Conic Scan-and-Cover algorithms for nonparametric topic modeling

Mikhail Yurochkin (moonfolk@umich.edu), Aritra Guha (aritra@umich.edu), XuanLong Nguyen (xuanlong@umich.edu)
Department of Statistics, University of Michigan

Introduction

- Parametric Topic modeling: Speed and accuracy at crossroads.
- Geometric Dirichlet Means (GDM) algorithm,[1] takes care of both, but needs assumption of number of topics.
- This work proposes algorithm for unknown number of topics.
- Accuracy comparable to Gibbs sampler at much faster run-time.

Geometric topic modeling

- Topic β_k : point in a $V-1$ dimensional probability simplex Δ^{V-1}. $B := \text{conv}(\beta_1, \ldots, \beta_K)$: K dimensional topic polytope.
- Each document corresponds to a point $p_m := (p_{m1}, \ldots, p_{mV})$ inside the polytope B.
- Topic proportions θ_m : barycentric coordinate vector of document m.

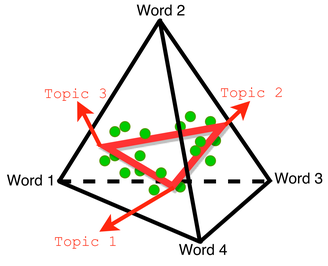

Figure 1: Topic and vocabulary polytopes

- Each $p_m = \sum_k \beta_k \theta_{mk}, m = 1, \ldots, M$ is a vector of cartesian coordinates of m-th document's multinomial probabilities.
- Each document $w_m \sim Multinomial(p_m, N_m)$, $N_m \in \mathbb{N}$: number of words in document m.
- The model is equivalent to the LDA[2] with individual words to topic label assignments marginalized out.

Conic Scan-and-Cover (CoSAC)

- Iteratively pick the farthest point from the center estimate. Let its direction from the center be v.
- Construct a cone around v, for some angle with cosine distance ω.
- Remove all the data in this cone.

Proposition 1. *For choice of ω in some range (ω_1, ω_2), a complete coverage of the topic polytope is achievable with cones only, such that each cone contains exactly one topic vertex.*

- Stop when $\forall\ m\ , \|\tilde{p}_m\|_2 < \mathcal{R}$.

Proposition 2. *For choice of ω in some range (ω_3, ω_2), we can choose a sphere of radius R along with cones, with each cone containing exactly one topic vertex, such that $\omega_3 < \omega_1$.*

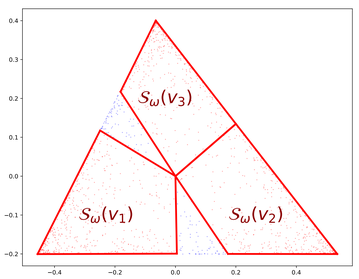

Figure 2: Incomplete coverage using cones

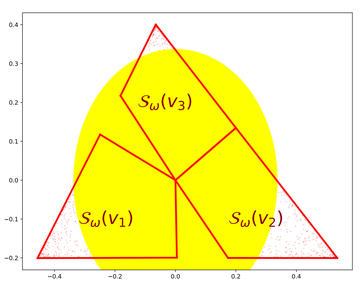

Figure 3: Coverage using cones and sphere

- If cone contains less than λ portion of data points, ignore cone.

Proposition 3. *For $\omega \in (\omega_\lambda, \omega_4)$, each cone around a topic contains at least λ proportion of documents.*

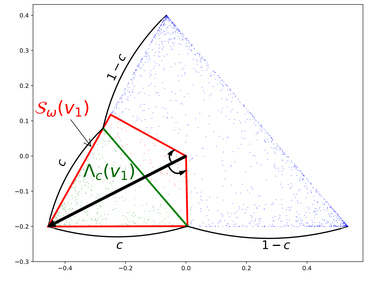

Figure 4: Cap $\Lambda_c(v_1)$ and cone $\mathcal{S}_\omega(v_1)$.

Theorem 1. *Under the LDA model, with the word to topic assignments marginalized out, the minimum matching distance between estimated and true topics $\to 0$ almost surely.The estimated number of topics also equals the true number of topics almost surely.*

We found the choices $\omega = 0.6$, $\lambda = 0.001$ and $\mathcal{R}$ to be median of $\{\|\tilde{p}_1\|_2, \ldots, \|\tilde{p}_M\|_2\}$ to be robust in practice and agreeing with our theoretical results.

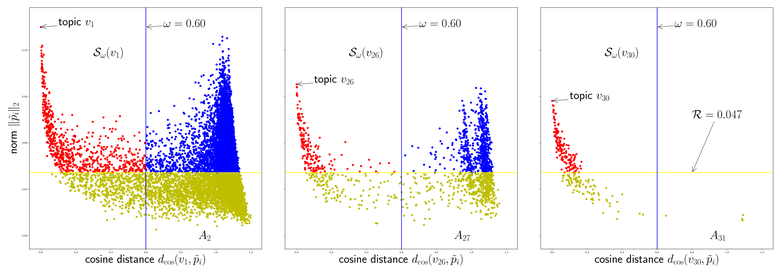

Figure 5: Iterations 1, 26, 30 of the Algorithm. Red: documents in the cone $\mathcal{S}_\omega(v_k)$; Blue: documents in the active set A_{k+1} for next iteration. Yellow: documents $\|\tilde{p}_m\|_2 < \mathcal{R}$.

Mean-shifting Algorithm

- Farthest point is no longer a robust topic estimate when number of words per document is small.
- Use Mean-shifting to shift topic estimate to high-density region. For each iteration,
  - Pick the farthest point, from the polytope center estimate.
  - Construct a cone corresponding to some ω in the direction of this point.
  - Compute the weighted mean of all documents in this cone.
  - Take this as the new direction of topic estimate and repeat the above steps until convergence.
  - Once converged, choose the farthest projected point in the converged direction as the topic estimate.
- Mean-shifting leads to reduced variance of topic direction estimate due to averaging over data residing in the cone.

Experimental Results

CoSAC for documents and cscRecoverKL (Recover KL with anchor words found using CoSAC approach) without access to true K, versus popular parametric topic modeling approaches (trained with true K): Stochastic Variational Inference (SVI), Collapsed Gibbs sampler, RecoverKL and GDM.

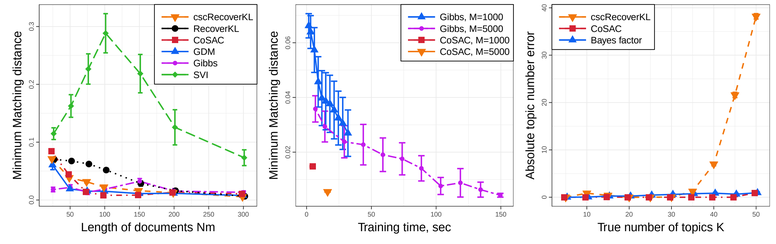

Figure 6: Minimum matching Euclidean distance for (a) varying length of documents, (b) Running times for varying corpora size; (c) Estimation of number of topics.

CoSAC on NYTimes articles. 130,000 documents and 5320 unique words.

	cscRecoverKL	HDP	Gibbs LDA	Gibbs CoSAC
K	27	221 ± 5	80	159
Perplexity	2603	1477 ± 1.6	1520 ± 1.5	1568
Time	37 min	35 hours	5.3 hours	19 min

Cooking	*Cloning*	*Antitrust*	*LGBT*	*Voting*
cup	cell	zzz_microsoft	gay	ballot
minutes	stem	window	lesbian	zzz_al_gore
tablespoon	research	company	right	election
add	human	software	sex	votes
teaspoon	scientist	case	marriage	recount
pepper	cloning	system	group	zzz_florida
oil	patient	operating	couples	court
sugar	disease	computer	sexual	vote
butter	phones	antitrust	partner	voter
pan	researcher	court	issue	count

References

1. Yurochkin & Nguyen, Geometric Dirichlet means for algorithm for topic inference, NIPS,2016.
2. Blei, Ng & Jordan, Latent Dirichlet Allocation,J. Mach. Learn. Res.,2003.