Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models

Mattias Villani
(but really Måns Magnusson and Leif Jonsson)

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

February 4, 2017

Overview

- **Topic models**
- **Inference** in topic models
- **PC-LDA** - a fast sparse parallel Gibbs sampler for topic models

Topics in Science

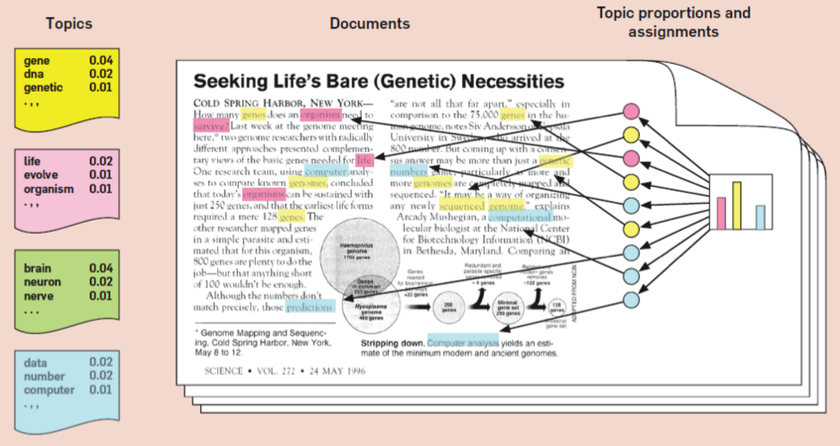


Figure: Example: learned topics from 17 000 articles in *Science* . (Blei et al., 2010)

Topic models in practical work

- Analyzing topics/**summarizing documents**
- Using topics as explanatory variables in other models
- Information retrieval tasks
- **Documentation similarities** - suggest documents
- Computer vision
- ...

The graphical model

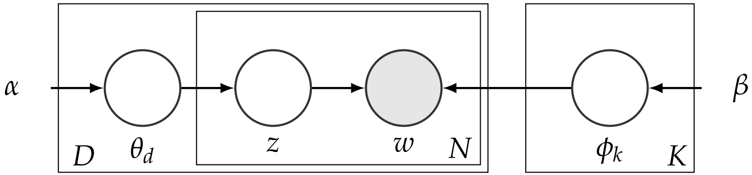


Figure: The LDA model

Bayesian learning

- We want use the words ($\mathbf{w}$) to learn:
  - The **topics**: Φ - a $K \times V$ matrix (V is the vocabulary size).
  - The **topic proportions**: Θ - a $D \times K$ matrix.
  - The **topic indicators**: $\mathbf{z}$ - a vector of length $N \cdot D$.
- **Posterior distribution** for the topic model

$$p(\mathbf{z}, \Theta, \Phi | \mathbf{w}) \quad = \quad \frac{p(\mathbf{z}, \Theta, \Phi | \mathbf{w}) \cdot p(\mathbf{z}, \Theta, \Phi)}{p(\mathbf{w})}$$

- Posterior distribution is complex.
- Explore it by simulating $\mathbf{z}$, Θ and Φ from $p(\mathbf{z}, \Theta, \Phi | \mathbf{w})$.
- **Gibbs sampling** (MCMC).

Collapsed Gibbs sampling for topic models

- Integrating out (**collapsing**) Θ and Φ:

$$p(\mathbf{z}|\mathbf{w}) \quad = \quad \int\int p(\mathbf{z},\Theta,\Phi|\mathbf{w}) \cdot p(\mathbf{z},\Theta,\Phi) d\Phi d\Theta$$

- The **collapsed Gibbs sampler** (Griffiths and Steyvers, 2004)

$$p(z_i = k|w_i, \mathbf{z}_{\neg i}) = \underbrace{\frac{n^{(w)}_{k,v_i} + \beta}{n^{(w)}_{k,\cdot} + V\beta}}_{type-topic\ (\Phi)} \cdot \underbrace{(n^{(d)}_{k,d_i} + \alpha)}_{topic-doc\ (\Theta)}$$

where $n^{(w)}$ and $n^{(d)}$ are matrices with counts.

- **Serial sampler**:
  - Sample z_1 given all other z
  - Sample z_2 given all other z
  - and so on for every word in the corpus ...
- Every z draw is $O(K)$
- Sloooooooow.

Big data - big models - big headache

- Big corpuses today (Yuan et al., 2015):

Dataset	V	N	D
NYTimes	101K	99M	300K
PubMed	140K	737M	8.2M
BingWebC	1M	200B	1.2B

- How to handle **big** corpuses:
  - Parallelism
  - Improve algorithm speed

Parallel Gibbs samplers for topic models

- Integrating out (collapsing) **both** Θ and Φ makes all z dependent.
- **AD-LDA** (Newman et al., 2009) parallelizes with respect to documents. Ignores the dependence. Approximate!

- Integrating out only Θ (partially collapsed) makes the z dependent within a document, but
  - Documents are independent
  - Topics are independent
  - We can **parallelize** with respect to documents! **PC-LDA** (Magnusson et al., 2015).

- Θ is $D \times K$ and grows fast with corpus size.
- Φ is $D \times K$ and grows slowly with corpus size.
- **PC-LDA scales well** with corpus size.

The partially collapsed sampler (PC-LDA)

- Sample
$$\mathbf{z}_1, ..., \mathbf{z}_D | \mathbf{w}, \Phi$$
in parallel over documents.

- Sample
$$\phi_1, ..., \phi_K | \mathbf{z}, \mathbf{w}$$
in parallel over topics.

- Extra tricks:
  - Walker-Alias method (Li et al., 2014) and sparsity in $n^{(d)}$ (a document talks about a small set of topics)
  - Cashed Marsaglia gamma sampling (for Φ) (Marsaglia and Tsang, 2000)
  - Job stealing

AD-LDA is not quite right

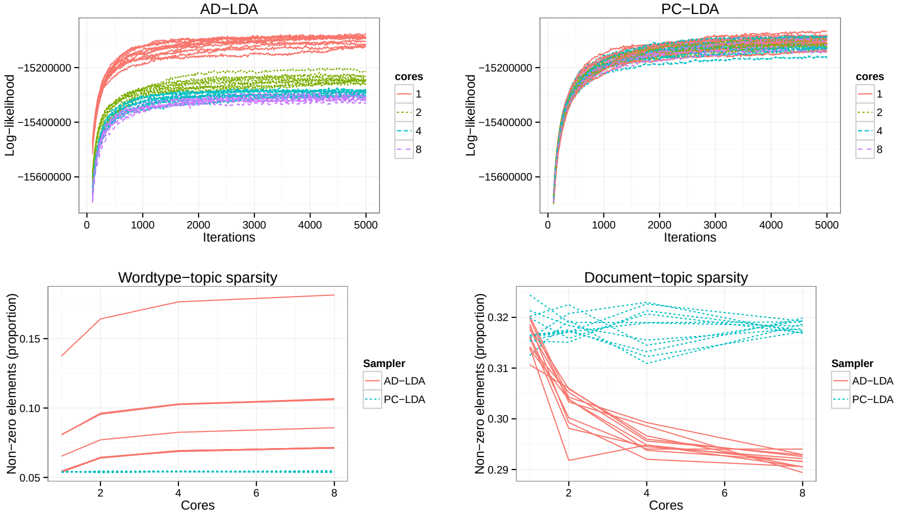
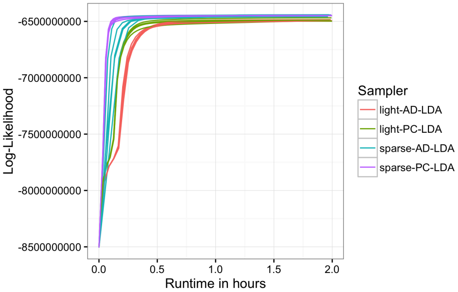


Figure: Speedup of PC-LDA and sparse AD-LDA Magnusson et al. (2015)

PubMed - 10, 100 and 1000 topics



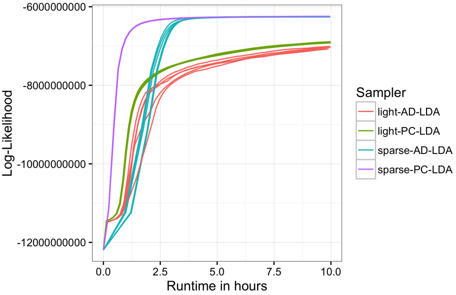


Figure: Inference in big models Magnusson et al. (2015)

Wikipedia and NY Times - 100 topics on 16 cores

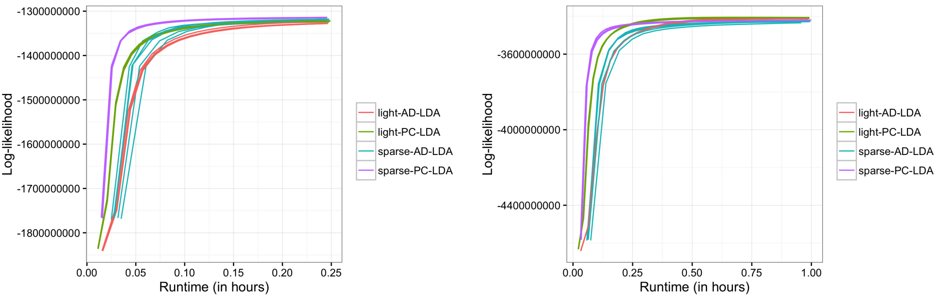


Figure: Wikipedia (left) and New York Times (right). Magnusson et al. (2015)

Summary of findings

- Approximate distributed LDA can lead to the wrong model
- Parallelizing topic models using partially collapsed sampling
  - fast
  - can handle big corpuses
  - can model Φ
  - is not necessarily less efficient
  - is correct
  - seems to explore the posterior better

References

Blei, D., Carin, L., Dunson, D., Nov. 2010. Probabilistic Topic Models. IEEE Signal Processing Magazine, 77–84. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5563111`

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. …academy of Sciences of the United …. URL `http://www.pnas.org/content/101/suppl.1/5228.short`

Li, A. Q., Ahmed, A., Ravi, S., Smola, A. J., 2014. Reducing the sampling complexity of topic models. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 891–900.

Magnusson, M., Jonsson, L., Villani, M., Broman, D., 2015. Parallelizing lda using partially collapsed gibbs sampling. arXiv preprint arXiv:1506.03784.

Marsaglia, G., Tsang, W. W., Sep. 2000. A simple method for generating gamma variables. ACM Trans. Math. Softw. 26 (3), 363–372. URL `http://doi.acm.org/10.1145/358407.358414`

Newman, D., Asuncion, A., Smyth, P., Welling, M., 2009. Distributed