Scalable inference of topic evolution via models for latent geometric structures

[1,2]Mikhail Yurochkin (mikhail.yurochkin@ibm.com), [3]Zhiwei Fan (zhiwei@cs.wisc.edu), [1] Aritra Guha (aritra@umich.edu),
[3] Paraschos Koutris (paris@cs.wisc.edu), [1] XuanLong Nguyen (xuanlong@umich.edu)
[1] *Department of Statistics, University of Michigan ,[2] IBM Research AI ,[3]Department of Computer Science, University of Wisconsin-Madison*

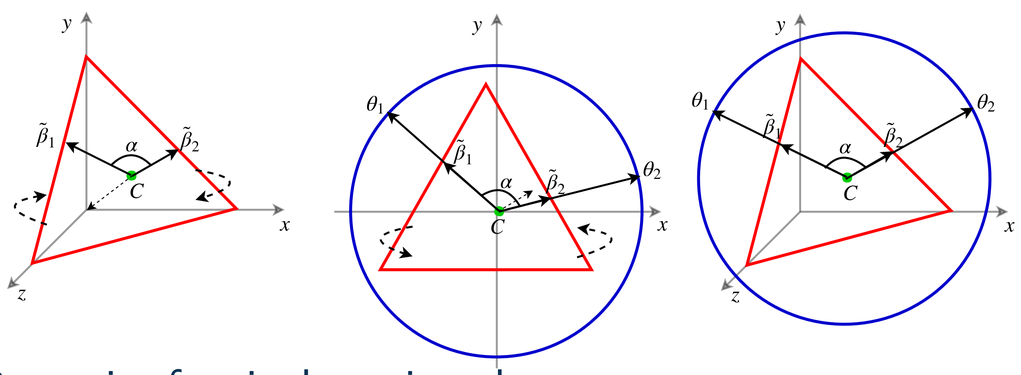
Overview

- series of Bayesian nonparametric models in increasing levels of complexity :
  - simpler model: topic polytope evolving over time
  - full model: temporal dynamics of topic polytope collection from multiple corpora
- scalable approximate inference algorithms suitable for online and distributed settings via the use of one-pass MAP estimates

Introduction

- The Dynamic Topic Models (DTM) [Blei and Lafferty, 2006]:
  - lack of scalability
  - inefficient joint modeling at each time point and topic evolution over time
- solution: decoupling the two phases of inference.

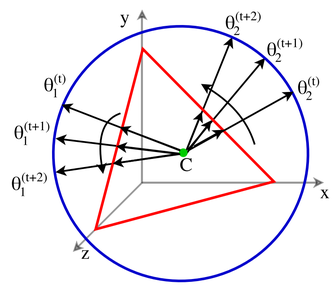
Isometric embedding of topic polytope on sphere



Dynamics for single topic polytope

Available metadata: $\{v_k^{(t)}\}_{t,k}$, topic estimates at each time t

$$
\begin{aligned}
Q &= \sum q_i \delta_{\theta_i} | \gamma_0, H \sim \text{BP}(\gamma_0, H) \\
\theta_i &:= \{\theta_i^{(t)}\}_{t=1}^T \sim H \\
\theta_i^{(t)} | \theta_i^{(t-1)} &\sim \text{vMF}(\theta_i^{(t-1)}, \tau_0) \text{ for } t \leq T, \\
\theta_i^{(0)} &\sim \text{vMF}(\cdot, 0) - \text{uniform on } \mathbb{S}^{V-2} \\
\mathcal{T}^{(t)} &:= \sum_i b_i^{(t)} \delta_{\theta_i^{(t)}}, \; b_i^{(t)} | q_i \sim \text{Bern}(q_i), \\
v_k^{(t)} | \mathcal{T}^{(t)} &\sim \text{vMF}(\mathcal{T}_k^{(t)}, \tau_1) \text{ for } k = 1, \ldots, K
\end{aligned}
$$



Streaming dynamic matching problem

- assign topics, $v_k^{(t)}$ at stage t to previously discovered topics, $\theta_i^{(t-1)}$ or attribute to new topics.
- Cost function for assigning topics (L_t number of topics at stage t, $m_i^{(t)}$ topic occupancy upto stage t)

$$
C_{ik}^{(t)} = \begin{cases} \|\tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}\|_2 - \tau_0 + \log \frac{m_i^{(t-1)}}{t - m_i^{(t-1)}}, & \text{if } i \text{ is a previous topic} \\ \tau_1 + \log \frac{\gamma_0}{t} - \log(i - L_{t-1}), & \text{if } i \text{ is a new topic} \end{cases}
$$

- Solution:

$$
\begin{cases} \frac{\tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}}{\|\tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}\|_2}, & \text{if new topic } k \text{ is assigned to previously discovered topic } i \\ v_k^{(t)}, & \text{if topic } k \text{ is a new topic} \\ \theta_i^{(t-1)} & \text{if topic is dormant at } t \end{cases}
$$

Topics learned by SDM algorithm: EJC data

- The Early Journal Content dataset years $1665 - 1922$, aggregated to single time-point for SDM
- $400k$ scientific articles, vocabulary 4516 words.

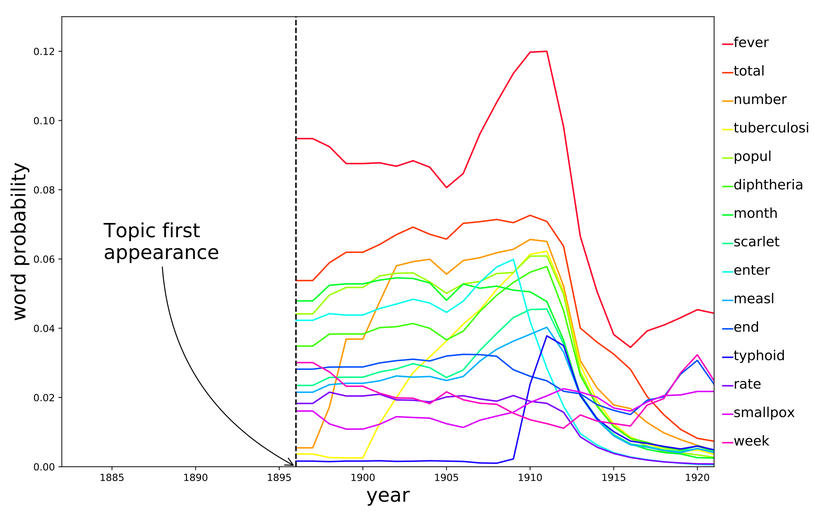


Figure 1: SDM *Epidemics*: evolution of top 15 words

Distributed Matching

- Groups $j = 1, \ldots, J$

$$
\begin{aligned}
Q &= \sum q_i \delta_{\theta_i} | \gamma_0, H \sim \text{BP}(\gamma_0, H) \\
\mathcal{T}_j | Q &\sim \text{BeP}(Q), \text{ then } \mathcal{T}_j := \textstyle\sum_{i=1} b_{ji} \delta_{\theta_i}, \text{ where } b_{ji} | q_i \sim \text{Bern}(q_i), \; \forall i \\
v_{jk} | \mathcal{T}_j &\sim \text{vMF}(\mathcal{T}_{jk}, \tau_1) \text{ for } k = 1, \ldots, K_j, \text{ where } K_j := \text{card}(\mathcal{T}_j)
\end{aligned}
$$

- Cost of topic assignment

$$
\begin{cases} \tau_1 \|v_{jk} + \sum_{-j,i,k} B_{-jik} v_{-jk}\|_2 - \tau_1 \| \sum_{-j,i,k} B_{-jik} v_{-jk}\|_2 + \log \frac{m_{-ji}}{J - m_{-ji}}, & \text{if } i \text{ is old topic} \\ \tau_1 + \log \frac{\gamma_0}{J} - \log(i - L_{-j}), & \text{if } i \text{ is new topic,} \end{cases}
$$

$-j$: groups excluding group j, L_{-j} : number of global topics before group j, m_{-ji} : number of i^{th} topic occurence before group j, B_{-jik} : topic assignment for k^{th} topic in group j to i^{th} global topic.

- Topic estimates $\theta_i = \frac{\sum_{j,k} B_{jik} v_{jk}}{\| \sum_{j,k} B_{jik} v_{jk} \|_2}$.

Topic 9 in *The Virginia Law Register*
court case law state ani suprem circuit jurisdict opinion said judg order right appeal time unit judgment action counti question

Topic 54 in The North American Review
court law ani judg case justic time trial onli state juri befor new unit suprem gener public constitut power mani

Global Topic 61
court law case state ani unit time act new onli suprem justic gener par question power befor right jurisdict shall

Topic 16 in *The Yale Law Journal*
court case law state jurisdict suprem unit decis ani act question defend rule constitut power new right journal befor action

Topic 18 in *Columbia Law Review*
court case law state jurisdict rule decis ani right new question act statut review onli unit feder constitut gener power

Figure 2: DM *Law*: matched topics from journals

Streaming dynamic distributed matching

- Global topic estimates $\theta_i^{(t)} = \frac{\tau_1 \sum_{j,k} B_{jik}^{(t)} v_{jk}^{(t)} + \tau_0 \theta_i^{(t-1)}}{\|\tau_1 \sum_{j,k} B_{jik}^{(t)} v_{jk}^{(t)} + \tau_0 \theta_i^{(t-1)}\|_2}$.
- $B^{(t)}$ is assignment matrix at time t.
- At t, use CoSAC for noisy topic estimates of each group in parallel.

Performance Comparison

	Perplexity	Time	Topics	Cores	Perplexity	Time	Topics	Cores
SDM	**1179**	22min	125	1	1254	2.4hours	182	1
DM	1361	5min	125	20	1260	**15min**	182	20
SDDM	1241	**2.3min**	103	20	**1201**	20min	238	20
DTM	1194	56hours	100	1	NA	¿72hours	100	1
SVB	1840	3hours	100	20	1219	29.5hours	100	20
CoSAC	1191	51min	132	1	1227	4.4hours	173	1

Table 1: Modeling topics of EJC —— Modeling Wikipedia articles

Choice of parameters

- τ_0 controls rate of topic dynamics of the SDM and SDDM, where smaller values imply higher dynamics rate
- Parameter τ_1 controls variance of local topics, when this variance is small, the model will tend to identify local topics as new global topics more often
- γ_0 affects the probability of new topic discovery

SDM parameters

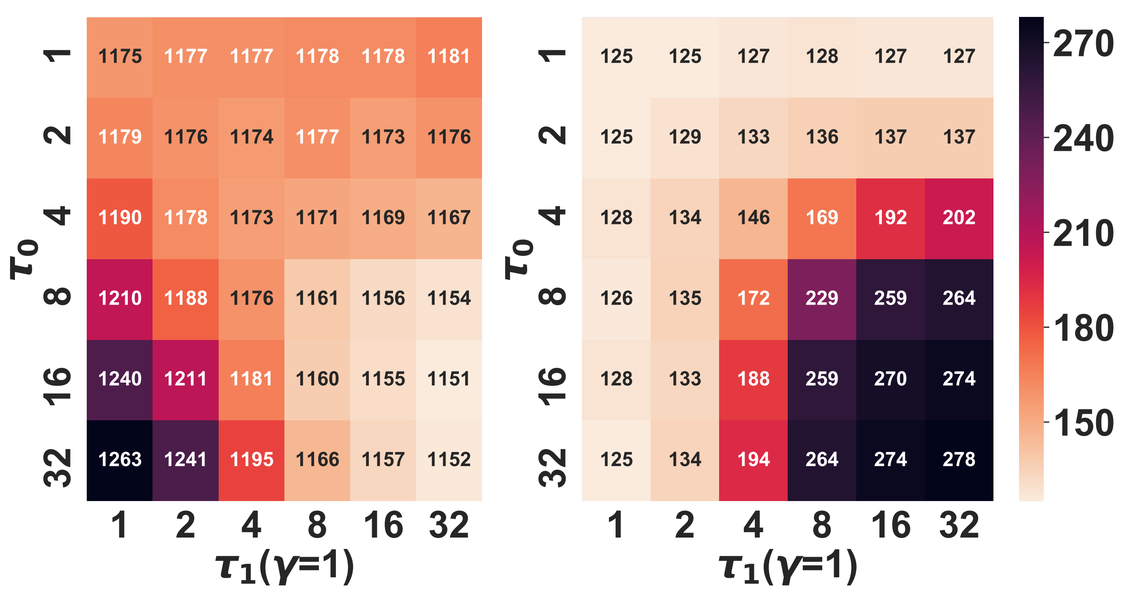


Figure 3: EJC perplexity

Figure 4: # of topics in EJC