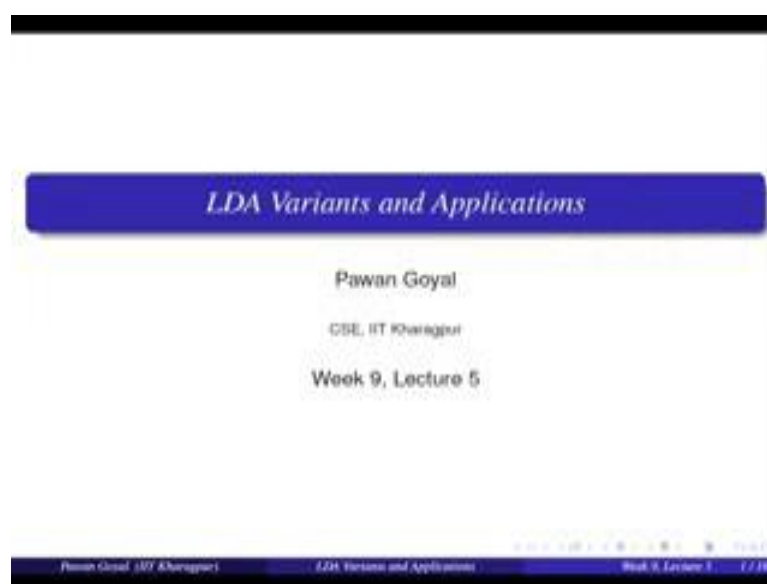


Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 45
LDA Variants and Applications – II

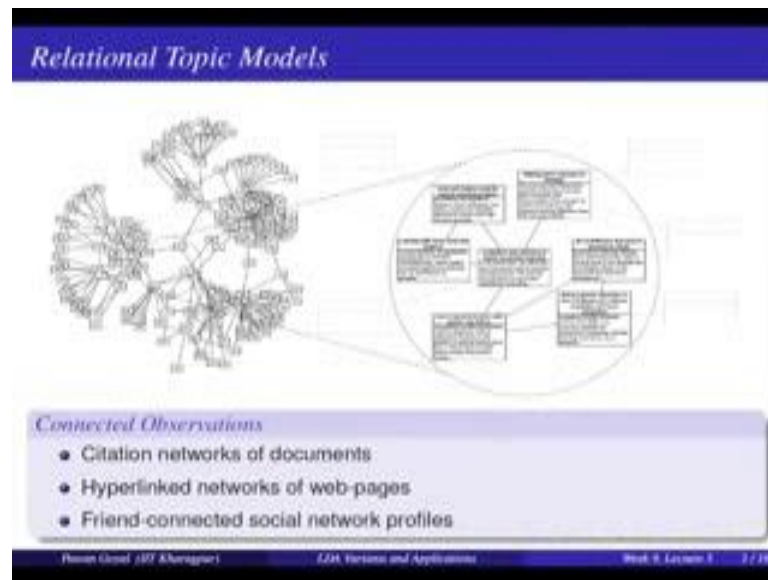
So, welcome back for the final lecture of this week. So, we are talking about different LDA variants and what are their applications. And we talked about three different variations like code rate topic models, dynamo topic models and sewage topic models. And we saw how they can model different assumptions we can also be used to pair the response as observation inside the model to make nicer topic models.

(Refer Slide Time: 00:44)



So, we now see in this lecture, we will also see some two other variants. So, one is called relational topic models, another is some sort of nonparametric Bayesian models, and you will see what sort of applications they can be used for.

(Refer Slide Time: 00:59)



So, what is the idea of relational topic models see? So, when you are talking about data, so many a times this is simple text data, so where they are not related to each other sometimes this is like a network also, so where different observations or different data points are connected together by some sort of graphic structure. So, for example, think about the scientific articles. So, as such you can think of them as separate different, different articles, this is one document, another document, different observations in fit a topic model there, or you can think of this as these are documents that are also connected.

What is the connection between the scientific documents with a citation network one paper might cite another paper. So, if it will be citing another paper, there is a link between them. So, you are seeing these observations are also connected. Similarly, think about the web page one web page might give a hyperlink to another web page. So, again these are observations that are connected. Then you can talk about friends then they are connected friends in the social network profile with the profile is kind of the document and the connection is like the so, and you are trying to model the connection.

(Refer Slide Time: 02:15)

Relational Topic Models

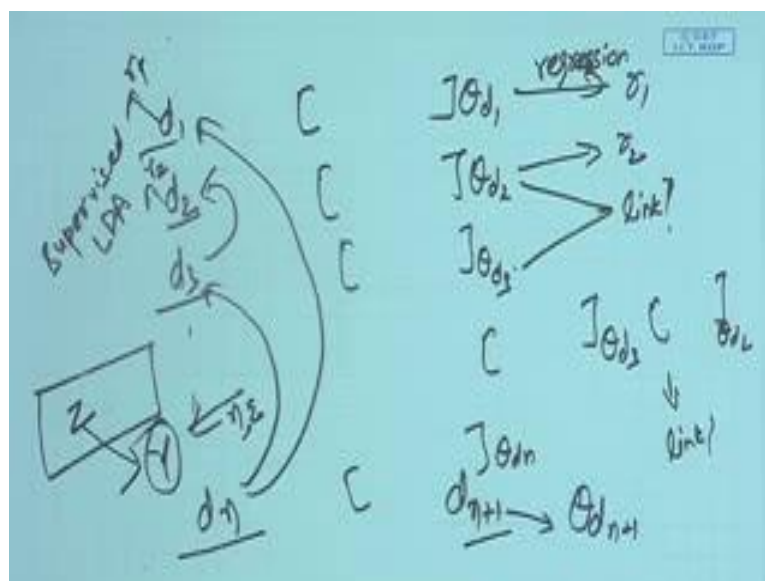
- LDA needs to be adapted to a model of content and connection
- RTMs find hidden structure in both types of data

Prasen Ghosal (IIT Kharagpur) LDA Variants and Applications Week 9, Lecture 3 3 / 16

So, now what we are seeing here that can be adapt our LDA model where such that it can take care of not only the content, but also of this link that is also contained in connection both can this be handled by the topic models. And that is where you have this variation called a relational topic models or RTMs. So, relational topic models try to take care of this variation that I have the content as well as a connection.

So, before going to the model how we will try to do that. Suppose, we did not have this relational topic models what we are trying to predict here, I have this whole set of observations, now which two observations will be connected to each other. So, let us say I am talking about the scientific articles, when a person is writing a new article, what are the papers he will cite how will I model this. So, I will think of it as a learning problem, I will say ok, I have a dataset I know there are lot of papers and some papers writing another papers.

(Refer Slide Time: 03:27)



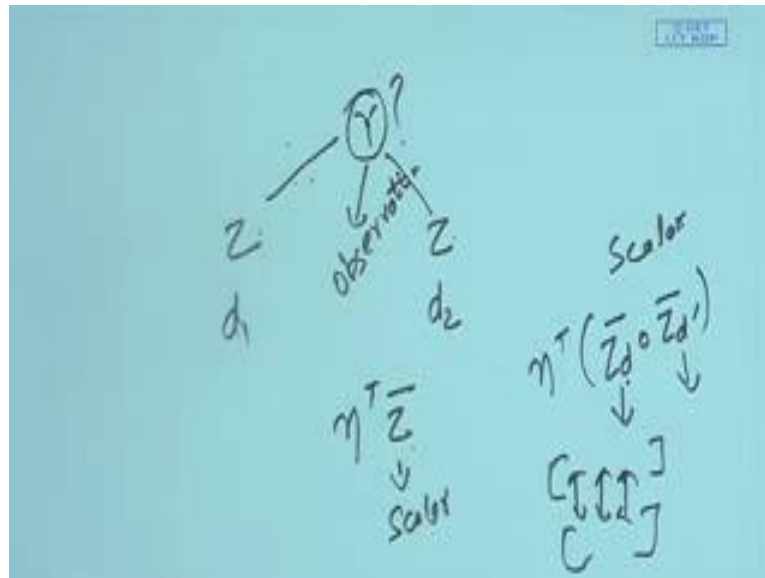
So, I will say ok, I have observations document d_1 , d_2 , d_3 and so on d_n . I know d_3 is citing d_2 , I know d_n is citing d_3 , d_n is citing d_1 and so on. I have these observations right. And I want to find out when a new able d_{n+1} comes in what will it cite, how will I solve this problem is suppose using LDA. What I will do? I will first in LDA over this corpus and I will find out this topic distributions θ_{d_1} θ_{d_2} up to θ_{d_n} . You find that out then I try to model given that this is θ_{d_2} , θ_{d_3} , will they link together or whether d_3 will link to d_2 . How will I model this?

Again you can use a regression here right. So, you can say given θ_{d_3} and θ_{d_2} predict the link, yes or no. So, I have some positive examples wherever there is link and negative examples there were there is no link. At run time, when I get a new document using LDA, I get $\theta_{d_{n+1}}$, now I try to combine it with d_1 to d_n find out which one it is more likely to linked that is one way it can be handled. Now, does that remind your something, so that will remind if the now last topic that we discussed that was supervised LDA.

Remember there also we had the same problem, we said each document has a response. So, each document might have a response r_1 r_2 and so on. So, we said ok one way could be take θ_{d_1} and run a regression from θ_{d_1} to r_1 , θ_{d_2} to r_2 that was one possibility LDA plus regression. But we said we can do something smarter than that that is we take it as an observation within my model, and we say with each document

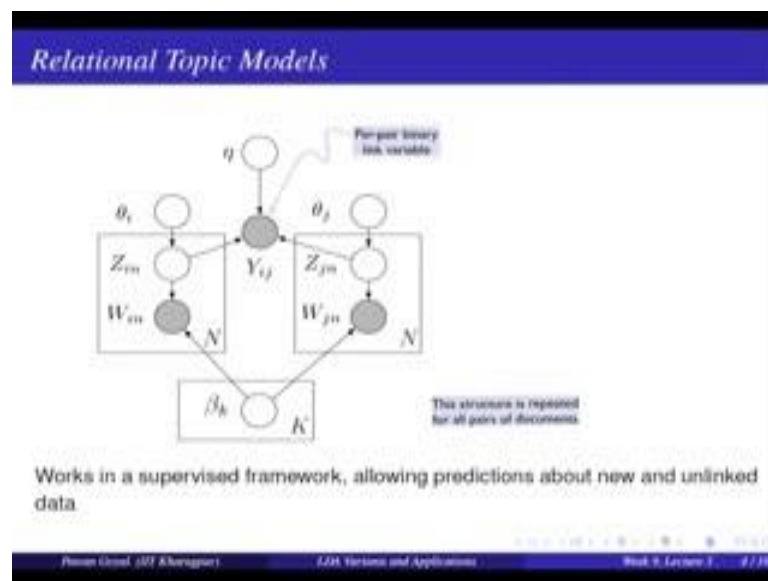
with z , you are also sampling your response variable with some η and σ . Now, can we use the same idea in this relational topic model also? So, the right now the difference is that we are working with the pairs.

(Refer Slide Time: 06:01)



So, what we can do here? We can say that I have document d_1 , I have document d_2 . Now, this will have some z , this will have some z . Now, can I model whether they will link to each other? Given their individual topics and this can now be in observation inside my model. So, this is idea, this is a simple and direct extension of this supervised LDA that can I now use pairs and model the response as my another observation sorry model the connection as my another observation.

(Refer Slide Time: 06:41)



So, this is how it is done. So, you are having the same topic models sorry same model for so beta k the topic probabilities over different observations, but now you look at different pairs. So, let us look at these pairs document d i and d z. So, they have theta i, theta j variable. So, what your modeling given the z i n and z j n, their individual topic probabilities whether they will link to each other. So, there it is per pair this is a binary link variable; and this is indirect analogous to what we written in the case sewage topic models this eta times this into this plus some variance. So, this can also allow prediction about new and unlinked data.

(Refer Slide Time: 07:31)

Link Prediction Task using RTM

Given a new document, which documents is it likely to link to?

Markov chain Monte Carlo convergence diagnostics: A comparative review
Minorization conditions and convergence rates for Markov chain Monte Carlo
 Rates of convergence of the Hastings and Metropolis algorithms
Possible biases induced by MCMC convergence diagnostics
 Bounding convergence time of the Gibbs sampler in Bayesian image restoration
 Self representative Markov chain Monte Carlo
 Auxiliary variable methods for Markov chain Monte Carlo with applications
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation
 Diagnosing convergence of Markov chain Monte Carlo algorithms

RTM allows for such predictions

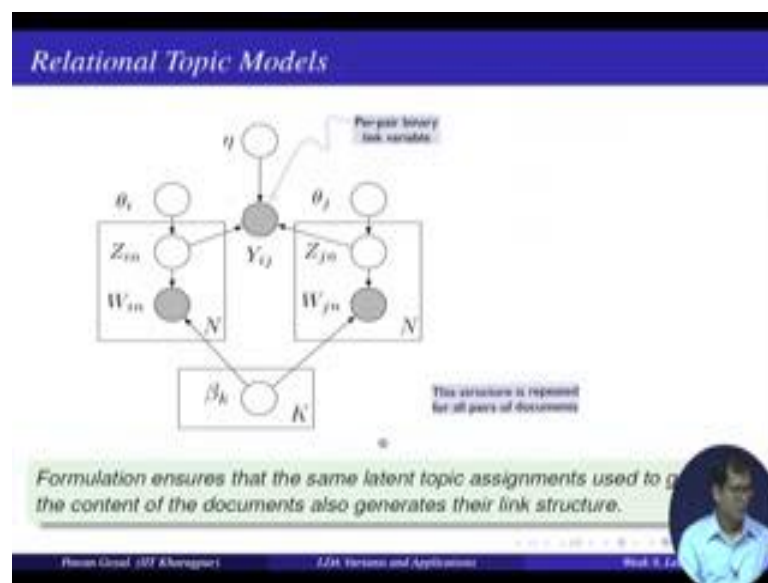
- links given the new words of a document
- words given the links of a new document

Phonon Cloud (MIT Edinburgh) LDA Variants and Applications Week 9, Lecture 3 9 / 18

So, what are the examples, so how it will be used to something like this? Suppose, I have a top paper Markov chain, Monte Carlo convergence diagnostics a comparative review. So, and I want to find out what are the other papers it will link to or I can say suppose I am reading this paper, what are the other paper that are relevant to it like a recommendation problem. So, I can use this idea let what is the other papers that it will link to win my RTM model. So, here is some examples that this is what your RTM model gives and these are actually important papers that this document actually linked to.

So, RTM in general can allow the predictions that given a document with new words what are the document it will link to this is one thing that can be done. Also suppose you give a link that this document gives links to another document, what will be the approximate distribution of the document that also can be found using my RTM model. So, links given the new words of the document and words given the links of a new document both can be handled.

(Refer Slide Time: 08:34)



So, coming back to the model, this formulation what it ensures is that the same latent topic assignment is used to generate the content of the document and it also generate their linked structure. So, you are not separately learning the linked structure, you are learning it at the same time when you are learning your topic distributions, so that is

interesting in this RTM model. So, from this Z_i and Z_j , you are trying to predict what will be your Y_{ij} .

(Refer Slide Time: 09:10)

RTM: Generative Model

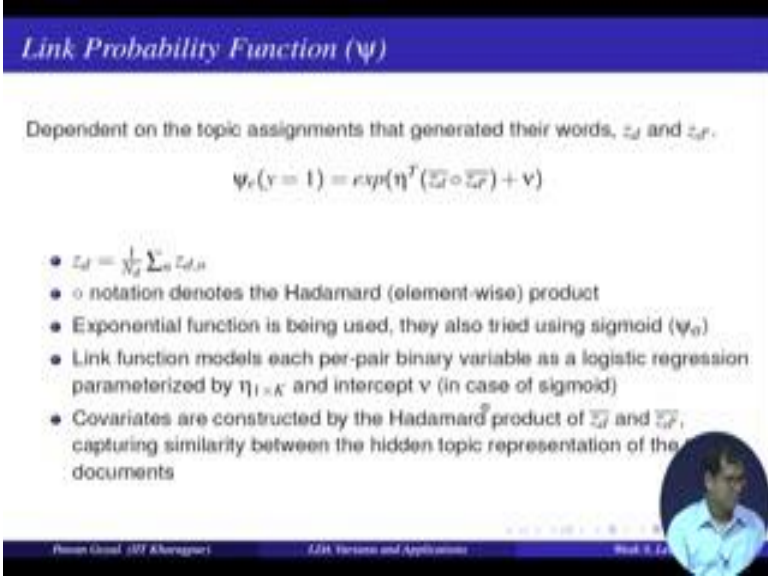
1. For each document d :
 - (a) Draw topic proportions $\theta_d | \alpha \sim \text{Dir}(\alpha)$.
 - (b) For each word $w_{d,n}$:
 - i. Draw assignment $z_{d,n} | \theta_d \sim \text{Mult}(\theta_d)$.
 - ii. Draw word $w_{d,n} | z_{d,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{d,n}})$.
2. For each pair of documents d, d' :
 - (a) Draw binary link indicator

$$y | z_d, z_{d'} \sim \psi(\cdot | z_d, z_{d'}).$$

Pranav Goyal (IIT Kharagpur) LDA: Variants and Applications Week 5, Lecture 3 3 / 18

Now, here is one problem. So, this is your simple model. So, this is what your LDA model is draw your topic proportions then for each word draw the topic assignment and the draw the word. Now, for the relational part, for each pair document d and d' prime what you are doing you are trying to sample the variable y , whether they will link or not. And this depends on the z_d and $z_{d'}$ prime, and what is the function it is a ψ of given z_d $z_{d'}$ prime.

(Refer Slide Time: 09:43)



Link Probability Function (ψ)

Dependent on the topic assignments that generated their words, z_d and $z_{d'}$.

$$\psi_v(y = 1) = \exp(\eta^T (\bar{z}_d \circ \bar{z}_{d'}) + v)$$

- $z_d = \frac{1}{N_d} \sum_n z_{d,n}$
- \circ notation denotes the Hadamard (element-wise) product
- Exponential function is being used, they also tried using sigmoid (ψ_σ)
- Link function models each per-pair binary variable as a logistic regression parameterized by $\eta_{1 \times K}$ and intercept v (in case of sigmoid)
- Covariates are constructed by the Hadamard product of \bar{z}_d and $\bar{z}_{d'}$, capturing similarity between the hidden topic representation of the documents

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 9, Lecture 9

And this is model like this. This is as an exponential function over eta transpose a Hadamard product over Z_d and $Z_{d'}$. Now, what do I mean by that? So, remember in the sewage LDA module also we did eta transpose Z bar, and this gave me a scalar - two vectors you are multiplying this giving you a scalar, you have the same dimensions. Right now here also you have eta transpose, same dimension has the number of topics and you are multiplying it with Z_d bar Hadamard product with $Z_{d'}$ bar. Now, what are Z_d and $Z_{d'}$? Z_d is what are the topic probability is doc document d and this is in document d' again their vectors. Hadamard properties is nothing but element wise product, you multiply this element with this element this element with this element and so on and then you add it, and multiply with eta transpose. So, this will again give me a scalar and plus you might add some sort of bias or valence.

(Refer Slide Time: 10:56)

Link Probability Function (ψ)

Dependent on the topic assignments that generated their words, z_d and $z_{d'}$.

$$\psi_e(y=1) = \exp(\eta^T (\bar{z}_d \circ \bar{z}_{d'}) + v)_e$$

- $z_d = \frac{1}{N_d} \sum_n z_{d,n}$
- \circ notation denotes the Hadamard (element-wise) product
- Exponential function is being used, they also tried using sigmoid (ψ_σ)
- Link function models each per-pair binary variable as a logistic regression parameterized by $\eta_{1=K}$ and intercept v (in case of sigmoid)
- Covariates are constructed by the Hadamard product of \bar{z}_d and $\bar{z}_{d'}$, capturing similarity between the hidden topic representation of the two documents

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 9, Lecture 3 8 / 18

This is a bias here. So, Z_d is nothing but the topics of probabilities in each document use it for the pair and get a Hadamard product of those and plus with bias term gives you whether the document is likely to these two pairs are likely to link or not right, eta transpose and this total thing. And in the paper they also use the sigmoid function apart from the exponential function. And link function models each per pair binary variable as a logistic regression and this is by eta and nu.

(Refer Slide Time: 11:44)

Inference: How many links to model

- One can fix $y_{d_1, d_2} = 1$ whenever a link is observed between d_1 and d_2 and set $y_{d_1, d_2} = 0$ otherwise

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 9, Lecture 3 9 / 18

So, this is your relational topic model. What is one problem with this model? So, for each pair of documents you have to define their response that is whether they are linking together or not. So, why d_1, d_2 is it 1 or 0, whenever document is giving a link to another document, you can always say to 1. But whenever document is not giving a link to another document, it is not necessary that they are not related at all, it may be that he is not giving the link or the link might occur in future there are many other things depending on the data points. So, one thing is that whenever there is a link you take it as 1, whenever there is no link you take it as an unobserved variables. So, I do not know if there is a link or not, so that also helps you in being computationally efficient, it avoids getting you so many different pairs of document.

(Refer Slide Time: 12:41)

Inference: How many links to model

- One can fix $y_{d_1, d_2} = 1$ whenever a link is observed between d_1 and d_2 and set $y_{d_1, d_2} = 0$ otherwise
- Problem with that approach is that the absence of a link cannot be construed as evidence for $y_{d_1, d_2} = 0$
- So, in these cases, these links are treated as unobserved variables
- Also provides a significant computational advantage

In large social networks like Facebook, the absence of a link between two people doesn't necessarily mean that they are not friends.

Prasen Ghosh (IIT Kharagpur) LDA: Variants and Applications Week 5, Lecture 3 3 / 18

So, that is in especially in the case of social networks, whenever there is an absence of link you cannot take it as if y_{d_1, d_2} is equal to 0. Example also you can see from your like Facebook profile. So, you are taking the profile as your documents, and you are finding out if this person will become a friend of another person. So, even if they are not friends at this time it might happen that they will be become friends in future. So, absence of the link cannot be taken for that their response should be 0. So, it can be taken as in an unobserved variable that is a better solution.

(Refer Slide Time: 13:18)

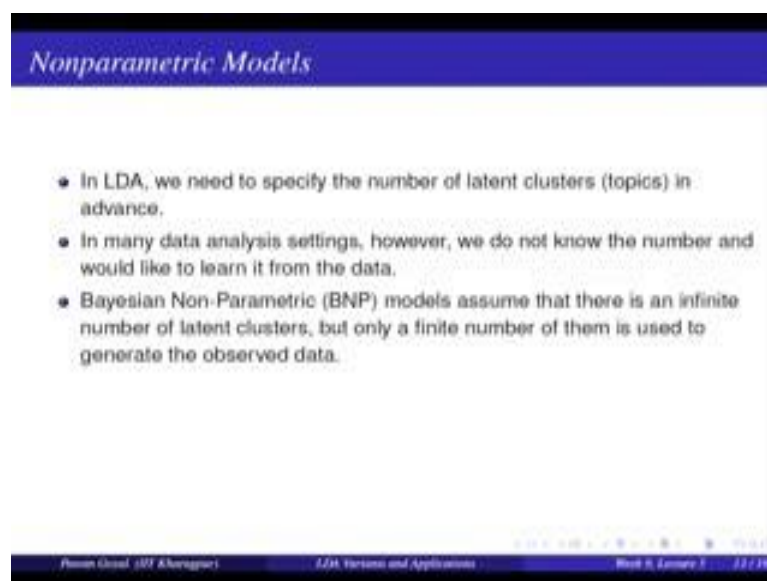
Predicting links from documents	
<ul style="list-style-type: none"> Markov chain Monte Carlo convergence diagnostics: A comparative review Minimization conditions and convergence rates for Markov chain Monte Carlo Rates of convergence of the Hastings and Metropolis algorithms Possible biases induced by MCMC convergence diagnostics Bounding convergence time of the Gibbs sampler in Bayesian image restoration Self-regenerative Markov chain Monte Carlo Auxiliary variable methods for Markov chain Monte Carlo with applications Rate of Convergence of the Gibbs Sampler by Gaussian Approximation Diagnosing convergence of Markov chain Monte Carlo algorithms Exact Bound for the Convergence of Metropolis Chains Self-regenerative Markov chain Monte Carlo Minimization conditions and convergence rates for Markov chain Monte Carlo Gibbs-Markov models Auxiliary variable methods for Markov chain Monte Carlo with applications Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models Mediating instrumental variables A qualitative framework for probabilistic inference Adaptation for Self-Regenerative MCMC 	<ul style="list-style-type: none"> RTM (6) LDA + Regression

Given a new document, which documents is it likely to link to?

Prasen Ghosh (IIT Kharagpur) LDA: Variants and Applications Slide 6, Lecture 3 18 / 18

So, this is done and then so they tested whether this words better than the model of LDA plus regression. This was the person that we saw that you take the topics from document d 1 document d 2 and learn a regression over there that is one option. Another is you take the pair which are linking as an observed variable and also use it for learning your topics inside the model itself. And that they found gives much better performance than LDA plus regression model for this task given any document what is the document it will linked. So, this is a other examples like for this document competitive environments evolved better solutions for complex task documents like coevolving high level representations were coming out to be at first position in the RTM model, but were not coming out from LDA plus regression model. So, this was something interesting that there saw from this model.

(Refer Slide Time: 14:16)



Nonparametric Models

- In LDA, we need to specify the number of latent clusters (topics) in advance.
- In many data analysis settings, however, we do not know the number and would like to learn it from the data.
- Bayesian Non-Parametric (BNP) models assume that there is an infinite number of latent clusters, but only a finite number of them is used to generate the observed data.

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 6, Lecture 3 13 / 16

Now, finally, we will wrap up this week by some small some simple discussion on non-parametric Bayesian models. So, what do we see in the case of LDA models? So, in LDA, what we are having we are defining that this corpus consists of certain topics; and then each document I find out what are the topic assignment is what I find out what is the topic assignment for that word. So, what is one problem here, I have to tell a priori how many topics that are there in the data, but that may not be an easy parameter to always give right, you may not know should I use 10 topics, 20 topics, 100 topics in this data.

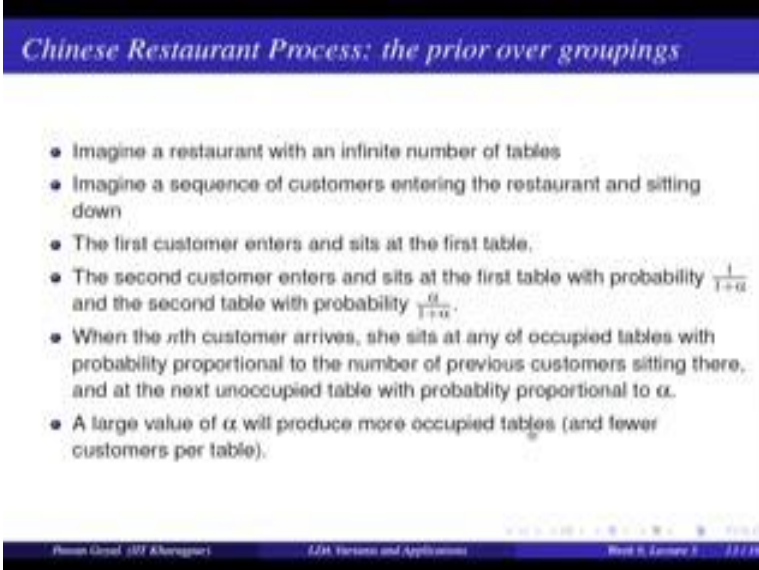
So, can my model itself find out what will be the number of topics as such and that is where we come to this topic of non-parametric Bayesian. So, this parameter goes away you do not have this parameter of number of topics and anymore. So, in many data analysis settings, we will not know what is the number of topics a priori. So, in Bayesian non-parametric models, what we assume is that there is an so we are not fixing number of parameters, but that also means my data can have any number of parameters that means, any number of topics. So, I have to a start with the assumption that there are infinite number of topics, and my data can take any finite number of them. So, it can take 100, 500, 30 whatever it means.

So, in generally I have an infinite number of topics, so that is what is interesting about it assumes that there are there is an infinite number of latent clusters or topics, but only a finite number of them is used to generate the observed data. But you do not give it as an

input to your model how many such questions we need this is what model on its own comes up when it sees the observations.

So, the posterior provides the distribution over the number of clusters, the assignment of data to clusters and the parameters of each cluster all these are different, different parameters of your LDA model that can be learned from the Bayesian non-parametric models. So, we will see so there are actually many different variations of this Bayesian non-parametric models also known as hierarchical Dirichlet processes because this is a very wide topic we will not cover this topic in detail in this course. We will just give you some intuition that how do you come up with this infinite topics in your setting, and how does it model any finite number of topics in the data even if you are starting with infinite topics. So, we will talk about one particular setting that is a Chinese restaurant process that can be used for modeling or grouping your observation into some finite number of groups the number is not told a priori. So, what is this process, it is quite interesting.

(Refer Slide Time: 17:21)



Chinese Restaurant Process: the prior over groupings

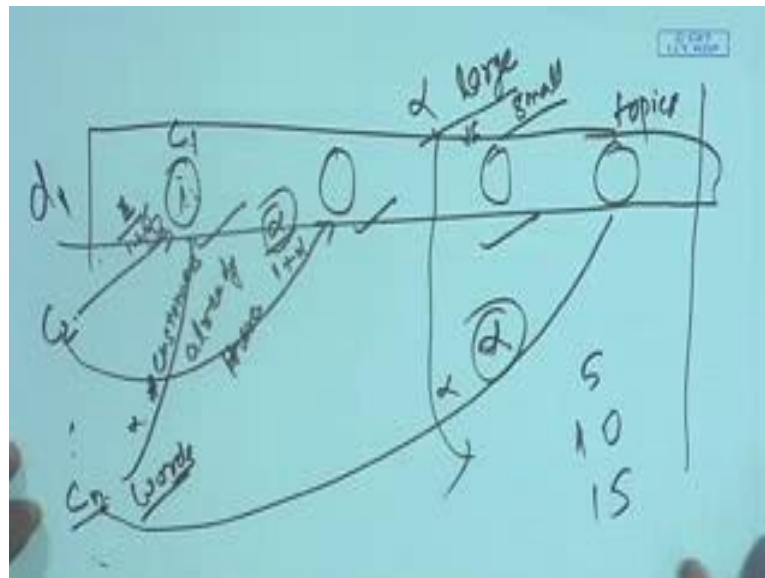
- Imagine a restaurant with an infinite number of tables
- Imagine a sequence of customers entering the restaurant and sitting down
- The first customer enters and sits at the first table.
- The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$ and the second table with probability $\frac{\alpha}{1+\alpha}$.
- When the n th customer arrives, she sits at any of occupied tables with probability proportional to the number of previous customers sitting there, and at the next unoccupied table with probability proportional to α .
- A large value of α will produce more occupied tables (and fewer customers per table).

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 6, Lecture 3 18 / 18

So, the Chinese restaurant process, so we are talking about the restaurant. So, we have saying that suppose this is a restaurant where there are infinite number of tables when you see the here the term infinite immediately what comes to your mind, these are the latent clusters. So, I have infinite tables in that restaurant. And then there are customers that are coming up. So, we are like observations. So, we are fitting the observation to top to different clusters.

Now, the model says when the customers are coming in the restaurant, how are they seated on the tables. Now, one thing that it says is that the tables are of infinite capacity. So, they can handle each table can take as many customers as you want. So, there is no limit on the number of customers that can sit on a table. So, now how do we assign customers to the tables in this process and that is where so it says ok.

(Refer Slide Time: 18:20)



So, I have some tables infinite number of tables. So, customer one is comes in and sits at the first table, table 1 fine. So, customer 1 is no problem. Now, the customer 2 comes in. So, idea is that customers 2 now sits on table 1 with the probability $1 / (1 + \alpha)$, and sits on a new table with a probability $\alpha / (1 + \alpha)$ here c is add adds up to 1. So, it can sit at either table 1 or table 1. Now, when a new customer comes in certain points of C_n , it can sit of suppose this table, this table, this table is occupied; it can sit either at the occupied tables or add a new table. So, what they say? It can sit on any of the occupied tables with a probability proportional to number of customers already there, and on a new table, with the probability proportional to α .

So, remember that is what we did here this is proportional to 1; this is proportional to α , and then we normalize we get one divided by $1 + \alpha$, α divided by $1 + \alpha$. At any given a point of time, you will have certain configuration different tables will have different number of customers. So, new customer comes in, it can sit of any other table with a probability proportional to the number of customers already there or it

can choose a new table and that is how you can fill in any number of customers on n number of tables. So, you will see at the end of this process, you can have any number of tables. So, random process you might have 5 tables filled, 10 tables filled, 15 tables filled. So, you can have any number of clusters, this does not have to be defined a priori. From your sampling, you can choose any number of topics or clusters

Now, one thing what will be the effect of the parameter alpha. Suppose, your alpha is large versus alpha is a small. If your alpha is large, what will happen? If alpha is large when a customer comes in, there is a large probability that will choose a new table so that means, with large alpha you can get large number of topics. So, is it gives you some idea. If you want more number of topics and each topic having small number of observations, then you will choose a large alpha, but if you choose a smaller alpha, you will probably choose a small number of topics because if the customer will have a higher probability of sitting at any of the tables that are previously occupied. So, that will be the effect of these alpha.

So, coming back, so first customer enters and sits at the first table; second customer enters and sits at the first table with the probability of $\frac{1}{1 + \alpha}$ and the or the next unoccupied table with the probability $\frac{\alpha}{1 + \alpha}$. And when the another customer arrives, she sits at any of the occupied tables with a probability proportional to the number of previous customers already seated on the table, and at the next unoccupied table with the probability proportional to alpha, and that is how I can distribute all the customers in a restaurant. And we saw that if I have a large value of alpha, I will have more occupied tables and fewer customers per table. So, we saw that.

(Refer Slide Time: 22:17)

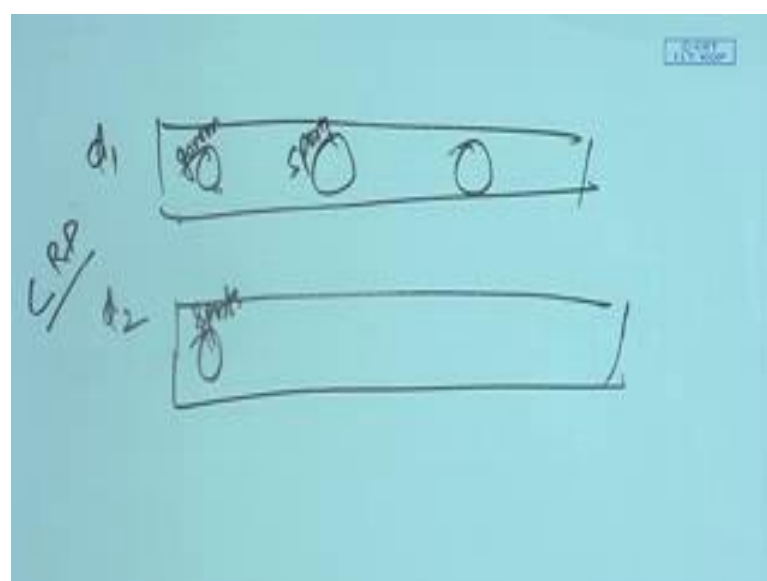
How to model a corpus

- CRP helps in obtaining a random partition as the sequence of customers sitting at tables in a restaurant.
- Tables can be thought of as 'topics' and customers as 'word' in the restaurant (document)
- CRP, however, does not model the entire corpus.
- For that, we extend CRP to a set of restaurants, each for one document.
- This model is known as Chinese Restaurant Franchise (CRF)

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 6, Lecture 3 18 / 19

Now, it looks interesting, if I have a document I can model. So, these tables as my topics and these customers as my words and I can say which words are assigned to what topics. But how do I model my whole corpus you see I can model my one this can be thought of as my one document d_1 I know what are the topics, and what are the customers. But suppose I have a corpus right that, so generally I wanted for my corpus. So, corpus means I will have multiple such documents. Now, one thing would be I fill it independently for each document, but if I do that there is no correlation between the assignment here in d_1 and d_2 .

(Refer Slide Time: 23:10)



So, it might have happened that 1, I have first table some words are filling in here like government, table two some words like a sports something. And document d 2 the word is sports might start coming here and there will be no correlation between topics here in this document, topics in this document. So, I cannot run this Chinese restaurant process independently for different documents that way I will not be able to have a correlation between these. So, what is actually done?

So, we saw that Chinese restaurant process it helps in obtaining a random partition as the sequence of customers sitting at tables in a restaurant. So, we get a random partition of words into various topics. Now, we also sit tables can be thought of as topics and customers as word in the restaurant or restaurant can be thought of as document. However, it cannot model the entire corpus for that we can extend CRP to a set of restaurants, so that is can be extended to defining each document as a separate restaurant and this is called Chinese restaurant franchise. So, now, we are going from Chinese restaurant process to Chinese restaurant franchise. So, this franchise has a lot of restaurant and that is why we will try to connect the different topics. So, let us see how do we connect the different topics.

(Refer Slide Time: 24:39)



Chinese Restaurant Franchise (CRF)

- Customer in the j th restaurant sits at tables in the same manner as in CRP, and this is done independently in the restaurants.
- But then, how do we achieve coupling among restaurants?
- The coupling is achieved by a franchise-wide menu.
- The first customer to sit at a table in a restaurant chooses a dish from the menu and all subsequent customers who sit at that table inherit that dish.
- Dishes are chosen with probability proportional to the number of tables (franchise-wide) which have previously served that dish.

Prasen Ghosal (IIT Kharagpur) LDA: Variants and Applications Week 6, Lecture 3 13 / 16

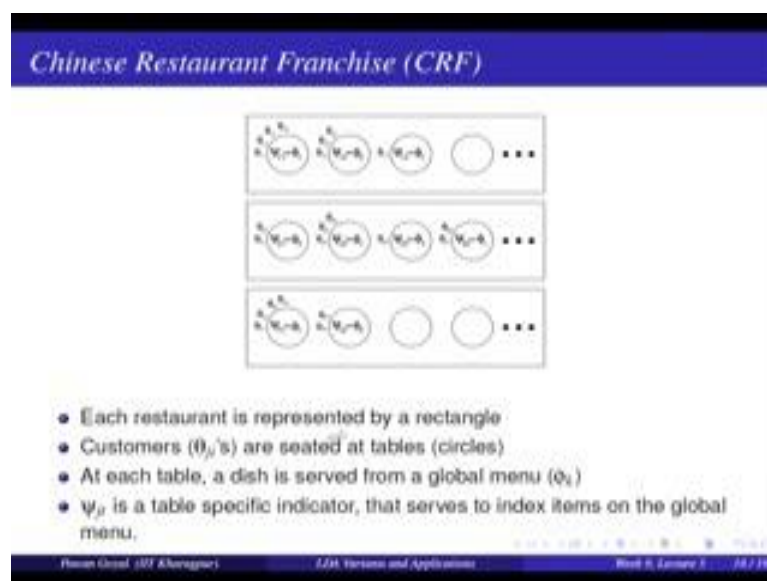
So, again so this is what is similar to CRP. So, customer in the j th restaurant sits at tables in the same manner as in CRP. So, customer comes in the restaurant, any of the restaurant now it sees which tables already occupied how many customers are there, the

probability proportional to the number of customers is sits at that table, and the probability proportional to α it sits at the next unoccupied table, so that remains the same as CRP.

Now, where is this connection among topics coming and that is coming in from the nice idea about the franchise that is let us say that they just franchise wide menu. So, now, you have a new term of menu, there is a certain dishes that are there in all the all the restaurants and that is franchise wide. So, whenever a customer comes at a given table, it orders a dish from that menu, now whoever comes in next, we will have the same dish whoever customer comes in next we will share the same dish. Now, this dish is shared throughout the restaurants. So, now, this dish at a table acts as the topic of that document. So, now, the dishes are same across the across the different restaurants and that is why you can model it the same topics.

So, how do we achieve coupling among the restaurants. So, this is achieved by a franchise wide menu. So, what we say that the first customer to sit at a table in a restaurant chooses a dish from the menu, and all the customers was sitting at that table will share the same menu. And dishes are chosen with a probability proportional to the number of tables which already have that served that dish. So, if this dishes serve at many different tables across the franchise then this will have a higher probability, so that way you are also giving you prior probability on choosing a particular dish on a given table.

(Refer Slide Time: 26:51)



So, it will look something like that. So, what you are seeing? You are seeing three different restaurants in this franchise 1, 2 and 3. Now, this ψ_{11} tells for the restaurant one what is the dish at table 1. So, this is ϕ_1 this is from the global menu of dishes ϕ_1, ϕ_2 and so on it is a global menu. The second table has ϕ_2 , third table has ϕ_1 ; that means, more than one tables can order the same dish that is possible in that model. So, why because you see, when a customer comes in it will first choose the table. So, it can either to choose one of the table that is occupied on your a right table, then when you choose the table it chooses one of the dishes so that means, it can the same dish can be chosen at more than one tables. So that means, this table also the same dish was chosen and so on.

Next, second restaurant, first table choose dish ϕ_3 , second table ϕ_1 , third table ϕ_3 and so on. So, it has certain different are choices and then the third restaurant first table choose ϕ_1 and second table choose ϕ_2 and so on. Now, you also seeing the customer assignments. So, in this first restaurant customers are coming in $\theta_{11}, \theta_{12}, \theta_{13}$ and so on; θ_{11} sits a table 1; θ_{12} sits a table 2, θ_{13} at table 1, θ_{14} table 2, θ_{18} at table 1, θ_{15}, θ_{16} here, θ_{17} here and so on. So, there are certain assignments that are given to these customers.

Similarly, here θ_{21}, θ_{22} sit here and $\theta_{23}, \theta_{24}, \theta_{26}$ sit here. Now, similar to LDA what we will have when a customer chooses a dish, it will see what the other customers have

chosen that is where they will be coupling that the words will be assigned similar sort of topics across the different restaurants. So, you are seeing each restaurant is modeled by a rectangle, each customer are seated at the tables, and each table you are serving a dish from a global menu. So, this is a table indicator. So, by modeling this, now you will know that what are the assignments of your different words to different topics in all these documents; interesting thing is that you do not have to tell how many topics do you need a priori, what is the model can itself learn from the observation.

So, what we have seen? We have seen ok, you have an LDA model you can use it for very nice applications like finding out similarity between documents, words, but suppose you want to use some rich assumption like topics are correlated changing over time. So, you can modify that model that is why it is it has a lot of then the model is very, very strong, it can use many different variations. Then you say ok, I this is unsupervised model, but suppose I want to model some responses with that that is what is the rating of this text or how many likes I can take it as an observation variable inside response variable and learn my topics accordingly that it is sewage topic model. I can go further and say which two documents are connected together that I can again measure the link as an observation, so link can be as a response or observation link two pair of documents. This was by relational topic models.

And then we said ok, suppose we do not want to use any parameters how many how many topics etcetera then you go to Bayesian non-parametric. And again there a lot of different models we talked about in very briefly about Chinese restaurant process and Chinese restaurant franchise, but they are the other variations also that can be used. And depending on your application, you can choose one of the other models and they are tools available that will allow you to model any of this. So, that brings us to the end of this ninth week where we discussed a lot about topic models, and also about semantics, token semantics.

From the next thing, we will start talking about different applications. So, we will talk about entity linking information extraction in the next week, and there we will go about other applications in the subsequent weeks.

Thank you and see you in the next week.