Independent Factor Topic Models

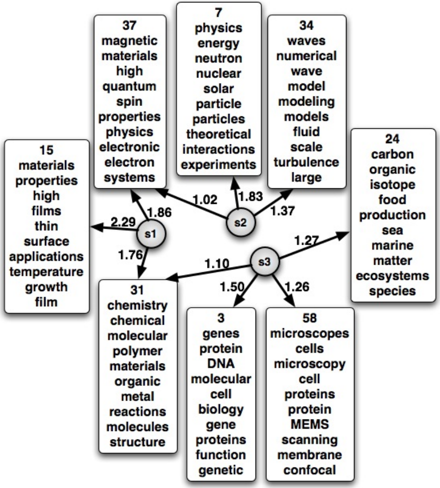
Duangmanee (Pew) Putthividhya
Hagai T. Attias
Sri Nagarajan

Institute for Neural Computation
UC San Diego, La Jolla, CA

June 16, 2009

Overview

- Statistical topic models that can capture topic correlations.
- **Exploratory Analysis**: navigate through document collections using topic graphs.



Outline

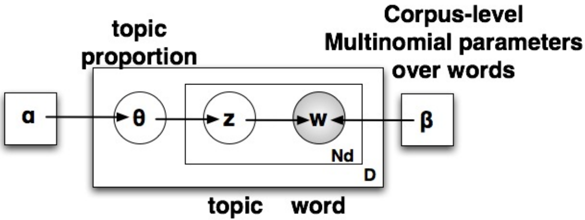
Related Work
- LDA, PAM, and CTM

Proposed Model
- Independent Factor Topic Models (IFTM)
- Gaussian Source Distribution
- Non-Gaussian Source Distribution
- Approximate Inference

Experimental Results
- Correlated Topics Visualization
- Model Comparison

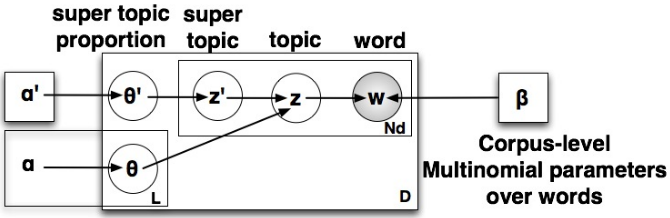
Latent Dirichlet Allocation (LDA)
(Blei, Ng, Jordan 2003)

- A document is modeled as a mixture of topics: $\theta \sim \text{Dir}(\alpha)$.
- Under Dirichlet, the components θ_i and θ_j are modeled as nearly independent.



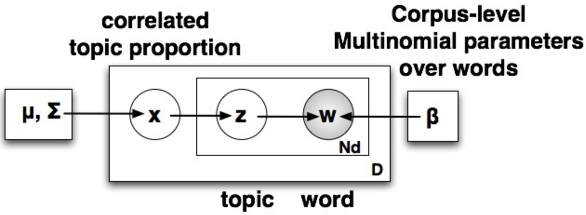
Pachinko Allocation Models (PAM)
(Li, McCallum 2006)

- Each super-topic models a distribution over topics.
- Under different super-topics, different sets of topics are more likely to co-occur.



Correlated Topic Models (CTM)
(Blei, Lafferty 2006)

- Proportion of topics $\theta \sim$ logistic Normal(μ, Σ)
  - sample $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$.
  - topic proportion: $\theta_k = \frac{e^{x_k}}{\sum_j e^{x_j}}$.
- Σ_{ij} encodes correlations between topic *i* and topic *j*.



Motivation

Problem with CTM:

- scale poorly with K.
- variational inference for CTM is slow especially for large K.

Solution: Low-rank covariance.

- Factor Analysis (FA) to constrain covariance structure
- Linear latent variable model for dimensionality reduction:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mu + \mathbf{n}, \text{ where } \begin{cases} \mathbf{s} \sim \mathcal{N}(\mathbf{s}; 0, I_L). \\ \mathbf{n} \sim \mathcal{N}(\mathbf{n}; 0, \Lambda). \end{cases}$$

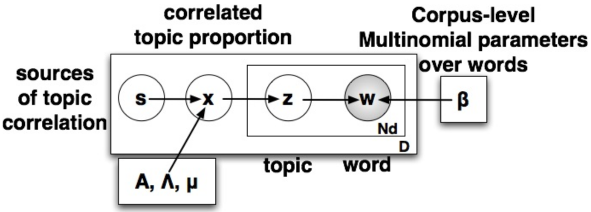
- Parameterize covariance parameter $\Sigma = \mathbf{A}\mathbf{A}^\top + \Lambda^{-1}$.
- Reduce the number of parameters $\mathcal{O}(K^2)$ to $\mathcal{O}(KL)$.

Independent Factor Topic Models (IFTM)

For each document:

- sample $\mathbf{s} \sim p(\mathbf{s})$.
- sample $\mathbf{x} \sim p(\mathbf{x}|\mathbf{s}, \mathbf{A}, \Lambda, \mu) = \mathcal{N}(\mathbf{x}; \mathbf{A}\mathbf{s} + \mu, \Lambda)$.
- Topic proportion: $\theta_k = \frac{e^{x_k}}{\sum_j e^{x_j}}$.
- For each word:
  - sample $z_n \sim \text{Mult}(\theta)$.
  - sample $w_n \sim \text{Mult}(\beta_{z_n})$.

- L sources
- K topics
- T words
- Model parameters: $\beta_{T \times K}, A_{K \times L}, \Lambda_{K \times 1}, \mu_{K \times 1}$



Non-Gaussian Source Distribution

Problems with Gaussian sources:

- **A** is often not very interpretable.
- Non-identifiability associated with rotations.
  - **A** and **AQ** give the same likelihood, for any arbitrary rotation matrix **Q**.

Solution: sparse source prior.

- Laplacian distribution: $p(\mathbf{s}) \propto \prod_{l=1}^{L} \exp(-|s_l|)$.
- **A** becomes interpretable and can be uniquely determined up to a scale factor.

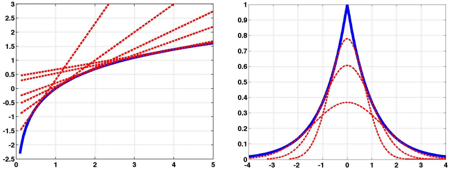
Approximate Inference

Variational mean-field approximation:

- Approximate $p(\mathbf{Z}, \mathbf{x}, \mathbf{s}|\mathbf{W}) \approx \prod_n q(z_n) q(\mathbf{x}) q(\mathbf{s})$.
- Find $q(\cdot)$ that minimizes $KL(q||p)$.

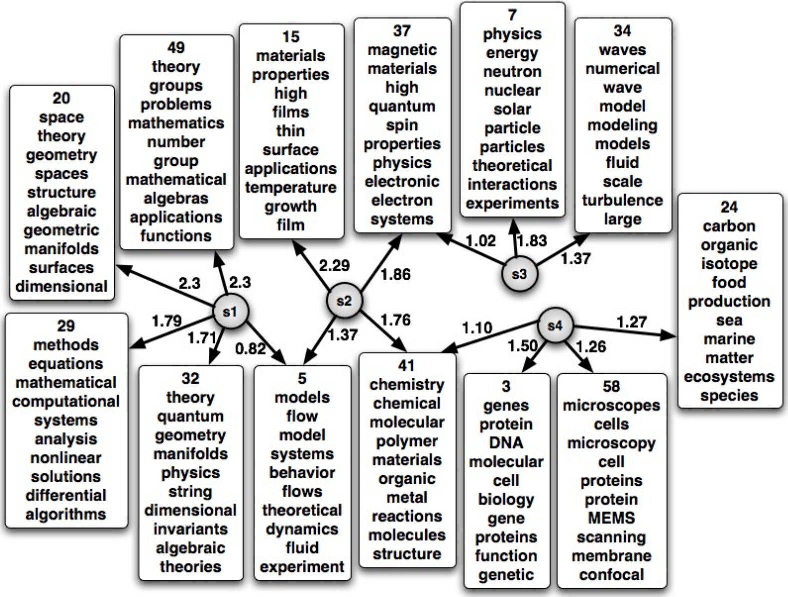
Convex variational Bound:

- $-\log \sum_j e^{x_j} \geq -\log \xi - \frac{1}{\xi} \sum_j e^{x_j} + 1$.
- $-|s|^\nu \geq -(1 - \frac{\nu}{2})\eta^\nu - \frac{\nu}{2\eta^{2-\nu}} s^2 \; , \; 0 < \nu \leq 2$.

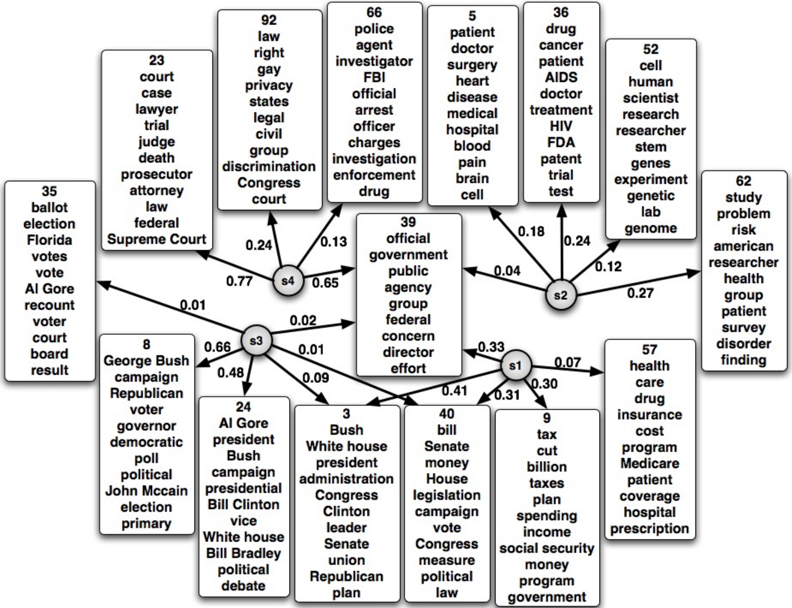


Fast 1-d Newton-Raphson update for $q(\mathbf{x}) = \prod_k q(x_k)$.

Topics from 4,000 NSF Abstracts



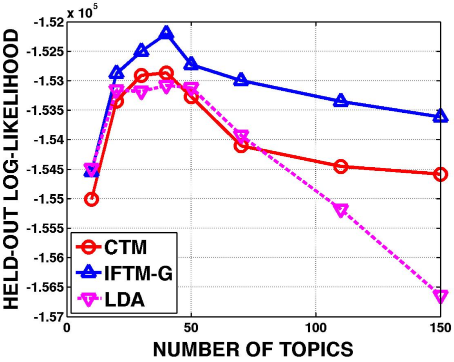
Topics from 15,000 Articles from NYTimes



Model Comparison I

NSF Abstracts

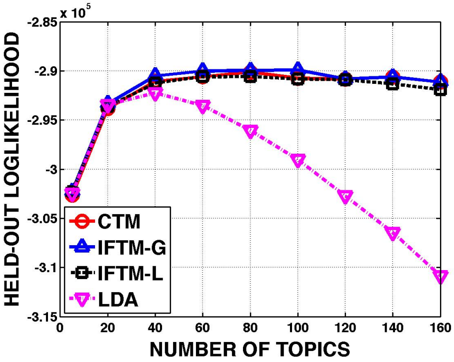
- D = 1,500 documents
- T = 2,938 words
- Average 108 words/document



Model Comparison II

NSF Abstracts

- D = 4,000 documents
- T = 5,261 words
- Average 120 words/document



Computational Savings

Training time (in hours) as K increases.

K	LDA	IFTM-G	IFTM-L	CTM
20	0.546	0.878	1.177	3.648
40	1.108	1.833	4.033	13.973
60	2.795	3.861	7.733	22.551
80	3.651	8.705	13.030	43.156
100	4.147	9.296	18.599	53.376
120	4.840	13.836	19.963	65.446
140	6.521	17.340	22.946	70.833
160	10.173	20.287	25.523	90.900
	$\times$**0.45**	$\times$**1**	$\times$**1.5**	$\times$**5**

Summary

Key ideas:

- IFTM learns sparse patterns of correlations between topics.
- Performs competitively with CTM as measured by held-out log-likelihood.
- 3-5 Times speed gain compared to CTM.

Limitation

- Need to specify the number of sources L.

Future Work:

- Automatically determine the number of hidden sources L by putting a prior over $\mathbf{A}$.
- Explore other source distribution $p(\mathbf{s})$.