# Machine Learning

Justin Grimmer

Associate Professor Department of Political Science University of Chicago

January 24th, 2018

#### Clustering

Document → One Cluster

Doc 1

Doc 2

Doc 3

:

Doc N

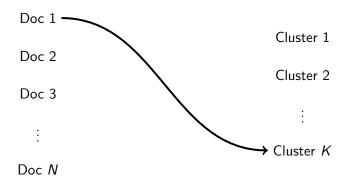
Cluster 1

Cluster 2

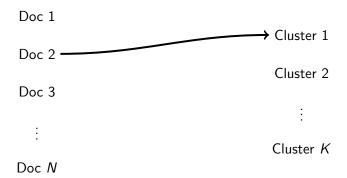
:

Cluster K

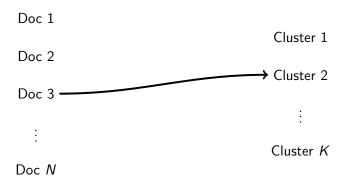
#### Clustering



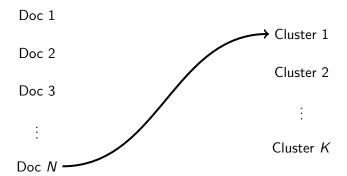
#### Clustering



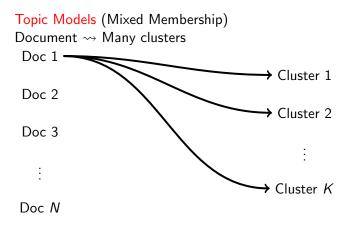
#### Clustering



#### Clustering



```
Topic Models (Mixed Membership)
Document → Many clusters
 Doc 1
                                        Cluster 1
 Doc 2
                                        Cluster 2
 Doc 3
                                       Cluster K
Doc N
```



# A Statistical Highlighter (With Many Colors)

#### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes

other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center

lecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Titol genes
Genes
In Dectresciat, Viviaty land. Comparing all
Genes
In common
23 genes

Discoppiame
genes
Ge

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.



<sup>\*</sup> Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Latent Dirichlet Allocation: an admixture model

Latent Dirichlet Allocation: an admixture model Intuition: each unit is a mixture of latent components

Latent Dirichlet Allocation: an admixture model Intuition: each unit is a mixture of latent components Mixture Model

$$egin{array}{ll} m{\pi} & \sim & \mathsf{Dirichlet}(m{lpha}) \ & au_i & \sim & \mathsf{Multinomial}(1,m{\pi}) \ m{x}_i | au_{ik} = 1 & \sim & \mathsf{Distribution}(m{ heta}_k) \end{array}$$

Latent Dirichlet Allocation: an admixture model Intuition: each unit is a mixture of latent components Mixture Model

$$egin{array}{ll} m{\pi} & \sim & \mathsf{Dirichlet}(m{lpha}) \ & au_i & \sim & \mathsf{Multinomial}(1,m{\pi}) \ m{x}_i | au_{ik} = 1 & \sim & \mathsf{Distribution}(m{ heta}_k) \end{array}$$

#### AdMixture Model

$$egin{array}{ll} m{\pi_i} & \sim & \mathsf{Dirichlet}(m{lpha}) \\ & au_{im} & \sim & \mathsf{Multinomial}(1,m{\pi}) \\ m{x_{im}} | au_{imk} = 1 & \sim & \mathsf{Distribution}(m{ heta}_k) \end{array}$$

Latent Dirichlet Allocation: an admixture model Intuition: each unit is a mixture of latent components Mixture Model

$$egin{array}{ll} m{\pi} & \sim & \mathsf{Dirichlet}(m{lpha}) \ & au_i & \sim & \mathsf{Multinomial}(1,m{\pi}) \ m{x}_i | au_{ik} = 1 & \sim & \mathsf{Distribution}(m{ heta}_k) \end{array}$$

#### AdMixture Model

$$egin{array}{ll} m{\pi_i} & \sim & \mathsf{Dirichlet}(m{lpha}) \\ & au_{im} & \sim & \mathsf{Multinomial}(1,m{\pi}) \\ m{x_{im}} | au_{imk} = 1 & \sim & \mathsf{Distribution}(m{ heta}_k) \end{array}$$

- Consider document i, (i = 1, 2, ..., N).

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

<sup>\*</sup>Notice: this is a different representation than a document-term matrix.  $x_{im}$  is a number that says which of the J words are used. The difference is for clarity and we'll this representation is closely related to document-term matrix

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

$$\pi_i | \alpha \sim \mathsf{Dirichlet}(\alpha)$$

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

$$m{\pi}_i | m{lpha} \ \sim \ \mathsf{Dirichlet}(m{lpha})$$
  $m{ au}_{im} | m{\pi}_i \ \sim \ \mathsf{Multinomial}(1, m{\pi}_i)$ 

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

$$egin{array}{ll} m{\pi}_i | m{lpha} & \sim & \mathsf{Dirichlet}(m{lpha}) \ m{ au}_{im} | m{\pi}_i & \sim & \mathsf{Multinomial}(1, m{\pi}_i) \ m{ imes}_{im} | m{ heta}_k, au_{imk} = 1 & \sim & \mathsf{Multinomial}(1, m{ heta}_k) \end{array}$$

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

$$m{ heta}_k \sim \mathsf{Dirichlet}(\mathbf{1})$$
  $m{\pi}_i | m{lpha} \sim \mathsf{Dirichlet}(m{lpha})$   $m{ au}_{im} | m{\pi}_i \sim \mathsf{Multinomial}(1, m{\pi}_i)$   $m{x}_{im} | m{ heta}_k, au_{imk} = 1 \sim \mathsf{Multinomial}(1, m{ heta}_k)$ 

- Consider document i, (i = 1, 2, ..., N).
- Suppose there are  $M_i$  total words and  $\mathbf{x}_i$  is an  $M_i \times 1$  vector, where  $\mathbf{x}_{im}$  describes the  $m^{\text{th}}$  word used in the document\*.

$$egin{array}{ll} oldsymbol{ heta}_k & \sim & \mathsf{Dirichlet}(\mathbf{1}) \ lpha_k & \sim & \mathsf{Gamma}(lpha,eta) \ oldsymbol{\pi}_i | oldsymbol{lpha} & \sim & \mathsf{Dirichlet}(oldsymbol{lpha}) \ oldsymbol{ au}_{im} | oldsymbol{\pi}_i & \sim & \mathsf{Multinomial}(1,oldsymbol{\pi}_i) \ oldsymbol{x}_{im} | oldsymbol{ heta}_k, au_{imk} = 1 & \sim & \mathsf{Multinomial}(1,oldsymbol{ heta}_k) \end{array}$$

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \propto p(\alpha)p(\boldsymbol{\pi}|\alpha)p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})$$

$$egin{array}{ll} p(m{\pi},m{T},m{\Theta},lpha|m{X}) & \propto & p(m{lpha})p(m{\pi}|m{lpha})p(m{T}|m{\pi})p(m{X}|m{ heta},m{T}) \ & \propto & p(m{lpha})\prod_{i=1}^N \left[p(m{\pi}_i|m{lpha})\prod_{m=1}^{M_i}p(m{ au}_{im}|m{\pi})p(m{x}_{im}|m{ heta}_k, au_{imk}=1)
ight] \end{array}$$

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ p(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right]$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{x}_{imj}} \right]^{\tau_{ikm}} \right]$$

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ p(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right]$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{x}_{imj}} \right]^{\tau_{ikm}} \right]$$

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ p(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right]$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{x}_{imj}} \right]^{\tau_{ikm}} \right]$$

Together the model implies the following posterior:

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ p(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right]$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M_{k}} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{\chi}_{imj}} \right]^{\tau_{ikm}} \right]$$

Optimization:

Together the model implies the following posterior:

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ p(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right]$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M_{k}-1} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{x}_{imj}} \right]^{\tau_{ikm}} \right]$$

#### Optimization:

- Variational Approximation → Find "closest" distribution

Together the model implies the following posterior:

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ p(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right]$$

$$\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M_{k}} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{x}_{imj}} \right]^{\tau_{ikm}} \right]$$

#### Optimization:

- Variational Approximation → Find "closest" distribution
- Gibbs sampling \sim MCMC algorithm to approximate posterior

Together the model implies the following posterior:

$$\rho(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \propto \rho(\boldsymbol{\alpha}) \rho(\boldsymbol{\pi} | \boldsymbol{\alpha}) \rho(\boldsymbol{T} | \boldsymbol{\pi}) \rho(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
\propto \rho(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \rho(\boldsymbol{\pi}_{i} | \boldsymbol{\alpha}) \prod_{m=1}^{M_{i}} \rho(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) \rho(\boldsymbol{x}_{im} | \boldsymbol{\theta}_{k}, \boldsymbol{\tau}_{imk} = 1) \right] \\
\propto \rho(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[ \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k})}{\prod_{k=1}^{K} \Gamma(\alpha_{k})} \prod_{k=1}^{K} \prod_{m=1}^{M_{k}} \prod_{k=1}^{K} \left[ \pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{\boldsymbol{x}_{imj}} \right]^{\tau_{ikm}} \right] \\$$

#### Optimization:

- Variational Approximation → Find "closest" distribution
- Gibbs sampling \infty MCMC algorithm to approximate posterior

Described in the slides appendix

# Running a Topic Model with STM

to the STM Code

#### Why does this work → Co-occurrence

Where's the information for each word's topic?

#### Why does this work → Co-occurrence

Where's the information for each word's topic? Reconsider document-term matrix

Where's the information for each word's topic? Reconsider document-term matrix

	$Word_1$	$Word_2$		$Word_J$
Doc <sub>1</sub>	0	1		0
$Doc_2$	2	0		3
:	:	:	٠	:
DocN	0	1		1

Where's the information for each word's topic? Reconsider document-term matrix

	$Word_1$	Word <sub>2</sub>		ر Word
Doc <sub>1</sub>	0	1		0
$Doc_2$	2	0		3
:	:	÷	٠	:
$Doc_{N}$	0	1		1

Inner product of Documents (rows):  $\mathbf{Doc}_{i}^{'}\mathbf{Doc}_{l}$ 

Where's the information for each word's topic? Reconsider document-term matrix

	$Word_1$	Word <sub>2</sub>		ر Word
Doc <sub>1</sub>	0	1		0
$Doc_2$	2	0		3
:	:	:	٠	:
$Doc_{\mathcal{N}}$	0	1		1

Inner product of Documents (rows):  $\mathbf{Doc}_{i}^{'}\mathbf{Doc}_{l}$ 

Inner product of Terms (columns):  $\mathbf{Word}_{j}^{'}\mathbf{Word}_{k}$ 

Where's the information for each word's topic? Reconsider document-term matrix

	$Word_1$	Word <sub>2</sub>		Word <sub>J</sub>
Doc <sub>1</sub>	0	1		0
$Doc_2$	2	0		3
:	:	:	٠	:
$Doc_{\mathcal{N}}$	0	1		1

Inner product of Documents (rows):  $\mathbf{Doc}_{i}^{'}\mathbf{Doc}_{l}$ 

Inner product of Terms (columns):  $\mathbf{Word}_{j}'\mathbf{Word}_{k}$ 

Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)

Where's the information for each word's topic? Reconsider document-term matrix

	$Word_1$	Word <sub>2</sub>		Word <sub>J</sub>
Doc <sub>1</sub>	0	1		0
$Doc_2$	2	0		3
÷	:	:	٠	÷
$Doc_{\mathcal{N}}$	0	1		1

Inner product of Documents (rows):  $\mathbf{Doc}_{i}^{'}\mathbf{Doc}_{l}$ 

Inner product of Terms (columns):  $\mathbf{Word}_{j}'\mathbf{Word}_{k}$ 

Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)

Latent Semantic Analysis: Reduce information in matrix using linear

algebra (provides similar results, difficult to generalize)

Where's the information for each word's topic? Reconsider document-term matrix

	$Word_1$	Word <sub>2</sub>		Word <sub>J</sub>
Doc <sub>1</sub>	0	1		0
$Doc_2$	2	0		3
÷	:	:	٠	:
$Doc_{N}$	0	1		1

Inner product of Documents (rows):  $\mathbf{Doc}_{i}^{'}\mathbf{Doc}_{l}$ 

Inner product of Terms (columns):  $\mathbf{Word}_{i}^{'}\mathbf{Word}_{k}$ 

Allows: measure of correlation of term usage across documents

(heuristically: partition words, based on usage in documents)

Latent Semantic Analysis: Reduce information in matrix using linear

algebra (provides similar results, difficult to generalize)

Biclustering: Models that partition documents and words simultaneously

Why does this work → Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \propto p(\alpha) p(\boldsymbol{\pi} | \alpha) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

Why does this work $\rightsquigarrow$  Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\pi, T, \Theta, \alpha | X) \propto p(\alpha)p(\pi | \alpha)p(T | \pi)\underbrace{p(X | \theta, T)}_{1}$$

1)  $\theta \leadsto$  Greater weight on terms that occur together

Why does this work $\rightsquigarrow$  Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\pi, T, \Theta, \alpha | X) \propto p(\alpha)p(\pi | \alpha) \underbrace{p(T | \pi)}_{2} \underbrace{p(X | \theta, T)}_{1}$$

- 1)  $\theta \leadsto$  Greater weight on terms that occur together
- 2)  $\pi \leadsto \text{Greater weight on indicators that appear more regularly}$

Why does this work $\rightsquigarrow$  Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \propto p(\alpha) \underbrace{p(\boldsymbol{\pi} | \alpha)}_{3} \underbrace{p(\boldsymbol{T} | \boldsymbol{\pi})}_{2} \underbrace{p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

- 1)  $\theta \leadsto$  Greater weight on terms that occur together
- 2)  $\pi \leadsto \text{Greater weight on indicators that appear more regularly}$
- 3)  $\alpha \leadsto \mathsf{Emphasis}$  on  $\pi$  with greater weight

### Validation → Topic Intrusion

#### Discussed several validations

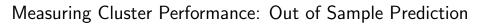
- Labeling paragraphs
  - Identify separating words automatically
  - Label topics manually (read!)
- Statistical methods
  - 1) Entropy
  - 2) Exclusivity
  - 3) Cohesiveness
- Experiment Based Methods
  - Word intrusion → topic validity
  - Topic intrusion → model fit

### Validation → Topic Intrusion

- 1) Ask research assistant to read paragraph
- 2) Construct experiment
  - For the document, select top three topics
  - Select a fourth topic
  - Show participant, ask her/him to identify intruder

Higher identification → topics are a better model of text

How well does our model perform?



How well does our model perform? $\leadsto$  predict new documents?

How well does our model perform? → predict new documents? Problem

How well does our model perform? → predict new documents? Problem → in sample evaluation leads to overfit.

How well does our model perform?→ predict new documents? Problem→ in sample evaluation leads to overfit. Solution→ evaluate performance on held out data with perplexity

- Prediction → One Task

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it?

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

- Compare perplexity with human based evaluations

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy → measure quality in topics and clusters

(Roberts, et al 2017

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy → measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al 2017 Forthcoming)

- Prediction → One Task
- Do we care about it? → Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves"):

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy → measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al 2017 Forthcoming)
- Experiments: measure topic and cluster quality

Mathematical approaches

Mathematical approaches → suppose we can capture quality with numbers assumes we're in the model → including text representation

Mathematical approaches → suppose we can capture quality with numbers assumes we're in the model → including text representation

Humans → read texts

Mathematical approaches→ suppose we can capture quality with numbers assumes we're in the model→ including text representation Humans→ read texts Humans→ use cluster output

Mathematical approaches → suppose we can capture quality with numbers assumes we're in the model → including text representation

Humans → read texts

Humans → use cluster output

Do humans think the model is performing well?

Mathematical approaches → suppose we can capture quality with numbers assumes we're in the model → including text representation

Humans → read texts

Humans → use cluster output

Do humans think the model is performing well?

1) Topic Quality

Mathematical approaches → suppose we can capture quality with numbers assumes we're in the model → including text representation

Humans → read texts

Humans → use cluster output

Do humans think the model is performing well?

- 1) Topic Quality
- 2) Cluster Quality

- 1) Take *M* top words for a topic
- 2) Randomly select a top word from another topic
  - 2a) Sample the topic number from I from K-1 (uniform probability)
  - 2b) Sample word j from the M top words in topic l
  - 2c) Permute the words and randomly insert the intruder:
    - List:

test = 
$$(v_{k,3}, v_{k,1}, v_{l,j}, v_{k,2}, v_{k,4}, v_{k,5})$$

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

bowl, flooding, olympic, olympics, nfl, coach

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

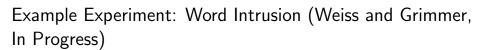
bowl, flooding, olympic, olympics, nfl, coach

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

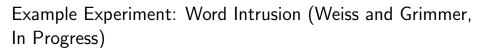
stocks, investors, fed, guns, trading, earning

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

stocks, investors, fed, guns, trading, earning



Higher rate of intruder identification → more exclusive/cohesive topics



Higher rate of intruder identification → more exclusive/cohesive topics

Deploy on Mechanical Turk

- Why is Japan revising its constitution?
- IR question: why is Japan now willing to engage militaristic foreign action?
- One explanation: election reform in 1993, changed electoral incentives
- To answer well: characterize campaigns across 50 + years
  - That sounds hard
  - That sounds impossible
- Determined (relentless) data collection
- Latent Dirichlet Allocation (on japanese texts)

#### Japanese Elections:

- Election Administration Commission runs elections  $\rightarrow$  district level

- Election Administration Commission runs elections  $\rightarrow$  district level
- Required to submit manifestos for all candidates to National Diet

#### Typical Manifesto:



- Election Administration Commission runs elections  $\rightarrow$  district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009

- Election Administration Commission runs elections  $\rightarrow$  district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level

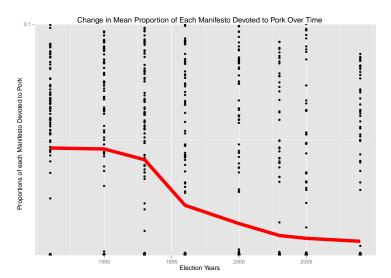
- Election Administration Commission runs elections  $\rightarrow$  district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm

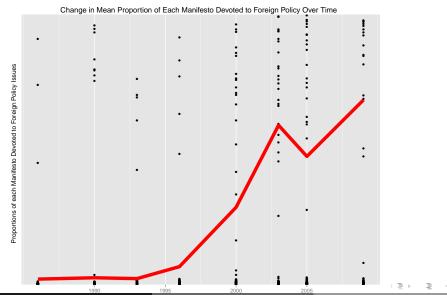
- Election Administration Commission runs elections  $\rightarrow$  district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM

- Election Administration Commission runs elections  $\rightarrow$  district level
- Required to submit manifestos for all candidates to National Diet
- Collected from 1950- 2009
  - Available only at district level
  - Until: 2009 national library made texts available on microfilm
- Collected from microfilm, hand transcribed (no OCR worked), used a variety of techniques to create a TDM
- Harder for Japanese

- Applies Vanilla LDA (using R Code )
- Output: topics (with Japanese characters)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic
改革	年金	推進	Z ·	政治	日本
郵政	P	整備	政策	改革	国
民営	廃止	図る	地域	国民	外交
小泉	改革	つとめる	まち	企業	国家
構造	JR St.	社会	鹿児島	自民党	社会
政府	実現	対策	全力	日本	国民
官	無駄	振興	選挙	共産党	保障
推進	日本	充実	国政	献金	安全
民	増税	促進	作り	金権	地域
自民党	削減	安定	横浜	党	拉致
日本	一元化	確立	対策	選挙	経済
制度	政権	企業	中小	禁止	守る
民間	子供	実現	発電	憲法	問題
年金	地域	中小	推進	腐敗	北朝魚
実現	ひと	育成	エネルギー	団体	教育
進める	サラリーマン	制度	企業	X	責任
断行	制度	政治	声	ソ連	力
地方	議員	地域	実現	守る	創る
止める	金	福祉	活性	平和	安心
保障	民主党	事業	自民党	H	目指す
財政	年間	改革	地方	反対	誇り
作る	一掃	確保	尽くす	真	憲法
贊成	郵政	強化	商店	是正	可能
社会	道路	教育	いかす	一掃	道
国民	交代	施設	全国	悪政	未来
公務員	社会保険庁	生活	政党	抜本	ひと
カ	月額	支援	ひと	定数	再生
経済	手当	環境	支援	政党	将来
<b>E</b>	談合	発展	経済	金丸	解決
安心	支援	施策	福祉	改悪	基本
	Padusing Wester I Dublic				
Postal privatization	Reducing Wasteful Public Spending	Pork for the District	Policies for the district	Political Reform	Natio





### REPRESENTATIONAL STYLE IN CONGRESS

What Legislators Say and Why It Matters

#### JUSTIN GRIMMER



#### Example 2: Automated Literature Reviews

Recall: literature reviews are hard to conduct LDA: developed (in part) to help structure JSTOR database Use JSTOR's research service to obtain data to analyze

Question: How do scholars use classic text: Home Style

Analysis: all articles that cite Home Style in JSTOR's data

#### Example 2: Automated Literature Reviews

#### Output: topic estimates

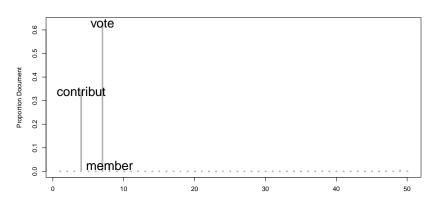
- Obtain  $\log \theta_k$  from model
- One method to summarize a topic:
  - $\exp(\log \theta_k)$  (select 10-20 biggest words)
  - $\exp(\log \theta_k)$  Average<sub> $j \neq k$ </sub>  $\exp(\log \theta_j)$  (select 10-20 biggest words)

#### Example 2: Automated Literature Reviews

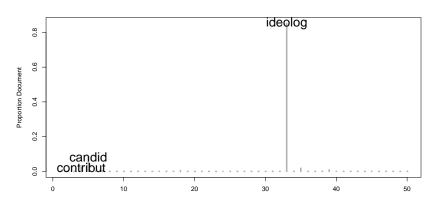
#### Example topics:

Label	Stems	Proportio
Life Style	member, district, attent, congress, time, cohort, retir	0.03
Comp.Home	constitu, mp, member, parti, role, local, british	0.02
Casework	casework, district, constitu, variabl, staff, congression, fiorina	0.03
Votes	vote,variabl,model,estim,measur,legisl,constitu	0.04
ld. Shirk	ideolog,vote,shirk,constitu,parti,senat,voter	0.03
C. letters	mail,govern, activ,respond,commun,offic	0.02

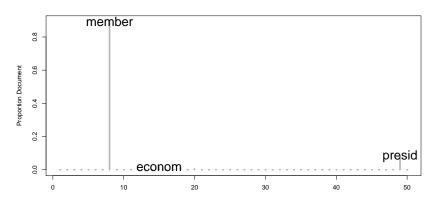
Wawro (2001) "A Panel Probit Analysis of Campaign Contributions and Roll Call Votes"



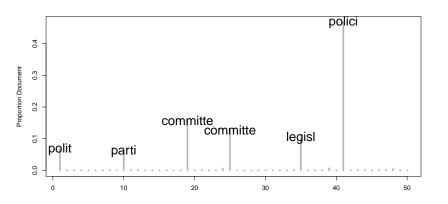
Bender (1996) "Legislator Voting and Shirking A Critical Review of the Literature"

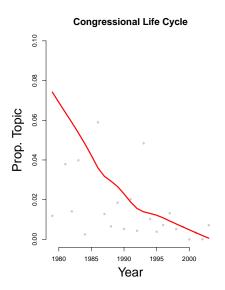


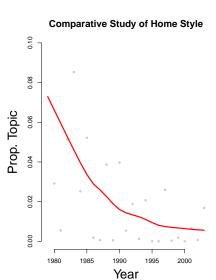
Parker (1980) "Cycles in Congressional District Attention"



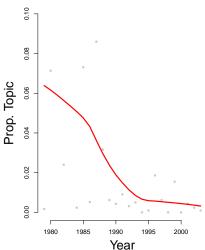
Shepsle (1985) "Policy Consequences of Government by Congressional Subcommittees"



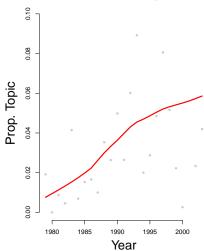


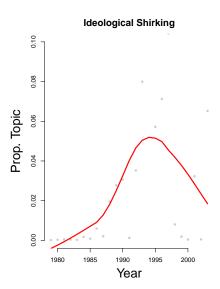


#### Casework and the Incumbency Advantage

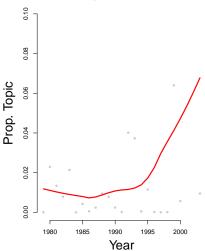


#### **Causes of Roll Call Voting Decisions**





#### **Biases in Congressional Communication**



# • The IMPRESSION of INFLUENCE

Legislator Communication, Representation, and Democratic Accountability

JUSTIN GRIMMER SEAN J. WESTWOOD SOLOMON MESSING What legislators claim (Grimmer, Westwood, Messing 2014)

### What legislators claim (Grimmer, Westwood, Messing 2014) → LDA credit claiming press releases

Labels Key Words Proportion

	8	
Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08

"Dave Camp announced today that he was able to secure \$2.5 million for widening M-72 from US-31 easterly 7.2 miles to Old M-72. The bill will now head to the Senate for consideration...We have two more hurdles to clear to make sure the money is in the bill when it hits the President's desk: a vote in the Senate and a conference committee" (Camp, 2005)

	<del> </del>	
Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08

"Congressman Doc Hastings has boosted federal funding for work on the Columbia Basin water supply for next year. Hastings has added \$400,000 for work on the Odessa Subaquifer, which when combined with the funding in the President's budget request, totals \$1 million for Fiscal Year 2009"... "Hastings' funding for the Odessa Subaquifer and Potholes Reservoir was included in the Fiscal Year 2009 Energy and Water Appropriations bill which was approved today by the full House Appropriations Committee. (Hastings, 2008)"

	<u> </u>	
Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire, grant, department, program, fire fighters	0.08

"Maurice Hinchey (D-NY) today announced that the West Endicott Fire Company has been awarded a \$17,051 federal grant to purchase approximately 10 sets of protective clothing, as well as radio equipment and air packs for its volunteer firefighters" (Hinchey, 2008)

	<u> </u>	
Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire, grant, department, program, fire fighters	0.08

"Congressman Pete Visclosky today announced that the Crown Point Fire Department will receive a \$16,550 Department of Homeland Security (DHS) grant to purchase a modular portable video system" (Visclosky, 2008)

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08
Stimulus	recovery,funding,jobs,information, act,	0.06

Labels	Key Words	Proportion
Requested appropriations	bill,funding,house,million,appropriations	0.08
Fire department grants	fire,grant,department,program,firefighters	0.08
Stimulus	recovery,funding,jobs,information, act,	0.06
Transportation	transportation, project, airport, transit, million	0.06

### Correlated Topic Models

Dirichlet distribution → Assumes negative covariance between topics Logistic Normal Distribution → Allows some positive covariance between topics

$$egin{array}{lll} oldsymbol{ heta}_k & \sim & \mathsf{Dirichlet}(\mathbf{1}) \ oldsymbol{\eta}_i | oldsymbol{\mu}, oldsymbol{\Sigma} & \sim & \mathsf{Multivariate} \; \mathsf{Normal}(oldsymbol{\mu}, oldsymbol{\Sigma}) \ oldsymbol{\pi}_i & = & \dfrac{\exp{(\eta_i)}}{\sum_{k=1}^K \exp{(\eta_{ik})}} \ oldsymbol{ au}_{im} | oldsymbol{\pi}_i & \sim & \mathsf{Multinomial}(1, oldsymbol{\pi}_i) \ oldsymbol{x}_{im} | oldsymbol{ heta}_k, au_{imk} = 1 & \sim & \mathsf{Multinomial}(1, oldsymbol{ heta}_k) \end{array}$$

## Appendix: Estimating LDA

- 1) Variational Approximation
- 2) Collapsed Gibbs Sampling

Basic set up

Basic set up

Call  $q(\pi, \theta, T, \alpha)$  an arbitrary distribution the approximating distribution.

Basic set up

Call  $q(\pi, \theta, T, \alpha)$  an arbitrary distribution the approximating distribution.

Recall  $p(m{\pi}, m{ heta}, m{T}, m{lpha} | m{X})$  is the posterior

Basic set up

Call  $q(\pi, \theta, T, \alpha)$  an arbitrary distribution the approximating distribution.

Recall  $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$  is the posterior

Our goal is to make  $q(\pi, \theta, T)$  as close as possible to  $p(\pi, \theta, T, \alpha | X)$ .

Basic set up

Call  $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$  an arbitrary distribution the approximating distribution. Recall  $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$  is the posterior

Our goal is to make  $q(\pi, \theta, T)$  as close as possible to  $p(\pi, \theta, T, \alpha | X)$ .

$$q(\pi, \theta, T, \alpha)^* = \arg\min_{q(\pi, \theta, T, \alpha)} \mathsf{KL}(q(\pi, \theta, T, \alpha)) || p(\pi, \theta, T, \alpha | X)))$$

Basic set up

Call  $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$  an arbitrary distribution the approximating distribution. Recall  $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$  is the posterior

Our goal is to make  $q(\pi, \theta, T)$  as close as possible to  $p(\pi, \theta, T, \alpha | X)$ .

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\alpha)^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\alpha)} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\alpha)) || p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\alpha|\boldsymbol{X})))$$

KL is the Kullback-Leibler Divergence between  $q(\pi, \theta, T, \alpha)$  and  $p(\pi, \theta, T, \alpha | X)$ .

Basic set up

Call  $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$  an arbitrary distribution the approximating distribution. Recall  $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$  is the posterior

Our goal is to make  $q(\pi, \theta, T)$  as close as possible to  $p(\pi, \theta, T, \alpha | X)$ .

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})) || p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X})))$$

KL is the Kullback-Leibler Divergence between  $q(\pi, \theta, T, \alpha)$  and  $p(\pi, \theta, T, \alpha | X)$ .

$$\mathsf{KL}(q||p) \ = \ -\sum_{\boldsymbol{\mathcal{T}}} \iint q(\boldsymbol{\mathcal{T}}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\boldsymbol{\mathcal{T}}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha} | \boldsymbol{X})}{q(\boldsymbol{\mathcal{T}}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha})} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta}$$

Basic set up

Call  $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})$  an arbitrary distribution the approximating distribution. Recall  $p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$  is the posterior

Our goal is to make  $q(\pi, \theta, T)$  as close as possible to  $p(\pi, \theta, T, \alpha | X)$ .

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})) || p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X})))$$

KL is the Kullback-Leibler Divergence between  $q(\pi, \theta, T, \alpha)$  and  $p(\pi, \theta, T, \alpha | X)$ .

$$\mathsf{KL}(q||p) = -\sum_{\boldsymbol{T}} \iint q(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha} | \boldsymbol{X})}{q(\boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\alpha})} \right\} d\boldsymbol{\pi} d\boldsymbol{\theta}$$

KL-divergence measures dissimilarity between two distributions.

4 D > 4 B > 4 E > 4 E > 9 Q C

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* \ = \ \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* \ = \ \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* \ = \ \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q then,  $q(\pi, \theta, T, \alpha)^* = p(\pi, \theta, T, \alpha | X)$ 

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q then,  $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha})^* = p(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha} | \boldsymbol{X})$ Simplifying Assumption:  $q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{\alpha}) \equiv q(\boldsymbol{\pi})q(\boldsymbol{\theta})q(\boldsymbol{T})q(\boldsymbol{\alpha})$ .

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q then,  $q(\pi, \theta, T, \alpha)^* = p(\pi, \theta, T, \alpha | X)$ Simplifying Assumption:  $q(\pi, \theta, T, \alpha) \equiv q(\pi)q(\theta)q(T)q(\alpha)$ . Sufficient to make inference tractable

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q then,  $q(\pi, \theta, T, \alpha)^* = p(\pi, \theta, T, \alpha | X)$ Simplifying Assumption:  $q(\pi, \theta, T, \alpha) \equiv q(\pi)q(\theta)q(T)q(\alpha)$ . Sufficient to make inference tractable!

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q then,  $q(\pi, \theta, T, \alpha)^* = p(\pi, \theta, T, \alpha | X)$ Simplifying Assumption:  $q(\pi, \theta, T, \alpha) \equiv q(\pi)q(\theta)q(T)q(\alpha)$ .

Sufficient to make inference tractable!

So, how do we minimize KL-divergence with respect to

$$q(\boldsymbol{\pi})q(\boldsymbol{\theta})q(\boldsymbol{T})q(\boldsymbol{\alpha})$$
?

Variational Approximation

$$q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})^* = \arg\min_{q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})} \mathsf{KL}(q(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha})||p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{\alpha}|\boldsymbol{X}))$$

No assumptions about q then,  $q(\pi, \theta, T, \alpha)^* = p(\pi, \theta, T, \alpha | X)$ 

Simplifying Assumption: 
$$q(\pi, \theta, T, \alpha) \equiv q(\pi)q(\theta)q(T)q(\alpha)$$
.

Sufficient to make inference tractable!

So, how do we minimize KL-divergence with respect to

$$q(\boldsymbol{\pi})q(\boldsymbol{\theta})q(\boldsymbol{T})q(\boldsymbol{\alpha})$$
?

We solve an equivalent maximization problem

 $\log p(\mathbf{Y})$ 

$$\log p(\mathbf{Y}) = \log \sum_{\mathbf{T}} \iint p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha) d\theta d\pi$$

$$\log p(\mathbf{Y}) = \log \sum_{\mathbf{T}} \iint p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha) d\theta d\pi$$
$$= \log \sum_{\mathbf{T}} \iint \frac{q(\pi, \theta, \mathbf{T}, \alpha)}{q(\pi, \theta, \mathbf{T}, \alpha)} p(\mathbf{X}, \mathbf{T}, \pi, \theta) d\theta d\pi$$

$$\log p(\mathbf{Y}) = \log \sum_{\mathbf{T}} \iint p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha) d\theta d\pi$$

$$= \log \sum_{\mathbf{T}} \iint \frac{q(\pi, \theta, \mathbf{T}, \alpha)}{q(\pi, \theta, \mathbf{T}, \alpha)} p(\mathbf{X}, \mathbf{T}, \pi, \theta) d\theta d\pi$$

$$\geq \sum_{\mathbf{T}} \iint q(\pi, \theta, \mathbf{T}, \alpha) \log \left\{ \frac{p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha)}{q(\pi, \theta, \mathbf{T}, \alpha)} \right\} d\pi d\theta$$

$$\log p(\mathbf{Y}) = \log \sum_{\mathbf{T}} \iint p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha) d\theta d\pi$$

$$= \log \sum_{\mathbf{T}} \iint \frac{q(\pi, \theta, \mathbf{T}, \alpha)}{q(\pi, \theta, \mathbf{T}, \alpha)} p(\mathbf{X}, \mathbf{T}, \pi, \theta) d\theta d\pi$$

$$\geq \sum_{\mathbf{T}} \iint q(\pi, \theta, \mathbf{T}, \alpha) \log \left\{ \frac{p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha)}{q(\pi, \theta, \mathbf{T}, \alpha)} \right\} d\pi d\theta$$

$$\sum_{\mathbf{T}} \iint q(\pi, \theta, \mathbf{T}, \alpha) \log \left\{ \frac{p(\mathbf{X}, \mathbf{T}, \pi, \theta, \alpha)}{q(\pi, \theta, \mathbf{T}, \alpha)} \right\} d\pi d\theta = \mathcal{L}(q)$$

$$\log p(\mathbf{Y}) =$$

$$\log p(\mathbf{Y}) = \mathsf{KL}(q||p) + \mathcal{L}(q)$$

$$\underbrace{\log p(\mathbf{Y})}_{\mathsf{Fixed Number}} = \mathsf{KL}(q||p) + \mathcal{L}(q)$$

$$\underbrace{\log p(m{Y})}_{\mathsf{Fixed Number}} = \mathsf{KL}(q||p) + \mathcal{L}(q)$$

If  $\mathcal{L}(q)$  increases

$$\underbrace{\log p(Y)}_{\mathsf{Fixed Number}} = \mathsf{KL}(q||p) + \mathcal{L}(q)$$

If  $\mathcal{L}(q)$  increases then  $\mathsf{KL}(q||p)$  must decrease.

$$\underbrace{\log p(\mathbf{Y})}_{\mathsf{Fixed Number}} = \mathsf{KL}(q||p) + \mathcal{L}(q)$$

If  $\mathcal{L}(q)$  increases then  $\mathsf{KL}(q||p)$  must decrease. Choose q to maximize  $\mathcal{L}(q)$ 

$$\underbrace{\log p(\mathbf{Y})}_{\mathsf{Fixed Number}} = \mathsf{KL}(q||p) + \mathcal{L}(q)$$

If  $\mathcal{L}(q)$  increases then  $\mathsf{KL}(q||p)$  must decrease. Choose q to maximize  $\mathcal{L}(q)$  equivalent to minimizing  $\mathsf{KL}(q||p)$ .

Iterative algorithm to maximize  $\mathcal{L}(q)$ .

Iterative algorithm to maximize  $\mathcal{L}(q)$ . Initialize  $q(\pi)^{\text{old}}$ ,  $q(\theta)^{\text{old}}$ ,  $q(\mathcal{T})^{\text{old}}$ , and  $q(\alpha^{\text{old}})$ 

Iterative algorithm to maximize  $\mathcal{L}(q)$ . Initialize  $q(\pi)^{\mathrm{old}}$ ,  $q(\theta)^{\mathrm{old}}$ ,  $q(\mathcal{T})^{\mathrm{old}}$ , and  $q(\alpha^{\mathrm{old}})$  Find  $q(\pi)^{\mathrm{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\mathrm{old}}$ , and  $q(\mathcal{T})^{\mathrm{old}}$  constant

Iterative algorithm to maximize  $\mathcal{L}(q)$ . Initialize  $q(\pi)^{\text{old}}$ ,  $q(\theta)^{\text{old}}$ ,  $q(\mathcal{T})^{\text{old}}$ , and  $q(\alpha^{\text{old}})$  Find  $q(\pi)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{old}}$ , and  $q(\mathcal{T})^{\text{old}}$  constant Find  $q(\theta)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\mathcal{T})^{\text{old}}$ , and  $q(\pi)^{\text{new}}$  constant

Iterative algorithm to maximize  $\mathcal{L}(q)$ . Initialize  $q(\pi)^{\text{old}}$ ,  $q(\theta)^{\text{old}}$ ,  $q(\mathcal{T})^{\text{old}}$ , and  $q(\alpha^{\text{old}})$  Find  $q(\pi)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{old}}$ , and  $q(\mathcal{T})^{\text{old}}$  constant Find  $q(\theta)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\mathcal{T})^{\text{old}}$ , and  $q(\pi)^{\text{new}}$  constant Find  $q(\mathcal{T})^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{new}}$ , and  $q(\pi)^{\text{new}}$  constant

Iterative algorithm to maximize  $\mathcal{L}(q)$ . Initialize  $q(\pi)^{\text{old}}$ ,  $q(\theta)^{\text{old}}$ ,  $q(\mathcal{T})^{\text{old}}$ , and  $q(\alpha^{\text{old}})$  Find  $q(\pi)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{old}}$ , and  $q(\mathcal{T})^{\text{old}}$  constant Find  $q(\theta)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\mathcal{T})^{\text{old}}$ , and  $q(\pi)^{\text{new}}$  constant Find  $q(\mathcal{T})^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{new}}$ , and  $q(\pi)^{\text{new}}$  constant Assume  $q(\alpha)$  is degenerate, maximization step

Iterative algorithm to maximize  $\mathcal{L}(q)$ . Initialize  $q(\pi)^{\text{old}}$ ,  $q(\theta)^{\text{old}}$ ,  $q(\mathcal{T})^{\text{old}}$ , and  $q(\alpha^{\text{old}})$  Find  $q(\pi)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{old}}$ , and  $q(\mathcal{T})^{\text{old}}$  constant Find  $q(\theta)^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\mathcal{T})^{\text{old}}$ , and  $q(\pi)^{\text{new}}$  constant Find  $q(\mathcal{T})^{\text{new}}$  to max  $\mathcal{L}(q)$ - holding  $q(\theta)^{\text{new}}$ , and  $q(\pi)^{\text{new}}$  constant Assume  $q(\alpha)$  is degenerate, maximization step Guaranteed convergence:  $\mathcal{L}(q)$  is convex in  $q(\pi)$ ,  $q(\theta)$ , and  $q(\mathcal{T})$ 

Finding  $q(\pi)^{\text{new}}$ .

Finding  $q(\pi)^{\text{new}}$ .

$$\mathcal{L}(q) = \int q(\boldsymbol{\pi})^{\text{new}} \underbrace{\left\{ \sum_{\boldsymbol{T}} \int \log p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\pi}) q(\boldsymbol{T}, \boldsymbol{\alpha})^{\text{old}} q(\boldsymbol{\theta})^{\text{old}} d\boldsymbol{\theta} \right\}}_{\boldsymbol{E}_{\boldsymbol{T}, \boldsymbol{\theta}} [\log p(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta})]} d\boldsymbol{\pi}$$

$$- \int q(\boldsymbol{\pi})^{\text{new}} \log q(\boldsymbol{\pi})^{\text{new}} d\boldsymbol{\pi} + \text{constants}$$

Finding  $q(\pi)^{\text{new}}$ .

$$\mathcal{L}(q) = \int q(\pi)^{\text{new}} \underbrace{\left\{ \sum_{\mathbf{T}} \int \log p(\mathbf{X}, \mathbf{T}, \theta, \pi) q(\mathbf{T}, \alpha)^{\text{old}} q(\theta)^{\text{old}} d\theta \right\}}_{\mathsf{E}_{\mathbf{T}, \theta}[\log p(\mathbf{Y}, \mathbf{T}, \pi, \theta)]} d\pi$$

$$- \int q(\pi)^{\text{new}} \log q(\pi)^{\text{new}} d\pi + \text{constants}$$

$$\log \tilde{p}(\boldsymbol{\pi}) = \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\alpha})] + \mathsf{constants}$$

$$= \int q(\pi)^{\mathsf{new}} \log \left( \frac{\tilde{p}(\pi)}{q(\pi)^{\mathsf{new}}} \right) d\pi$$

$$= \int q(\pi)^{\mathsf{new}} \log \left( \frac{\tilde{p}(\pi)}{q(\pi)^{\mathsf{new}}} \right) d\pi$$

$$= -\mathsf{KL}(q(\pi)^{\mathsf{new}} || \tilde{p}(\pi))$$

$$= \int q(\pi)^{\text{new}} \log \left( \frac{\tilde{p}(\pi)}{q(\pi)^{\text{new}}} \right) d\pi$$
$$= -\text{KL}(q(\pi)^{\text{new}} || \tilde{p}(\pi))$$

At a maximum when  $q(\pi)^{\text{new}} = \tilde{p}(\pi)$ 

$$= \int q(\pi)^{\text{new}} \log \left( \frac{\tilde{p}(\pi)}{q(\pi)^{\text{new}}} \right) d\pi$$

$$= -\text{KL}(q(\pi)^{\text{new}} || \tilde{p}(\pi))$$

At a maximum when  $q(\pi)^{\mathsf{new}} = \widetilde{\rho}(\pi)$  Equivalently,

$$\log q(\boldsymbol{\pi})^{\mathsf{new}} = \log \tilde{p}(\boldsymbol{\pi})$$
$$= \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\alpha})] + \mathsf{constants}$$

$$= \int q(\pi)^{\text{new}} \log \left( \frac{\tilde{\rho}(\pi)}{q(\pi)^{\text{new}}} \right) d\pi$$
$$= -\text{KL}(q(\pi)^{\text{new}} || \tilde{\rho}(\pi))$$

At a maximum when  $q(\pi)^{\mathsf{new}} = \widetilde{p}(\pi)$  Equivalently,

$$\begin{array}{rcl} \log q(\boldsymbol{\pi})^{\mathsf{new}} &=& \log \tilde{p}(\boldsymbol{\pi}) \\ &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\alpha})] + \mathsf{constants} \end{array}$$

Or,

$$= \int q(\pi)^{\text{new}} \log \left( \frac{\tilde{\rho}(\pi)}{q(\pi)^{\text{new}}} \right) d\pi$$
$$= -\text{KL}(q(\pi)^{\text{new}} || \tilde{\rho}(\pi))$$

At a maximum when  $q(\pi)^{\mathsf{new}} = \tilde{p}(\pi)$  Equivalently,

$$\log q(\boldsymbol{\pi})^{\mathsf{new}} = \log \tilde{p}(\boldsymbol{\pi})$$

$$= \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\alpha})] + \mathsf{constants}$$

Or,

$$q(\boldsymbol{\pi})^{\text{new}} = \frac{\exp\left\{\mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\alpha})]\right\}}{\int \exp\left\{\mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\alpha})]\right\}d\boldsymbol{\pi}}$$

Algorithm Initialize  $q(\pi)^{\mathrm{old}}$ ,  $q(\theta)^{\mathrm{old}}$ , and  $q(\mathbf{\textit{T}})^{\mathrm{old}}$ 

Algorithm Initialize  $q(\pi)^{\text{old}}$ ,  $q(\theta)^{\text{old}}$ , and  $q(T)^{\text{old}}$   $\log q(\pi)^{\text{new}} = \mathsf{E}_{T,\theta}[\log p(\boldsymbol{X},T,\theta,\pi,\alpha)] + \text{constants}$ 

4 D > 4 D >

Initialize  $q(\pi)^{\mathrm{old}}$ ,  $q(\theta)^{\mathrm{old}}$ , and  $q(T)^{\mathrm{old}}$ 

$$\log q(\pi)^{\text{new}} = \mathsf{E}_{T,\theta}[\log p(X,T,\theta,\pi,\alpha)] + \text{constants}$$
  
 $\log q(\theta)^{\text{new}} = \mathsf{E}_{T,\pi}[\log p(X,T,\theta,\pi,\alpha)] + \text{constants}$ 

Initialize  $q(\boldsymbol{\pi})^{\text{old}}$ ,  $q(\boldsymbol{\theta})^{\text{old}}$ , and  $q(\boldsymbol{T})^{\text{old}}$ 

$$\begin{array}{lll} \log q(\boldsymbol{\pi})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \\ \log q(\boldsymbol{\theta})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \\ \log q(\boldsymbol{T})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{\theta},\boldsymbol{\pi}}[\log p(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \end{array}$$

Initialize  $q(\boldsymbol{\pi})^{\mathrm{old}}$ ,  $q(\boldsymbol{\theta})^{\mathrm{old}}$ , and  $q(\boldsymbol{T})^{\mathrm{old}}$ 

$$\begin{array}{lll} \log q(\boldsymbol{\pi})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \\ \log q(\boldsymbol{\theta})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \\ \log q(\boldsymbol{T})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{\theta},\boldsymbol{\pi}}[\log p(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \end{array}$$

All expectations over approximating distribution.

Initialize  $q(\pi)^{\mathrm{old}}$ ,  $q(\theta)^{\mathrm{old}}$ , and  $q(T)^{\mathrm{old}}$ 

$$\begin{array}{lll} \log q(\boldsymbol{\pi})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\theta}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \\ \log q(\boldsymbol{\theta})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{T},\boldsymbol{\pi}}[\log p(\boldsymbol{X},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \\ \log q(\boldsymbol{T})^{\mathsf{new}} &=& \mathsf{E}_{\boldsymbol{\theta},\boldsymbol{\pi}}[\log p(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\alpha})] + \mathsf{constants} \end{array}$$

All expectations over approximating distribution. To compute, rely on factorization in posterior

#### Variational Approximation Update Steps

Carrying out the expectations, we obtain the following forms → derivation not assumption

$$q( au_{im}) = ext{Multinomial}(1, au_{im})$$
 $q( heta_k) = ext{Dirichlet}( extstyle{\eta}_k)$ 
 $q( au_i) = ext{Dirichlet}( extstyle{\gamma}_i)$ 

## Update for $q(\boldsymbol{\tau}_{im})$

Consider document i, word m, topic k

$$r_{imk} \propto \exp\left(I(x_{im} = j)E[\log \eta_{kj}] + E[\log \gamma_{ik}]\right)$$

$$\propto \exp\left(I(x_{im} = j)\left[\Psi(\eta_{kj}) - \Psi(\sum_{l=1}^{J} \eta_{kl})\right] + \Psi(\gamma_{ik}) - \Psi(\sum_{m=1}^{K} \gamma_{im})\right)$$

where  $\Psi(\cdot)$  is the digamma function (the derivative of the gamma function)

## Update for $q(\eta_k)$

Consider word j and topic k, then,

$$\eta_{jk} \propto 1 + \sum_{i=1}^{N} \sum_{m=1}^{M_i} r_{imk} x_{im}$$

# Update for $q(\gamma_i)$

Consider word topic k and document i

$$\gamma_{ik} \propto \alpha_k + \sum_{m=1}^{M_i} r_{imk}$$

Update for  $\alpha$ )

Fast Newton-Raphson Algorithm