Topic Models

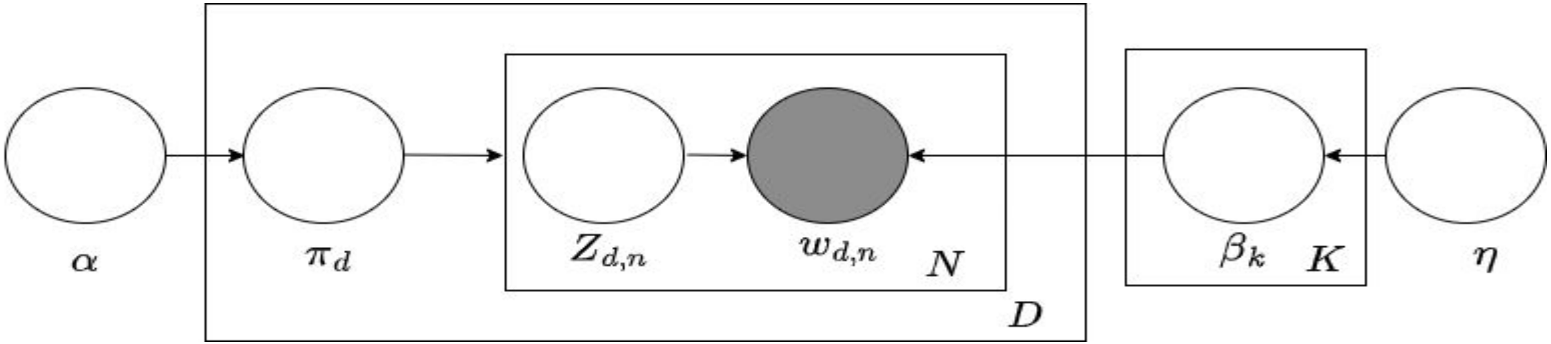
Lecture 3
Data Analysis
E0 259 - Fall 2022
Vikram Srinivasan
Indian Institute of Science, needl.ai

Why This Model?

- In PLSA, essentially modeling each document in the training set comes from a point distribution over topics
- Hence for new unseen documents, there is no way to have a generative model
- LDA addresses this by having a generative model for the topic distribution of a document (essentially instead of a point, it is a distribution over the simplex).
- This gives it way more flexibility.
- Still the parameter space is large, how do we estimate it efficiently?

LDA Inference

- α is a hyper parameter
- We need to infer:
  - Per **word** topic assignment Z (Multinomial)
  - Per **document** topic distribution π (Dirichlet - simplex with K dimensions)
  - Per **topic** word distribution β (Dirichlet - simplex with |V| dimensions)



Computing the Hidden Variable Distributions

$$p(\beta, \pi, \mathbf{Z}, \mathbf{W}) = \prod_{\mathbf{i=1}}^{\mathbf{K}} \mathbf{p}(\beta_\mathbf{i}) \prod_{\mathbf{i=1}}^{\mathbf{D}} \mathbf{p}(\pi_\mathbf{d}) \\ (\prod_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{p}(\mathbf{Z}_{\mathbf{d,n}}|\pi_\mathbf{d})\mathbf{p}(\mathbf{w}_{\mathbf{d,n}}|\beta, \mathbf{z}_{\mathbf{d,n}}))$$

Joint probability distribution from graphical model

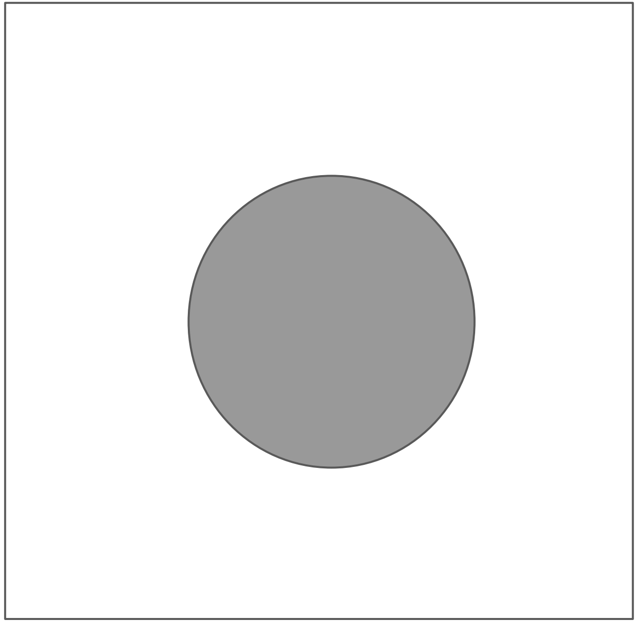
$$p(\beta, \pi, \mathbf{Z}|\mathbf{W}) = \frac{\mathbf{p}(\beta,\pi,\mathbf{Z},\mathbf{W})}{\mathbf{p}(\mathbf{W})}$$

$$p(\beta, \pi, \mathbf{Z}|\mathbf{W}) = \frac{\mathbf{p}(\beta,\pi,\mathbf{Z},\mathbf{W})}{\mathbf{p}(\mathbf{W})}$$

Posterior Distribution

Sampling Techniques

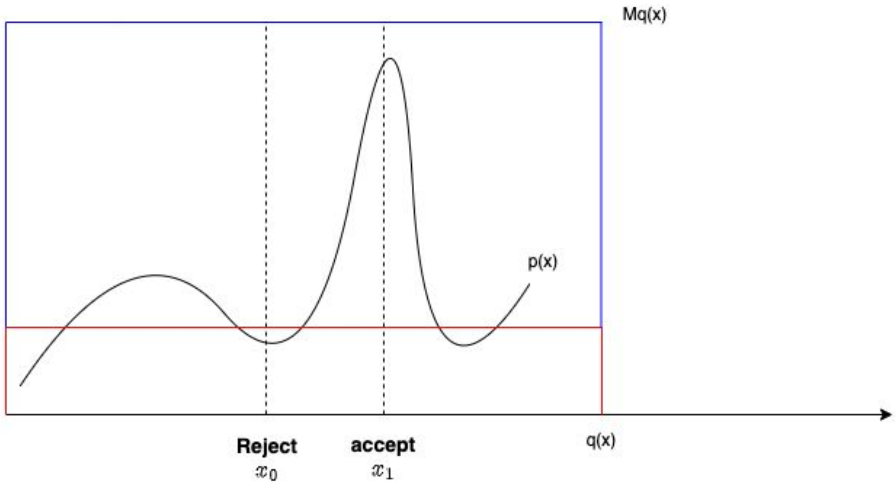
- How do we approximate complex multi dimensional distributions?
- Monte Carlo methods
- Sample from the distribution and estimate E[f(X)], where X is drawn from some arbitrary distribution?



E.g. estimate area of circle.

Rejection Sampling

- Want to approximate some complex distribution p(x)
- Want to sample high probability events more often.
- How do we know which are high probability events?
- Sample from a uniform distribution q(x)
- Accept all samples such that 0 <= p(x) <= Mq(x)



Importance Sampling

- Some values of f(X) may be unlikely and have very large values
- Expected values gets biased by these samples.
- Standard Monte Carlo doesn’t capture these well.
- Draw samples from some approximate distribution q
- Assign higher probability to “important” values
- Down weight them in sample averages

$$E[f(X)] = \frac{1}{N} \sum_{i=1}^{N} \frac{p(X_i)}{q(X_i)} f(X_i)$$

Issues with Importance and Rejection Sampling

- Rejection sampling - rejects too many samples in high dimensions
- Importance sampling - has high variance in high dimensions

Markov Chain Monte Carlo Methods

- Why Markov chain based sampling?
- If chain is regular, then converges to stationary distribution
  - Regular => >0 probability to go from any state to another state k hops away
- Allows for sampling from complex high dimensional distributions

Gibbs Sampling

- Consider T20 World Cup
- England in Group A and India in Group B
- Probabilities of each qualifying for Semi Finals is given below
- How do you sample from the distribution to get accurate estimates of P(I | E) and P(E | I)

India/England	Qualify (1)	Knocked Out (0)
Qualify (1)	0.1	0.4
Knocked Out (0)	0.2	0.3

Gibbs Sampling (contd.)

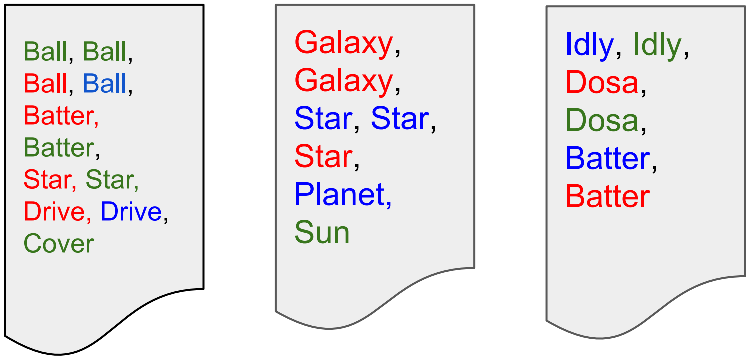
- Iterative process (the Markov comes from here).
- For t = 1:T:
  - $E^t \sim P(E \mid I^{t-1})$
  - $I^t \sim P(I \mid E^t)$

- In general, if we have a multivariate distribution $(X_1, X_2, \dots, X_n)$, then the sampling works as follows:
- For t = 1:T:
  - For i 1:n:
  - $X_i^t \sim P(X_i \mid X_1^t, \dots, X_{i-1}^t, X_{i+1}^{t-1}, \dots, X_n^{t-1})$

Dirichlet Distribution - Recall

$$
\begin{aligned}
p(\beta, \pi, \mathbf{Z}, \mathbf{W}) = \prod_{\mathbf{i}=1}^{\mathbf{K}} \overset{\text{Dirichlet}}{\mathbf{p}(\beta_{\mathbf{i}})} \prod_{\mathbf{i}=1}^{\mathbf{D}} \overset{\text{Dirichlet}}{\mathbf{p}(\pi_{\mathbf{d}})} \\
\left(\prod_{\mathbf{n}=1}^{\mathbf{N}} \underset{\text{Multinomial}}{\mathbf{p}(\mathbf{Z}_{\mathbf{d},\mathbf{n}} | \pi_{\mathbf{d}})} \underset{\text{Multinomial}}{\mathbf{p}(\mathbf{w}_{\mathbf{d},\mathbf{n}} | \beta, \mathbf{z}_{\mathbf{d},\mathbf{n}})}\right)
\end{aligned}
$$

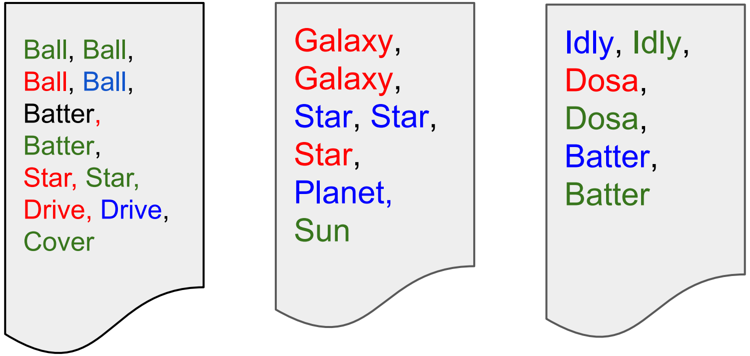
Gibbs Sampling for LDA - Example



Cricket Astrophysics Cooking

- Two Goals:
  - For each word in document, figure out which topic it belongs to.
  - For each document, figure out mixture of topics.

Gibbs Sampling for LDA



Cricket

Astrophysics

Cooking

- Pick a word in a document - say “Batter” in Document 1. What Topic does it belong to?
- Consider only Document 1, how frequently do Topic 1, 2 and 3 appear in Document 1?
- Answer: 5, 3 and 2.
- “Batter” should more likely be same as frequently occurring Topics in Document 1

Gibbs Sampling for LDA (contd.)

Ball, Ball, Ball, Ball, Batter, Batter, Star, Star, Drive, Drive, Cover

Cricket

Galaxy, Galaxy, Star, Star, Star, Planet, Sun

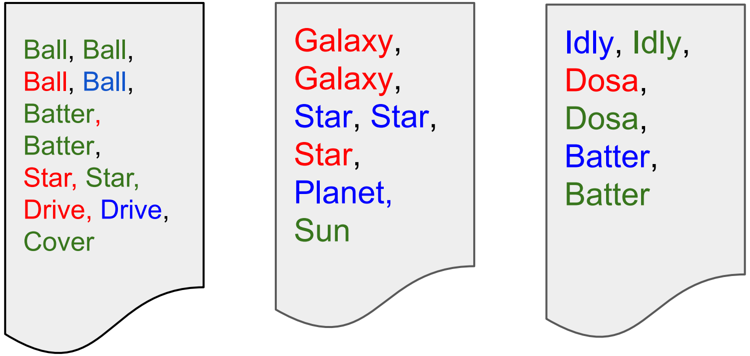
Astrophysics

Idly, Idly, Dosa, Dosa, Batter, Batter

Cooking

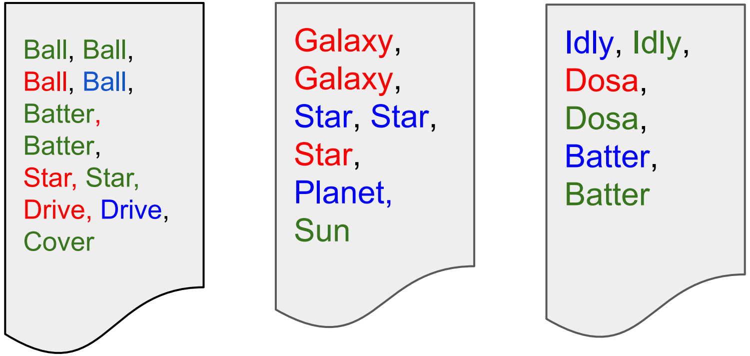
- What is the Topic associated with Batter across all Documents?
- Answer: 2, 0, 1

Gibbs Sampling for LDA (Contd.)



- Batter in Document 1: 5, 1 and 3.
- Batter across Documents: 2, 0, 1
- Assign green with probability = 5*2/(5*2 + 1*0 + 3*1) = 10/13

Gibbs Sampling for LDA (Contd.)



Cricket - 6/11
Astrophysics - 3/11
Cooking - 2/11

Astrophysics - 3/7
Cooking - 3/7
Cricket - 1/7

Cricket - 3/7
Cooking - 2/7
Astrophysics - 1/7

- Assign topic distribution to each document based on colors of words in document
- Keep iterating

Gibbs Sampling - Formally

- Recall:

Dirichlet Dirichlet

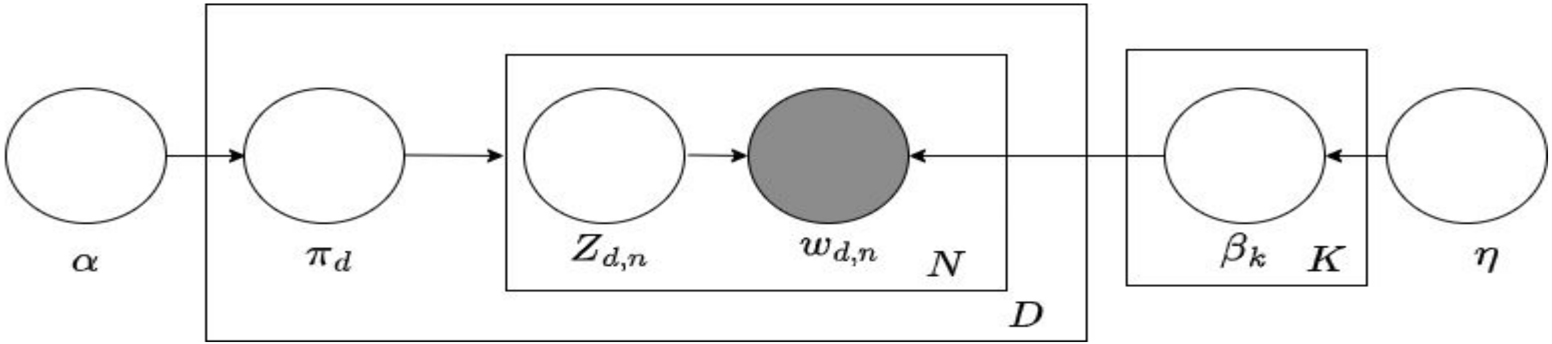
$$p(\beta, \pi, \mathbf{Z}, \mathbf{W}) = \prod_{\mathbf{i=1}}^{\mathbf{K}} \mathbf{p}(\beta_\mathbf{i}) \prod_{\mathbf{i=1}}^{\mathbf{D}} \mathbf{p}(\pi_\mathbf{d}) \left(\prod_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{p}(\mathbf{Z}_{\mathbf{d,n}}|\pi_\mathbf{d})\mathbf{p}(\mathbf{w}_{\mathbf{d,n}}|\beta, \mathbf{z}_{\mathbf{d,n}})\right)$$

Multinomial Multinomial

Gibbs Sampling (contd.)

- Define a $|V|XK$ matrix, C^V
- $C^V_{v,j}$, number of times word v is assigned to topic j, excluding current word v under consideration.
- Define a $|D|XK$ matrix, C^D
- $C^D_{d,j}$, fraction of words in d assigned to topic j, excluding current word v under consideration.

Gibbs Sampling (contd.)



$$p(z_v = j | z_{-v}, \{v, d\}) = \frac{C^V_{v,j} + \eta_v}{\sum_{v' \in V} C^V_{v',j} + |V|\eta_v} * \frac{C^D_{d,j} + \alpha_j}{\sum_{d' \in D} C^D_{d',j} + |D|\alpha_j}$$

- Add dirichlet parameter to avoid 0 values
- Dirichlet parameter is prior to multinomial

Similarities

- Document - Document
  - Use KL divergence between topic distribution of 2 documents to cluster/compare similarity between documents.
- Query - Document

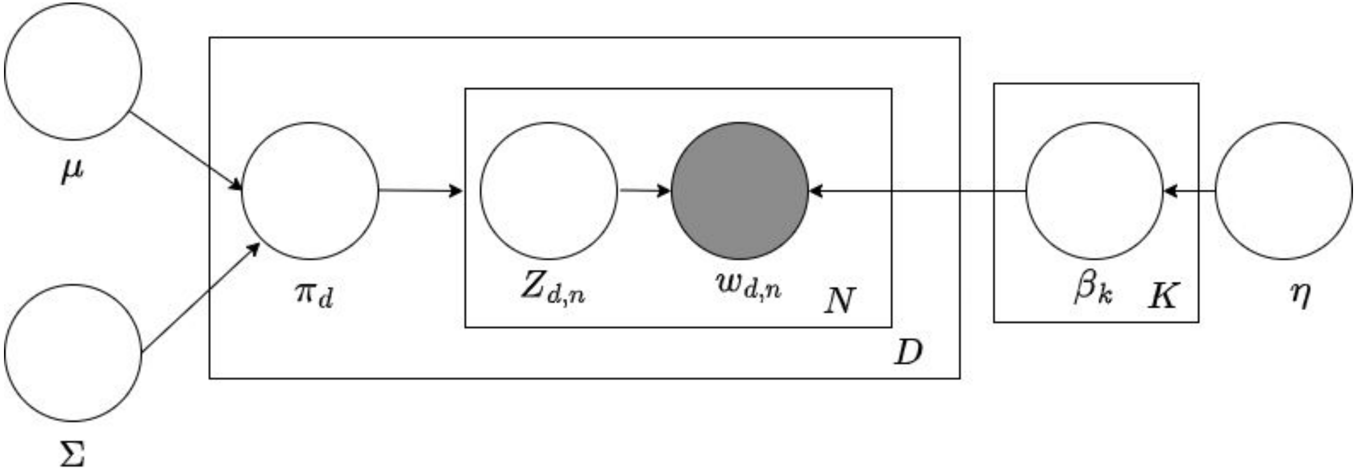
$$
\begin{aligned}
p(q|d) &= \prod_{w \in q} p(w|d) \\
&= \prod_{w \in q} \sum_{j \in K} p(w|z=j)p(z=j|d)
\end{aligned}
$$

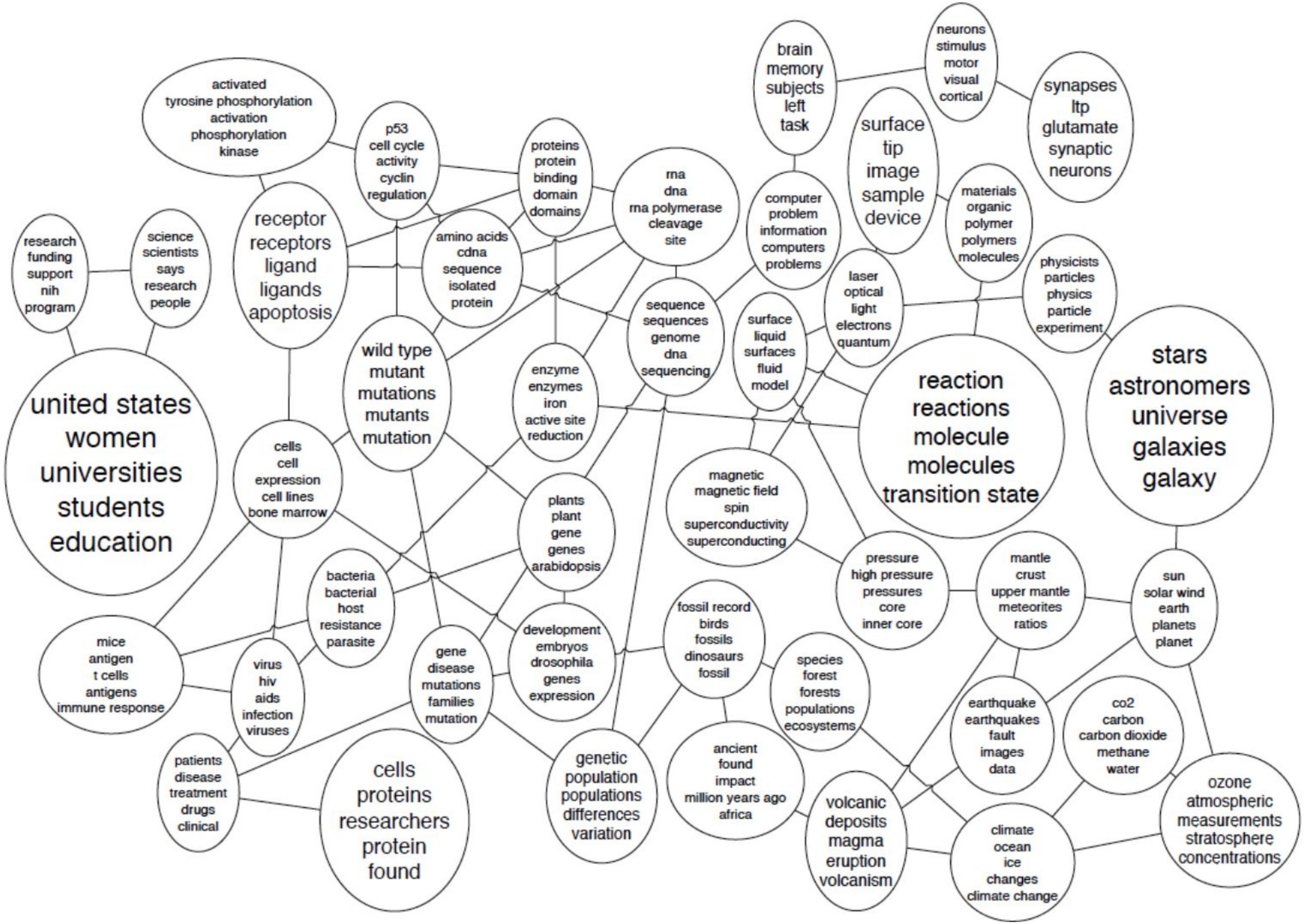
Correlated Topic Models

- The Dirichlet model assumes topics are independent of each other.
- Typically topics are correlated.
  - E.g. Topic on macroeconomics maybe correlated with topic on geo-politics
- How do we model such correlated topic models?
- Slight alteration to the graphical model does the trick

Correlated Topic Models

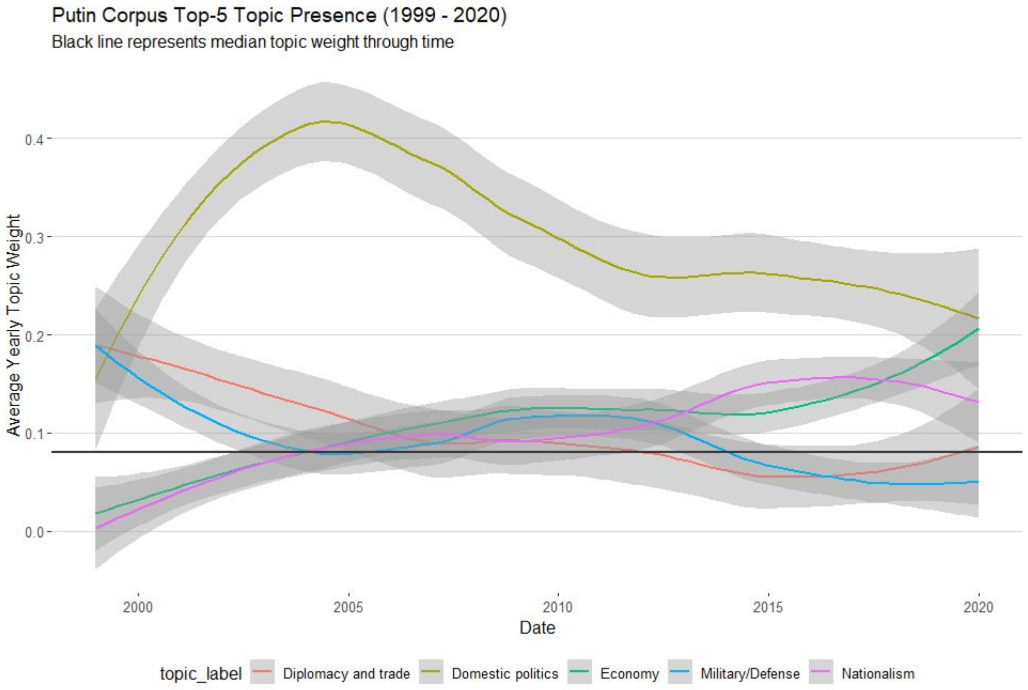
- Model the topic mixture as a multivariate normal ~ $N_k(\mu, \Sigma)$





Blei et. al
Annals of Applied Statistics, 2007

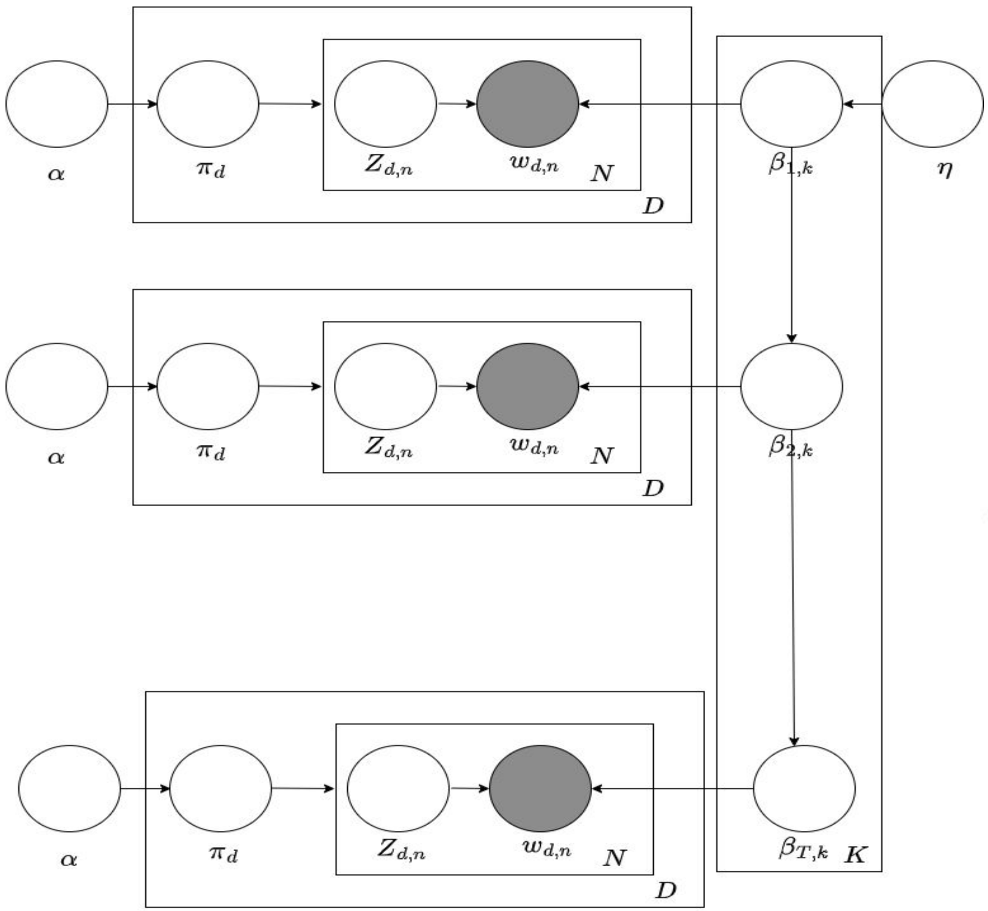
Dynamic Topic Models



https://medium.com/the-die-is-forecast/topic-modeling-as-osint-exploring-russian-presidential-speech-topics-over-time-ad6018286d37

Dynamic Topic Models

- Divide time into discrete chunks of time duration L (e.g. a year, a decade etc.)
- Do topic modeling on each corpus in that time duration L
- Assume topics evolve slowly
- Word topic distribution at time tL, depends on word topic distribution at (t-1)L



$$\beta_{k,t}|\beta_{k,t-1} \sim N(\beta_{k,t-1}, \sigma^2 I)$$