the number of topics it gives you a much better correlation and the LDA model.

(Refer Slide Time: 33:12)



So what did you see here? So it enables model based regression where the predictor variable is a text document. So you did not have to run regression separately that is run inside the model; inside the model itself pure sampling your response by using a regression model.

Now, it can be use wherever LDA is used in an unsupervised fashion. So you can use it with images music etcetera wherever the data is paired with some sort of response variable. So you can also say that LDA is some sort of supervised dimension reduction technique, wherever is a LDA is unsupervised technique right. You are seeing the response and by seeing that you are modeling your dimensional direction. So that is about using LDA. So there are some other variants also for topic models. So we will see some of those like the relation topic models and some simple intuition about the nonparametric basic models in the next lecture.

Thank you.