UPPSALA UNIVERSITET

Probabilistic Topic Models

Måns Magnusson
Department of Statistics
Uppsala University

June 11, 2024

UPPSALA UNIVERSITET

Section 1

Introduction

UPPSALA UNIVERSITET

Why topic models?

- Study semantic themes in a corpus (or topics)

Why topic models?

- Study semantic themes in a corpus (or topics)
- Exploratory (unsupervised) analysis

Why topic models?

- Study semantic themes in a corpus (or topics)
- Exploratory (unsupervised) analysis
- (Relatively) simple statistical models

Why topic models?

- Study semantic themes in a corpus (or topics)
- Exploratory (unsupervised) analysis
- (Relatively) simple statistical models
- Transparent models with statistical guarantees

Why topic models?

- Study semantic themes in a corpus (or topics)
- Exploratory (unsupervised) analysis
- (Relatively) simple statistical models
- Transparent models with statistical guarantees
- Extended in a large number of ways

Probability models for text: Multinomial

- Let θ be the probability of drawing the word type k. e.g.

  $$\theta = (\text{monkey} = 0.001, \text{the} = 0.03, ..., \text{Norrköping} = 0.0001)$$

Probability models for text: Multinomial

- Let θ be the probability of drawing the word type k. e.g.

  $\theta = (\text{monkey} = 0.001, \text{the} = 0.03, ..., \text{Norrköping} = 0.0001)$

- We are interested in a probability model for word (tokens), namely

$$w \sim p(w|\theta)$$

Probability models for text: Multinomial

- Let θ be the probability of drawing the word type k. e.g.

  $\theta = (\text{monkey} = 0.001, \text{the} = 0.03, \ldots, \text{Norrköping} = 0.0001)$

- We are interested in a probability model for word (tokens), namely

$$w \sim p(w|\theta)$$

- The Multinomial/Categorical distribution, where

$$p(w|\theta) = \theta_1^{w_1} \cdot \ldots \cdot \theta_K^{w_K}$$

where $\sum^K \theta_k = 1$

UPPSALA UNIVERSITET

Probability models for text: Multinomial

- Let θ be the probability of drawing the word type k. e.g.

  $\theta = (\text{monkey} = 0.001, \text{the} = 0.03, ..., \text{Norrköping} = 0.0001)$

- We are interested in a probability model for word (tokens), namely

$$w \sim p(w|\theta)$$

- The Multinomial/Categorical distribution, where

$$p(w|\theta) = \theta_1^{w_1} \cdot ... \cdot \theta_K^{w_K}$$

  where $\sum^K \theta_k = 1$

- A probabilistic unigram language model

Probability models for text: Dirichlet

- We are interested in a probability model for word probabilities (θ), namely

$$\theta \sim p(\theta|\alpha)$$

- The Dirichlet, where

$$p(\theta|\alpha) = \frac{\prod^K \Gamma(\alpha_k)}{\Gamma(\sum^K \alpha_k)} \prod^K \theta_k^{\alpha_k},$$

and $\sum^K \theta_k = 1$, $\alpha > 0$ and Γ is the gamma function.

UPPSALA UNIVERSITET

Probability models for text: Dirichlet

- We are interested in a probability model for word probabilities (θ), namely

$$\theta \sim p(\theta|\alpha)$$

- The Dirichlet, where

$$p(\theta|\alpha) = \frac{\prod^K \Gamma(\alpha_k)}{\Gamma(\sum^K \alpha_k)} \prod^K \theta_k^{\alpha_k},$$

and $\sum^K \theta_k = 1$, $\alpha > 0$ and Γ is the gamma function.

- The Dirichlet distribution generates "probability distributions", e.g.

$$\theta_1 = (0.019, 0.021, ..., 0.0002)$$
$$\theta_2 = (0.012, 0.019, ..., 0.0001)$$
$$...$$

Estimating parameters

- Maximum likelihood for word type v is estimated as

$$\hat{\theta}_{v,MLE} = \frac{n_v}{\sum_v^V n_v},$$

where n_w are the sufficient statistics (or word counts).

UPPSALA UNIVERSITET

Estimating parameters

- Maximum likelihood for word type v is estimated as

$$\hat{\theta}_{v,MLE} = \frac{n_v}{\sum_v^V n_v},$$

where n_w are the sufficient statistics (or word counts).

- The Dirichlet is a conjugate prior for the Multinomial

Estimating parameters

- Maximum likelihood for word type v is estimated as

$$\hat{\theta}_{v,MLE} = \frac{n_v}{\sum_v^V n_v},$$

where n_w are the sufficient statistics (or word counts).

- The Dirichlet is a conjugate prior for the Multinomial
- Using Bayes theorem, we get the posterior

$$p(\theta|w,\alpha) = \frac{p(w|\theta)p(\theta|\alpha)}{p(w)}.$$

Estimating parameters

- Maximum likelihood for word type v is estimated as

$$\hat{\theta}_{v,MLE} = \frac{n_v}{\sum_v^V n_v},$$

where n_w are the sufficient statistics (or word counts).

- The Dirichlet is a conjugate prior for the Multinomial
- Using Bayes theorem, we get the posterior

$$p(\theta|w,\alpha) = \frac{p(w|\theta)p(\theta|\alpha)}{p(w)}.$$

- Using the Multinomial (likelihood) and Dirichlet (prior), we get

$$\theta \sim p(\theta|w,\alpha) = \mathrm{Dir}(\alpha + \mathbf{n}_v),$$

where θ is a vector of length V and the prior hyperparameter α can be seen as a smoothing constant.

Estimating parameters

- Maximum likelihood for word type v is estimated as

$$\hat{\theta}_{v,MLE} = \frac{n_v}{\sum_v^V n_v},$$

where n_w are the sufficient statistics (or word counts).

- The Dirichlet is a conjugate prior for the Multinomial
- Using Bayes theorem, we get the posterior

$$p(\theta|w,\alpha) = \frac{p(w|\theta)p(\theta|\alpha)}{p(w)}.$$

- Using the Multinomial (likelihood) and Dirichlet (prior), we get

$$\theta \sim p(\theta|w,\alpha) = \mathrm{Dir}(\alpha + \mathbf{n}_v),$$

where θ is a vector of length V and the prior hyperparameter α can be seen as a smoothing constant.
- This can be used in more elaborate models

Example

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{w}_3$	money	bank	soccer	money	

MLE:

$$\hat{\theta}_{\text{bank},MLE} = \frac{n_v}{\sum_v^V n_v} = \frac{3}{12} = 0.25\,,$$

Posterior:

$$p(\theta_{\text{boat},\ldots,\text{bank}}|w,\alpha) = \text{Dir}(\alpha + (2, \ldots, 3))\,,$$

Section 2

Topic Models

UPPSALA UNIVERSITET

The Latent Dirichlet Allocation model

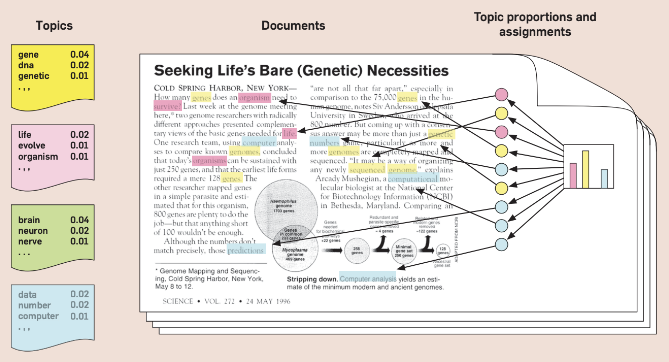


Figure: The intuitions behind latent Dirichlet allocation (Blei, 2012)

UPPSALA UNIVERSITET

- Introduction
  - Recap: Probability distributions for text
- Topic Models
  - Inference
  - Model Evaluation
  - Model Practicalities
- Example: Constraining topic models
- Example: Structural Topic Models

Latent Dirichlet Allocation

1. For each k in 1 ... K:
   1.1 $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each document d in 1 ... D:
   2.1 $\theta_d \sim \text{Dirichlet}(\alpha)$
   2.2 For each word i:
      2.2.1 $z_{id} \sim \text{Categorical}(\theta_d)$
      2.2.2 $w_{id} \sim \text{Categorical}(\phi_{z_{id}})$

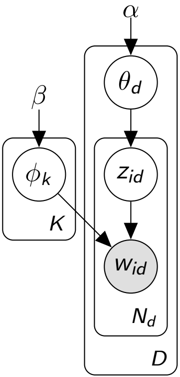


Figure: Probabilistic model for Latent Dirichlet Allocation (LDA)

Example of parameters $\mathbf{z}$, Θ and Φ

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	2	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\Phi =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	0.35	0.35	0.05	0.05	0.15	0.05
Topic 2	0.025	0.025	0.45	0.45	0.025	0.025
Topic 3	0.025	0.025	0.025	0.025	0.45	0.45

$\Theta =$

	Topic 1	Topic 2	Topic 3
doc 1	0.96	0.02	0.02
doc 2	0.3	0.2	0.5
doc 3	0.05	0.35	0.6

A small topic model example

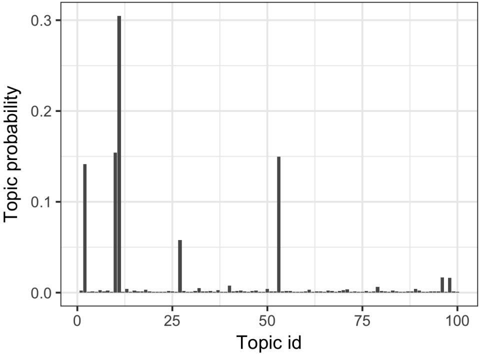
Closing arguments were heard yesterday in the Federal bankruptcy fraud trial of Stephen J. Sabbeth, whose legal problems have raised doubts about his ability to continue as leader of the Nassau County Democratic Party.

Mr. Sabbeth is charged with trying to conceal $750,000 from his bank creditors by hiding the money in a secret account in his wife's maiden name, rather than use it to pay creditors when his lumber business went into bankruptcy 10 years ago.

– The New York Times 25th of Febuary 1999

UPPSALA UNIVERSITET

Estimated topic distribution $E(\theta_d)$



Most probable word type by topic

Topic	Top words (by ϕ_{kv})
2	party election voters campaign democratic
10	bank banks loans loan insurance savings
11	trial prison jury prosecutors convicted guilty
53	investigation inquiry documents investigators

Table: Most probable words in topic 2, 10, 11 and 53.

UPPSALA UNIVERSITET

Analytical use of topic models

- We have three parameters we can use for inference, Φ, Θ, and $\mathbf{z}$

UPPSALA UNIVERSITET

Analytical use of topic models

- We have three parameters we can use for inference, Φ, Θ, and $\mathbf{z}$

$$\Phi = \begin{array}{lcccccc} & \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ \text{Topic 1} & 0.35 & 0.35 & 0.05 & 0.05 & 0.15 & 0.05 \\ \text{Topic 2} & 0.025 & 0.025 & 0.45 & 0.45 & 0.025 & 0.025 \\ \text{Topic 3} & 0.025 & 0.025 & 0.025 & 0.025 & 0.45 & 0.45 \end{array}$$

Analytical use of topic models

- We have three parameters we can use for inference, Φ, Θ, and $\mathbf{z}$

$$\Phi = \begin{array}{lcccccc} & \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ \text{Topic 1} & 0.35 & 0.35 & 0.05 & 0.05 & 0.15 & 0.05 \\ \text{Topic 2} & 0.025 & 0.025 & 0.45 & 0.45 & 0.025 & 0.025 \\ \text{Topic 3} & 0.025 & 0.025 & 0.025 & 0.025 & 0.45 & 0.45 \end{array}$$

$$\Theta = \begin{array}{lccc} & \text{Topic 1} & \text{Topic 2} & \text{Topic 3} \\ \text{doc 1} & 0.96 & 0.02 & 0.02 \\ \text{doc 2} & 0.3 & 0.2 & 0.5 \\ \text{doc 3} & 0.05 & 0.35 & 0.6 \end{array}$$

Analytical use of topic models

- We have three parameters we can use for inference, Φ, Θ, and $\mathbf{z}$

$$\Phi = \begin{array}{lcccccc} & \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ \text{Topic 1} & 0.35 & 0.35 & 0.05 & 0.05 & 0.15 & 0.05 \\ \text{Topic 2} & 0.025 & 0.025 & 0.45 & 0.45 & 0.025 & 0.025 \\ \text{Topic 3} & 0.025 & 0.025 & 0.025 & 0.025 & 0.45 & 0.45 \end{array}$$

$$\Theta = \begin{array}{lccc} & \text{Topic 1} & \text{Topic 2} & \text{Topic 3} \\ \text{doc 1} & 0.96 & 0.02 & 0.02 \\ \text{doc 2} & 0.3 & 0.2 & 0.5 \\ \text{doc 3} & 0.05 & 0.35 & 0.6 \end{array}$$

$$\begin{array}{lccccc} \mathbf{z}_1 & 1 & 1 & 1 & & \\ \mathbf{w}_1 & \text{boat} & \text{shore} & \text{bank} & & \\ \mathbf{z}_2 & 2 & 1 & 1 & 3 & 3 \\ \mathbf{w}_2 & \text{Zlatan} & \text{boat} & \text{shore} & \text{money} & \text{bank} \\ \mathbf{z}_3 & 3 & 3 & 2 & 3 & \\ \mathbf{w}_3 & \text{money} & \text{bank} & \text{soccer} & \text{money} & \end{array}$$

Subsection 1

Inference

Inference methods

- We want to estimate the parameters $\mathbf{z}, \theta, \phi$ based on data $\mathbf{w}$

Inference methods

- We want to estimate the parameters $\mathbf{z}, \theta, \phi$ based on data $\mathbf{w}$
- Most commonly done using Bayesian methods by computing the joint posterior

$$p(\mathbf{z}, \theta, \phi | \mathbf{w}) = \frac{p(\mathbf{w} | \mathbf{z}, \theta, \phi) p(\mathbf{z}, \theta, \phi)}{p(\mathbf{w})}$$

Inference methods

- We want to estimate the parameters $\mathbf{z}, \theta, \phi$ based on data $\mathbf{w}$
- Most commonly done using Bayesian methods by computing the joint posterior

$$p(\mathbf{z}, \theta, \phi | \mathbf{w}) = \frac{p(\mathbf{w}|\mathbf{z}, \theta, \phi) p(\mathbf{z}, \theta, \phi)}{p(\mathbf{w})}$$

- $p(\mathbf{w})$ is intractable, so we use
  - MCMC/Gibbs sampling (Griffiths and Steyvers, 2004), or
  - Variational inference (Blei, Ng and Jordan, 2003).

Sufficient statistics

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	1	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	2	2	0	1	1	0
Topic 2	0	0	1	0	0	0
Topic 3	0	0	0	0	2	3

$\mathbf{n}_d =$

	Topic 1	Topic 2	Topic 3
doc 1	3	0	0
doc 2	3	0	2
doc 3	0	1	3

Inference methods

- Gibbs sampling (Griffiths and Steyvers, 2004)

$$p(z_i = k) \propto \theta_{d(i),k} \phi_{k,v(i)}$$

$$\theta_d \sim \mathrm{Dir}(\mathbf{n}_d + \alpha)$$

$$\phi_k \sim \mathrm{Dir}(\mathbf{n}_v + \beta)$$

where n_d and n_v are sufficient statistics for θ and ϕ

Inference methods

- Gibbs sampling (Griffiths and Steyvers, 2004)

$$p(z_i = k) \propto \theta_{d(i),k}\phi_{k,v(i)}$$

$$\theta_d \sim \mathrm{Dir}(\mathbf{n}_d + \alpha)$$

$$\phi_k \sim \mathrm{Dir}(\mathbf{n}_v + \beta)$$

  where n_d and n_v are sufficient statistics for θ and ϕ
- Run until convergence (commonly log-likelihood converges)
- Then the MCMC/Gibbs sampler generates draws from

$$p(\mathbf{z}, \theta, \phi | \mathbf{w})$$

Inference methods

- We can integrate out Θ and Φ

$$p(z_i = k|\mathbf{z}_{-i}) \propto \frac{n_{dk} + \alpha}{\sum^K (n_{dk} + \alpha)} \frac{n_{vk} + \beta}{\sum^V (n_{vk} + \beta)}$$
$$\propto (n_{dk} + \alpha) \frac{n_{vk} + \beta}{\sum^V (n_{vk} + \beta)}$$

Inference methods

- We can integrate out Θ and Φ

$$p(z_i = k | \mathbf{z}_{-i}) \propto \frac{n_{dk} + \alpha}{\sum^K (n_{dk} + \alpha)} \frac{n_{vk} + \beta}{\sum^V (n_{vk} + \beta)}$$
$$\propto (n_{dk} + \alpha) \frac{n_{vk} + \beta}{\sum^V (n_{vk} + \beta)}$$

- This is the collapsed Gibbs sampler for LDA

Collapsed Gibbs sampling

Let $\alpha = \beta = 0.5$, and we will sample $\mathbf{z}_2$:

$$p(z_i = k) \propto \left(\frac{2.5}{5.5}\frac{0.5}{1.5}, \frac{0.5}{5.5}\frac{0.5}{1.5}, \frac{2.5}{5.5}\frac{0.5}{1.5} \right)$$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	?	1	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	2	2	0	0	1	0
Topic 2	0	0	1	0	0	0
Topic 3	0	0	0	0	2	3

$\mathbf{n}_2 =$

	Topic 1	Topic 2	Topic 3
doc 2	2	0	2

UPPSALA UNIVERSITET

- Introduction
  - Recap: Probability distributions for text
- Topic Models
  - Inference
  - Model Evaluation
  - Model Practicalities
- Example: Constraining topic models
- Example: Structural Topic Models

Collapsed Gibbs sampling

Let $\alpha = \beta = 0.5$, and we will sample $\mathbf{z}_2$:

$$p(z_i = k) \propto \left(\frac{1.5}{5.5}\frac{1.5}{2.5}, \frac{0.5}{5.5}\frac{0.5}{2.5}, \frac{3.5}{5.5}\frac{0.5}{2.5} \right)$$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	**3**	?	1	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	**1**	2	0	0	1	0
Topic 2	0	0	1	0	0	0
Topic 3	0	0	0	**1**	2	3

$\mathbf{n}_2 =$

	Topic 1	Topic 2	Topic 3
doc 2	**1**	0	**3**

Collapsed Gibbs sampling

Let $\alpha = \beta = 0.5$, and we will sample $\mathbf{z}_2$:

$$p(z_i = k) \propto \left(\frac{1.5}{5.5}\frac{1.5}{2.5}, \frac{0.5}{5.5}\frac{0.5}{2.5}, \frac{3.5}{5.5}\frac{0.5}{2.5} \right)$$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	3	**1**	?	3	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	**2**	**1**	0	0	1	0
Topic 2	0	0	1	0	0	0
Topic 3	0	0	0	1	2	3

$\mathbf{n}_2 =$

	Topic 1	Topic 2	Topic 3
doc 2	**1**	0	**3**

UPPSALA UNIVERSITET

- Introduction
  - Recap: Probability distributions for text
- Topic Models
  - Inference
  - Model Evaluation
  - Model Practicalities
- Example: Constraining topic models
- Example: Structural Topic Models

Collapsed Gibbs sampling

Let $\alpha = \beta = 0.5$, and we will sample $\mathbf{z}_2$:

$$p(z_i = k) \propto \left(\frac{1.5}{5.5}\frac{0.5}{3.5}, \frac{0.5}{5.5}\frac{0.5}{3.5}, \frac{3.5}{5.5}\frac{2.5}{3.5} \right)$$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	3	1	**3**	?	3
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	2	1	0	0	1	0
Topic 2	0	0	1	0	0	0
Topic 3	0	**1**	0	1	2	**2**

$\mathbf{n}_2 =$

	Topic 1	Topic 2	Topic 3
doc 2	1	0	**3**

Collapsed Gibbs sampling

Let $\alpha = \beta = 0.5$, and we will sample $\mathbf{z}_2$:

$$p(z_i = k) \propto \left(\frac{1.5}{5.5}\frac{1.5}{3.5}, \frac{0.5}{5.5}\frac{0.5}{3.5}, \frac{3.5}{5.5}\frac{1.5}{3.5} \right)$$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	3	1	3	**3**	?
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	2	1	0	0	1	0
Topic 2	0	0	1	0	0	0
Topic 3	0	1	0	1	**1**	**3**

$\mathbf{n}_2 =$

	Topic 1	Topic 2	Topic 3
doc 2	1	0	**3**

UPPSALA UNIVERSITET

Collapsed Gibbs sampling

Let $\alpha = \beta = 0.5$, and we will sample $\mathbf{z}_2$:

$$p(z_i = k) \propto (-, -, -)$$

$\mathbf{w}_1$	boat	shore	bank		
$\mathbf{z}_1$	1	1	1		
$\mathbf{w}_2$	Zlatan	boat	shore	money	bank
$\mathbf{z}_2$	3	1	3	3	**1**
$\mathbf{w}_3$	money	bank	soccer	money	
$\mathbf{z}_3$	3	3	2	3	

$\mathbf{n}_v =$

	boat	shore	soccer	Zlatan	bank	money
Topic 1	2	1	0	0	**2**	0
Topic 2	0	0	1	0	0	0
Topic 3	0	1	0	1	1	3

$\mathbf{n}_2 =$

	Topic 1	Topic 2	Topic 3
doc 2	**2**	0	3

Other inference methods

- Mean-Field Variational inference
  (Blei, Ng and Jordan, 2003, Blei, 2013)

Other inference methods

- Mean-Field Variational inference (Blei, Ng and Jordan, 2003, Blei, 2013)
- Stochastic EM (Zaheer et al., 2015)

$$\hat{\theta}_d = \arg\max_{\theta \in \Theta} p(\theta, \phi | \mathbf{z}, \mathbf{w}) = \frac{n_{dk} + \alpha}{\sum^K (n_{dk} + \alpha)}$$

$$\hat{\phi}_k = \arg\max_{\phi \in \Phi} p(\theta, \phi | \mathbf{z}, \mathbf{w}) = \frac{n_{vk} + \beta}{\sum^V (n_{vk} + \beta)}$$

$$p(z_i = k | \mathbf{w}, \Theta, \Phi) \propto \hat{\theta}_{d(i),k} \cdot \hat{\phi}_{k,v(i)}$$

Subsection 2

Model Evaluation

Evaluating Topic Models

- Parameter inspection (top words, relevance words, document topic distributions by other variables)

UPPSALA UNIVERSITET

Evaluating Topic Models

- Parameter inspection (top words, relevance words, document topic distributions by other variables)
- Documents with high topic proportion: look at data

Evaluating Topic Models

- Parameter inspection (top words, relevance words, document topic distributions by other variables)
- Documents with high topic proportion: look at data
- Estimating held-out log likelihood (Wallach et al, 2009)

UPPSALA UNIVERSITET

Evaluating Topic Models

- Parameter inspection (top words, relevance words, document topic distributions by other variables)
- Documents with high topic proportion: look at data
- Estimating held-out log likelihood (Wallach et al, 2009)
- Estimating topic coherence (Mimno et al, 2011)

$$C(V^{(t)}) = \sum_{m}^{M} \sum_{l}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

where $V^{(t)}$ is the set of the M top words $v_1, ..., v_M$ and $D(v_m^{(t)}, v_l^{(t)})$ is the co-document frequency.

UPPSALA UNIVERSITET

Subsection 3

Model Practicalities

How do we define a document?

- The definition of a document matters
- Book, chapter, paragraph, speech, ...
- What to choose and why?

How do we define a document?

- The definition of a document matters
- Book, chapter, paragraph, speech, ...
- What to choose and why?

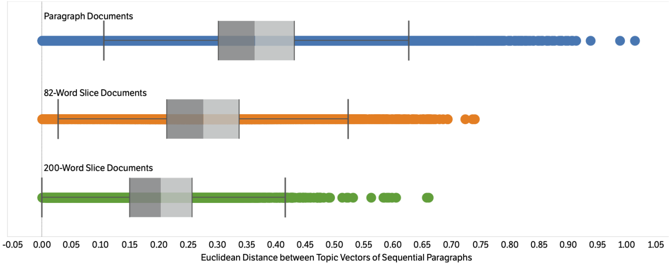


Figure: Algee-Hewitt et al (2015), Fig. 6.1

Choosing K

- We can estimate the optimal K. Is this a good idea?

Choosing K

- We can estimate the optimal K. Is this a good idea?
- Alternative: Think of K as the resolution of a map
- Evaluate when the relevant part of the model is good for the use case

Section 3

Example: Constraining topic models

Why constraining?

- Topic modeling is difficult (today)

Why constraining?

- Topic modeling is difficult (today)
- The standard LDA is fully unsupervised - we might want to measure specific topics.

UPPSALA UNIVERSITET

Why constraining?

- Topic modeling is difficult (today)
- The standard LDA is fully unsupervised - we might want to measure specific topics.
- A simple approach is constraining topic models

Why constraining?

- Topic modeling is difficult (today)
- The standard LDA is fully unsupervised - we might want to measure specific topics.
- A simple approach is constraining topic models
- Idea: Use the prior $p(\theta)$ to a priori define topics and documents

UPPSALA UNIVERSITET

Example of constraining Φ

Constraining topic 2 to be a soccer topic:

$$\Phi = \begin{array}{lcccccc} & \text{boat} & \text{shore} & \text{soccer} & \text{Zlatan} & \text{bank} & \text{money} \\ \text{Topic 1} & 0.35 & 0.35 & 0 & 0 & 0.15 & 0.05 \\ \text{Topic 2} & 0.025 & 0.025 & 0.45 & 0.45 & 0.025 & 0.025 \\ \text{Topic 3} & 0.025 & 0.025 & 0 & 0 & 0.45 & 0.45 \end{array}$$

Constraining topic 3 only exist in document 3:

$$\Theta = \begin{array}{lccc} & \text{Topic 1} & \text{Topic 2} & \text{Topic 3} \\ \text{doc 1} & 0.96 & 0.04 & 0 \\ \text{doc 2} & 0.6 & 0.4 & 0 \\ \text{doc 3} & 0.05 & 0.35 & 0.6 \end{array}$$

UPPSALA UNIVERSITET

Pro- and con of constraining topic models

- Pros
  1. Reproducible
  2. Transparent
  3. Easier to diagnose problems
  4. Can adapt the model to the research problem at hand

UPPSALA UNIVERSITET

Pro- and con of constraining topic models

- Pros
  1. Reproducible
  2. Transparent
  3. Easier to diagnose problems
  4. Can adapt the model to the research problem at hand
- Cons
  1. Open area of research
  2. Can break/work poorly in some settings

The Swedish migration discourse 1945-2020

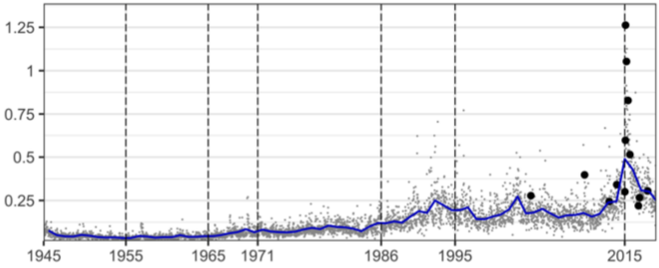


Figure: The number of immigrants to Sweden (A) and the saliency of immigration in the Swedish public discourse (B) (Hurtado Bodell et al., in print)

UPPSALA UNIVERSITET

Section 4

Example: Structural Topic Models

Why Structural Topic Models?

- You have document-level covariates you want to include

Why Structural Topic Models?

- You have document-level covariates you want to include
- STM estimate covariate effects of document-topics distribution (θ) by covariates X and topic-word distribution (ϕ) by covariates Y

UPPSALA UNIVERSITET

Covariate effects on θ

- Modeling θ is done as

$$\theta_d \sim \text{LogisticNormal}(\mathbf{x}_d \gamma, \Sigma)\,,$$

where $\mathbf{x}_d \in \mathbb{R}^P$, $\gamma \in \mathbb{R}^{P\times(K-1)}$, and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$

Covariate effects on θ

- Modeling θ is done as

$$\theta_d \sim \text{LogisticNormal}(\mathbf{x}_d\gamma, \Sigma)\,,$$

  where $\mathbf{x}_d \in \mathbb{R}^P$, $\gamma \in \mathbb{R}^{P\times(K-1)}$, and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$
- Interpretation of γ, the effect of a document covariate in using the topic.
- Note!
  1. Without covariates, it reduces to a correlated topic model

Covariate effects on θ

- Modeling θ is done as

$$\theta_d \sim \text{LogisticNormal}(\mathbf{x}_d \gamma, \Sigma)\,,$$

  where $\mathbf{x}_d \in \mathbb{R}^P$, $\gamma \in \mathbb{R}^{P\times(K-1)}$, and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$
- Interpretation of γ, the effect of a document covariate in using the topic.
- Note!
  1. Without covariates, it reduces to a correlated topic model
  2. Without covariates and $\Sigma = I$, it reduces to a standard topic model (ish)

Covariate effects on θ

- Modeling θ is done as

$$\theta_d \sim \text{LogisticNormal}(\mathbf{x}_d\gamma, \Sigma)\,,$$

  where $\mathbf{x}_d \in \mathbb{R}^P$, $\gamma \in \mathbb{R}^{P\times(K-1)}$, and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$
- Interpretation of γ, the effect of a document covariate in using the topic.
- Note!
  1. Without covariates, it reduces to a correlated topic model
  2. Without covariates and $\Sigma = I$, it reduces to a standard topic model (ish)
  3. The parameters γ are interpreted to a reference category

Covariate effects on ϕ

- Modeling ϕ is done as

$$\phi_{k,v,y} \propto \exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v}\right)$$

where $y \in \{1, ..., A\}$, m_v is the marginal rate for word type v, $\kappa^{(t)}_{k,v} \in \mathbb{R}^{K \times V}$, $\kappa^{(c)}_{y_d,v} \in \mathbb{R}^{A \times V}$, and $\kappa^{(i)}_{y_d,k,v} \in \mathbb{R}^{A \times K \times V}$.

Covariate effects on ϕ

- Modeling ϕ is done as

$$\phi_{k,v,y} \propto \exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v}\right)$$

  where $y \in \{1, ..., A\}$, m_v is the marginal rate for word type v, $\kappa^{(t)}_{k,v} \in \mathbb{R}^{K\times V}$, $\kappa^{(c)}_{y_d,v} \in \mathbb{R}^{A\times V}$, and $\kappa^{(i)}_{y_d,k,v} \in \mathbb{R}^{A\times K\times V}$.
- The effect of y_d on the word usage

Covariate effects on ϕ

- Modeling ϕ is done as

$$\phi_{k,v,y} \propto \exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v}\right)$$

  where $y \in \{1, ..., A\}$, m_v is the marginal rate for word type v, $\kappa^{(t)}_{k,v} \in \mathbb{R}^{K \times V}$, $\kappa^{(c)}_{y_d,v} \in \mathbb{R}^{A \times V}$, and $\kappa^{(i)}_{y_d,k,v} \in \mathbb{R}^{A \times K \times V}$.
- The effect of y_d on the word usage
- Note!
  1. β is used instead of ϕ in the paper

Covariate effects on ϕ

- Modeling ϕ is done as

$$\phi_{k,v,y} \propto \exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v}\right)$$

  where $y \in \{1, ..., A\}$, m_v is the marginal rate for word type v, $\kappa^{(t)}_{k,v} \in \mathbb{R}^{K \times V}$, $\kappa^{(c)}_{y_d,v} \in \mathbb{R}^{A \times V}$, and $\kappa^{(i)}_{y_d,k,v} \in \mathbb{R}^{A \times K \times V}$.
- The effect of y_d on the word usage
- Note!
  1. β is used instead of ϕ in the paper
  2. A Laplace (sparsity prior) is used to shrink κ parameters toward zero

Example

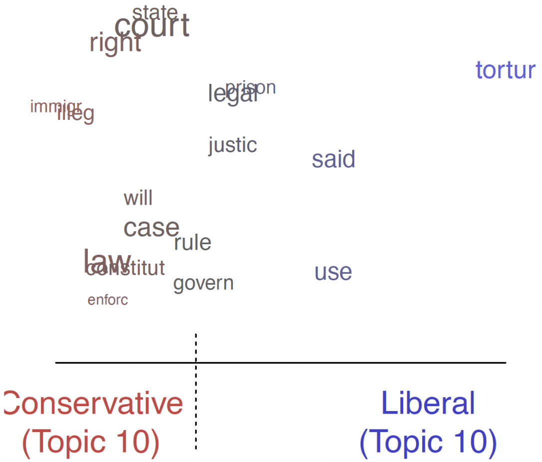


Figure: Difference between party word use (Roberts et al., 2019)

UPPSALA UNIVERSITET

Inference

- STM (standard) is implemented using variational inference

Inference

- STM (standard) is implemented using variational inference
- Variational inference tends to underestimate the uncertainty of the parameters (Wang and Blei, 2018)