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 \( \beta_{1:K} \), a \( K \)-vector \( \mu \) and a \( K \times K \) covariance matrix \( \Sigma \):

1. Draw \( \eta_d | \{ \mu, \Sigma \} \sim \mathcal{N}(\mu, \Sigma) \).
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 \( W_{d,n} | \{ z_{d,n}, \beta_{1:K} \} \) from \( \text{Mult}(\beta_{z_{d,n}}) \),

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

\[
\theta = f(\eta) = \frac{\exp\{\eta\}}{\sum_i \exp\{\eta_i\}}.
\]
Computation with CTM

• Posterior Inference with variational methods.

  3.1. Posterior inference with variational methods. Given a document \( w \) and a model \( \{\beta_{1:K}, \mu, \Sigma\} \), the posterior distribution of the per-document latent variables is

  \[
  p(\eta, z|w, \beta_{1:K}, \mu, \Sigma) = \frac{p(\eta|\mu, \Sigma) \prod_{n=1}^{N} p(z_n|\eta) p(w_n|z_n, \beta_{1:K})}{\int p(\eta|\mu, \Sigma) \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|\eta) p(w_n|z_n, \beta_{1:K}) d\eta},
  \]

• 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.

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

- Use the variational distribution for inference.

\[
q(\eta_{1:K}, z_{1:N} | \lambda_{1:K}, \nu^2_{1:K}, \phi_{1:N}) = \prod_{i=1}^{K} q(\eta_i | \lambda_i, \nu^2_i) \prod_{n=1}^{N} q(z_n | \phi_n).
\]
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 \( q \))

\[
\hat{\beta}_i \propto \sum_d \phi_{d,i} n_d,
\]

\[
\hat{\mu} = \frac{1}{D} \sum_d \lambda_d,
\]

\[
\hat{\Sigma} = \frac{1}{D} \sum_d I v_d^2 + (\lambda_d - \hat{\mu})(\lambda_d - \hat{\mu})^T,
\]
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.
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

\[
E[d(\theta_i, \theta_j)] = E_q \left[ \sum_k (\sqrt{\theta_{ik}} - \sqrt{\theta_{jk}})^2 \right] = 2 - 2 \sum_k E_q[\sqrt{\theta_{ik}}]E_q\left[\frac{\theta_{jk}}{\sqrt{\theta_{jk}}}\right]
\]

• Comparison to LDA
  – Test on different number of documents.
  – 10-fold cross validation
  – Predict words based on the word already seen. (Perplexity)
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          | RELATED TOPICS                                                                 | RELATED DOCUMENTS                                                                 |
|---------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| genetic       | genetic population populations data dna                                       | “On the Probability of Matching DNA Fingerprints” (1992)                         |
| population    | fossil species evolution birds evolutionary                                   | “Experimental Tests of the Roles of Adaptation, Chance, and Evolution” (1995)   |
| populations   | male males female females species                                              | “Forensic DNA Tests and Hardy-Weinberg Equilibrium” (1993)                      |
| data          | life colonies insect larvae queens                                             | “Genes, Environment, and Personality” (1994)                                    |
| dna           | gene disease human chromosome cancer                                          | “The Utility of DNA Typing in Forensic Work” (1991)                              |
| evolution     | sequence dna genome sequences genes                                           | “No Excess of Homozygosity at Loci Used for DNA Fingerprinting” (1994)           |
| variation     | human humans spain homo chimpanzees                                           | “Statistical Evaluation of DNA Fingerprinting: A Critique of the MRC Report” (1993) |
| differences   |                                                                                 | “The Genetic Basis of Complex Human Behaviors” (1994)                           |
| studies       |                                                                                 | “Of Genes and Genomes” (1991)                                                    |
| evolutionary  |                                                                                 | “Gene Trees and the Origins of Inbred Strains of Mice” (1994)                    |
| analysis      |                                                                                 | “Sources of Human Psychological Differences: The Minnesota Twins Reared Apart” (1990) |
| different     |                                                                                 | “Historical Genetics: The Parentage of Chardonnay, Gamay Wine Grapes of Northeastern France” (1999) |
| two           |                                                                                 | “Balancing Selection at Allozyme Loci in Oysters: Implications” (1999)          |
| genes         |                                                                                 |                                                                                 |
| genetics      |                                                                                 |                                                                                 |
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.
