The Discrete Infinite Logistic Normal Distribution for Mixed-Membership Modeling

John Paisley
Chong Wang
David Blei

Presented by Xiaoxue Li

Introduction

- Mixed membership models： model relational data, characterized by grouped observations generated by a mixture of latent distributions over the observation space -- originally designed for topic models
- HDP-LDA: models shared ‘atoms’ among documents, an infinite number of statistically independent topics. Little about correlations of topics in group level distribution
- Discrete Infinite Logistic Normal distribution – DILN, as hierarchical Bayesian nonparametric prior to model correlations between the occurrences of latent components

Gamma Process Construction of the HDP

- Hierarchical representation of Dirichlet Process

$$G \sim \mathrm{DP}(\alpha G_0), \quad G'_m \overset{iid}{\sim} \mathrm{DP}(\beta G),$$

$$\theta_n^{(m)} \sim G'_m, \quad X_n^{(m)} \sim f(\theta_n^{(m)}).$$

- In a two-level HDP of topic modeling:

  Top level

$$G = \sum_{k=1}^{\infty} V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{\eta_k},$$

$$V_k \overset{iid}{\sim} \mathrm{Beta}(1, \alpha), \quad \eta_k \overset{iid}{\sim} G_0.$$

Gamma Process Construction of the HDP

- Second level

$$G'_m = \sum_{k=1}^{\infty} \frac{Z_k^{(m)}}{\sum_{j=1}^{\infty} Z_j^{(m)}} \delta_{\eta_k},$$

$$Z_k^{(m)} \stackrel{ind}{\sim} \mathrm{Gamma}(\beta p_k, 1);$$

$$p_k := V_k \prod_{j=1}^{k-1} (1 - V_j)$$

completely random measure

Discrete Infinite Logistic Normal

- Latent features imbued with location vectors, "close" features tend to co-occur more often than those that are "far apart"

- Top level $G \sim \mathrm{DP}(\alpha G_0 \times L_0)$.
- Second level

$$G_m^{DP} \sim DP(\beta G), \quad w^{(m)}(\ell) \sim GP(\mathbf{m}(\ell), \mathbf{K}(\ell, \ell')).$$

scale the group-level DP by the exponentiated GP,

$$G'_m(\{\eta, \ell\}) \propto G_m^{DP}(\{\eta, \ell\}) \exp\{w^{(m)}(\ell)\}.$$

Normalized Gamma Representation

Top-Level DP

$$V_k \stackrel{iid}{\sim} Beta(1, \alpha), \quad \eta_k \stackrel{iid}{\sim} G_0, \quad \ell_k \stackrel{iid}{\sim} L_0$$

$$\downarrow$$

$$G = \sum_{k=1}^{\infty} V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{\{\eta_k, \ell_k\}}$$

think of ℓ_k as the location of atom k.

Normalized Gamma Representation

<u>Group-Level Distributions</u>

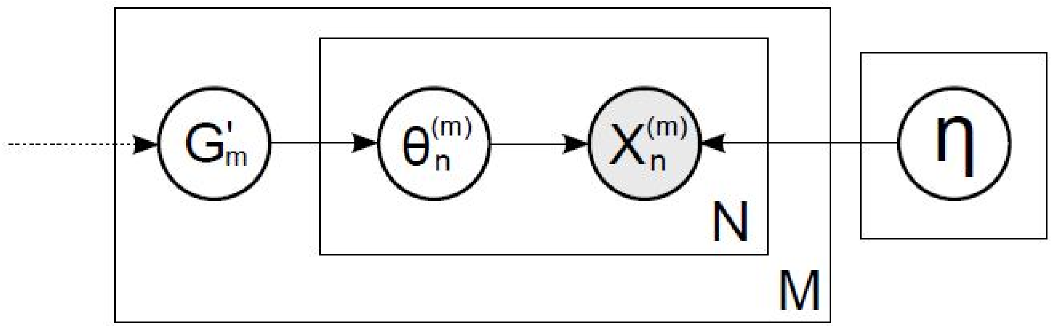
$$Z_k^{(m)} \sim Gamma(\beta p_k, e^{-w_k^{(m)}}), \quad w^{(m)} \overset{iid}{\sim} GP(\mathbf{m}, \mathbf{K})$$

$$\downarrow$$

$$G'_m = \sum_{k=1}^{\infty} \frac{Z_k^{(m)}}{\sum_{j=1}^{\infty} Z_j^{(m)}} \delta_{\eta_k}$$

If $w^{(m)} = 0$, then this is a representation of the HDP.

DILN Topic Model



Given G'_m, words in document m are generated according to

$$X_n^{(m)} \sim Mult(\theta_n^{(m)}), \quad \theta_n^{(m)} \overset{iid}{\sim} G'_m \, .$$

For inference, we introduce a latent indicator $C_n^{(m)}$, such that

$$\theta_n^{(m)} = \eta_{C_n^{(m)}} .$$

Variational Inference for DILN

- We use variational inference to learn the approximate posterior of the DILN model [Jordan *et al.*, 1999].
  - Mean-field variational inference uses a factorized q distribution to approximate the true posterior of a model's parameters.
  - Searches for the parameters of q that minimize the KL divergence between q and the true posterior.
- In a DILN topic model, the hidden variables are

  Document level: $\mathbf{Z}$, $\mathbf{w}$, $\mathbf{C}$

  Corpus level: $\boldsymbol{\eta}$, $\mathbf{V}$, $\mathbf{m}$, $\mathbf{K}$, α, β
- Note: We learn $\mathbf{K}$ directly, rather than latent locations, ℓ_k. This leads to a fast, closed-form update.

Variational Inference for DILN

- Inference note: For each group, we use the lower bound

$$-\mathbb{E}_Q\left[\ln \sum_{k=1}^{T} Z_k\right] \geq -\ln \xi - \frac{\sum_{k=1}^{T} \mathbb{E}_Q[Z_k] - \xi}{\xi}.$$

- Results in analytical updates for $q(Z_k) = Gamma(Z_k | a_k, b_k)$,

$$\begin{aligned} a_k &= \beta p_k + \sum_{n=1}^{N} \phi_n(k), \\ b_k &= \mathbb{E}_Q[\exp\{-w_k\}] + \frac{N}{\xi}. \end{aligned}$$

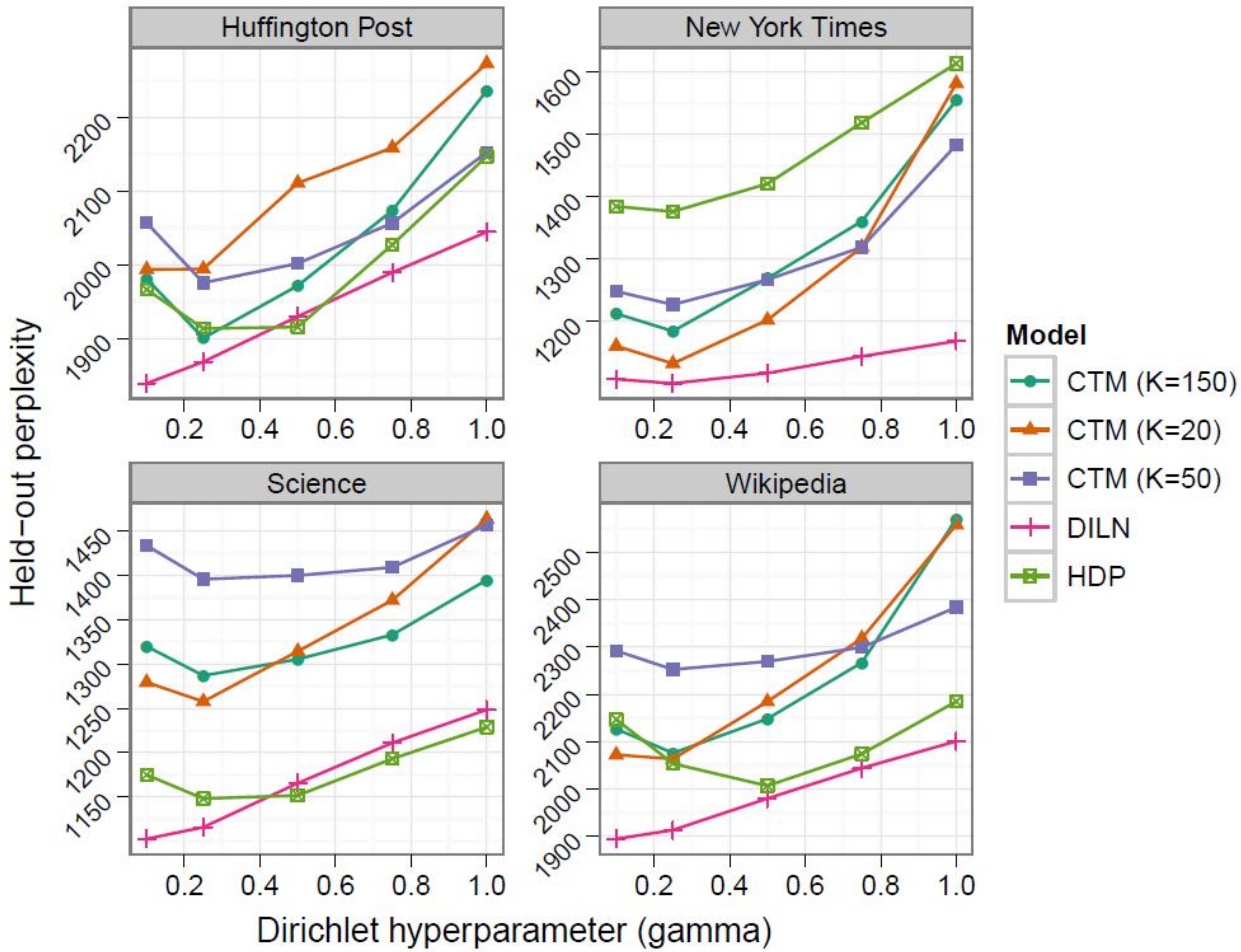
- If $w_k = 0$, this is a new inference algorithm for HDPs.

Experiments

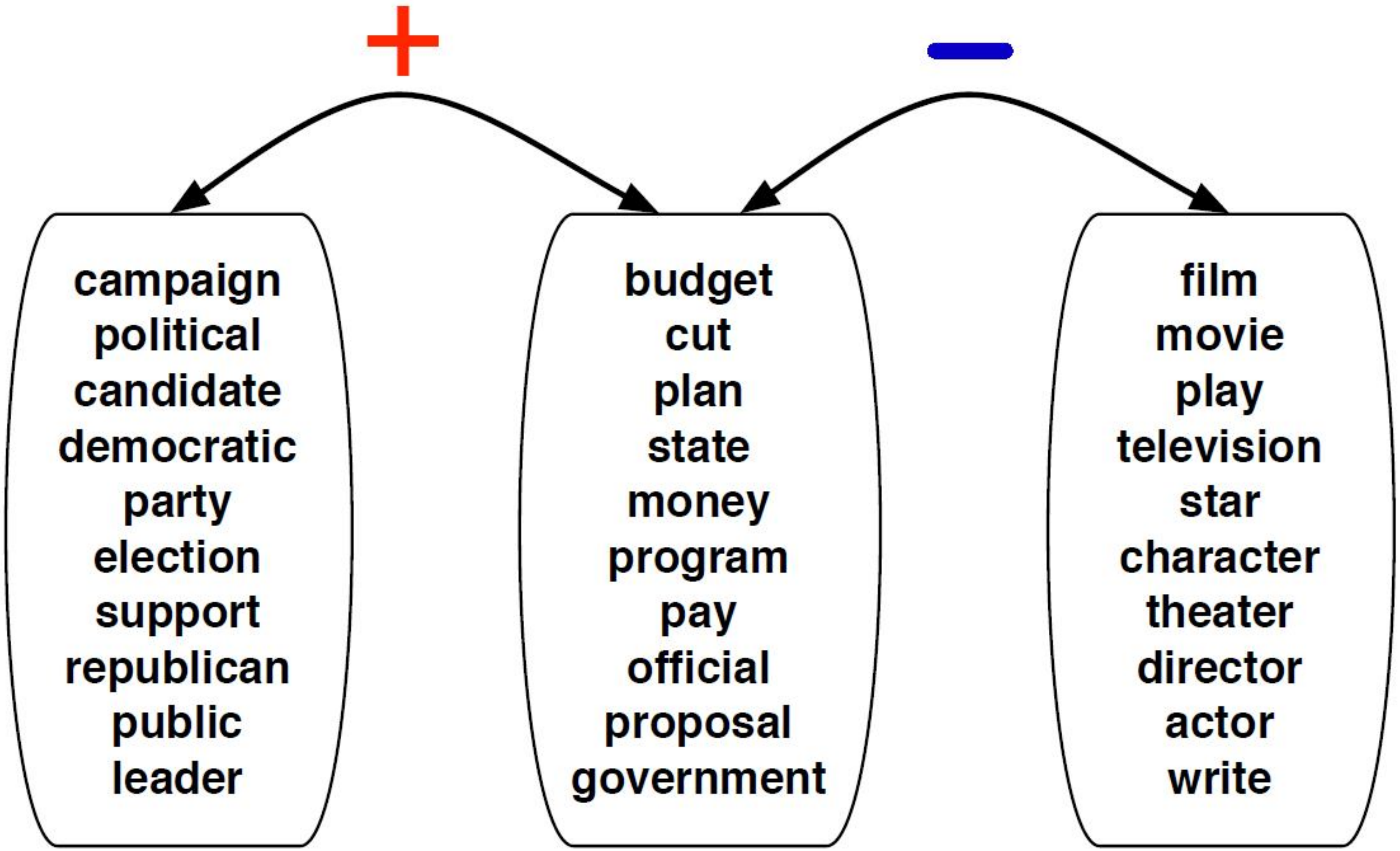
- Four text corpora: the Huffington Post, the New York Times, Science and Wikipedia, compared with HDP and CTM
- Partition a test document into two halves. Learn document-specific parameters on one half and predict the other half.
- then calculate the per-word perplexity

$$\text{perplexity} = \exp\left\{\frac{-\ln p(X_{\text{half2}} | X_{\text{half1}})}{N}\right\}.$$

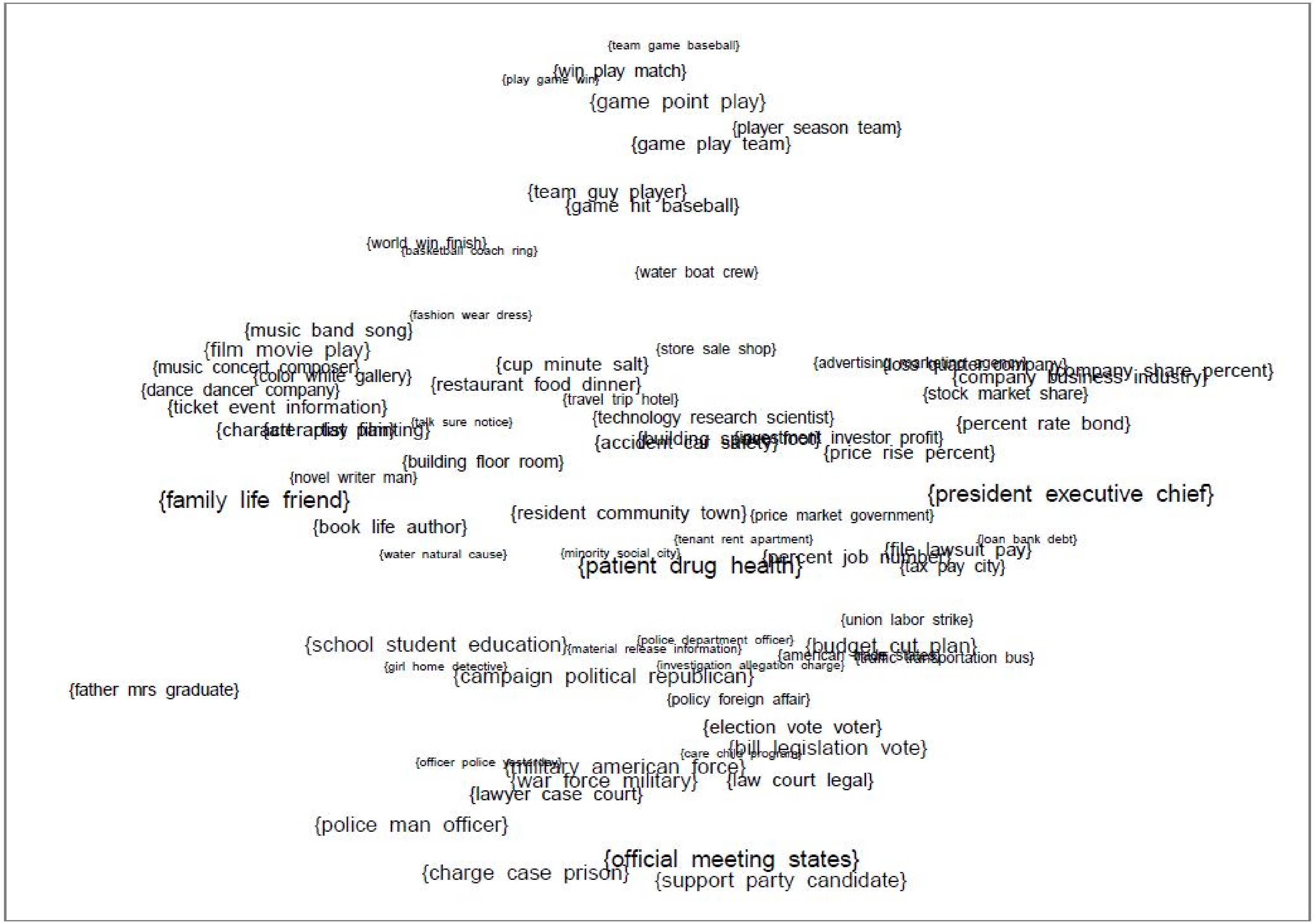
Experiments



Experiments



Experiments



Question and comment