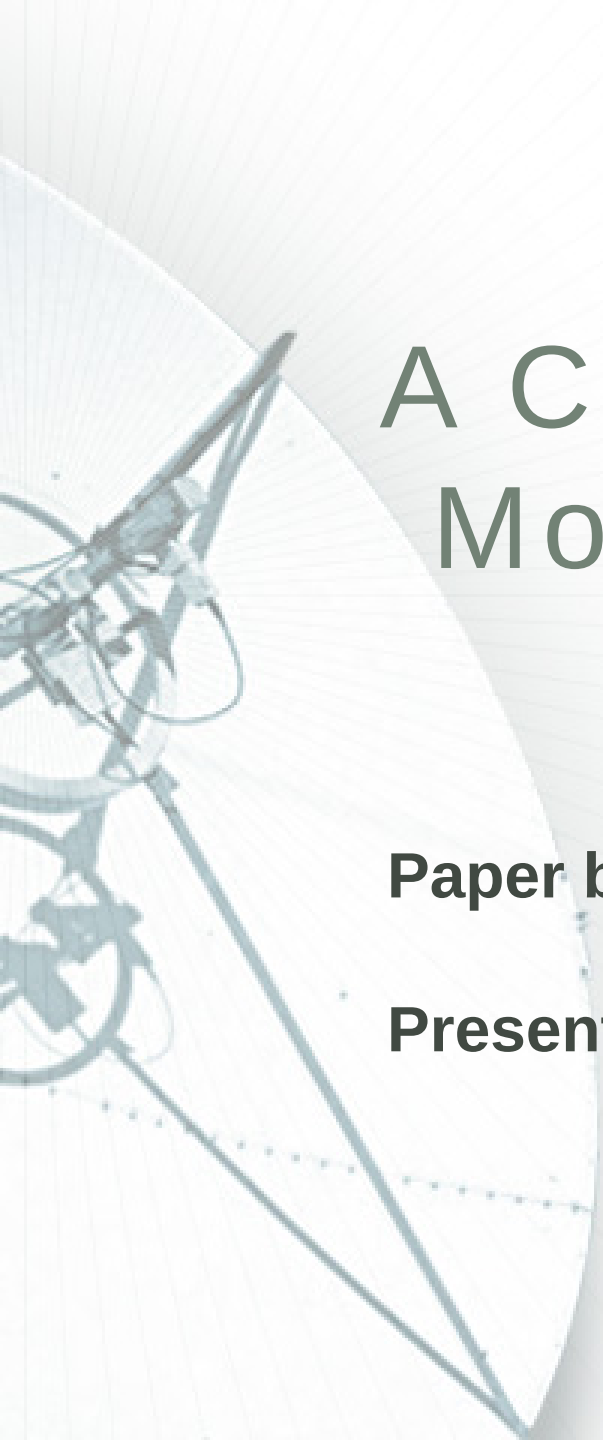


A Correlated Topic Model of *Science*

Paper by:

David M. Blei, John D. Lafferty

Presented by:

Muhammad Aurangzeb Ahmad

Outline

- Motivation
- The Correlated Topic Model
- Experiments and Results
- Demonstration
- Conclusion
- Related Work

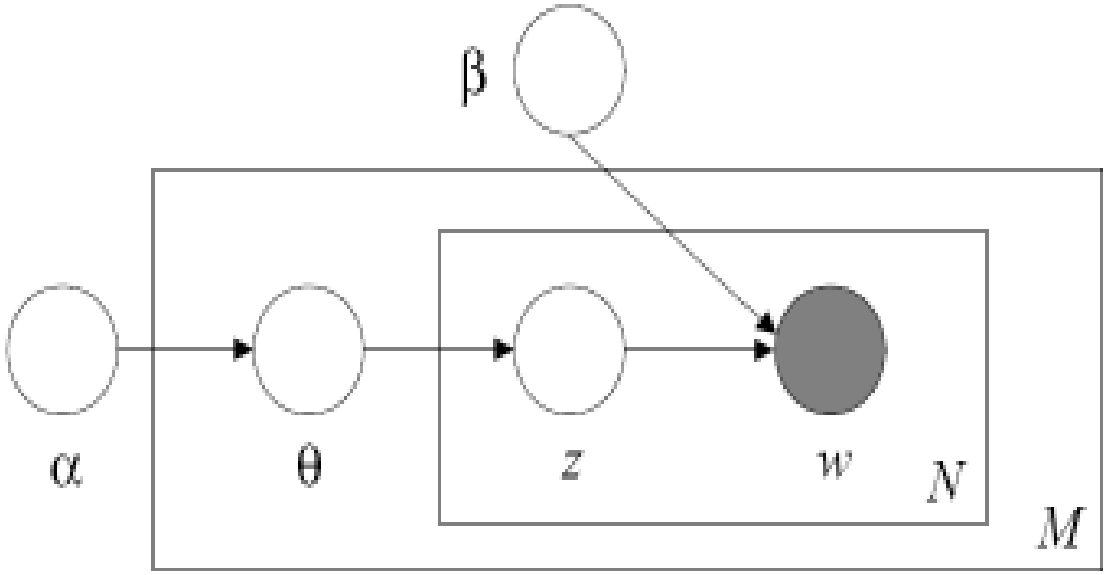
Motivation

- Immense amount of document collections available for the first time.
- There is a need to develop tools for browsing, searching and exploring such data collections.
- Goal: Develop models to extract structure from the data without any explicit “understanding” of the language.

LDA Revisited

- Assume that the words for each document arise from a mixture of topics
- Each topic is multinomial over a (fixed) vocabulary.
- Bag of words Assumption (Exchangability)
- Topics are Independent

LDA: Graphical Model



Limitation of LDA

- There may be cases where the topics are actually correlated with one another.
- Example: In the journal *Science an article about genetics is more likely to be also about health and diseases than it is likely to be about X-Ray astronomy.*
- *However models like LDA cannot model correlation between topics because of independence assumption.*
  - *A consequence of using Dirichlet*

CTM Model

- Solution: Replace the Dirichlet with another distribution – lognormal distribution
- Implications:
  - The conjugacy is lost.
  - Gibbs Sampling can no longer be done
  - MCMC sampling is prohibitive due to scale and high dimensionality of the data
- Solution: Use Variational Inference

CTM: Basics

- Model the words for each document from a mixture model.
- Mixture components are shared by all the documents.
- Mixture proportions are document specific
- Each document can have multiple topics with different proportions.

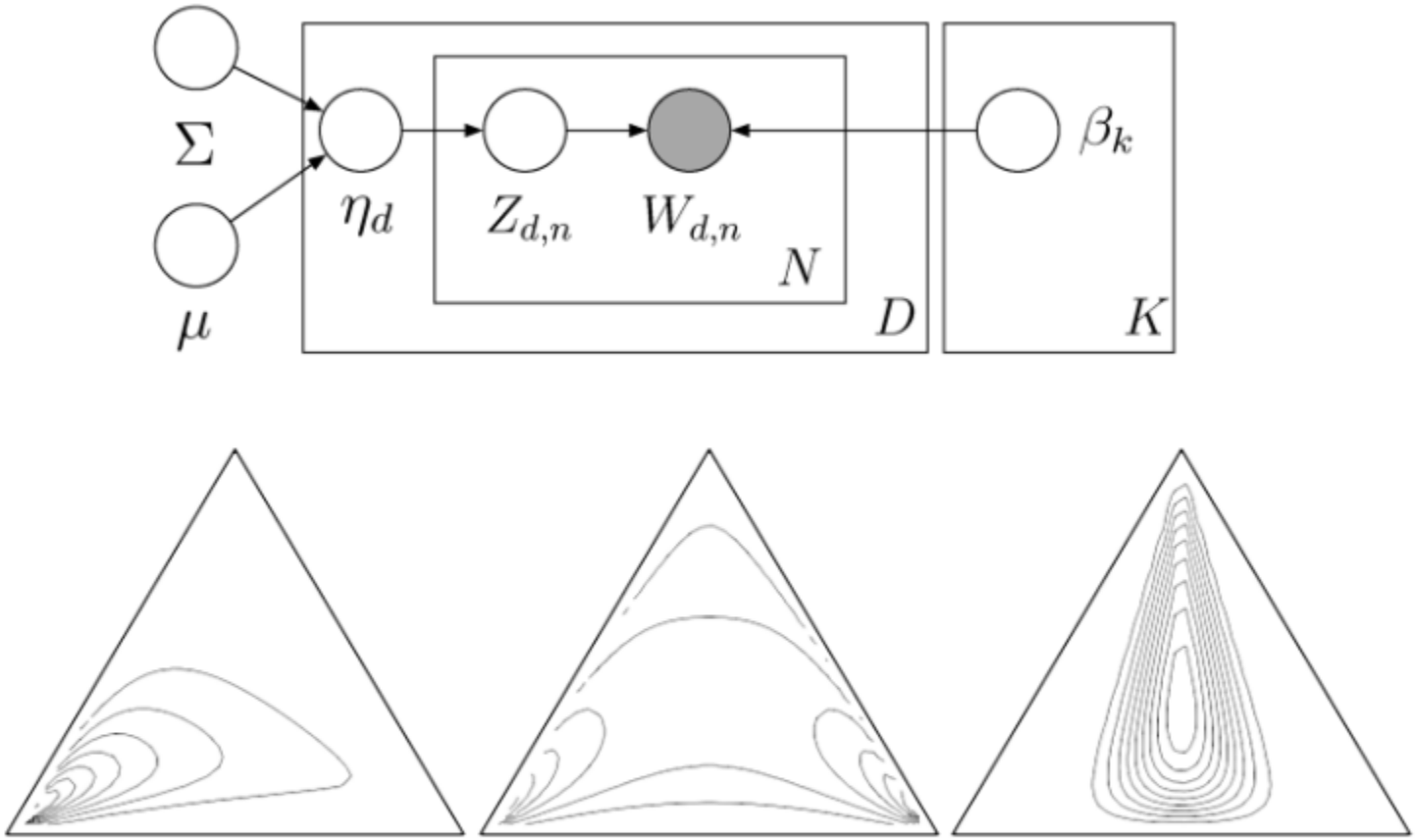


FIG. 1. Top: *Probabilistic graphical model representation of the correlated topic model. The logistic normal distribution, used to model the latent topic proportions of a document, can represent correlations between topics that are impossible to capture using a Dirichlet.* Bottom: *Example densities of the logistic normal on the* 2*-simplex. From left*: *diagonal covariance and nonzero-mean, negative correlation between topics* 1 *and* 2*, positive correlation between topics* 1 *and* 2.

The Generative Model

Specifically, the correlated topic model assumes that an N-word document arises from the following generative process. Given topics $\boldsymbol{\beta}_{1:K}$, a K-vector $\boldsymbol{\mu}$ and a $K \times K$ covariance matrix $\boldsymbol{\Sigma}$:

1. Draw $\eta_d | \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. For $n \in \{1, \ldots, N_d\}$:

(a) Draw topic assignment $Z_{d,n} | \eta_d$ from $\text{Mult}(f(\eta_d))$.
(b) Draw word $W_{d,n} | \{z_{d,n}, \boldsymbol{\beta}_{1:K}\}$ from $\text{Mult}(\boldsymbol{\beta}_{z_{d,n}})$,

where $f(\eta)$ maps a natural parameterization of the topic proportions to the mean parameterization,

$$\boldsymbol{\theta} = f(\eta) = \frac{\exp\{\eta\}}{\sum_i \exp\{\eta_i\}}. \tag{1}$$

Computation with CTM

- Posterior Inference with variational methods.

3.1. *Posterior inference with variational methods.* Given a document $\boldsymbol{w}$ and a model $\{\boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, the posterior distribution of the per-document latent variables is

$$p(\boldsymbol{\eta}, z|\boldsymbol{w}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\eta}) p(w_n|z_n, \boldsymbol{\beta}_{1:K})}{\int p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|\boldsymbol{\eta}) p(w_n|z_n, \boldsymbol{\beta}_{1:K}) d\boldsymbol{\eta}},$$

- Problem:
  - Computing the integral is intractable
  - The distribution of topic proportions is not conjugate to the distribution of topic assignments.

Implications of non-conjugacy

- Cannot use MCMC sampling techniques developed for computing with Dirichlet-based mixed membership model.
- These methods are based on Gibbs Sampling where the conjugacy between the latent variables allows us to compute posteriors).
- Alternate solution: Use another tailored Metropolis-Hastings algorithm but this has the problem of speed and convergence.

Variational Inference in CTM

- Variational Inference Revisited: Optimize free parameters over a distribution over the latent variables so that the distributions' KL Divergence is close to the true posterior.
- Using Jensen's inequality to bound the log probability of a document.

$$\log p(w_{1:N}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}) \geq \mathrm{E}_q[\log p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})] + \sum_{n=1}^{N} \mathrm{E}_q[\log p(z_n|\boldsymbol{\eta})] + \sum_{n=1}^{N} \mathrm{E}_q[\log p(w_n|z_n, \boldsymbol{\beta})] + \mathrm{H}(q),$$

- Use the variational distribution for inference.

$$q(\eta_{1:K}, z_{1:N}|\lambda_{1:K}, \nu^2_{1:K}, \boldsymbol{\phi}_{1:N}) = \prod_{i=1}^{K} q(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n|\boldsymbol{\phi}_n).$$

Parameter Estimation

- Goal: Given a collection of documents, do parameter estimation to maximize the likelihood of a corpus of documents as a function of topics and the multivariate Gaussian.
- The presence of latent structures precludes the use of marginal likelihood.
- Variational EM: In the E- step use the variational distribution instead of the posterior as normally done in EM.

Parameter Estimation (ii)

- In the M-step maximize the bound w.r.t. model parameters *i.e.,* maximize likelihood estimates of the topics and multivariate Gaussian. (Expectation is taken with respect to q)

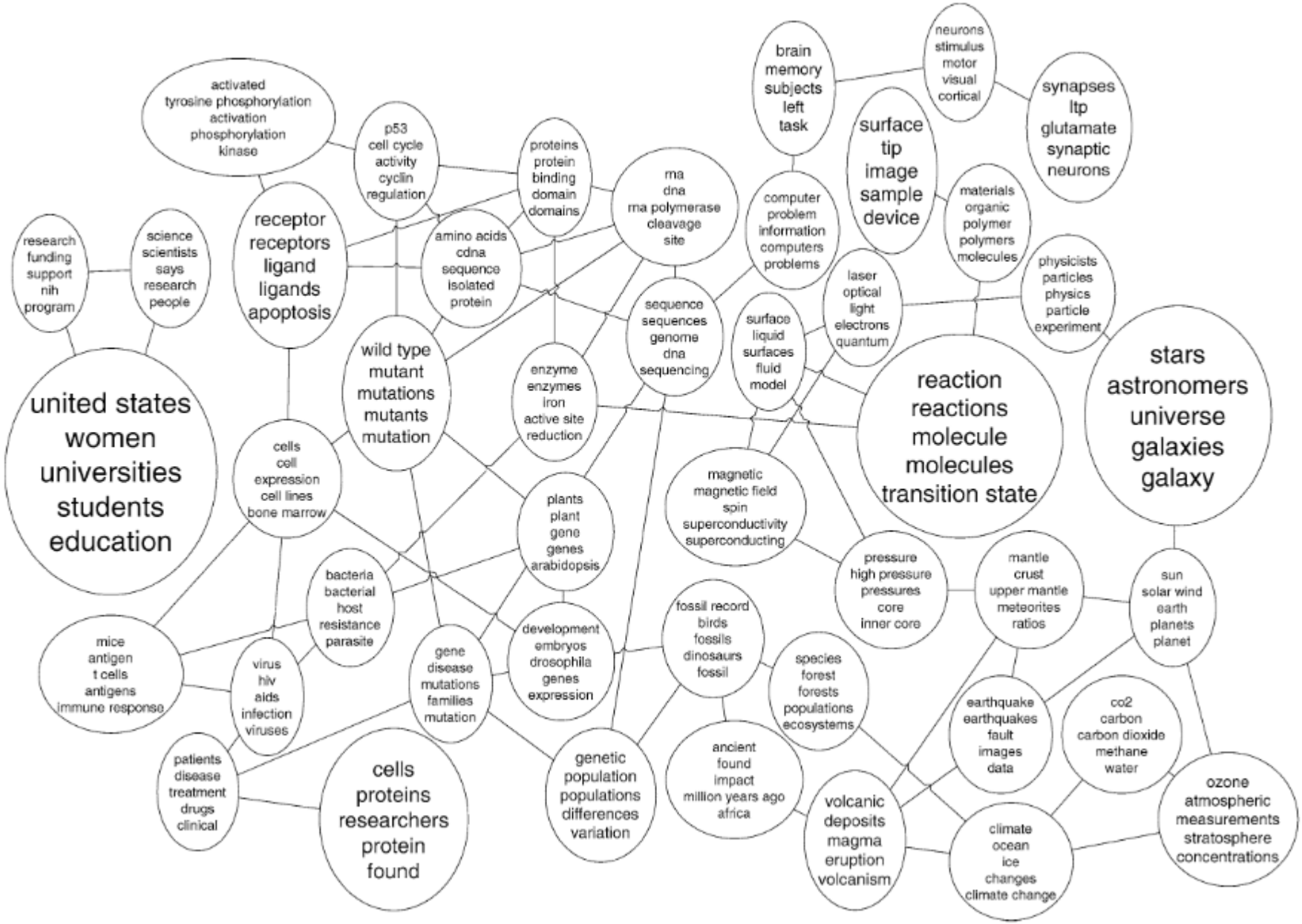
$$\widehat{\boldsymbol{\beta}}_i \propto \sum_d \phi_{d,i} \mathbf{n}_d,$$

$$\widehat{\boldsymbol{\mu}} = \frac{1}{D} \sum_d \boldsymbol{\lambda}_d,$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{D} \sum_d I v_d^2 + (\boldsymbol{\lambda}_d - \widehat{\boldsymbol{\mu}})(\boldsymbol{\lambda}_d - \widehat{\boldsymbol{\mu}})^T,$$

Topic Graphs

- Idea: Covariance can be used to visualize the relationship between topics
- Topic Graph: Nodes represent topics and edges represent correlation between topics.
- Problem: Control the sparsity of the graph.
- According to Meinshausen and Buhlmann the lasso (used for regularization) can be used to construct such a graph.
- In CTM, for a given document use the mean for the variational approximation as data.
- Regress each component onto another w/ L1.


activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
materials organic polymer polymers molecules
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
ma dna rna polymerase cleavage site
computer problem information computers problems
research funding support nih program
science scientists says research people
physicists particles physics particle experiment
laser optical light electrons quantum
sequence sequences genome dna sequencing
surface liquid surfaces fluid model
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
stars astronomers universe galaxies galaxy
reaction reactions molecule molecules transition state
united states women universities students education
cells cell expression cell lines bone marrow
plants plant gene genes arabidopsis
magnetic magnetic field spin superconductivity superconducting
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
bacteria bacterial host resistance parasite
fossil record birds fossils dinosaurs fossil
development embryos drosophila genes expression
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
species forest forests populations ecosystems
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change
ozone atmospheric measurements stratosphere concentrations

Experiments and Results

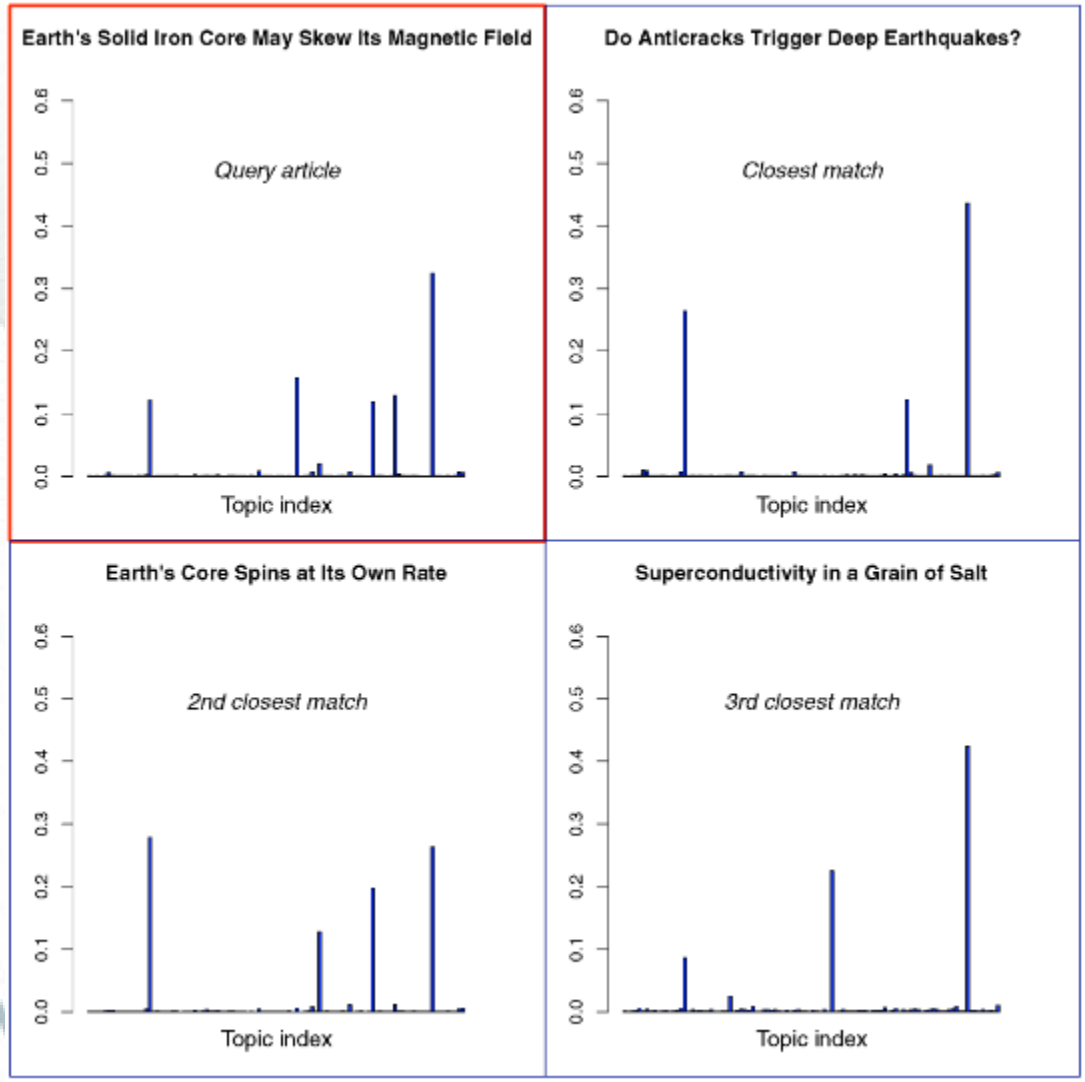
- Dataset Description
  - JSTOR: Documents dating back to 1600s.
  - 100 topic model on *Science* articles from 1990 to 1999
  - Vocabulary Size: 356,195 (pruned)
  - 16351 documents
  - 19,088 unique terms
  - 5.7 million words

Finding Similar Documents

- Use topic proportions to determine similarity between documents.
- Hellinger Distance

$$\mathrm{E}[d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)] = \mathrm{E}_q\left[\sum_k (\sqrt{\theta_{ik}} - \sqrt{\theta_{jk}})^2\right] = 2 - 2\sum_k \mathrm{E}_q[\sqrt{\theta_{ik}}]\mathrm{E}_q\left[\frac{\theta_{jk}}{\sqrt{\theta_{jk}}}\right]$$

- Comparison to LDA
  - Test on different number of documents.
  - 10-fold cross validation
  - Predict words based on the word already seen. (Perplexity)


Earth's Solid Iron Core May Skew its Magnetic Field
Query article
Topic index
Do Anticracks Trigger Deep Earthquakes?
Closest match
Topic index
Earth's Core Spins at Its Own Rate
2nd closest match
Topic index
Superconductivity in a Grain of Salt
3rd closest match
Topic index
0.0
0.1
0.2
0.3
0.4
0.5
0.6

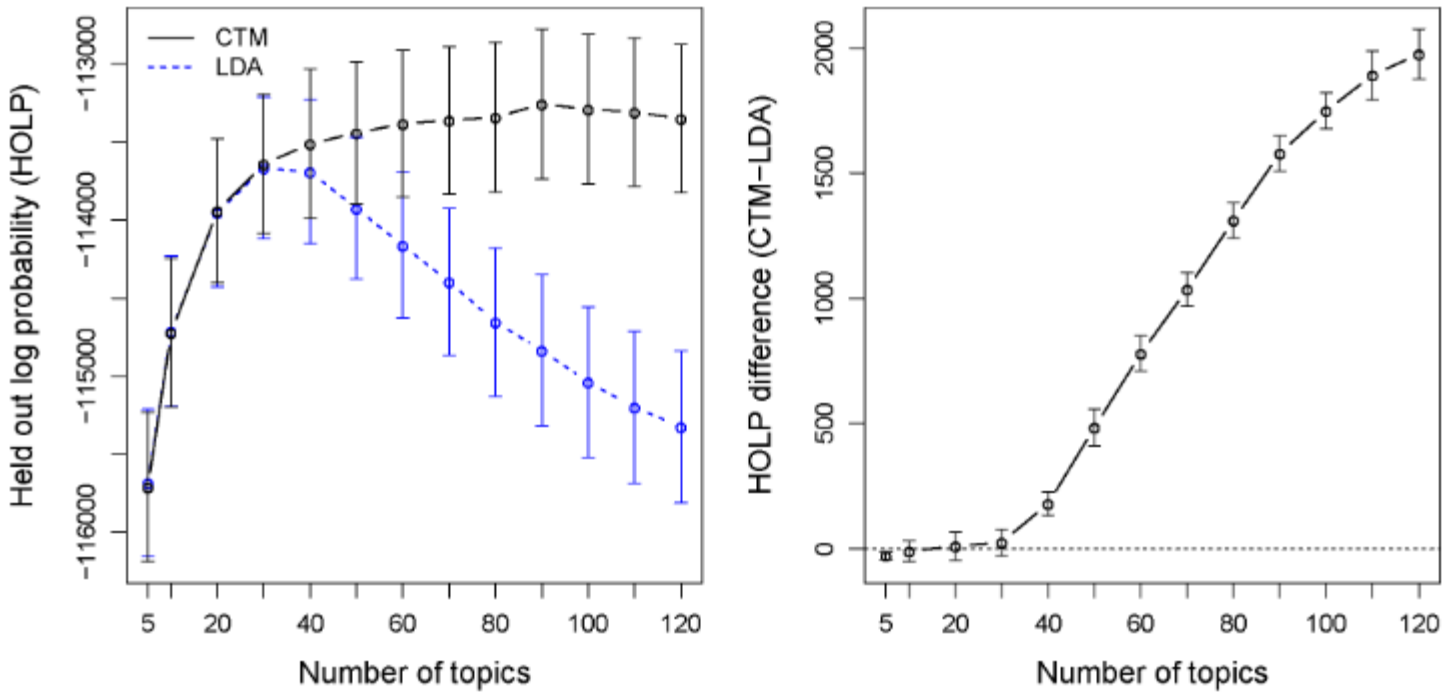


FIG. 4. (Left) *The* 10*-fold cross-validated held-out log probability of the* 1960 *Science corpus, computed by importance sampling. The CTM supports more topics than LDA. See figure at right for the standard error of the difference.* (Right) *The mean difference in held-out log probability. Numbers greater than zero indicate a better fit by the CTM.*

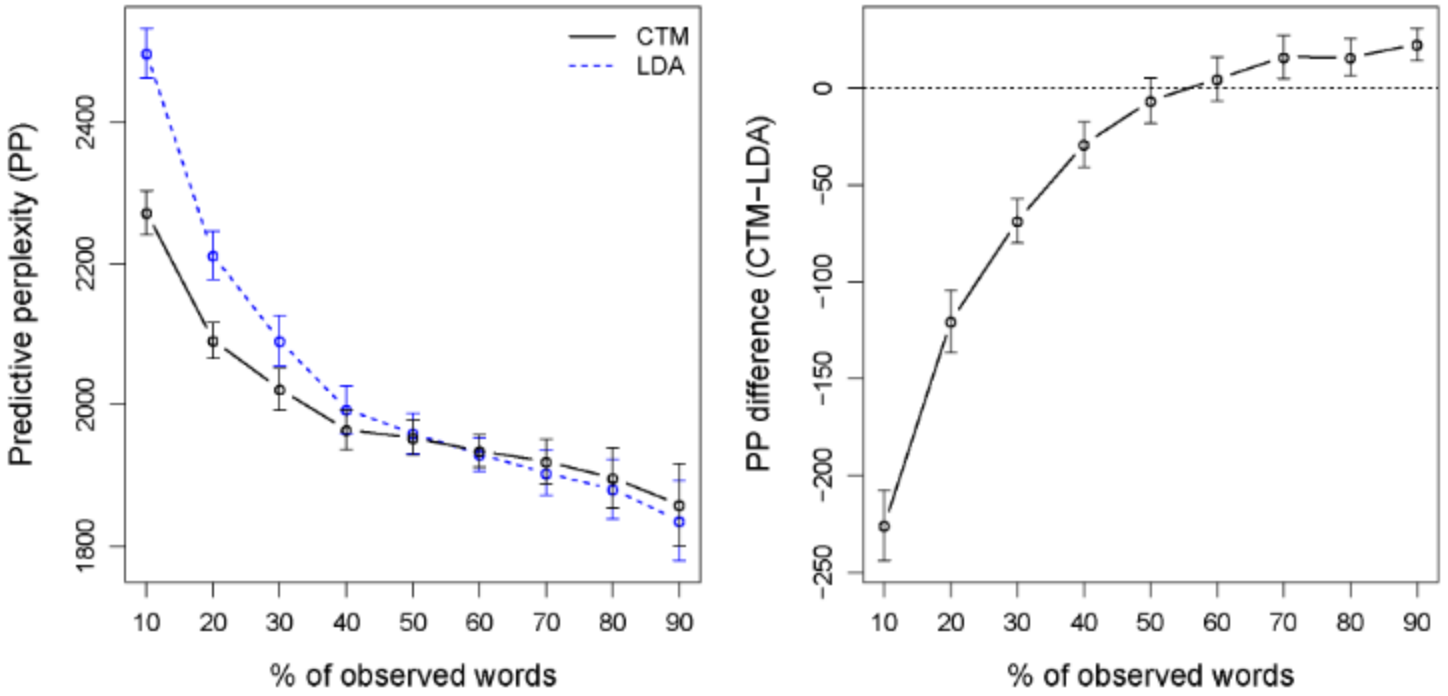


FIG. 5. (Left) *The* 10*-fold cross-validated predictive perplexity for partially observed held-out documents from the* 1960 *Science corpus* ($K = 50$). *Lower numbers indicate more predictive power from the CTM.* (Right) *The mean difference in predictive perplexity. Numbers less than zero indicate better prediction from the CTM.*

WORDS
genetic
population
populations
data
dna
evolution
variation
differences
studies
evolutionary
analysis
different
two
genes
genetics

RELATED TOPICS
genetic population populations data dna
fossil species evolution birds evolutionary
male males female females species
life colonies insect larvae queens
gene disease human chromosome cancer
sequence dna genome sequences genes
human humans spain homo chimpanzees

RELATED DOCUMENTS
"On the Probability of Matching DNA Fingerprints" (1992)
"Experimental Tests of the Roles of Adaptation, Chance, ar Evolution" (1995)
"Forensic DNA Tests and Hardy-Weinberg Equilibrium" (199
"Genes, Environment, and Personality" (1994)
"The Utility of DNA Typing in Forensic Work" (1991)
"No Excess of Homozygosity at Loci Used for DNA Fingerpri
"Statistical Evaluation of DNA Fingerprinting: A Critique of Report" (1993)
"The Genetic Basis of Complex Human Behaviors" (1994)
"Of Genes and Genomes" (1991)
"Gene Trees and the Origins of Inbred Strains of Mice" (199
"Sources of Human Psychological Differences: The Minneso Twins Reared Apart" (1990)
"Historical Genetics: The Parentage of Chardonnay, Gamay Wine Grapes of Northeastern France" (1999)
"Balancing Selection at Allozyme Loci in Oysters: Implicatio

Related Work

- **Pachinko Allocation Model:** (Li, McCallum '06) An LDA-style topic model that uses a DAG to capture arbitrary, nested and possibly sparse correlations among topics.
- **Non-Parametric PAM: (**Li, Blei, Andrew McCallum)
- ***Evolution of Topic Models: Nallapati et al.***

References