Sparse Parallel Training of Hierarchical Dirichlet Process Topic Models

Alexander Terenin
Imperial College London

Joint work with Måns Magnusson and Leif Jonsson

EMNLP 2020

November 16th–20th, 2020

HTTPS://AVT.IM/ · @AVT_IM

UPPSALA UNIVERSITET

Imperial College London

Aalto University

Topic Models

Latent Dirichlet allocation

DM Blei, AY Ng, MI Jordan - Journal of Machine Learning Research, 2003 - jmlr.org

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of …

☆ 99 Cited by 28439 Related articles All 92 versions »

Hierarchical Dirichlet processes

YW Teh, MI Jordan, MJ Beal, DM Blei - Journal of the American Statistical …, 2006 - JSTOR

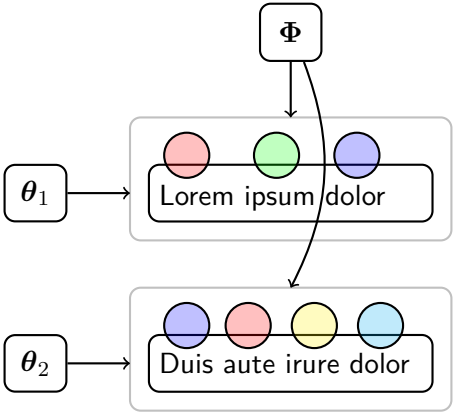
We propose the hierarchical Dirichlet process (HDP), a nonparametric Bayesian model for clustering problems involving multiple groups of data. Each group of data is modeled with a mixture, with the number of components being open-ended and inferred automatically by the …

☆ 99 Cited by 3853 Related articles All 82 versions »

The canonical topic models – everybody uses them!

This work: compute as fast, parallel, and principled as possible

Latent Dirichlet Allocation



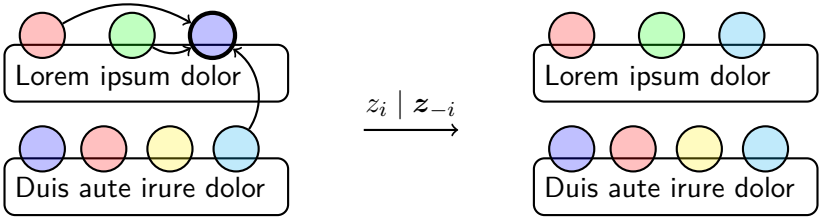
θ_d : document-topic probabilities

Φ : topic-word probabilities

z : topic indicators

w : word tokens

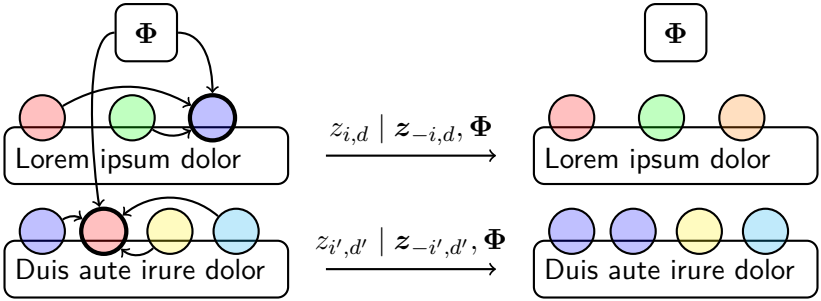
Sparse Fully Collapsed Gibbs Sampling



✓ Relatively fast

☒ Sequential
☒ Not fully sparse

Sparse Partially Collapsed Gibbs Sampling



✓ Massively parallel

☒ Still not fully sparse

M. Magnusson, L. Jonsson, M. Villani, and D. Broman. Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models. Journal of Computational and Graphical Statistics 27(2):449–463, 2018.

Pólya Urn LDA

Dirichlet

$$\boldsymbol{x} = \left[\frac{\gamma_1}{\sum_{i=1}^k \gamma_i}, .., \frac{\gamma_k}{\sum_{i=1}^k \gamma_i}\right] \qquad \rightarrow \qquad \boldsymbol{y} = \left[\frac{\tilde{\gamma}_1}{\sum_{i=1}^k \tilde{\gamma}_i}, .., \frac{\tilde{\gamma}_k}{\sum_{i=1}^k \tilde{\gamma}_i}\right]$$

$$\gamma_i \sim \mathrm{Gamma}(\varpi F_i, 1) \qquad\qquad \tilde{\gamma}_i \sim \mathrm{Poisson}(\varpi F_i)$$

Dense

Poisson Pólya Urn

Sparse

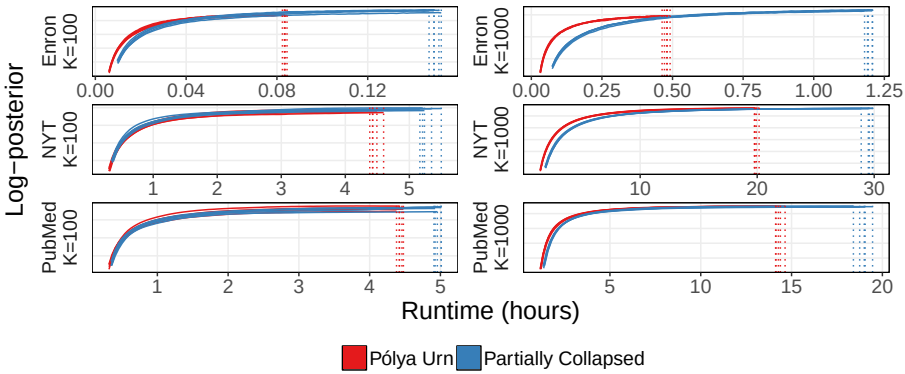
Theorem. Let $\boldsymbol{x} \sim \mathrm{Dir}(\varpi, \boldsymbol{F})$ and $\boldsymbol{y} \sim \mathrm{PPU}(\varpi, \boldsymbol{F})$. Then for all $\boldsymbol{F}$ we have $d_{\mathrm{LP}}(\boldsymbol{x}, \boldsymbol{y}) \to 0$ as $\varpi \to \infty$.

(d_{LP}: Levy-Prokhorov metric)

✓ Dense matrix $\boldsymbol{\Phi}$ becomes sparse

✓ Utilize all sources of sparsity

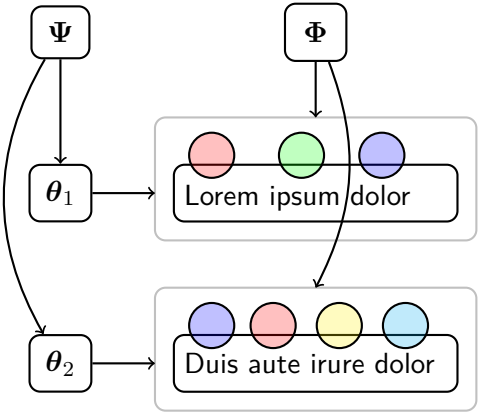
Performance



✓ Faster runtime with no discernible loss in topic quality

A. Terenin, M. Magnusson, L. Jonsson, and D. Draper. Pólya Urn Latent Dirichlet Allocation: a doubly sparse massively parallel sampler. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(7):1709–1719, 2019.

Hierarchical Dirichlet Process Topic Model



$\boldsymbol{\theta}_d$: document-topic probabilities

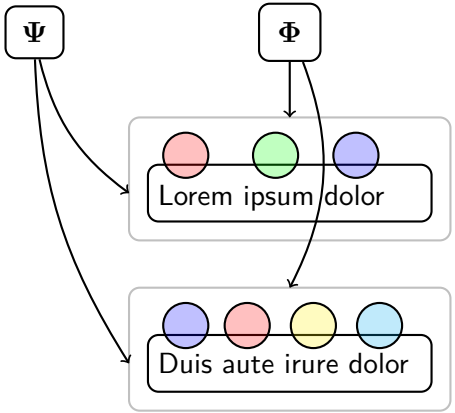
$\boldsymbol{\Phi}$: topic-word probabilities

$\boldsymbol{z}$: topic indicators

$\boldsymbol{w}$: word tokens

$\boldsymbol{\Psi}$: global topic probabilities

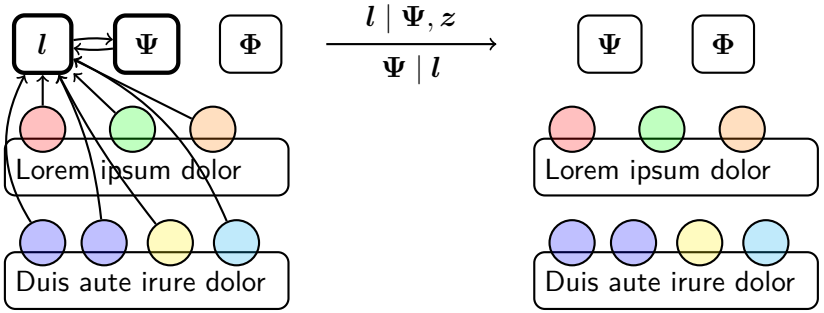
Global topic probabilities



Ψ: chosen to be a *Griffiths-Engen-McCloskey* (GEM) distribution

What's the posterior full conditional?

Sparse Partially Collapsed Gibbs Sampling



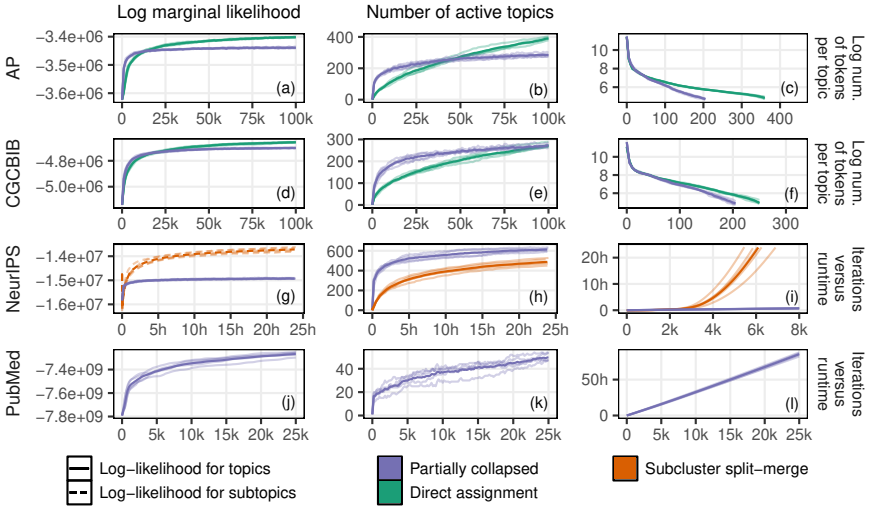
Theorem. $\Psi \mid l$ is given by

$$\Psi_k = \varsigma_k \prod_{i=1}^{k-1} (1 - \varsigma_i) \qquad \varsigma_k \sim \mathrm{B}(a_k^{(\Psi)}, b_k^{(\Psi)})$$

$$a_k^{(\Psi)} = 1 + l_k$$
$$b_k^{(\Psi)} = \gamma + \textstyle\sum_{i=k+1}^{\infty} l_i$$

and each l_k is a sum of certain binomial random variables.

Performance



✓ Sparsity and parallelism enable scalability to large data sets

Topics for PubMed

Large topics

47 322 709	40 229 486	34 685 122	30 795 166	30 707 144
care	age	model	cell	gene
health	risk	data	expression	protein
patient	children	system	growth	dna
medical	year	time	protein	expression
research	women	analysis	factor	sequence
clinical	patient	effect	receptor	genes
system	factor	test	kinase	rna
cost	population	field	beta	region

Topics for PubMed

Medium-size topics

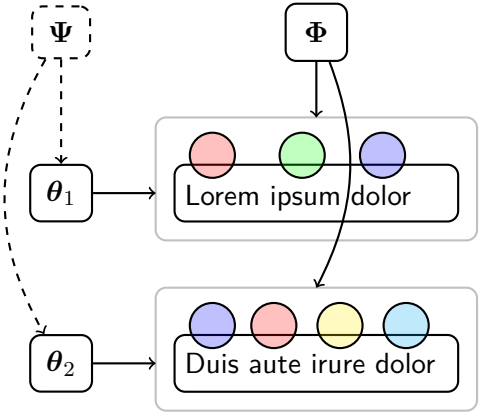
22 095 623	21 838 239	21 363 408	20 887 061	20 828 980
ca	structure	concentration	pregnancy	protein
effect	binding	degrees	level	binding
receptor	protein	samples	women	human
channel	reaction	liquid	hormone	antibodies
cell	acid	solution	day	acid
calcium	interaction	assay	fetal	alpha
concentration	compound	detection	infant	antibody
na	site	system	concentration	gel

Topics for PubMed

Small topics

16 136 164	14 201 063	13 706 016	13 191 158	13 105 245
effect	diet	patient	strain	protein
platelet	weight	disease	plant	membrane
induced	intake	gastric	growth	cell
oxide	food	asthma	acid	domain
rat	body	test	bacteria	binding
cell	effect	pylori	activity	receptor
endothelial	acid	arthritis	cell	lipid
activity	vitamin	chronic	species	membranes

Prior information: HDP vs. LDA

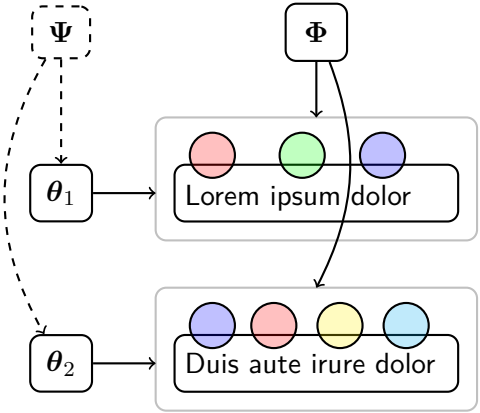


LDA: K topics with same number of tokens per topic *a priori*

HDP: ∞ topics with number of tokens per topic decreasing

Number of active topics: proxy for true number of topics

Prior information: HDP vs. LDA



Different prior specification yields different qualitative behavior

$$\Psi_k = \varsigma_k \prod_{i=1}^{\infty}(1-\varsigma_k) \qquad \varsigma_k \sim \mathrm{B}(1,\gamma) \qquad \text{vs.} \qquad \boldsymbol{\Psi} = \mathrm{U}(1,..,K)$$

Concluding remarks

Thank you for your attention!

HTTPS://AVT.IM/ · @AVT_IM

Imperial College London

UPPSALA UNIVERSITET

Aalto University

M. Magnusson, L. Jonsson, M. Villani, and D. Broman. Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models. Journal of Computational and Graphical Statistics 27(2):449–463, 2018.

A. Terenin, M. Magnusson, L. Jonsson, and D. Draper. Pólya Urn Latent Dirichlet Allocation: a doubly sparse massively parallel sampler. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(7):1709–1719, 2019.

A. Terenin, M. Magnusson, and L. Jonsson. Sparse Parallel Training of Hierarchical Dirichlet Process Topic Models. Conference on Empirical Methods in Natural Language Processing, 2020.