# 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*

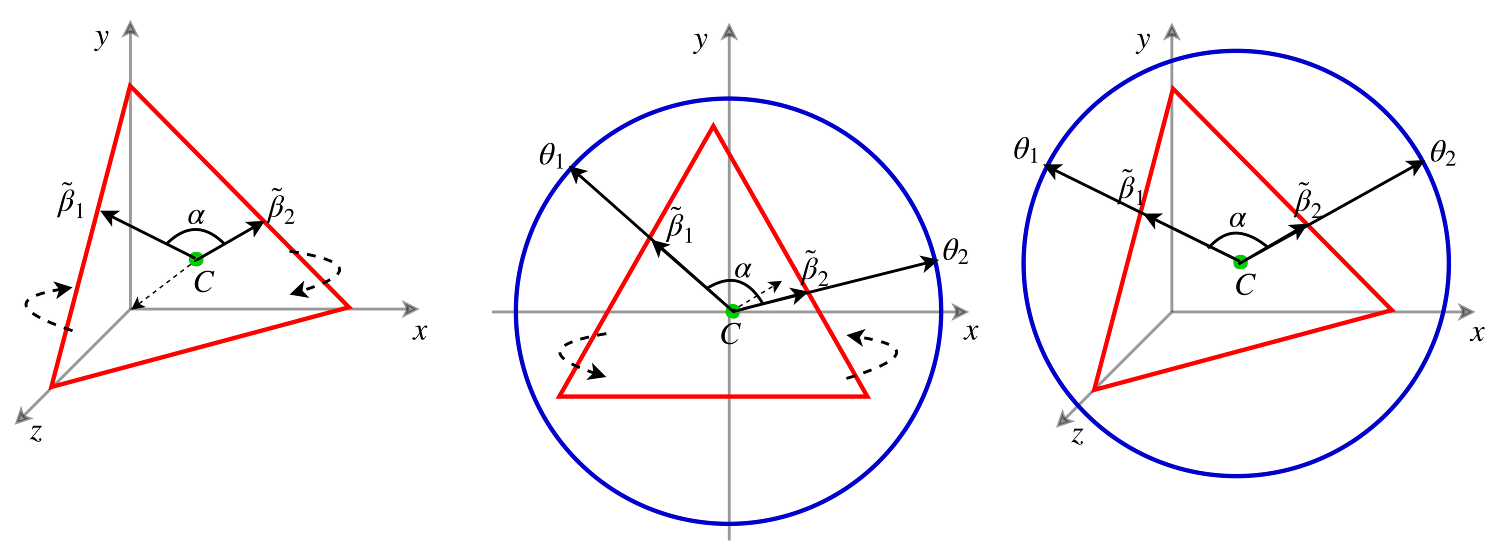
## Overview

- 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

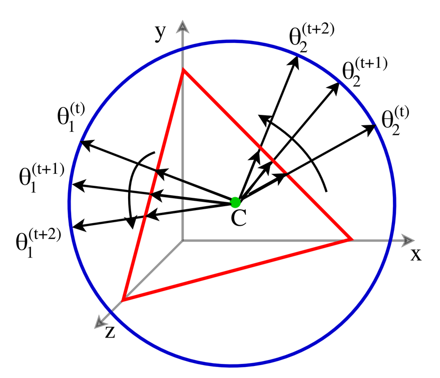
- The Dynamic Topic Models (DTM) [Blei and Lafferty, 2006]:
  - lack of scalability
  - inefficient joint modeling at each time point and topic evolution over time
- solution: decoupling the two phases of inference.

## Isometric embedding of topic polytope on sphere



## Dynamics for single topic polytope

$$
\begin{aligned}
Q &= \sum q_i \delta_{\theta_i} | \gamma_0, H \sim \mathrm{BP}(\gamma_0, H)\\
\theta_i &:= \{\theta_i^{(t)}\}_{t=1}^T \sim H\\
\theta_i^{(t)} | \theta_i^{(t-1)} &\sim \mathrm{vMF}(\theta_i^{(t-1)}, \tau_0) \text{ for } t = 1, \dots, T,\\
\theta_i^{(0)} &\sim \mathrm{vMF}(\cdot, 0) - \text{uniform on } \mathbb{S}^{V-2}
\end{aligned}
$$



## Posterior Contraction for mixing measures

**Wasserstein Metric:**

- **Definition:** $G, G_0$ probability measures. Coupling $\kappa$ is a joint distribution inducing marginals $G, G_0$.
- $r$-order Wasserstein distance:

$$
W_r(G, G_0) = \left[ \inf_\kappa \int \|\theta_1 - \theta_2\|^r d\kappa(\theta_1, \theta_2) \right]^{1/r}. \tag{1}
$$

- Interpretation: Cost of mass transfer between $G$ and $G_0$ if unit mass transfer between locations $x$ and $y$ is proportional to $\|x - y\|$.
- $W_r(G, G_0) \approx \omega_n \implies$ parameters converge at rate $\omega_n$, corresponding masses converge at rate $\omega_n^r$.

## Merge-Truncate-Merge (MTM) Algorithm

Suppose $W_r(G, G_0) = o_P(\omega_n)$, obtain the processed sample, $\tilde{G}$ as follows:

images/Initial.jpg images/Merge_1.jpg

images/Truncate.jpg images/final.jpg

(M1.) merge small atoms which are close to larger atoms (based on wt. $\omega_n$).

(T1.) identify smaller atoms further away from any merged larger clusters (based on distance $\omega_n^r$).

(M2.) identify midsize atoms which are closer to each other but less than cut-off weight and merge them with the closest large cluster (based on a combination of wt. and distance).

(T2.) truncate the smaller atoms identified before.

(R.) return $\tilde{G}$ =resultant mixing measure and $\tilde{k}$ = number of components of $\tilde{G}$.

**Theorem 1.** *For $G \sim \Pi = DPM$, $\Pi(\tilde{k} \neq k_0 | X_1, \dots, X_n) \to 0$ a.e., $p^n_{G_0}$. Also, $\Pi(W_r(\tilde{G}, G_0) \gtrsim \omega_n | X_1, \dots, X_n) \to 0$.*

## References


- A. Guha, N. Ho, and XL. Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling.Arxiv preprint Arxiv: 1901.05078, 2019.


## MTM Simulations

Gaussian: $\omega_n \sim (\frac{\log(\log(n))}{\log(n)})^{1/2}$.

images/500_normal.png images/500_large_spacing.png

## Strong identifiability

- classical 0-order identifiability:linear independence of functional kernel at different parameter values.
- $r$-order identifiability: linear independence upto $r$-order derivatives.
- Gaussian location-scale family is 1-order identifiable Student's t-distribution is 2-order identifiable.
- Under 1-order identifiability with exact fit, $h(p_G, p_{G_0}) \gtrsim W_1(G, G_0)$. **[Ho & Nguyen, 2016]**

## Integral Lipschitz property

$f$ satisfies,

- $\sum_{|\kappa|=r} \left| \left( \frac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta_1) - \frac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta_2) \right) \gamma^\kappa \right| \leq C(x) \|\theta_1 - \theta_2\|^\delta \|\gamma\|^r$ as a function of $\theta$ , where $C(\cdot)$ is integrable with respect to the Lebesgue measure.

## Adaptive Posterior Contraction

**Assumptions**

(P.1) $\Theta$ is compact and $f$ is 1-order identifiable and admits 1-order Lipschitz property.

(P.2) There exists $\epsilon_0 > 0$ such that $\int (p_{G_0}(x))^2 / p_G(x) d\mu(x) \leq M(\epsilon_0)$ as long as $W_1(G, G_0) \leq \epsilon_0$ for any $G \in \mathcal{O}_{k_0}$ where $M(\epsilon_0)$ depends only on $\epsilon_0$, $G_0$, and $\Theta$.

**Theorem 2.** *If $G_0$ has finite support and (P.1) and (P.2) hold,*

*(a) $\Pi(K \neq k_0 | X_1, \dots, X_n) \to 0$ a.s. $p_{G_0}$.*

*(b) $\Pi\left( G : W_1(G, G_0) \gtrsim \frac{(\log n)^{1/2}}{n^{1/2}} | X_1, \dots, X_n \right) \to 0$ in $p_{G_0}$probability.*

## Posterior Contraction: Misspecified regime

**Assumptions**

- Fit data with $p_G = G * f$ ($G$ and/or $f$ misspecified), support of $G$ in compact $\Theta$.
- $\exists$ $G_*$ with $k_*$ components s.t. $p_{G_*}$ minimizes KL-diveregence w.r.t. $p_{G_0, f_0} = G_0 * f_0$.
- The support of $G_0$, namely, $\mathrm{supp}(G_0)$ is a bounded subset of $\mathbb{R}^d$. Moreover, there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$
\sup_{x \in \mathbb{R}^d, \theta \in \Theta, \theta_0 \in \mathrm{supp}(G_0)} \frac{f(x|\theta)}{f_0(x|\theta_0)} 1_{\|x\|_2 \leq R} \leq C_1 \exp(C_0 R^\alpha).
$$

(a) $k_* = \infty$ for Laplace location family:

**Theorem 3.**

$$
\Pi\left( W_2(G, G_*) \lesssim exp\left( -\frac{m\tau(\alpha)}{2} \left( \frac{\log n - \log\log n}{2(d+2)} \right)^{1/\alpha} \right) \middle| X_1, \dots, X_n \right) \to 1
$$

*in $p_{G_0, f_0}$-probability for any positive constant $m < 4/(4 + 5d)$.*

*$\tau(\alpha) := \frac{\sqrt{2/(\lambda_{\max})}}{(\sqrt{2/(\lambda_{\min})} + \sqrt{2/(\lambda_{\max})} + C_0)^{1/\alpha}}$, $\lambda_{min}$ and $\lambda_{max}$ representing the minimum and maximum eigenvalues of the covariance matrix.*

(b) $k_* = \infty$ for Gaussian location family (with $\alpha = 2$):

**Theorem 4.**

$$
\Pi\left( W_2(G, G_*) \lesssim \left( \frac{\log\log n}{\log n} \right)^{1/2} \middle| X_1, \dots, X_n \right) \to 1
$$

*in $p_{G_0, f_0}$-probability.*

## Summary

- when number of components essential (clustering, topic models) use parametric prior.
- when only top few clusters are of interest: use non-parametric prior with post-processing.
- "one size does not fit all"- for clustering (parameter estimation) use heavy tailed kernels, for density estimation use smooth kernels -( Laplace >> Gaussian for parameter estimation).