Making Topic Models More Usable

CIKM 2015 Workshop

Topic Models: Post-Processing and Applications

Wray Buntine

Monash University
http://topicmodels.org

Monday 19th October, 2015

Overview

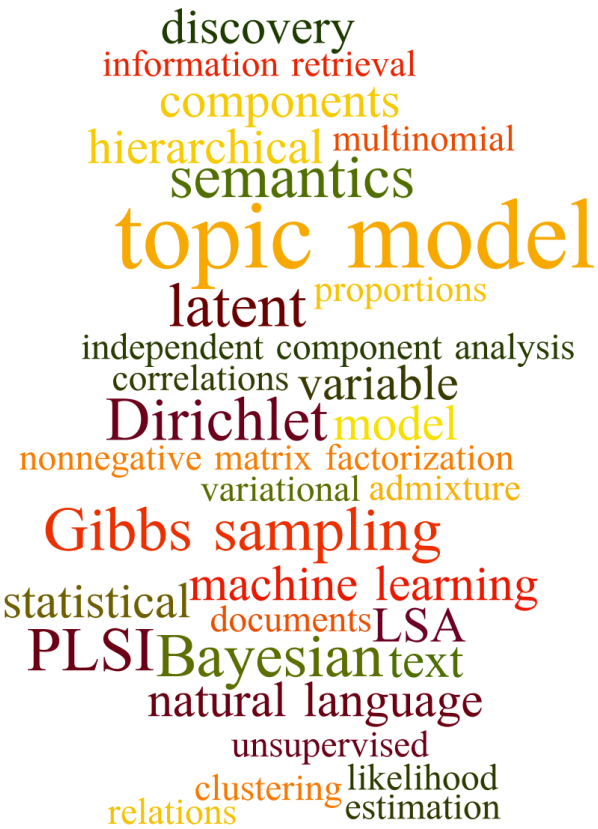
Have an improved topic model that:

- models prior topic proportions, allowing frequent and rare topics;
- damps down repeated words in documents like TF-IDF does;
- models the "background" distribution of words so topics are viewed as a departure from background;
- topics have a confidence parameter to help assess if "junk".

The resulting models:

- a factor larger in memory+time, but works multi-core;
- improves standard performance measures: test set perplexity and normalised PMI.

Outline



1 Topic Models Background

2 An Historical Perspective

3 Fully Nonparametric Topic Models

4 Alternatives

5 Conclusion

Topic Models Background

We will review essential background for later.

Component Models, Generally

image $\longrightarrow$

Prince, Queen, Elizabeth, title, son, ...	school, student, college, education, year, ...
John, David, Michael, Scott, Paul, ...	and, or, to , from, with, in, out, ...

text $\longrightarrow$

13 1995 accompany and(2) andrew at boys(2) charles close college day despite diana dr eton first for gayley harry here housemaster looking old on on school separation sept stayed the their(2) they to william(2) with year

Approximate faces/bag-of-words (RHS) with a linear combination of components (LHS).

Matrix Approximation View

$$\boldsymbol{W} \simeq \boldsymbol{L} * \boldsymbol{\Theta}$$

$$\overbrace{\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{pmatrix}}^{J \text{ words}} \left.\vphantom{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}\right\} \text{I documents} \simeq \overbrace{\begin{pmatrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{pmatrix}}^{K \text{ components}} * \overbrace{\begin{pmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{pmatrix}}^{J \text{ words}} \left.\vphantom{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}\right\} \text{K components}$$

Now:

$$\text{Learning} = \text{Representation} + \text{Model} + \text{Statistics} + \text{Optimisation}$$

So:

- what's the representation for the data in $\boldsymbol{W}$?
- how do we interpret "$\simeq$"?
- how do we regularise $\boldsymbol{\Theta}$ or put a prior in it?
- how do we treat the latent variables $\boldsymbol{L}$?

Matrix Approximation View, cont.

$$\boldsymbol{W} \simeq \boldsymbol{L} * \boldsymbol{\Theta}^T$$

$$\overbrace{\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{pmatrix}}^{J \text{ words}} \left.\vphantom{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}\right\} I \text{ documents} \simeq \overbrace{\begin{pmatrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{pmatrix}}^{K \text{ components}} * \overbrace{\begin{pmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{pmatrix}}^{J \text{ words}} \left.\vphantom{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}\right\} K \text{ components}$$

Data $\boldsymbol{W}$	Components $\boldsymbol{L}$	Error	Models
real valued	unconstrained	least squares	PCA and LSA
non-negative	non-negative	least squares	NMF, learning codebooks
non-neg int.	non-negative	cross-entropy	topic models
real valued	independent	small	ICA

Matrix Approximation View, cont.

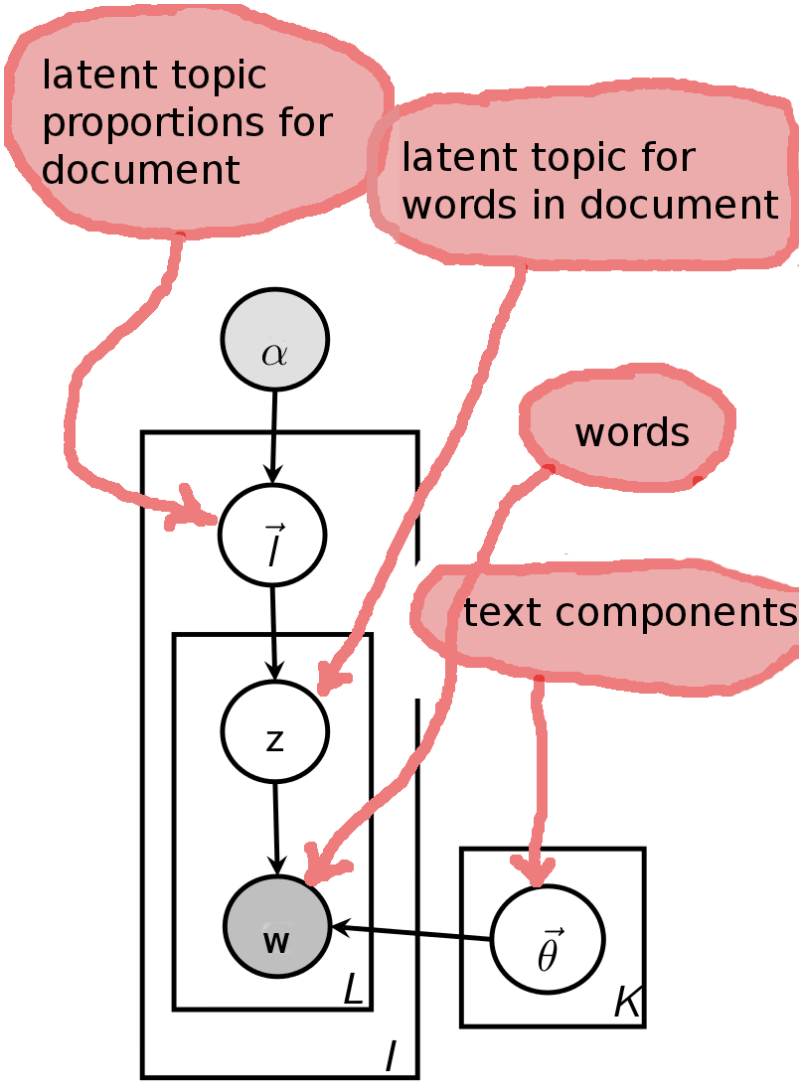
$$\boldsymbol{W} \simeq \boldsymbol{L} * \boldsymbol{\Theta}^T$$

$$\overbrace{\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{pmatrix}}^{J \text{ words}} \left.\vphantom{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}\right\} \text{I documents} \simeq \overbrace{\begin{pmatrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{pmatrix}}^{K \text{ components}} * \overbrace{\begin{pmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{pmatrix}}^{J \text{ words}} \left.\vphantom{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}\right\} \text{K components}$$

Data $\boldsymbol{W}$	Components $\boldsymbol{L}$	Error	Models
real valued	unconstrained	least squares	PCA and LSA
non-negative	non-negative	least squares	NMF, learning codebooks
non-neg int.	non-negative	cross-entropy	topic models
real valued	independent	small	ICA

Topic models are like other forms of independent component analysis.

LDA Topic Model



Each document i has a topic proportion vector:

$$\vec{l}_i \sim \text{Dirichlet}_K(\alpha) ,$$

Each topic k has a word probability vector:

$$\vec{\theta}_k \sim \text{Dirichlet}_J(\beta) ,$$

Generate topic id and word id for each place: for $i = 1, ..., I$, $l = 1, ..., L_i$:

$$z_{i,l} \sim \text{Discrete}\left(\vec{l}_i\right) ,$$

$$w_{i,l} \sim \text{Discrete}\left(\vec{\theta}_{z_{i,l}}\right) .$$

Learning Algorithms with Dirichlets in 1990's

Common text-book algorithms/methods in 1990s
machine-learning/statistics rely on Dirichlet distributions combined with:

- trees, tables,
- graphs, networks,
- context free grammars.

Algorithms on these combine simple Dirichlet methods with:

- model search, model averaging,
- EM and variational algorithms,
- Gibbs sampling,
- *etc.*

Learning Algorithms with Dirichlets in 1990's

Common text-book algorithms/methods in 1990s
machine-learning/statistics rely on Dirichlet distributions combined with:

- trees, tables,
- graphs, networks,
- context free grammars.

Algorithms on these combine simple Dirichlet methods with:

- model search, model averaging,
- EM and variational algorithms,
- Gibbs sampling,
- *etc.*

LDA was at the pinnacle of 1990s statistical machine learning.

Early Experiences: Wikipedia Model in 2006

We (Helsinki) did a 400 component topic of the one million (approx) Wikipedia articles in 2006.

NOUNS					
mythology	0.03337	God	0.02048	name	0.014747
goddess	0.012911	spirit	0.012639	legend	0.0087992
myth	0.0070882	demons	0.006807	Sun	0.0060099
Temple	0.0054717	deity	0.0054247	Bull	0.0051629
Dragon	0.0051379	Maya	0.0051243	King	0.00512
Sea	0.0049453	Norse	0.0044707	horse	0.0044592
symbol	0.0042196	animals	0.0040112	fire	0.0039879
hero	0.0038755	Romans	0.0038696	Apollo	0.0037588
VERBS					
called	0.034078	said	0.031081	see	0.029521
given	0.0269	associated	0.024591	according	0.021724
represented	0.020964	known	0.018896	could	0.017499
made	0.016952	depicted	0.01524	appeared	0.014662
ADJECTIVES					
Greek	0.091163	ancient	0.055393	great	0.02853
Egyptian	0.028071	Roman	0.025783	sacred	0.020446

Early Lessons from 2006

- topics based on grouping words, not always "themes"
  e.g. Catholic first names, numbers, mathematical symbols

Early Lessons from 2006

- topics based on grouping words, not always "themes"
  e.g. Catholic first names, numbers, mathematical symbols
- colocations or multiword terms a serious problem
  e.g. many names split
  e.g. the "new" topic

Early Lessons from 2006

- topics based on grouping words, not always "themes"
  e.g. Catholic first names, numbers, mathematical symbols
- colocations or multiword terms a serious problem
  e.g. many names split
  e.g. the "new" topic
- some topics have dual personalities
  e.g. the "Mother Teresa" topic (with top 10 words about Mother Teresa) would often be about maternal and matriarchical characteristics
  - some common uses of the topic not closely related to the top 10 words!

Early Lessons from 2006

- topics based on grouping words, not always "themes"
  e.g. Catholic first names, numbers, mathematical symbols
- colocations or multiword terms a serious problem
  e.g. many names split
  e.g. the "new" topic
- some topics have dual personalities
  e.g. the "Mother Teresa" topic (with top 10 words about Mother Teresa) would often be about maternal and matriarchical characteristics
    - some common uses of the topic not closely related to the top 10 words!
- junk topics not common but ever present

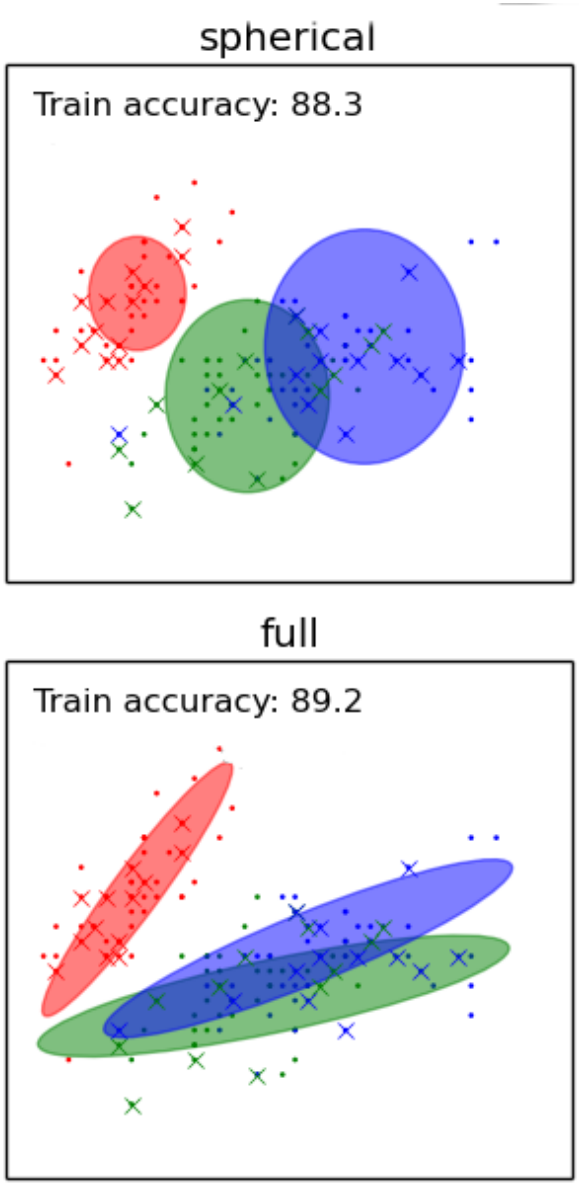
Early Lessons from 2006

- topics based on grouping words, not always "themes"
  e.g. Catholic first names, numbers, mathematical symbols
- colocations or multiword terms a serious problem
  e.g. many names split
  e.g. the "new" topic
- some topics have dual personalities
  e.g. the "Mother Teresa" topic (with top 10 words about Mother Teresa) would often be about maternal and matriarchical characteristics
    - some common uses of the topic not closely related to the top 10 words!
- junk topics not common but ever present
- topics seductively semantic, but quality of coherence varying
    - took smart empirical/NLP/IR folks to work on this

Early Lessons from 2006

- topics based on grouping words, not always "themes"
  e.g. Catholic first names, numbers, mathematical symbols
- colocations or multiword terms a serious problem
  e.g. many names split
  e.g. the "new" topic
- some topics have dual personalities
  e.g. the "Mother Teresa" topic (with top 10 words about Mother Teresa) would often be about maternal and matriarchical characteristics
    - some common uses of the topic not closely related to the top 10 words!
- junk topics not common but ever present
- topics seductively semantic, but quality of coherence varying
    - took smart empirical/NLP/IR folks to work on this
- topics pressured to be equiprobable

Equiprobable Topics



- note document topics are distributed with symmetric Dirichlet$_K$ (α)
- so document topics equiprobable:
  - some should be frequent, some rare!
- compare with Gaussian mixture model using spherical Gaussians (shown left)
  - empirically, when using PCA or ICA, we know components are rarely of equal size/proportion
- so we expect equiprobability to cause trouble for LDA

Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test”

Evaluation

David Lewis (Aug 2014) "topic models are like a Rorschach inkblot test"

Perplexity:

- measure of test set likelihood
  - *e.g.* "document completion," see Wallach, Murray, Salakhutdinov, and Mimno, *ICML* 2009
- it is not a bonafide evaluation task

Evaluation

David Lewis (Aug 2014) "topic models are like a Rorschach inkblot test"

Perplexity:

- measure of test set likelihood
  - *e.g.* "document completion," see Wallach, Murray, Salakhutdinov, and Mimno, *ICML* 2009
- it is not a bonafide evaluation task

PMI:

- measure of topic coherence
  - *e.g.* "average pointwise mutual information between all pairs of top 10 words in the topic," see Lau, Newman and Baldwin, *ACL*, 2014
- parallels development of measures of semantic relatedness from IR/NLP
- but at least it corresponds to a semi-realistic evaluation task

Evaluation

David Lewis (Aug 2014) "topic models are like a Rorschach inkblot test"

Perplexity:

- measure of test set likelihood
  - *e.g.* "document completion," see Wallach, Murray, Salakhutdinov, and Mimno, *ICML* 2009
- it is not a bonafide evaluation task

PMI:

- measure of topic coherence
  - *e.g.* "average pointwise mutual information between all pairs of top 10 words in the topic," see Lau, Newman and Baldwin, *ACL*, 2014
- parallels development of measures of semantic relatedness from IR/NLP
- but at least it corresponds to a semi-realistic evaluation task

In task:

- embedded in another performance task
- a real evaluation task with competition

Why Topic Models?

- *Topic models* discover hidden themes in text data to aid understanding.
  - when applied to tokens with semi-semantic interpretation;
  - doesn't work for low level signal data.

Why Topic Models?

- *Topic models* discover hidden themes in text data to aid understanding.
  - when applied to tokens with semi-semantic interpretation;
  - doesn't work for low level signal data.
- Recent research develops variations of topic models:
  - integrating multimedia, semi-structured data and network data;
  - higher performance, big data;
  - modelling sparsity, non-parametric methods;
  - extending for document summarisation.

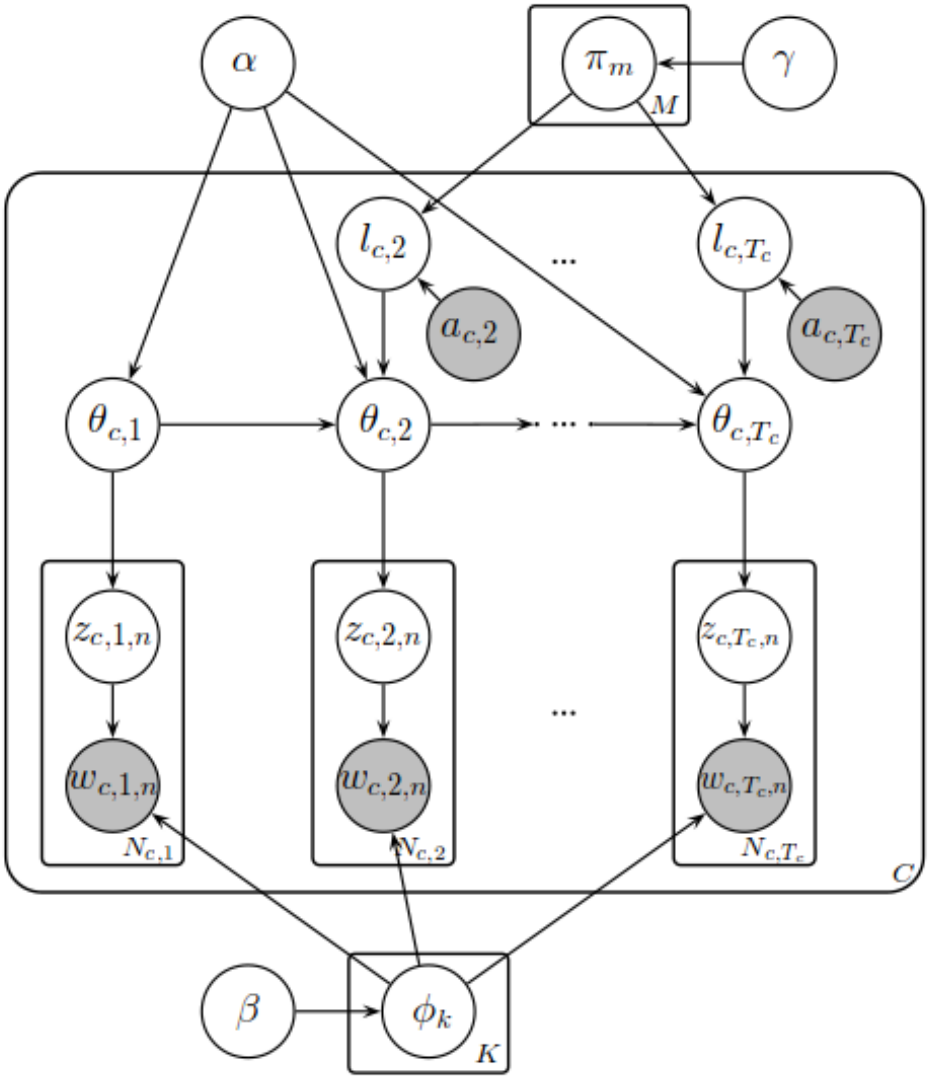
Why Topic Models?

- *Topic models* discover hidden themes in text data to aid understanding.
  - when applied to tokens with semi-semantic interpretation;
  - doesn't work for low level signal data.
- Recent research develops variations of topic models:
  - integrating multimedia, semi-structured data and network data;
  - higher performance, big data;
  - modelling sparsity, non-parametric methods;
  - extending for document summarisation.
- Some methods offer state-of-the-art performance on their task.
  - "Topic Segmentation with a Structured Topic Model", Du, Buntine and Johnson, *NAACL 2013*

Why Topic Models?

- *Topic models* discover hidden themes in text data to aid understanding.
  - when applied to tokens with semi-semantic interpretation;
  - doesn't work for low level signal data.
- Recent research develops variations of topic models:
  - integrating multimedia, semi-structured data and network data;
  - higher performance, big data;
  - modelling sparsity, non-parametric methods;
  - extending for document summarisation.
- Some methods offer state-of-the-art performance on their task.
  - "Topic Segmentation with a Structured Topic Model", Du, Buntine and Johnson, *NAACL 2013*
- Coherence remains a challenge.

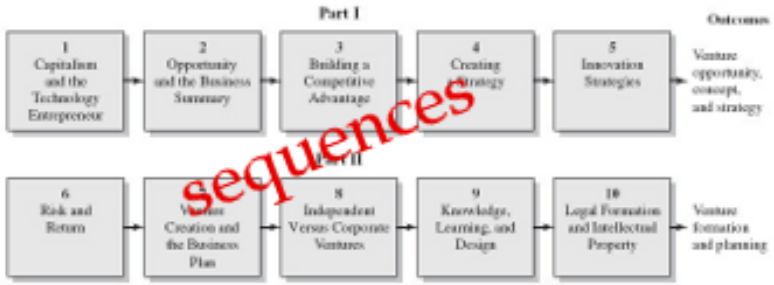
Example Extended Topic Model: SITS



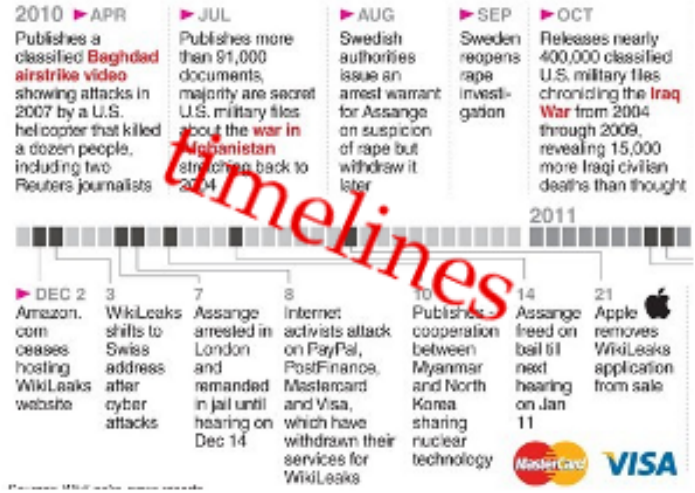
Speaker Identity for Topic Segmentation (SITS) model

- given text segments tagged with speaker identity
- assume topic shifts sometimes
- Nguyen, Boyd-Graber and Resnik, ACL 2012
- also have non-parametric version

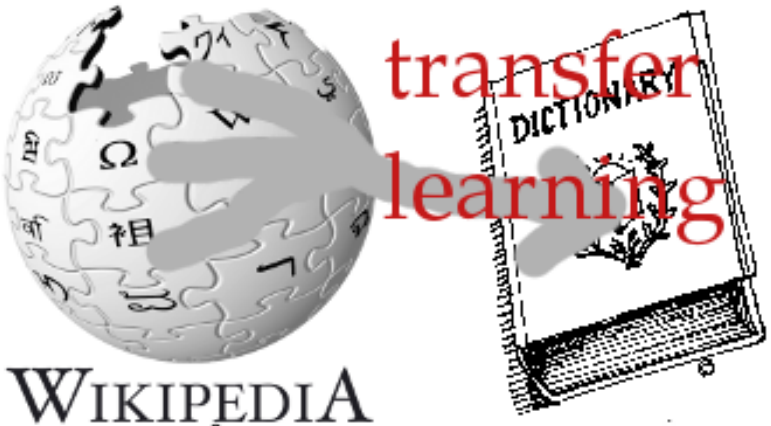
Open Territory for Applications/Extensions of TMs



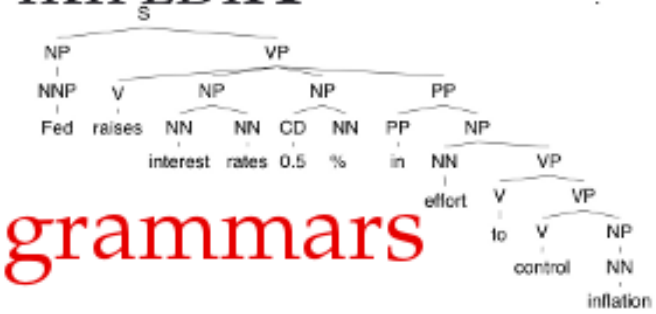
sequences



timelines



transfer learning

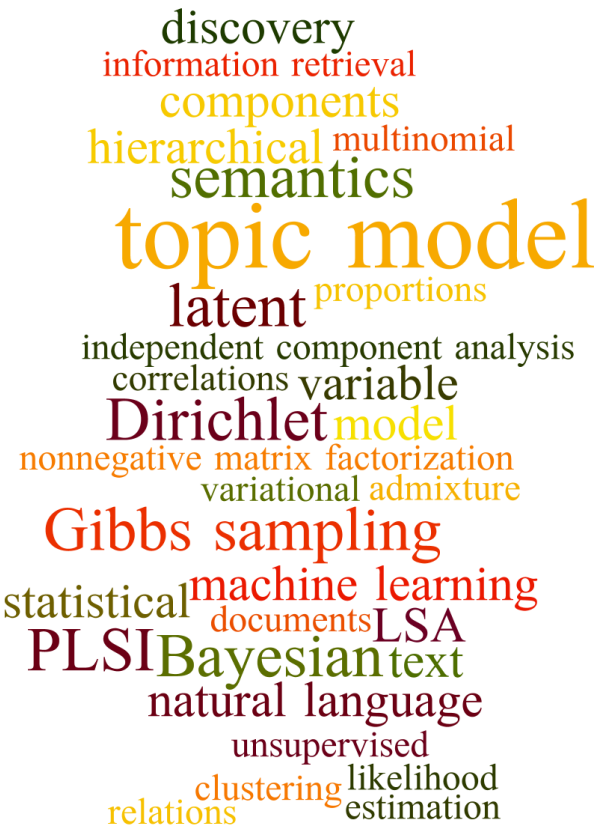


grammars

layouts


The Big Picture:
Document Summarisation to
alleviate Information Overload

Outline



1 Topic Models Background

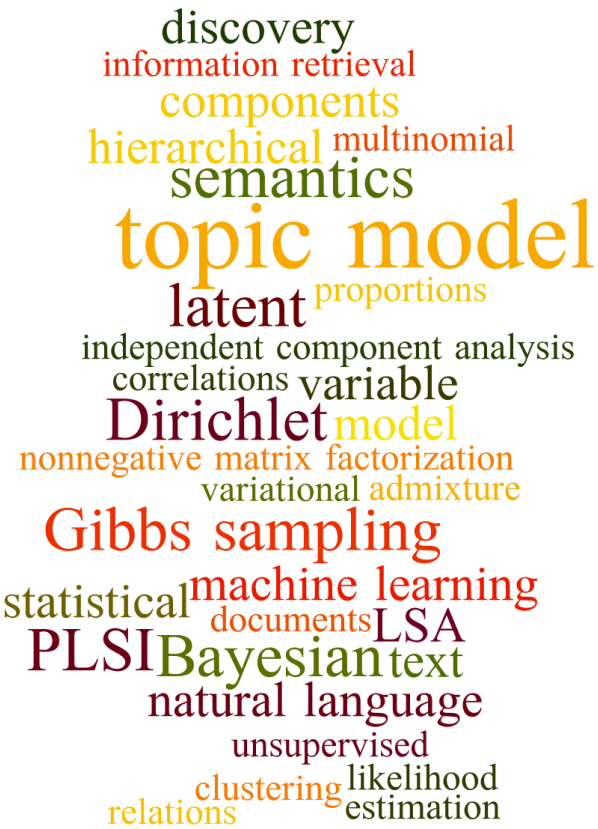
2 An Historical Perspective

3 Fully Nonparametric Topic Models

4 Alternatives

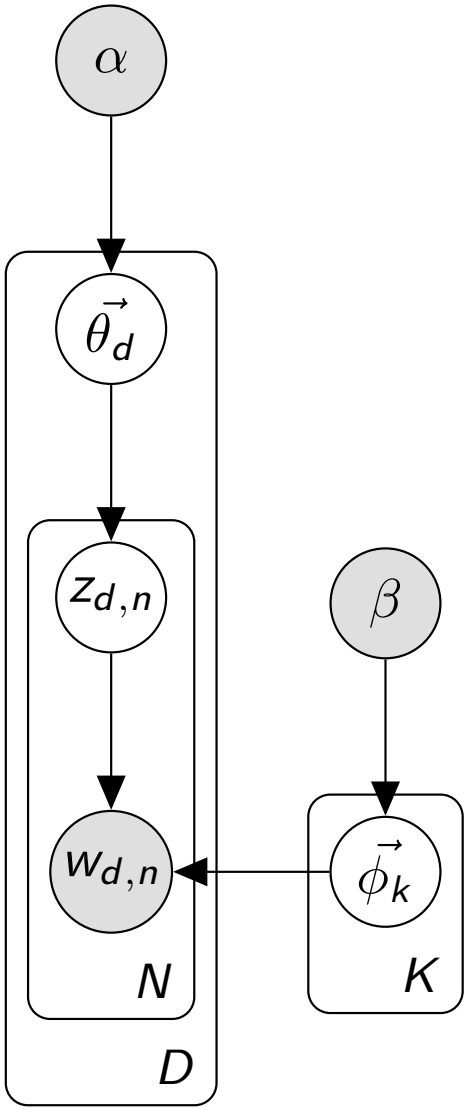
5 Conclusion

An Historical Perspective



Considering stops and starts leading to the current situation: why earlier non-parametric topic models didn't help.

Evolution of Models: LDA-Scalar (2003)



LDA-Scalar

original LDA

Estimating Priors

- a number of groups (independently) tried to replace α and β in LDA-scalar with vectors $\vec{\alpha}$ and $\vec{\beta}$.
- best known: "Rethinking LDA: Why priors matter," Wallach, Mimno, McCallum, *NIPS* 2009
- estimation techniques were primitive and only worked well for $\alpha \longrightarrow \vec{\alpha}$

i.e. similar attempts at $\beta \longrightarrow \vec{\beta}$ didn't work

- implemented by Mallet since 2008 as assymetric-symmetric LDA

Estimating Priors

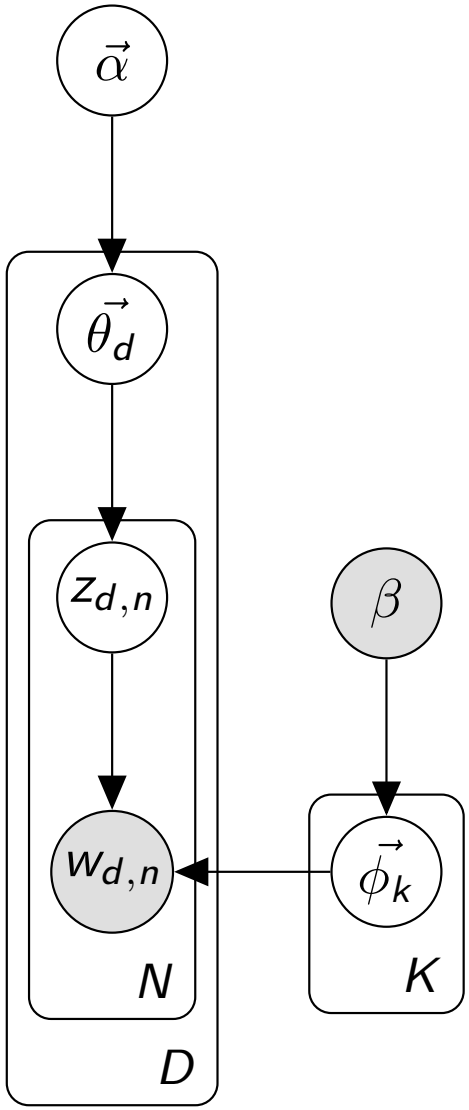
- a number of groups (independently) tried to replace α and β in LDA-scalar with vectors $\vec{\alpha}$ and $\vec{\beta}$.
- best known: "Rethinking LDA: Why priors matter," Wallach, Mimno, McCallum, *NIPS* 2009
- estimation techniques were primitive and only worked well for $\alpha \longrightarrow \vec{\alpha}$

i.e. similar attempts at $\beta \longrightarrow \vec{\beta}$ didn't work

- implemented by Mallet since 2008 as assymetric-symmetric LDA

With millions of model parameters and billions of data points (words), as a Bayesian it is lunacy not to introduce more hyper-parameters into LDA.

Evolution of Models: LDA-Vector (2008)



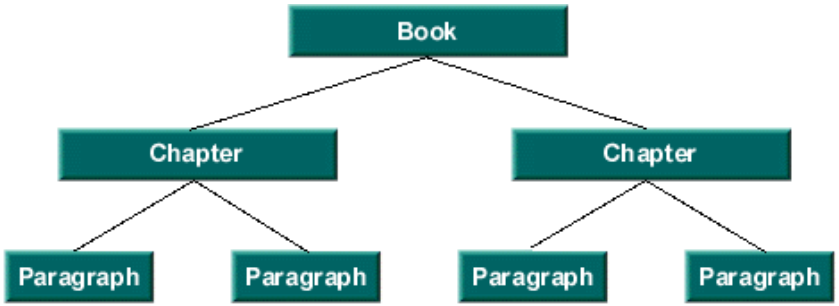
LDA-Vector

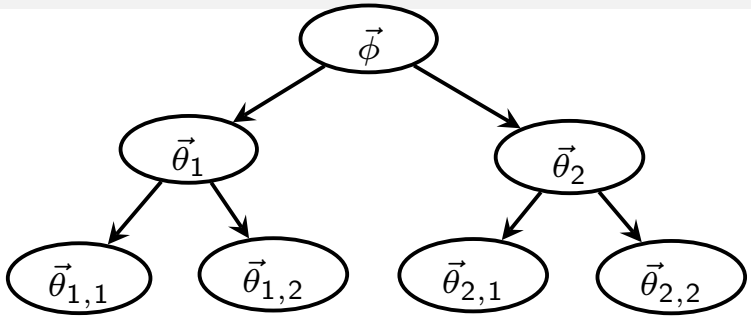
made α a vector and estimated it;

i.e., asymmetric Dirichlet prior of Wallach *et al.*;

is also truncated HDP-LDA (see next)

ASIDE: A Dirichlet Hierarchy





We might model a set of vocabularies/documents hierarchically:

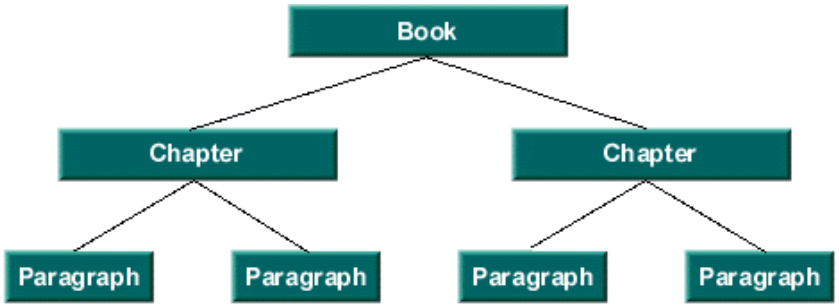
$$\vec{\theta}_1 \sim \text{Dirichlet}\left(\alpha_0 \vec{\phi}\right)$$

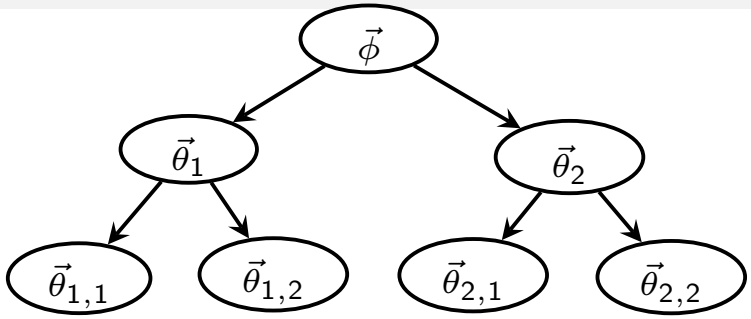
$$\vec{\theta}_{1,2} \sim \text{Dirichlet}\left(\alpha_1 \vec{\theta}_1\right)$$

But then we get difficult terms in our probabilities, like:

$$\prod_{v \in \mathit{Vocabulary}} \ldots \theta_{1,2,v}^{n_{1,2,v}+\alpha_1\theta_{1,v}-1} \theta_{1,v}^{n_{1,v}+\alpha_0\phi_v-1} \ldots$$

ASIDE: A Dirichlet Hierarchy





We might model a set of vocabularies/documents hierarchically:

$$\vec{\theta}_1 \sim \text{Dirichlet}\left(\alpha_0 \vec{\phi}\right)$$

$$\vec{\theta}_{1,2} \sim \text{Dirichlet}\left(\alpha_1 \vec{\theta}_1\right)$$

But then we get difficult terms in our probabilities, like:

$$\prod_{v \in \mathit{Vocabulary}} \ldots \theta_{1,2,v}^{n_{1,2,v}+\alpha_1\theta_{1,v}-1} \theta_{1,v}^{n_{1,v}+\alpha_0\phi_v-1} \ldots$$

Statistical estimation with these generally difficult and slow.

ASIDE: Non-Parametric Alternatives to the Dirichlet

- Teh introduced hierarchical non-parametric methods to the machine learning community:
  - Bayesian n-grams, 2006.
  - hierarchical Dirichlet Processes, *e.g.*, for LDA, 2006 with Jordan, Beal and Blei.
- Basic distributions are the Dirichlet Process (DP) and the Pitman-Yor Process (PYP)
- Hierarchical versions are the HDP and the HPYP.

ASIDE: Non-Parametric Alternatives to the Dirichlet

- Teh introduced hierarchical non-parametric methods to the machine learning community:
  - Bayesian n-grams, 2006.
  - hierarchical Dirichlet Processes, *e.g.*, for LDA, 2006 with Jordan, Beal and Blei.
- Basic distributions are the Dirichlet Process (DP) and the Pitman-Yor Process (PYP)
- Hierarchical versions are the HDP and the HPYP.

These allow variations of hierarchial Dirichlets to work.

ASIDE: Historical Context

1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*

2001-2003: Ishwaran and James develops and "translates" methods usable for machine learning.

2006: Teh develops hierarchical n-gram models using HPYPs.

2006: Teh, Jordan, Beal and Blei develop HDP, *e.g.* applied to LDA.

2006-2011: Chinese restaurant processes (CRPs) go wild!

- require dynamic memory in implementation,
- Chinese restaurant franchise,
- multi-floor Chinese restaurant process,
- *etc.*

ASIDE: Historical Context

1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*

2001-2003: Ishwaran and James develops and "translates" methods usable for machine learning.

2006: Teh develops hierarchical n-gram models using HPYPs.

2006: Teh, Jordan, Beal and Blei develop HDP, *e.g.* applied to LDA.

2006-2011: Chinese restaurant processes (CRPs) go wild!

- require dynamic memory in implementation,
- Chinese restaurant franchise,
- multi-floor Chinese restaurant process,
- *etc.*

Standard MCMC samplers for HCRPs require dynamic memory so are slow and costly! Don't use!

Non-Parametric Methods

- Michael Jordan's "thing" in the mid 2000's was Dirichlet Processes
- applied to topic models in: "Hierarchical Dirichlet Processes," Teh, Jordan, Beal, Blei, *JASA*, 2006.
  - a complex paper
  - over 2000 citations!
- based on the Chinese Restaurant Process (CRP) and the Hierarchical Chinese Restaurant Process (HCRP)

Non-Parametric Methods

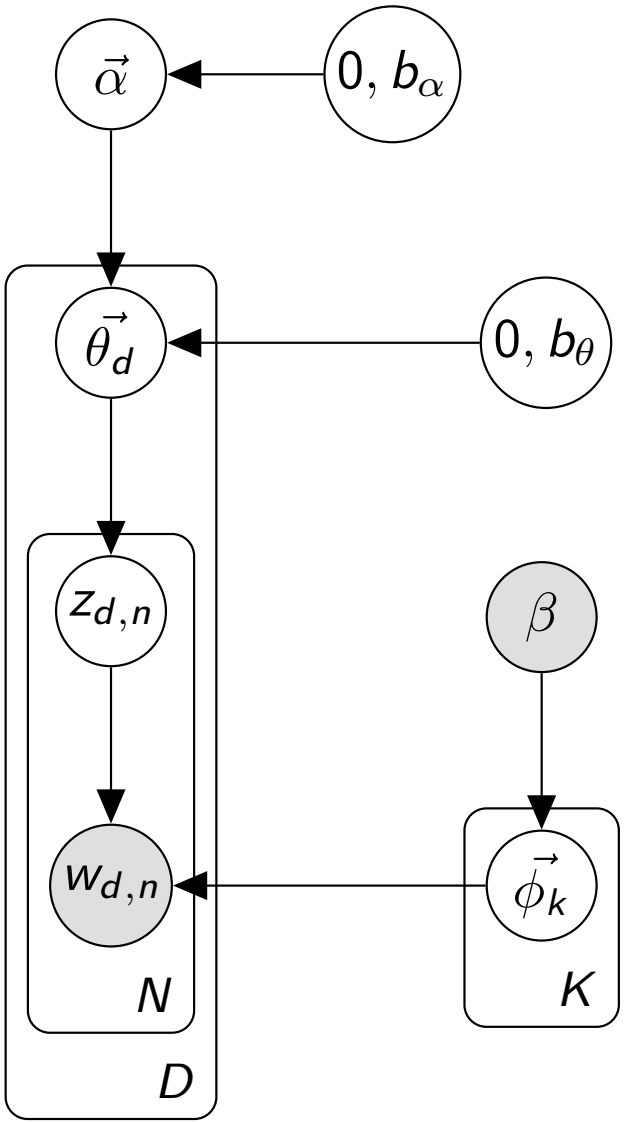
- Michael Jordan's "thing" in the mid 2000's was Dirichlet Processes
- applied to topic models in: "Hierarchical Dirichlet Processes," Teh, Jordan, Beal, Blei, *JASA*, 2006.
  - a complex paper
  - over 2000 citations!
- based on the Chinese Restaurant Process (CRP) and the Hierarchical Chinese Restaurant Process (HCRP)
- basic implementation in C+Matlib by Yeh Whye Teh in 2006

Non-Parametric Methods

- Michael Jordan's "thing" in the mid 2000's was Dirichlet Processes
- applied to topic models in: "Hierarchical Dirichlet Processes," Teh, Jordan, Beal, Blei, *JASA*, 2006.
  - a complex paper
  - over 2000 citations!
- based on the Chinese Restaurant Process (CRP) and the Hierarchical Chinese Restaurant Process (HCRP)
- basic implementation in C+Matlib by Yeh Whye Teh in 2006

Created huge academic race to speed-up the algorithm!

Evolution of Models: HDP-LDA (2006)



HDP-LDA

adds proper modelling of topic prior like Teh *et al.*

in principle $\vec{\alpha}$ is an infinite vector, so $K \to \infty$

Alternative Algorithms for HDP-LDA

- "Practical collapsed variational Bayes inference for hierarchical Dirichlet process," Sato, Kurihara, Nakagawa, *ACM SIGKDD*, 2012.
  - complex version of a variational algorithm motivated by a paper with Teh
- "Truly nonparametric online variational inference for hierarchical Dirichlet processes," Bryant, Sudderth, *NIPS*, 2012.
  - flattens the standard variational algorithm so it really works with a Dirichlet, not a Dirichlet process
- "Stochastic Variational Inference," Hoffman, Blei, Wang, Paisley, *JMLR*, 2013.
  - adds stochastic gradient descent to the standard variational algorithm

Alternative Algorithms for HDP-LDA

- "Practical collapsed variational Bayes inference for hierarchical Dirichlet process," Sato, Kurihara, Nakagawa, *ACM SIGKDD*, 2012.
  - complex version of a variational algorithm motivated by a paper with Teh
- "Truly nonparametric online variational inference for hierarchical Dirichlet processes," Bryant, Sudderth, *NIPS*, 2012.
  - flattens the standard variational algorithm so it really works with a Dirichlet, not a Dirichlet process
- "Stochastic Variational Inference," Hoffman, Blei, Wang, Paisley, *JMLR*, 2013.
  - adds stochastic gradient descent to the standard variational algorithm
- "Rethinking LDA: Why priors matter," Wallach, Mimno, McCallum, *NIPS*, 2009
  - is a truncated version of HDP-LDA; no one noticed until 2014!
  - substantially better than the above in computational and most statistical measures
    - though doesn't do split-merge of Bryant and Sudderth

Text and Burstiness

Original news article:

Women may only account for 11% of all Lok-Sabha MPs but they fared better when it came to representation in the Cabinet. Six women were sworn in as senior ministers on Monday, accounting for 25% of the Cabinet. They include Swaraj, Gandhi, Najma, Badal, Uma and Smriti.

Bag of words:

11% 25% Badal Cabinet(2) Gandhi Lok-Sabha MPs Monday Najma Six Smriti Swaraj They Uma Women account accounting all and as better but came fared for(2) in(2) include it may ministers of on only representation senior sworn the(2) they to were when women

NB. "Cabinet" appears twice! It is **bursty**.

Burstiness Theory

Aside: Burstiness and Information Retrieval (IR)

- burstiness and eliteness are concepts in information retrieval used to develop BM25 (*i.e.* dominant TF-IDF version)
- the two-Poisson model and the Pitman-Yor model can be used to justify some IR theory (Sunehag, 2007; Puurula, 2013)
- relationships not yet fully developed

Burstiness Theory

Aside: Burstiness and Information Retrieval (IR)

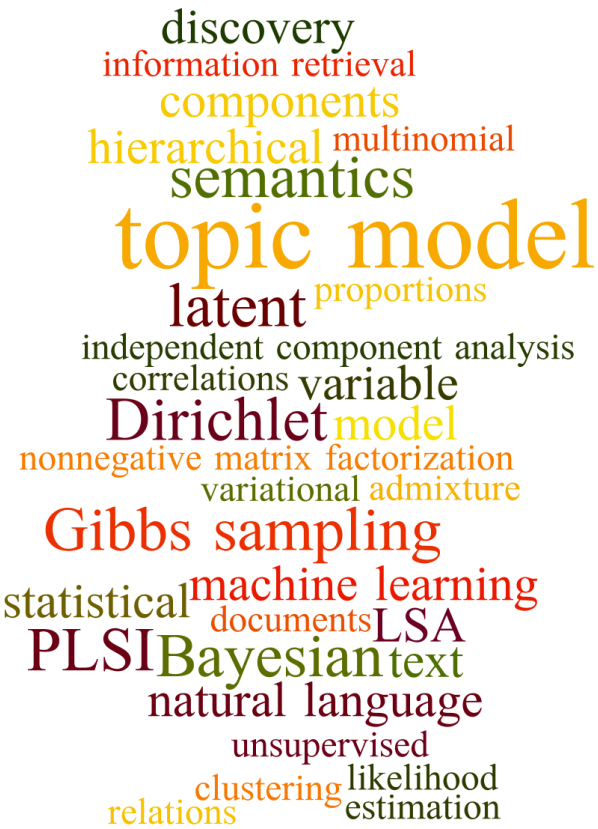
- burstiness and eliteness are concepts in information retrieval used to develop BM25 (*i.e.* dominant TF-IDF version)
- the two-Poisson model and the Pitman-Yor model can be used to justify some IR theory (Sunehag, 2007; Puurula, 2013)
- relationships not yet fully developed

"Accounting for burstiness in topic models," Doyle, Elkan *ICML*, 2009.

- added burstiness to topic models
- used a variational implementation that scaled badly

  e.g. barely able to handle 500 documents
- virtually ignored by others despite remarkable results!

Outline



1 Topic Models Background

2 An Historical Perspective

3 Fully Nonparametric Topic Models

4 Alternatives

5 Conclusion

Fully Nonparametric Topic Models

Describing our fully non-parametric topic models and how they partially improve co-herence.

Misunderstandings with the HDP and HPYP

- Previous implementations of HDP-LDA failed to realise the potential.
  - hierarchial CRPs: poor because of large dynamic memory requirements;
  - standard variational algorithms: poor because variational approximation on deep networks (the stick-breaking construction) is far too conservative;
  - community didn't know a superior approximation had been implemented (in Mallet) for 4 years.

Misunderstandings with the HDP and HPYP

- Previous implementations of HDP-LDA failed to realise the potential.
  - hierarchial CRPs: poor because of large dynamic memory requirements;
  - standard variational algorithms: poor because variational approximation on deep networks (the stick-breaking construction) is far too conservative;
  - community didn't know a superior approximation had been implemented (in Mallet) for 4 years.
- Jordan's *ICML* 2005 invited talk
  "Dirichlet Processes, Chinese Restaurant Processes, and all that"
  only stressed that DPs and PYPs helped pick "the right number of topics."
  - theory shows they only perform reasonably at this when their hyper-parameters are carefully estimated, the exception in the community rather than the rule;
  - the real power of HDPs and HPYPs is their efficient use in a hierarchical setting.

Efficient Methods for the Hierarchical DP/PYP

- Sampling table configurations for the hierarchical Poisson-Dirichlet process," Chen, Du and Buntine 2011.
  - insight gained from series of earlier implementation efforts by Lan Du.

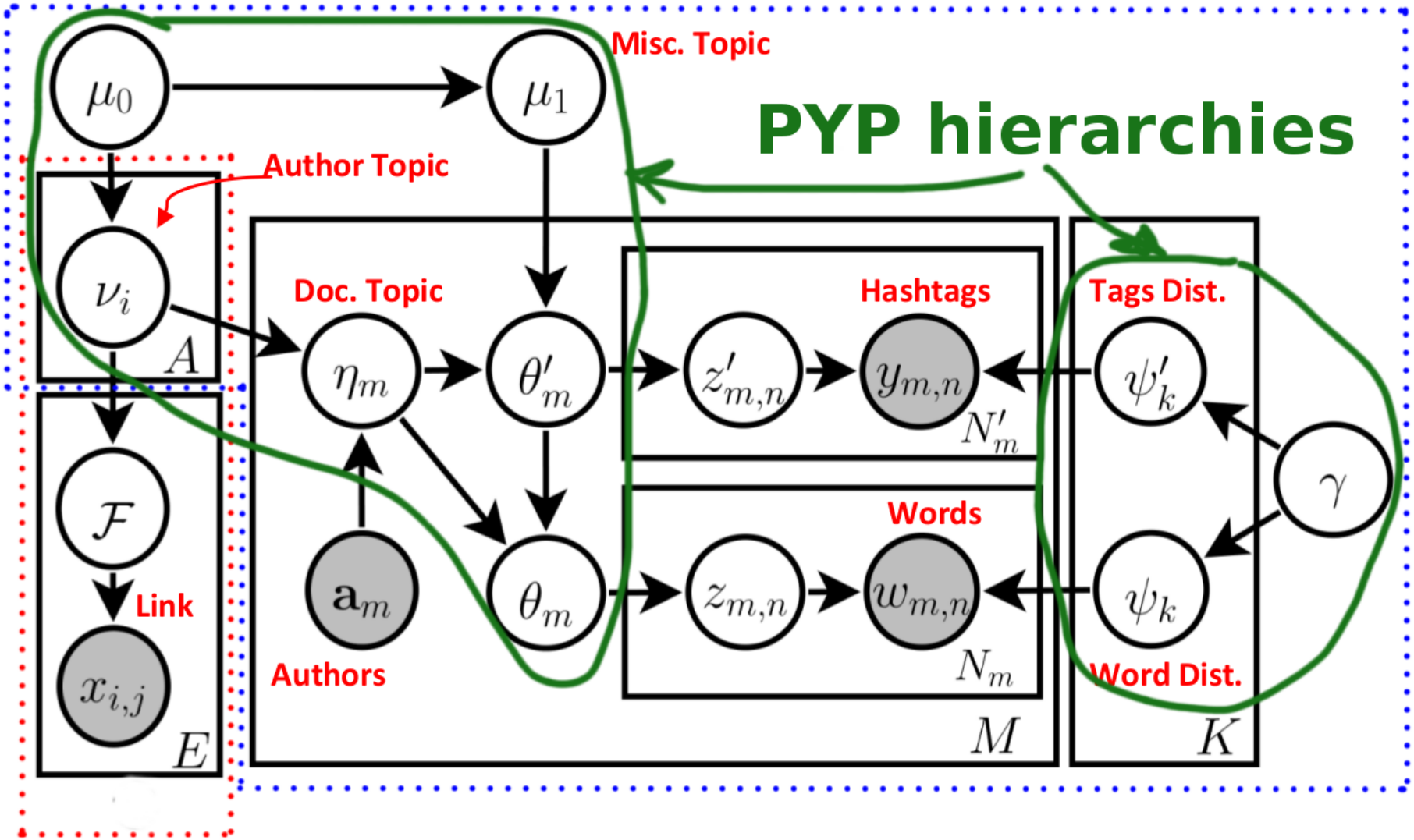
Efficient Methods for the Hierarchical DP/PYP

- Sampling table configurations for the hierarchical Poisson-Dirichlet process," Chen, Du and Buntine 2011.
  - insight gained from series of earlier implementation efforts by Lan Du.
- Leads to an efficient implementation with:
  - doubles memory and time of regular Gibbs (non-hierarchical) sampling
  - oftentimes a multi-core implementation good for upto 8 core
  - no dynamic memory.

Efficient Methods for the Hierarchical DP/PYP

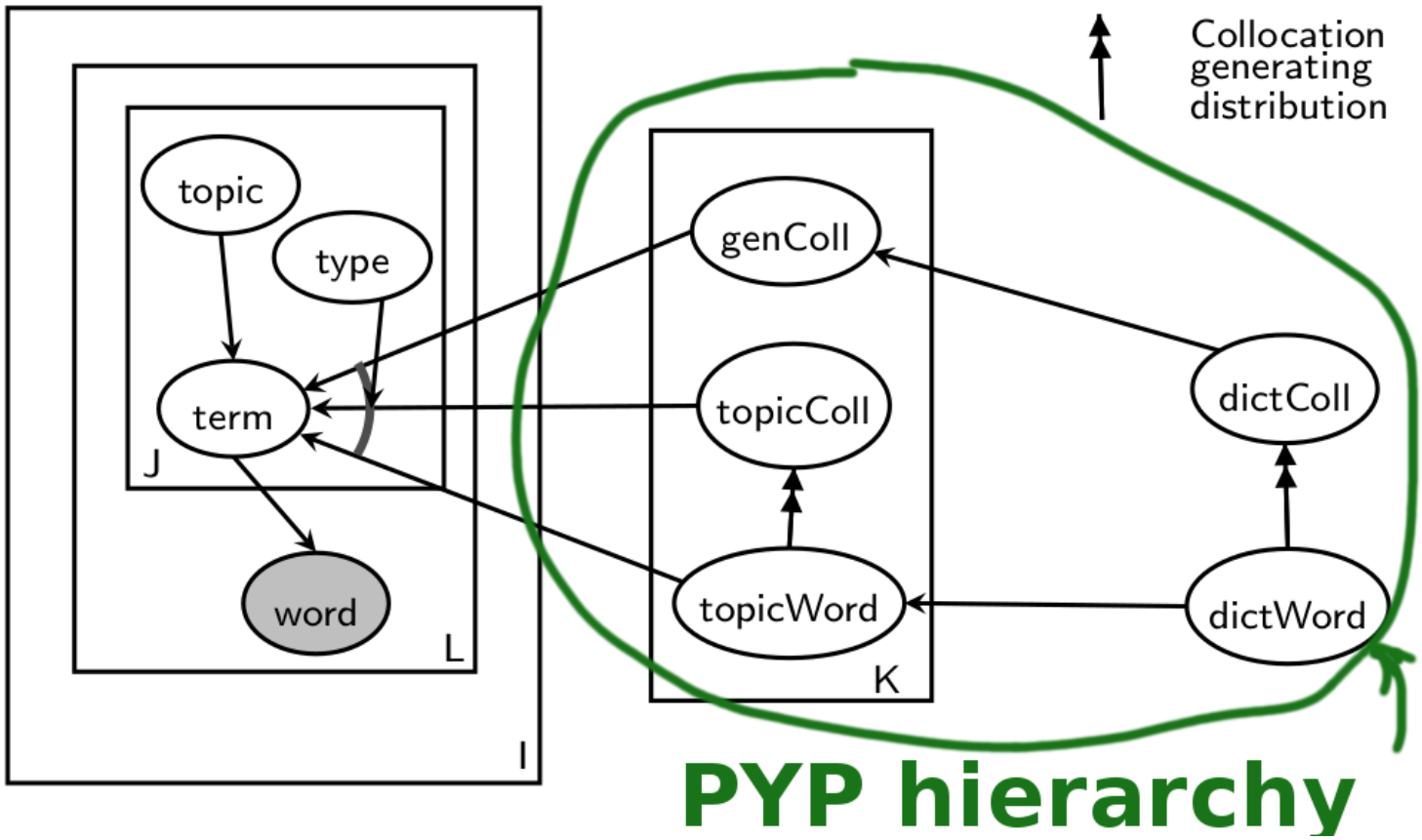
- Sampling table configurations for the hierarchical Poisson-Dirichlet process," Chen, Du and Buntine 2011.
  - insight gained from series of earlier implementation efforts by Lan Du.
- Leads to an efficient implementation with:
  - doubles memory and time of regular Gibbs (non-hierarchical) sampling
  - oftentimes a multi-core implementation good for upto 8 core
  - no dynamic memory.
- Substantially beats previous methods.
- Allows far more complex models.

ASIDE: Complex Model: Twitter-Network Topic Model



ASIDE: Complex Model: Collocation Topic Model

A fast version of Mark Johnson's Adaptor Grammars for Collocations, *ACL* 2010.



K = #topics, I = #docs, L = #words in doc, J = #terms in doc

ASIDE: Example Collocation Topics

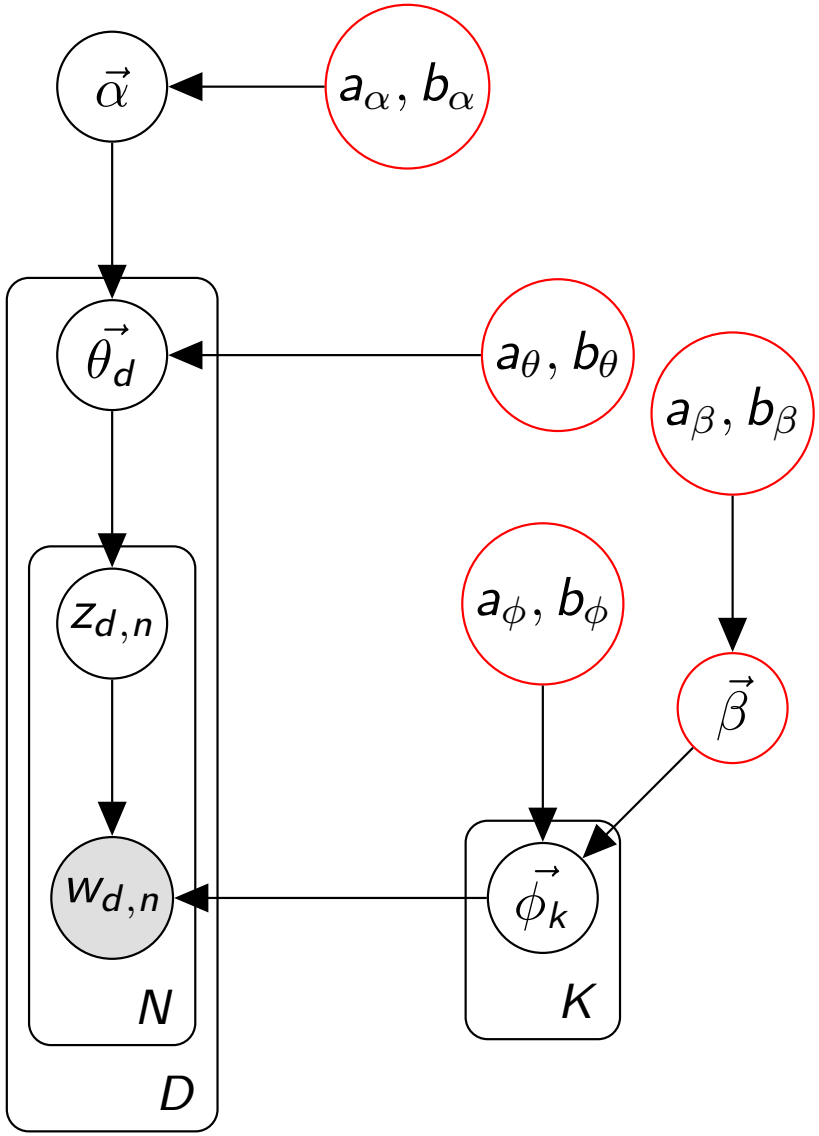
100 collocation topics built on text content from NIPS 1987-2012.
Legend: topical terms, general terms.

8: feature vector, texture, object recognition, image patches, input image, image, large scale, image segmentation, ..., random selection, ...

9: posterior distribution, monte carlo, generative model, graphical model, latent variables, gibbs sampling, joint distribution, prior distribution, hidden variables, log likelihood, ..., particles, ..., marginalizing, ..., mcmc, ... supplementary materials, evaluation metric

10: linear regression, regression problem, ridge regression, positive semidefinite, variable selection, ols, mars, cnf, regressors, data mining, ...

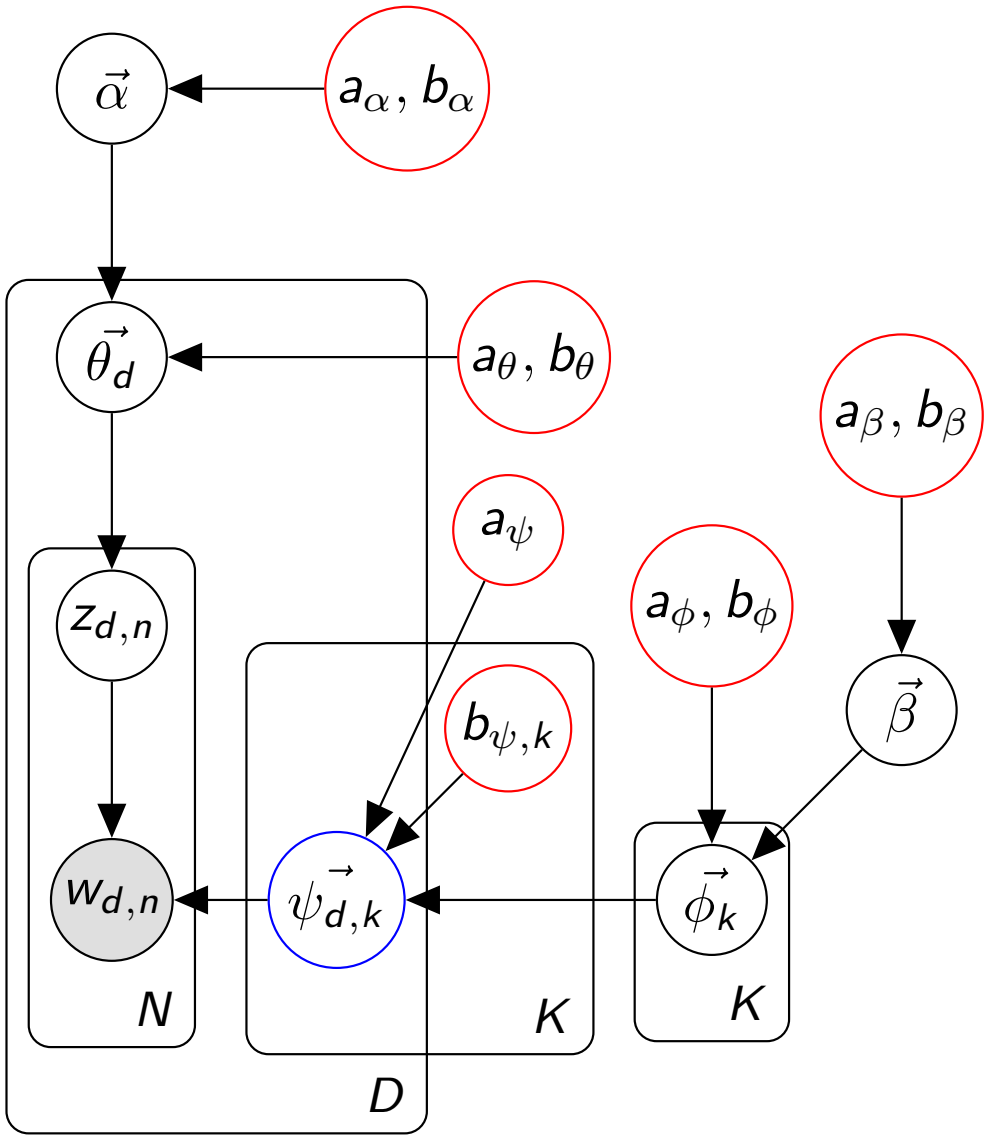
Evolution of Models: NP-LDA (2013)



NP-LDA

adds power law on word distributions like Sato and Nakagawa *KDD*, 2010 and estimation of background word distribution

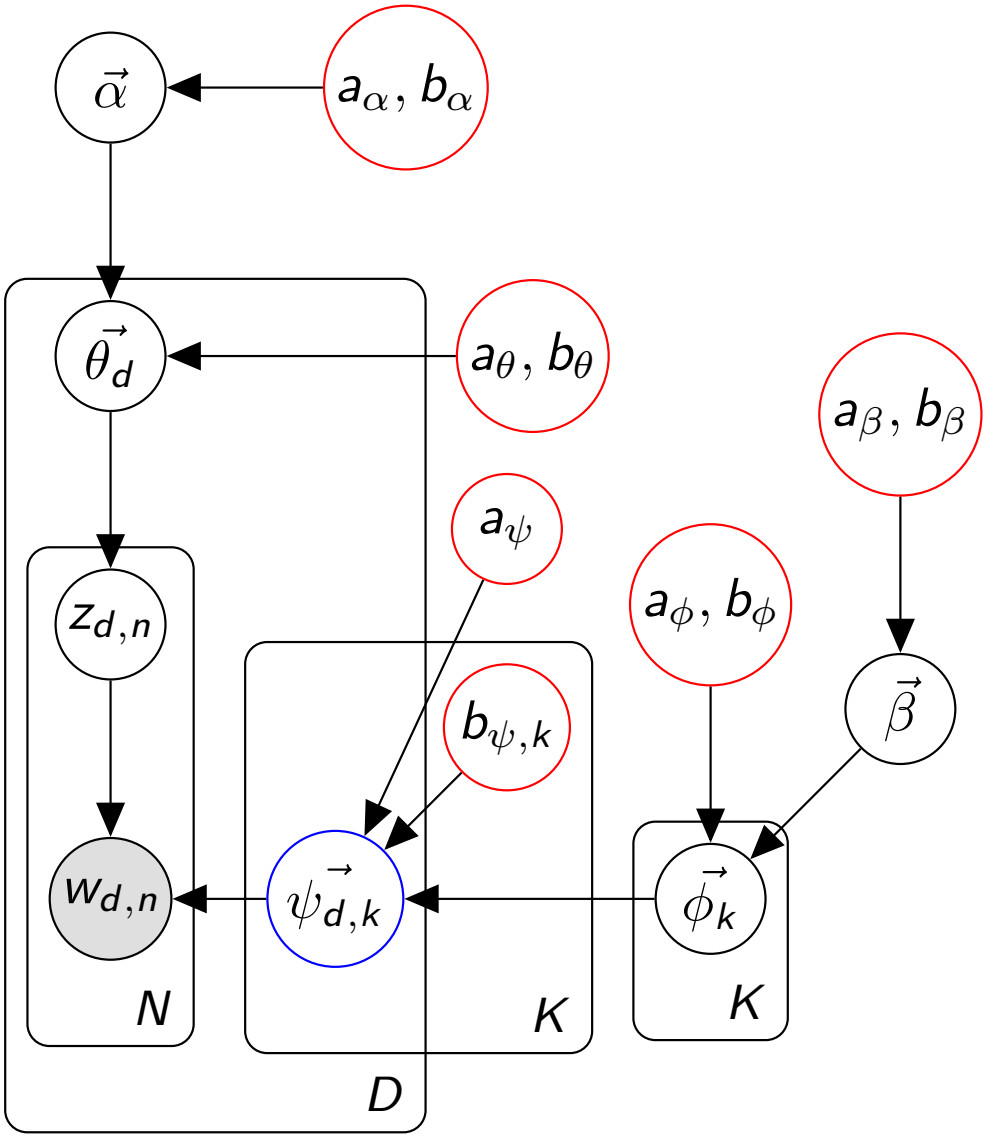
Evolution of Models: Bursty NP-LDA (2014)



NP-LDA with Burstiness

add's burstiness like Doyle and Elkan

Our Non-Parametric Topic Model



$\{\vec{\theta}_d : d\}$ = docu⊗topic matrix

$\{\vec{\phi}_k : k\}$ = topic⊗word matrix

$\{\vec{\psi}_{d,k} : d\}$ = document d's topic⊗word matrix

$\vec{\alpha}$ = prior for topic vec $\vec{\theta}_d$

$\vec{\beta}$ = prior for word vec $\vec{\phi}_k$

$b_{\psi,k}$ = confidence in $\vec{\phi}_k$

- Full fitting of priors, and their hyperparameters
- $\{\vec{\psi}_{d,k}\}$ = not stored so efficient

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Estimating parameters: whenever parameters cannot be reasonable set, we estimate them instead.

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Estimating parameters: whenever parameters cannot be reasonable set, we estimate them instead.

Fast-ish implementation: for full non-parametrics:

- in all, doubles memory and time of regular LDA Gibbs sampling
- multi-core implementation good for upto 8 core

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Estimating parameters: whenever parameters cannot be reasonable set, we estimate them instead.

Fast-ish implementation: for full non-parametrics:

- in all, doubles memory and time of regular LDA Gibbs sampling
- multi-core implementation good for upto 8 core

Burstiness: we developed a Gibbs sampler that acts as a front end to **any** LDA-style model with Gibbs:

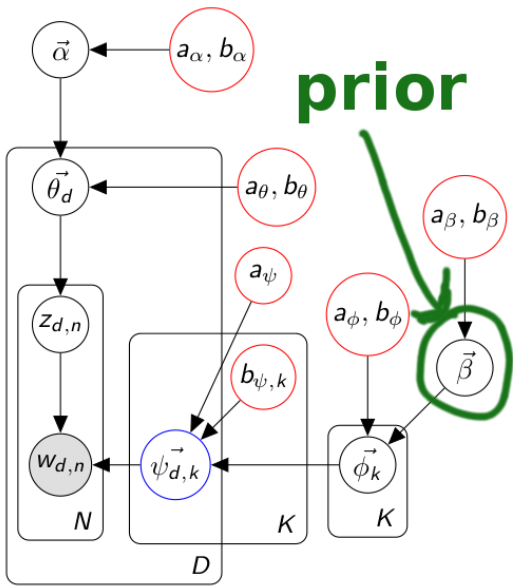
- implemented as a C function that calls the Gibbs sampler
- adds smallish memory (20%) and time (50%) overhead

Example Topics

2691 abstracts from JMLR vol. 1-11, about 60 words length (after removing stops). Built 1000 topics. Examples below contain "posterior".

#25, 0.42%	posteriori expectation likelihood-based analytically maximum likelihood maximization e-step posterior estimation map coming model *(top=14)*
#31, 0.40%	undirected graphical message-passing inference graphs directed posteriori junction graph clique intractable models cliques partition *(top=13)*
#38, 0.37%	dirichlet bayesian priors conjugate prior ill-posed posterior covariance gaussian infer serves distribution analytical distributions *(top=9)*
#58, 0.31%	latent variables discover posterior dependencies variable models modeling hidden parent unobserved part correlations constituent *(top=11)*
#95, 0.22%	particle kalman tracking filter filtering observer state appearance implement dynamics visual occlusion posterior multimodal *(top=5)*
#124, 0.19%	monte carlo chain markov jump mcmc chains reversible iterates mix proposal posterior problematic sampling *(top=2)*
#239, 0.11%	naive classifier bayes averaging logarithmic multiple-instance averaged posterior already counterpart considerably weakly classifications goal *(top=3)*
#678, 0.03%	nondeterministic posterior probabilities cancer hypotheses successively true deterministic distributions worst-case comprise discussed limiting *(top=2)*
#820, 0.02%	estimators constrains sparsely constraints cross-validated insufficient parameters made among posterior context brain correlated fast *(top=1)*

Understanding the Word Prior $\vec{\beta}$

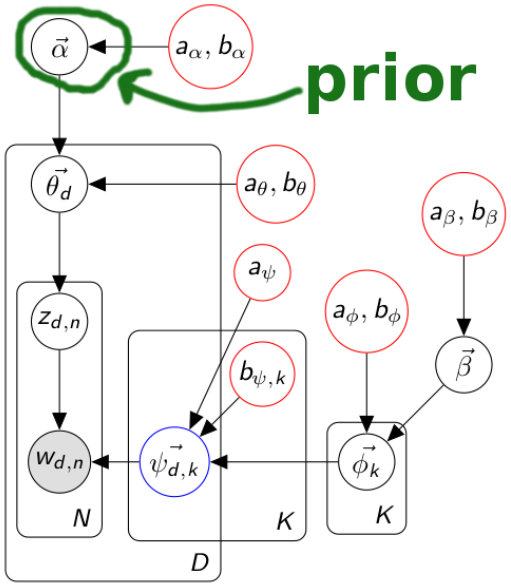


- word prior $\vec{\beta}$ corresponds to a background topic (with PYPs to make Zipfian)
- all topics $\vec{\phi}_k$ are variants of it
- "topical" words are reduced and "stop" words are increased compared to collection frequency
- word w importance for topic k can be measured as $\phi_{k,w}/\beta_w$

395 Reuters RCV1 news articles from 1996 containing "church".

background (most freq.)	the of to a in and 's was on for by telephone said sick fighting with as at is republican land shortly he remains difficult voice shown mary inside done travelled
topical (high df_w/β_w)	diana teresa missionaries russia parker elizabeth bowles camilla churchill winston harriman pamela her quoted princess pontiff prince navarro-valls dies kremlin averell parkinson
topic #1	'm else something n't everyone someone my stand i me like truth always really going do you ran know similar lover things look sun think 've

Understanding the Topic Prior $\vec{\alpha}$

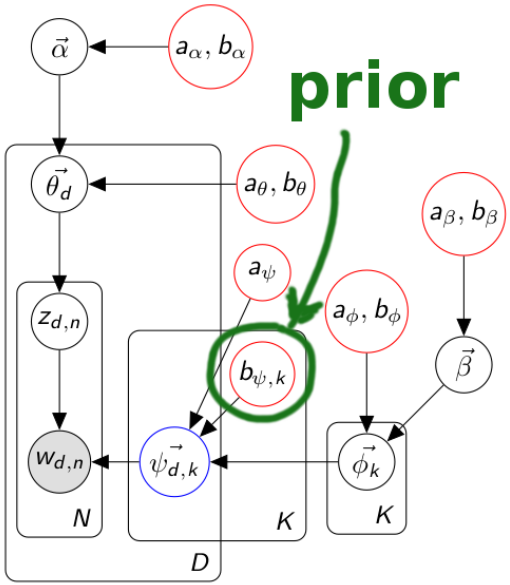


- topic prior $\vec{\alpha}$ corresponds to a expected occurence of topic
- all document topics $\vec{\theta}_d$ are variants of it
- uses Dirichlet processes

395 Reuters RCV1 news articles from 1996 containing "church".

#2 5.05%	quoted declined saying reports position further talks sought	#47 1.00%	nazi nazis germany german war jews recalled forces wartime
#4 2.28%	pope vatican navarro-valls pontiff solidarity 76-year-old	#48 1.85%	estimated sold york company estate island percent sell money
#5 1.90%	diana charles bowles parker prince camilla princess elizabeth	#49 0.98%	art works culture includes cultural st boris programme show

Understanding the Topic Confidence $b_{\psi,k}$



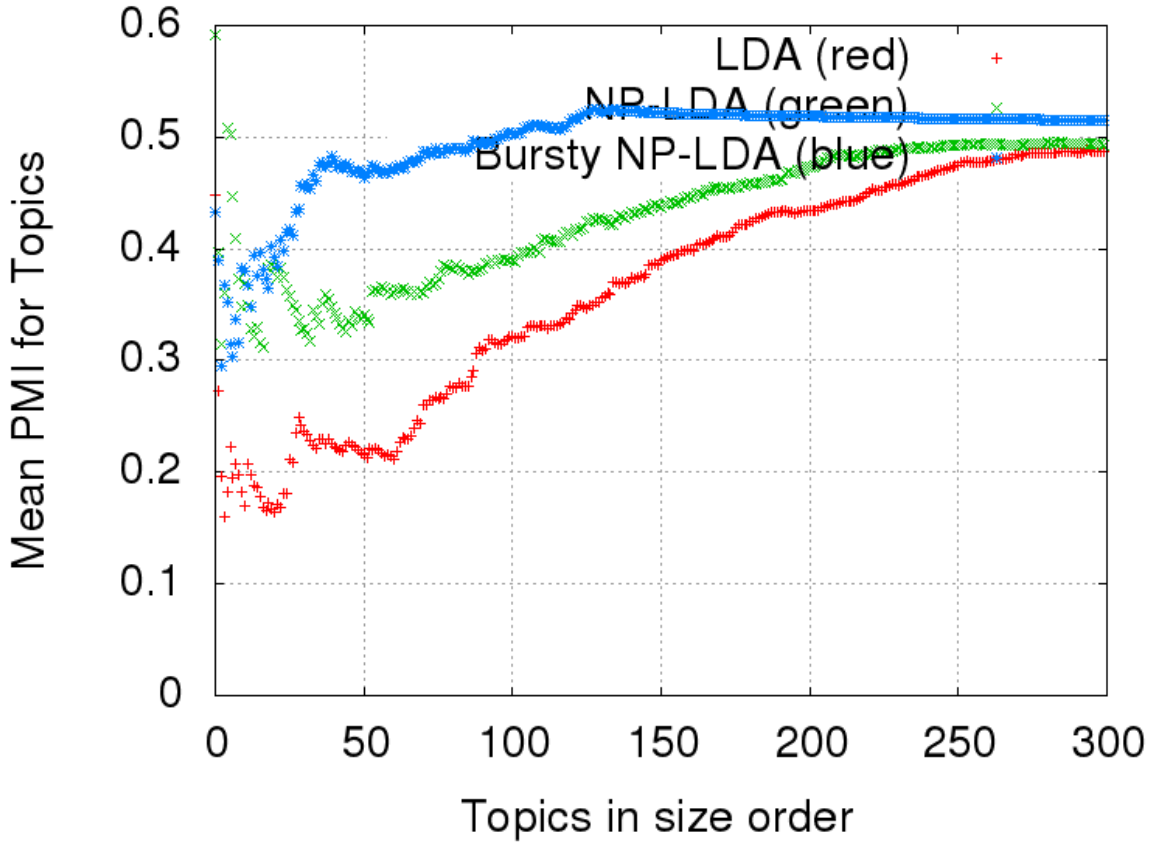
- topic confidence $b_{\psi,k}$ corresponds to a concentration parameter (inverse variance) when generating $\vec{\psi}_{d,k}$ from $\vec{\phi}_k$
- large values means $\vec{\psi}_{d,k}$ is a copy of $\vec{\phi}_k$
- low values means $\vec{\psi}_{d,k}$ differs greatly from $\vec{\phi}_k$

8616 Reuters RCV1 news articles from 1996 containing "person".

#2 2605	willingness guarantor caution definite absences disgruntled seriousness manoeuvring govern instability
#4 2271	detractors predecessors illustrious front-runner outsider credentials flair courteous woo self-effacing married
#9 2153	teresa missionaries woodlands birla pacemaker gutters nun calcutta

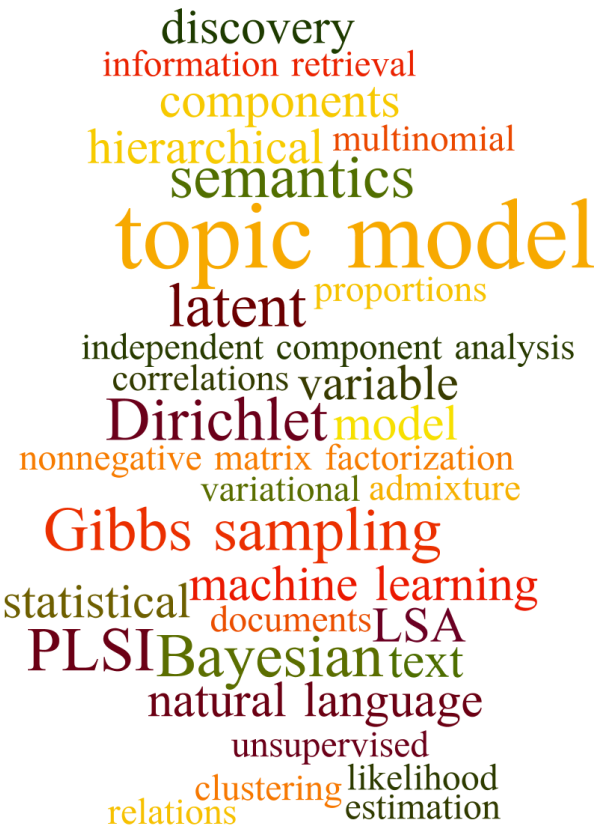
#199 3.57	royalties michelle job lopez earns hungarian-born eating credits
#200 2.92	penguin birthdays 1000 abc compiled wheel mausoleum 1800 provoke timetable budapest
#194 1.41	spa verdicts korzhakov beginnings burmese ethics betrayed blair fujimori heroin

NPMI on the NYT Health Data



71978 news articles from New York Times in the category "health"
NPMI is a progressive mean to show quality of initial topics

Outline



1 Topic Models Background

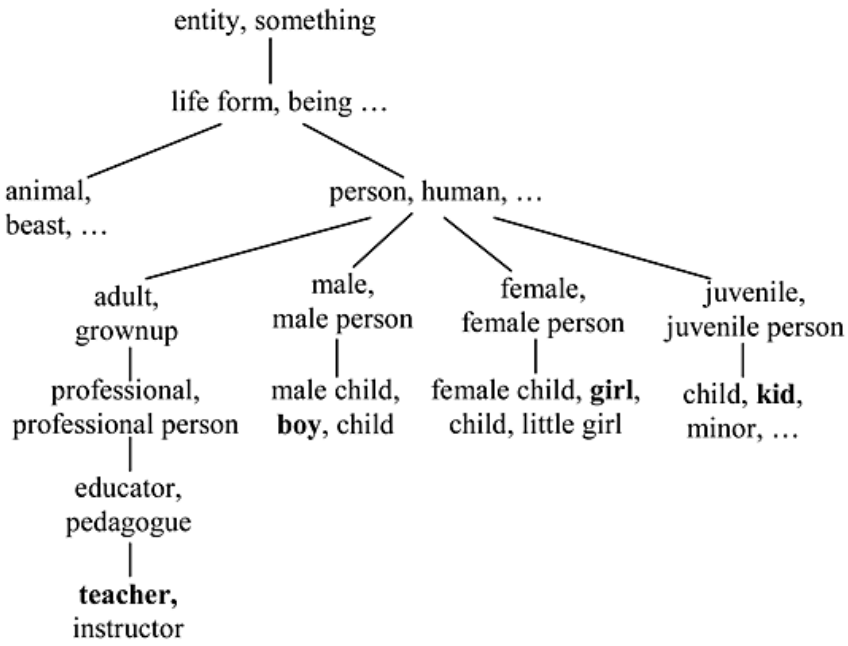
2 An Historical Perspective

3 Fully Nonparametric Topic Models

4 Alternatives

5 Conclusion

Non-Parametric Semantic Networks in Topic Models



- There are a large number of (near) hierarchical language resources: Wordnet, Freebase, *etc.*
- One can model word sets with distributional semantics using the language resources as the prior structure.
- HPYPs and deep neural networks can be used to model the distributional semantics.
  - topics would now be represented using distributional semantics so they are intrinsically coherent.

Non-Parametric Topic Models with Indian Buffet Process

- The Indian Buffet Process (IBP) is a non-parametric model that lets you selectively zero out features, thus zeroing some probabilities.
- Used to encourage sparsity in topics.
- Coupled with LDA analogously to our NP-LDA by "Latent IBP compound Dirichlet Allocation" by Archambeau, Lakshminarayanan, Bouchard, *IEEE Trans. PAMI* 2015

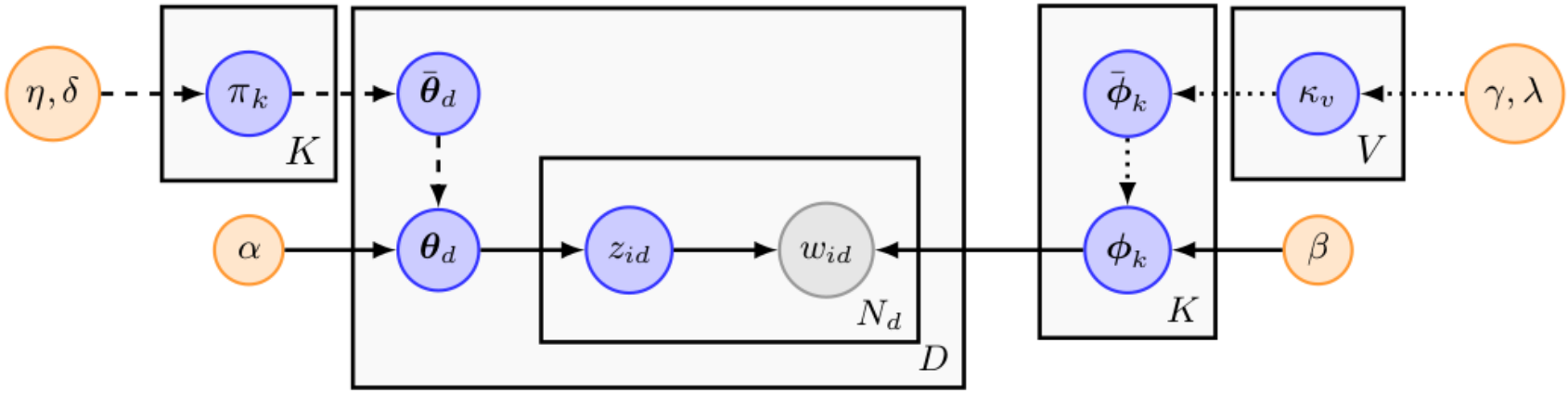


Figure 1: Graphical models for the different configurations (fixed K): LDA: solid arrows only, STM: solid + dotted arrows, FTM: solid + dashed arrows, LIDA: all arrows.

Outline

1 Topic Models Background

2 An Historical Perspective

3 Fully Nonparametric Topic Models

4 Alternatives

5 Conclusion

Conclusion

- Making topic models yield more coherent outputs is a great challenge problem:
  - we need empirical pull and theoretical push to make it work
- Theoretical push: deep learning algorithms (including Bayesian non-parametric methods) offer potential for embedding word semantics into the model.
- Applications and implications can have far reaching effects in related areas: especially document summarisation!

Download the software (`hca` from MLOSS or Github)