Experiments with Non-parametric Topic Models

Swapnil Mishra ¹ and Wray Buntine ²

 1 NICTA + ANU

²Monash University

Tuesday 26th August, 2014

We are hiring!



NP-TMs

Melbourne, world's most livable city, again





junior faculty in data science

Outline

- Background
- 2 The Non Parametric Topic Model
- 3 Experiments
- 4 Conclusion

Topic Models

- Topic Models discover hidden themes in text data to aid understanding
 - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003)
- recent research develops higher performance topic models

Topic Models

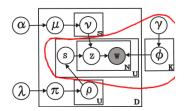
- Topic Models discover hidden themes in text data to aid understanding
 - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003)
- recent research develops higher performance topic models
- but why should you care?

Using Topic Models

topic models are the leading edge of a new wave of deep latent semantic models applied to real NLP tasks:

e.g., document segmentation, word sense disambiguation, facet discovery,

...

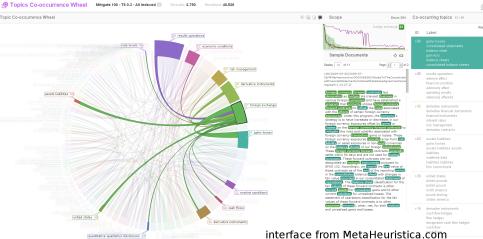


in the middle of this segmentation model is a topic model

• better topic models are important components

High Fidelity Topic Models and Visualisation

Can get 100's of topics from 1000's of documents providing real insight.



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** (see Doyle and Elkan 2009)

7 / 24

Previous Work

EXTENDING

COMPARING AGAINST

- "Hierarchical Dirichlet Processes," Teh, Jordan, Beal, Blei 2006.
- "Rethinking LDA: Why priors matter," Wallach, Mimno, McCallum, 2009.
- "Accounting for burstiness in topic models," Doyle and Elkan 2009.
- "Topic models with power-law using Pitman-Yor process." Sato and Nakagawa 2010
- Sampling table configurations for the hierarchical Poisson-Dirichlet process," Chen, Du and Buntine 2011.
- "Practical collapsed variational Bayes inference for hierarchical Dirichlet process," Sato, Kurihara, and Nakagawa 2012.
- "Truly nonparametric online variational inference for hierarchical Dirichlet processes," Bryant and Sudderth 2012.
- "Stochastic Variational Inference," Hoffman, Blei, Wang and Paisley 2013.

Better Sampling Methods for HDP and HPYP

Sampling for hierarchical Dirichlet Processes and Pitman-Yor Processes:

The Old: hierarchical Chinese Restaurant Processes (CRP) from Teh et al. 2006.

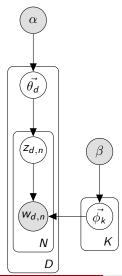
The New: block table indicator sampling from Chen, Du and Buntine 2011.

NP-TMs

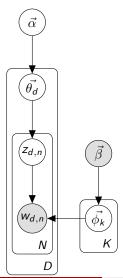
- requires no dynamic memory
- more rapid mixing so leads to better models
- more easily applied to more complex models
- demonstrated extensively on different problems!

Outline

- Background
- The Non Parametric Topic Model
 - Evolution of Models
 - Our Non-parametric Topic Model
- 3 Experiments
- 4 Conclusion



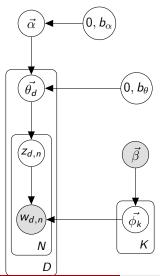
LDA- Scalar original LDA



LDA- Vector

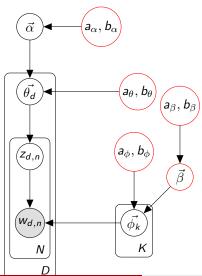
adds asymmetric Dirichlet prior like Wallach et al.; is also truncated HDP-LDA; implemented by Mallet as assymetric-symmetric LDA

Tuesday 26th August, 2014



HDP

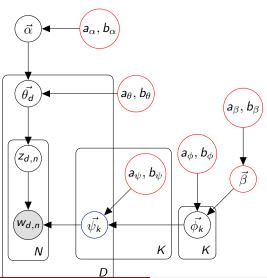
adds proper modelling of topic prior like Teh et al.



NP-LDA

adds power law on word distributions like Sato et al. and estimation of background word distribution

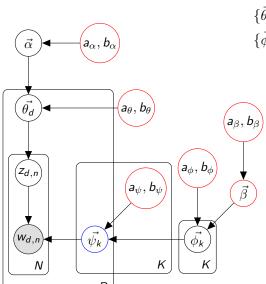
NP-TMs



NP-LDA with Burstiness

add's burstiness like Doyle and Elkan

Our Non-parametric Topic Model



 $\{ \vec{\theta_d} \} = \mathsf{document} \otimes \mathsf{topic} \; \mathsf{matrix}$ $\{ \vec{\phi_k} \} \; = \; \mathsf{topic} \otimes \mathsf{word} \; \mathsf{matrix}$

- Full fitting of priors on topic⊗word and document⊗topic matrices (red nodes).
- Topic \otimes word vectors $\vec{\phi}_k$ specialised to the document to yield $\vec{\psi}_k$.
- This models burstiness(blue node).

NP-TMs

Handling Burstiness

- proposed by Doyle and Elkan 2009
- used a slow variational method
- we developed a Gibbs sampler that acts as a front end to any LDA-style model with Gibbs, e.g.
 - dynamic topic models
 - bibliographic network models
- implemented as a C function that calls the Gibbs sampler
- adds smallish memory (20%) and time (30%) overhead

Outline

- Background
- 2 The Non Parametric Topic Model
- Experiments
 - Runtime
 - Performance
- 4 Conclusion

Runtime Characteristics

Alg.	mins.	Mb
LDA	11	630
Burst LDA	20	690
HDP-LDA	20	760
Burst HDP-LDA	30	850
NP-LDA	35	840
Burst NP-LDA	45	930
Online HDP	236	1800

Cycle times and memory requirements on the LA Times TREC 4 data.

- "Burst" is the burstiness version,
- "NP-LDA" is our default non-parametric version with full sampling of hyperparameters.
- "Online HDP" is (Wang, Paisley and Blei) in Python. Recent C++ version from Wang "faster".

Performance Metrics

Perplexity:

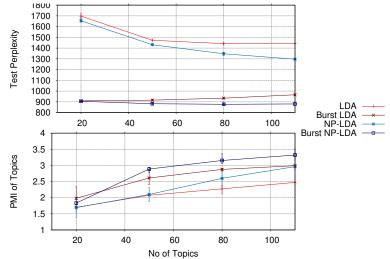
- measure of test set likelihood;
- equal to effective size of vocabulary;
- we use "document completion,"
- see Wallach, Murray, Salakhutdinov, and Mimno, 2009.

PMI:

- measure of topic coherence: "average pointwise mutual information between all pairs of the top 10 words in the topic"
- see Newman, Lau, Grieser, and Baldwin, 2010.

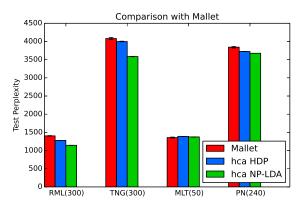
Performance on Reuters-21578 ModLewis Split

Training on 11314 news articles with vocabulary of 16994.



Comparison with Mallet

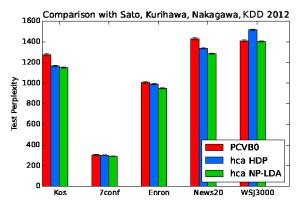
Asymmetric-symmetric option by Mimno implements Wallach's method for estimating $\vec{\alpha}$ since 2008. Great truncated HDP-LDA implementation.



NB. previous HDP-LDA work does not compare with AS-LDA in Mallet because they didn't realise it was effectively truncated HDP-LDA

Comparison with Sato, Kurihawa, Nakagawa, KDD 2012

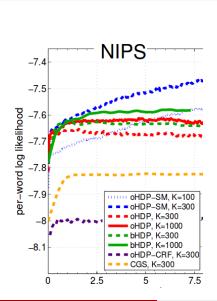
By our analysis, the best performing variational algorithm (without split-merge).

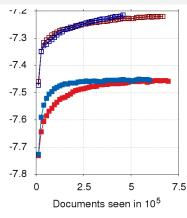


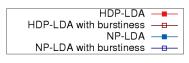
NB. WSJ3000 has 3000 docs and vocab with 30000 words!

Thanks to Issei Sato for providing the data.

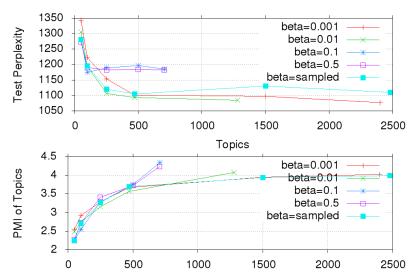
Comparison to Bryant+Sudderth (2012) on NIPS data







Effect of Hyperparameters on the Number of Topics



Outline

- Background
- 2 The Non Parametric Topic Model
- 3 Experiments
- Conclusion

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
 - most previous work failed to show this

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
 - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
 - now available, efficiently, for a broad variety of models

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
 - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
 - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
 - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
 - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)
- hyper-parameter adaptation is important for a high performance model

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
 - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
 - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)
- hyper-parameter adaptation is important for a high performance model
- speedup efficiency of 70%-75% has been achieved running with 6 parallel threads on multi-core machines using atomic operations

Thank You

Thank You Questions?

Grab our code from

```
https://github.com/wbuntine/topic-models or http://mloss.org/software/view/527/
```

- thanks to Issei Sato and David Mimno for data and discussion
- special thanks to next paper (Aaron Li et al.) for showing us how to speed up our algorthm again!

