Topic Models

Subrata Pal

Iowa State University

October 22, 2019

Intuition: Conference Dinner

There are k tables for a dinner.

- Scheme: I sit at a table with a **probability proportional to the number of people already sitting there**.

Intuition: Conference Dinner

There are k tables for a dinner.

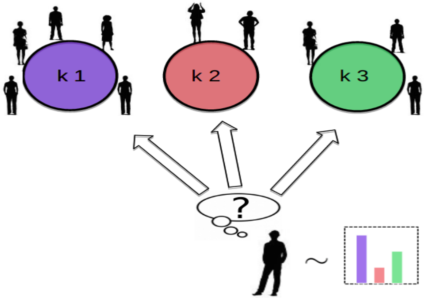
- Scheme: I sit at a table with a **probability proportional to the number of people already sitting there**.
- If everybody does the same and there are more and more people entering, the probabilities for choosing the tables converge.

Intuition: Conference Dinner

There are k tables for a dinner.

- Scheme: I sit at a table with a **probability proportional to the number of people already sitting there**.
- If everybody does the same and there are more and more people entering, the probabilities for choosing the tables converge.
- The scheme yields a sample of a *Dirichlet distribution* Parameters: initial number of participants at each table.

Conference Dinner (contd.)



Probabilities:

$\frac{5}{10}$ for k 1

$\frac{2}{10}$ for k 2

$\frac{3}{10}$ for k 3

Figure: Probability of going to k-th table is: $\left(\frac{n_k}{N}\right)$

So, $x_{N+1} \sim Mult(1, \boldsymbol{\theta}) \equiv Categorical(\boldsymbol{\theta})$.
where $\boldsymbol{\theta}$ consists of proportion of existing people at i-th place.

With an additional Parameter:

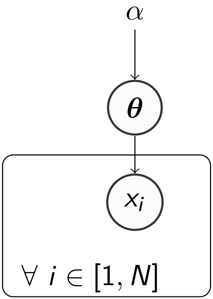
- We can make this probability/proportion 'smoother' by introducing another parameter α such that:
  Probability of going to k-th table is: $\left(\frac{n_k + \alpha}{N + K\alpha}\right)$.
- This can be re-phrased as a Hierarchical model as:

$$\boldsymbol{\theta} \sim Dirichlet(\alpha \mathbf{1})$$
$$x_{N+1} \sim Mult(1, \boldsymbol{\theta})$$

- Clearly, $\alpha = 0$ means it's the previous distribution. Greater value of α makes this like uniform distribution over all the tables.

Gibbs Sampling and Plate Notation:

The dependence can be graphically represented as:

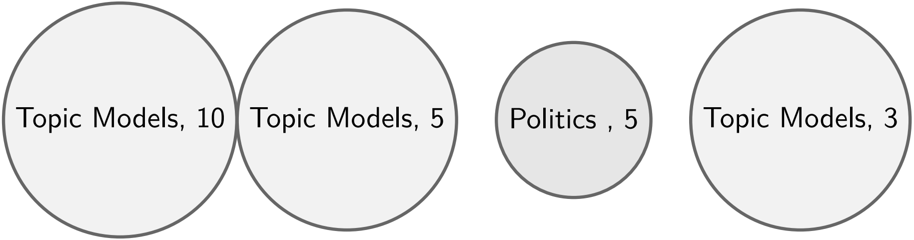


$$\boldsymbol{\theta} \sim \mathit{Dirichlet}(\alpha\mathbf{1})$$
$$x_{N+1} \sim \mathit{Mult}(1, \boldsymbol{\theta})$$

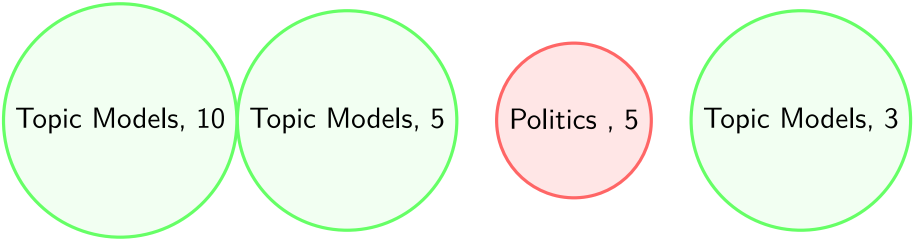
(We can do a Gibbs sampling for further inference.)

Wait!!

But, I choose table based on **number of people + the topic they talk about**.



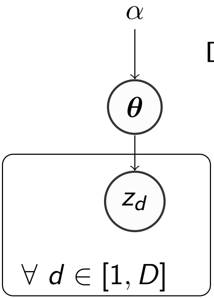
Wait!!



If I am not so interested in ‘Politics’, I may have a lower probability of choosing the table (mainly) containing ‘Politics’ as a topic even if there are same number of people in another table.

Simple Topic Model:

Generative Story-line is:



Draw a global distn over topics $\rightarrow \boldsymbol{\theta} \sim \textit{Dirichlet}(\alpha\mathbf{1})$

$$z_d \sim \textit{Mult}(1, \boldsymbol{\theta})$$

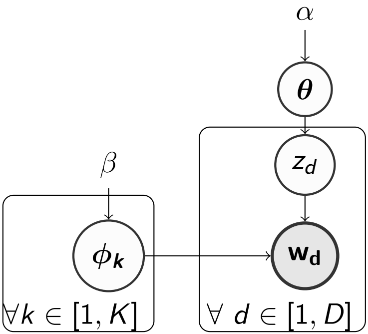
$$\uparrow$$

For each document d,

draw a topic

Simple Topic Model(contd.):

Generative Story-line is:



$$\boldsymbol{\theta} \sim Dirichlet(\alpha\mathbf{1})$$

$$z_d \sim Mult(1, \boldsymbol{\theta})$$

For each topic k, $\rightarrow$ $\phi_{\boldsymbol{k}} \sim Dirichlet(\beta\mathbf{1})$

draw a distribution over the vocabulary

$$\mathbf{w_d} \sim Mult(1, \phi_{\boldsymbol{z_d}})$$

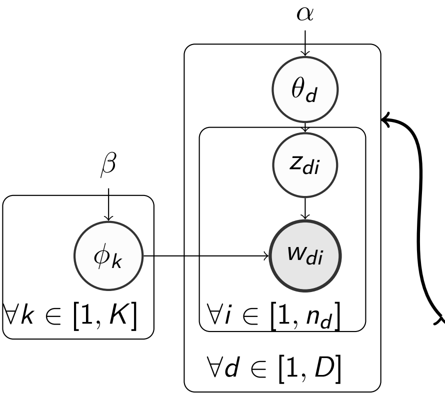
$\uparrow$

For each document d

draw the words $\mathbf{w_d}$ from the topic indexed by z_d

Latent Dirichlet Allocation(LDA):

Generative Story-line is:



$$\boldsymbol{\phi_k} \sim Dirichlet(\beta \mathbf{1})\ \forall k \in [1, K](\#topics)$$
$$\boldsymbol{\theta_d} \sim Dirichlet(\alpha \mathbf{1})\ \forall d \in [1, D](\#docs)$$
$$z_{di} \sim Mult(1, \boldsymbol{\theta_d})\ \forall i \in [1, n_d]\ \forall d \in [1, D]$$
$$w_{di} \sim Mult(1, \boldsymbol{\phi_{z_{di}}})\ \forall i \in [1, n_d]\ d \in [1, D]$$

Document Specific distribution over topics.

More complex models: Correlated Topic Models: CTM

To incorporate dependence structure inside the θ, logit-Normal distribution is used instead of Dirichlet distribution.
Only change from LDA is the **distribution of $\boldsymbol{\theta}$**'s.

$$\begin{aligned}
\phi_{\boldsymbol{k}} &\sim Dirichlet(\beta \mathbf{1}) \ \forall k \in [1, K](\# topics) \\
\eta &\sim N_{k-1}(\mu, \Sigma); \ \tilde{\eta}^T := c(\eta^T, 0) \\
\theta_k &= \frac{\exp \tilde{\eta}_k}{\sum_i \exp \tilde{\eta}_i} \\
z_{di} &\sim Mult(1, \boldsymbol{\theta_d}) \ \forall i \in [1, n_d] \ \forall d \in [1, D] \\
w_{di} &\sim Mult(1, \phi_{\boldsymbol{z_{di}}}) \ \forall i \in [1, n_d] \ d \in [1, D]
\end{aligned}$$

Thus in CTM topics are not independent, however note that only pairwise correlations are modeled, and the number of parameters in the covariance matrix grows as the square of the number of topics.

More complex models: Pachinko Allocation Model(PAM):

PAM connects words in V and topics in T with an arbitrary directed acyclic graph (DAG), where the leaves are words and topic nodes occupy the interior levels, having a distribution over its children.

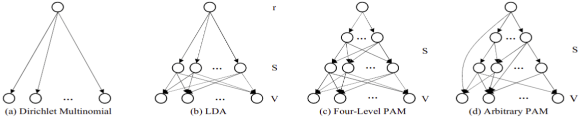


Figure: Model structures (a) Dirichlet Multinomial: For each document, a multinomial distribution over words is sampled from a single Dirichlet (b) LDA: This model samples a multinomial over topics for each document,and then generates words from the topics. (c) Four-Level PAM: A four-level hierarchy consisting of a root, a set of super-topics, a set of sub-topics and a word vocabulary. Both the root and the super-topics are associated with Dirichlet distributions, from which we sample multinomials over their children for each document. (d) PAM: An arbitrary DAG structure to encode the topic correlations. Each interior node is considered a topic and associated with a Dirichlet distribution

More complex models: Pachinko Allocation Model (contd.)

For a 4-level PAM, the first level topics are called super-topics, and the next level is called sub-topics.

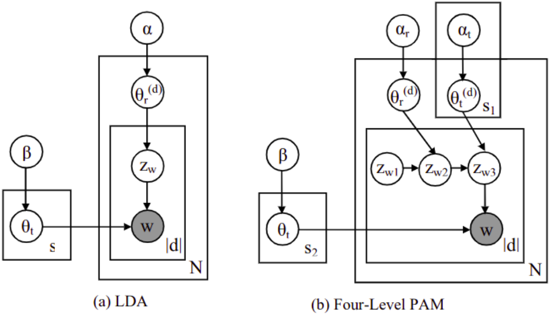
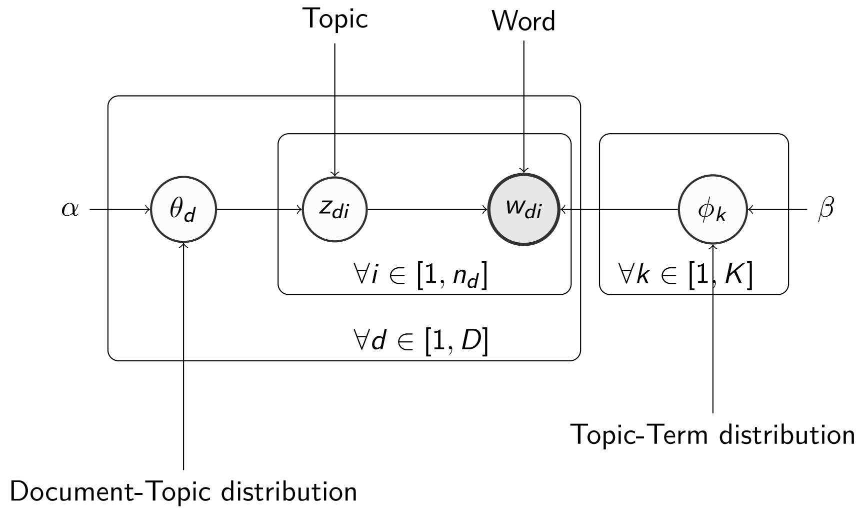


Figure: Graphical models for (a) LDA and (b) four-levelPAM

Other complex models

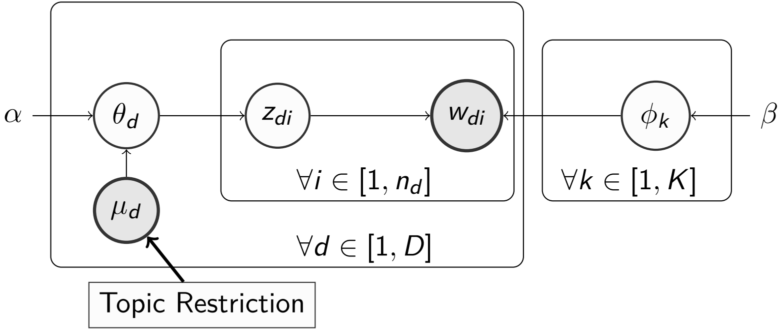
- Another extension is the hierarchical LDA (hLDA), where topics are joined together in a hierarchy by using the nested Chinese restaurant process, whose structure is learnt from data.
- Non-parametric extensions of LDA include the hierarchical Dirichlet process mixture model(It uses a Dirichlet process for each group of data, with the Dirichlet processes for all groups sharing a base distribution which is itself drawn from a Dirichlet process.), which allows the number of topics to be unbounded and learnt from data.
- Non-parametric versions of Pachinko allocations are also there.

Extensions for different scenarios: Recall LDA



Extensions for different scenarios: L-LDA

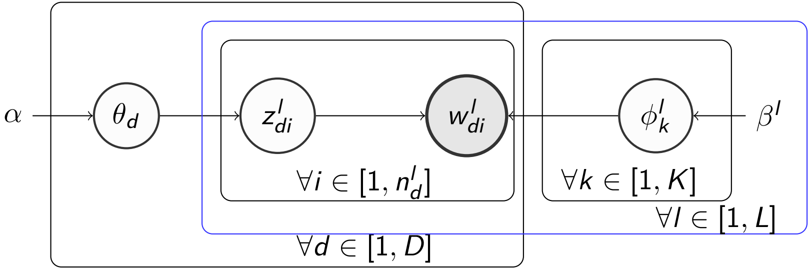
Supervised version of LDA. Data set is text document with multiple labels. Distribution over topics is restricted to document's labels.



e.g., Publications, labeled with a classification system; Tagged blog entries.

Extensions for different scenarios: Polylingual Topic Model(PLTM)

Documents are in several languages - but they should be loosely equivalent to each other. PLTM uses a *separate vocabulary for each language*, and each topic has a *word distribution for each language*.



e.g., Wiki articles in several languages; Documents that are annotated with a controlled vocabulary - creates topics for both natural language and the controlled vocabulary.

- Author-Topic Model:
  - Each author has a distribution over topics.
  - Generative story line is: For each word in a document - choose an author - choose a topic from that author's topic distribution - generate a word from that topic.

- Author-Topic Model:
  - Each author has a distribution over topics.
  - Generative story line is: For each word in a document - choose an author - choose a topic from that author's topic distribution - generate a word from that topic.
- Topics over Time
  - Topics generate both words and observed timestamps.
  - Jointly models word co-occurrences and localization in time.

- Author-Topic Model:
  - Each author has a distribution over topics.
  - Generative story line is: For each word in a document - choose an author - choose a topic from that author's topic distribution - generate a word from that topic.
- Topics over Time
  - Topics generate both words and observed timestamps.
  - Jointly models word co-occurrences and localization in time.
- Citation Influence Model
  - Estimates the weight of edges in a citation graph, i.e. the strength of influence one publication has on another.
  - Incorporates the aspects of topical innovation and topical inheritance via citations.

- Author-Topic Model:
  - Each author has a distribution over topics.
  - Generative story line is: For each word in a document - choose an author - choose a topic from that author's topic distribution - generate a word from that topic.
- Topics over Time
  - Topics generate both words and observed timestamps.
  - Jointly models word co-occurrences and localization in time.
- Citation Influence Model
  - Estimates the weight of edges in a citation graph, i.e. the strength of influence one publication has on another.
  - Incorporates the aspects of topical innovation and topical inheritance via citations.
- Variations on LDA have been used to automatically put natural images into categories, such as "bedroom" or "forest", by treating an image as a document, and small patches of the image as words; one of the variations is called Spatial Latent Dirichlet Allocation.

Some useful links:

http://www.cs.columbia.edu/~blei/topicmodeling.html

http://topicmodels.info/