Learning Topic Models: Identifiability and Estimation

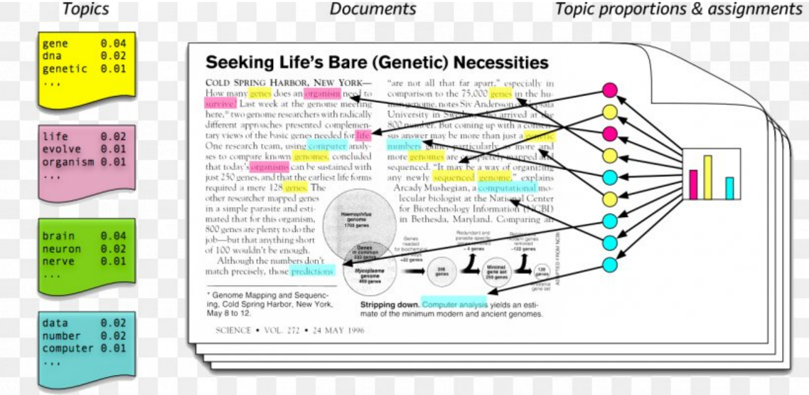
06/07/2023

Yinyin Chen, Shishuang He, Yun Yang, & Feng Liang

Department of Statistics
University of Illinois at Urbana-Champaign

1. Motivating Examples

Topic Models



Words in the i-th document:

$$X_1^{(i)}, X_2^{(i)}, \ldots, iid \sim \text{Multi}(\mathbf{u}_i)$$

$$\mathbf{u}_i = w_{i1}\mathbf{C}_1 + w_{i2}\mathbf{C}_2 + \cdots w_{iK}\mathbf{C}_K$$

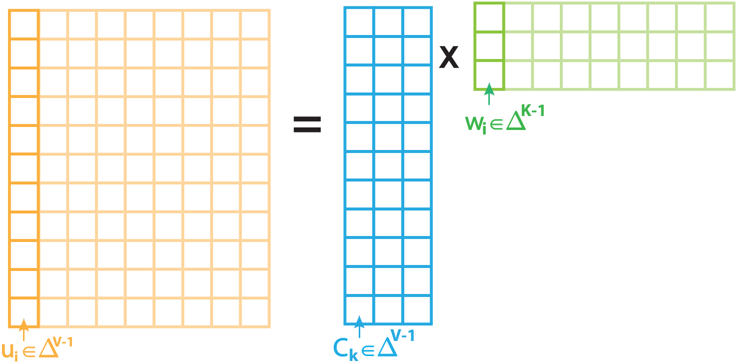
$\downarrow$ word distribution of a topic

$w_{i1}, \ldots, w_{iK}$: mixing weights

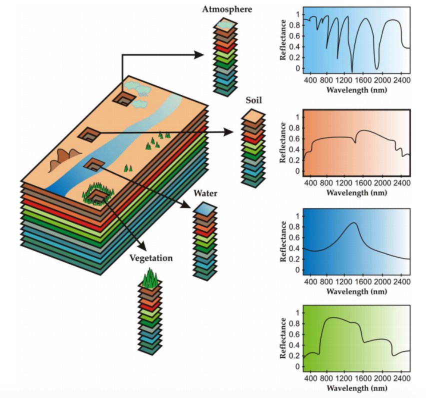
Matrix Factorization

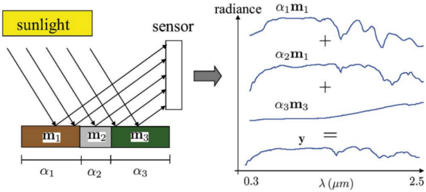
Modeling d documents of vocabulary size V by,

$$\underbrace{\mathbf{U}_{V\times d}}_{\text{true word frequency}} = \overbrace{\mathbf{C}_{V\times K}}^{\text{topic matrix}} \underbrace{\mathbf{W}_{K\times d}}_{\text{mixing weights}}.$$



Hyperspectral Unmixing





Observed spectral at pixel i:

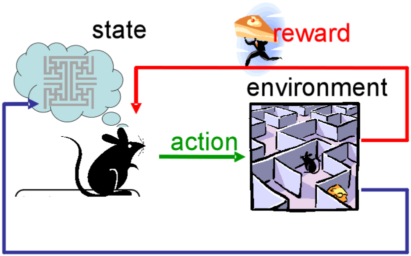
$$X_i \sim N(\mathbf{u_i}, \boldsymbol{\Sigma})$$

$$\mathbf{u}_i = w_{i1}\mathbf{C}_1 + w_{i2}\mathbf{C}_2 + \cdots w_{iK}\mathbf{C}_K$$

↓ spectral of a pure material

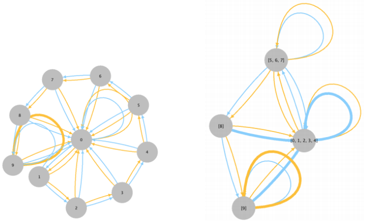
$w_{i1}, \ldots, w_{iK}$: mixing weights

State Aggregation in Reinforcement Learning



Computation of the reward function of each state and action pairs could fail with too many states. Therefore, it is crucial to aggregate the original state space into a more compact representation.

Most work in RL focuses on Markov Decision Processes (MDPs), in which an agent is assumed to move between different states following a Markov process.

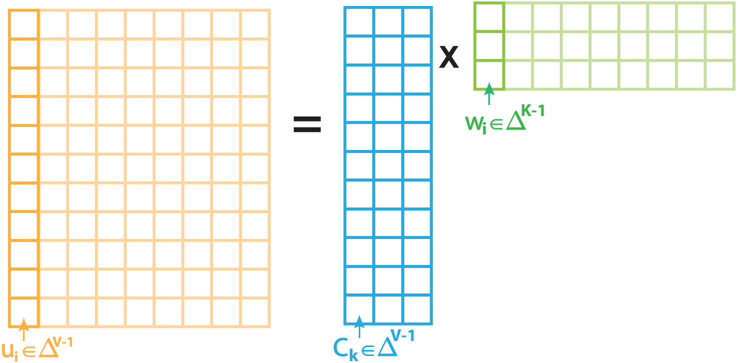


Soft State Aggregation

Soft state aggregation with K meta-states $\Longrightarrow$ topic model

[Singh et al., 1995]

$$\underbrace{\mathbb{P}(X_{t+1} = i | X_t = j)}_{\text{transition prob}} = \sum_{l=1}^{K} \overbrace{\mathbb{P}(X_{t+1} = i | Z_t = l)}^{\text{disaggregation}} \underbrace{\mathbb{P}(Z_t = l | X_t = j)}_{\text{aggregation}}$$



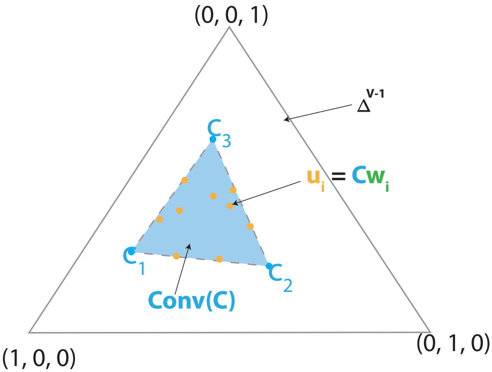
2. Overview

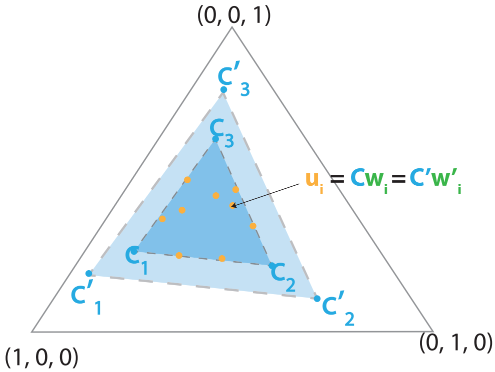
Overview

- Goal: given data from a model parameterized by $\mathbf{U} = \mathbf{C}\mathbf{W}$, recover the latent structure, i.e., the topic matrix $\mathbf{C}$.
- An obstacle to a rigorous analysis of estimation of $\mathbf{C}$: non-identifiability.
-
-

Geometric View

A toy example with $V = 3$ words and $K = 3$ topics.





One can find another pair of $(\mathbf{C}', \mathbf{W}')$ satisfying $\mathbf{U} = \mathbf{C}'\mathbf{W}' = \mathbf{C}\mathbf{W}$.

Overview

- Goal: given data from a model parameterized by $\mathbf{U} = \mathbf{C}\mathbf{W}$, recover the latent structure, i.e., the topic matrix $\mathbf{C}$.

- An obstacle to a rigorous analysis of estimation of $\mathbf{C}$: non-identifiability.

- Question 1: under what conditions, a topic model parameterized by $(\mathbf{C}, \mathbf{W})$ is identifiable (up to permutation)?

- Question 2: For an identifiable topic model, can we provide an estimator of $\mathbf{C}$ whose finite-sample error leads to the desired rate of convergence?

3. Prior Work

Literature Review

- The Bayesian approach
- The anchor-word approach
- The volume minimization approach

The Bayesian Approach

- In the Bayesian setting, the mixing weights, columns of $\mathbf{W}$, are assumed to be stochastically generated from a **known** distribution with full support over the simplex.
- Identifiability can be ensured under very mild conditions such as $\mathbf{C}$ being of full rank.
- Under this assumption, estimation accuracy has been studied in [Nguyen, 2015, Tang et al., 2014, Anandkumar et al., 2012, Anandkumar et al., 2014, Wang, 2019].
- Focus of this talk is the non-Bayesian setting, which is much more challenging.

The Separability Condition

[Donoho and Stodden, 2004, Arora et al., 2012, Recht et al., 2012, Ge and Zou, 2015, Ke and Wang, 2017] addressed identifiability via variants of the so-called **Separability Condition**.

- On rows of $\mathbf{C}$: every topic has an **“anchor word”** that only appears in that particular topic.

C

0.1	0	0
0	0.3	0
0	0	0.2

W

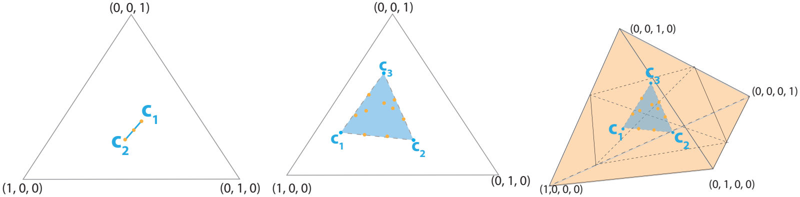
1	0	0						
0	1	0						
0	0	1						

- On columns of $\mathbf{W}$: vertices of Conv$(\mathbf{C})$ must be data points.

The Volume Minimization Approach (I)

Natural to focus on $\mathbf{C}$ whose convex hull has the smallest volume; i.e., Conv$(\mathbf{C})$ circumscribes the data as compactly as possible.

[Craig, 1994, Nascimento and Dias, 2005, Miao and Qi, 2007, Fu et al., 2015, Jang and Hero, 2019]



Definition 1 (Identifiability under Volume Minimization)

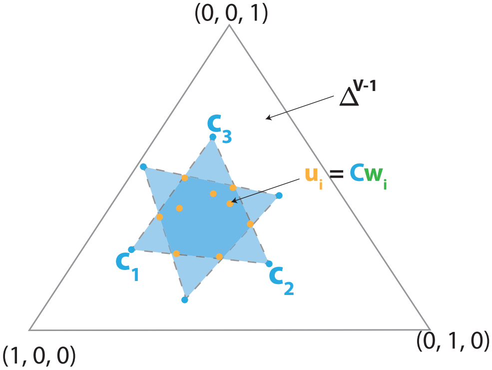
$(\mathbf{C}, \mathbf{W})$ is **identifiable**, if for any other $(\mathbf{C}', \mathbf{W}')$

$$\mathbf{C}\mathbf{W} = \mathbf{C}'\mathbf{W}' \text{ and } |\text{Conv}(\mathbf{C}')| \leq |\text{Conv}(\mathbf{C})|$$

if and only if $\mathbf{C}' = \mathbf{C}\mathbf{\Pi}$ for some permutation matrix $\mathbf{\Pi}$.

The Volume Minimization Approach (II)

Note that the minimum volume constraint alone still does not guarantee uniqueness.



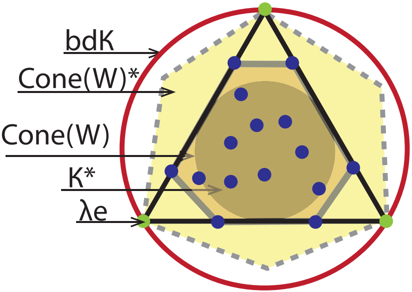
4. Identifiability

Sufficient Conditions for Identifiability

Theorem 2 (Identifiability)

If $\mathbf{W} = [\mathbf{w}_1, \cdots \mathbf{w}_d]$ is Sufficiently Scattered (SS) and $Rank(\mathbf{C}) = K$, then $(\mathbf{C}, \mathbf{W})$ is identifiable.

Sufficiently Scattered (I)



(Left: projection on simplex Δ^2.)

$$\begin{aligned}\mathcal{K} &= \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 \leq 1\} \\ bd\mathcal{K} &= \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = 1\} \\ \mathcal{K}^* &= \text{dual cone of } \mathcal{K}\end{aligned}$$

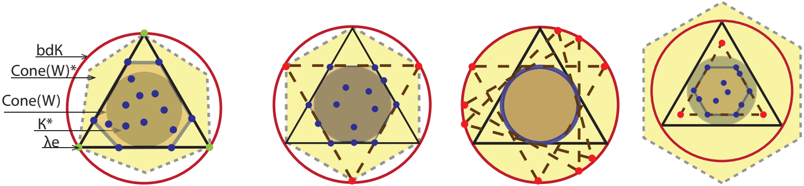
A useful fact of **dual cones**: $\mathcal{A}$ and $\bar{\mathcal{A}}$ are convex cones.

$$\mathcal{A} \subseteq \bar{\mathcal{A}} \Longleftrightarrow \bar{\mathcal{A}}^* \subseteq \mathcal{A}^*.$$

Sufficiently Scattered (II)

The matrix $\mathbf{W}$ is SS if it satisfies that:

1. Conv$(\mathbf{W})^* \subseteq \mathcal{K}$ (or equivalently, $\mathcal{K}^* \subseteq$ Conv$(\mathbf{W})$)
2. Conv$(\mathbf{W})^* \bigcap bd\mathcal{K} = \{\lambda \boldsymbol{e}_j : \lambda \geq 0, j = 1, ..., K\}$, where $bd\mathcal{K}$ denotes the boundary of $\mathcal{K}$.



(a) SS (b) not SS (c) not SS (d) not SS

Proof Sketch

Suppose $\mathbf{C}\mathbf{W} = \bar{\mathbf{C}}\bar{\mathbf{W}}$. Then $\mathbf{C} = \bar{\mathbf{C}}\mathbf{B}$, where

$$\mathbf{B}_{K\times K} = \bar{\mathbf{W}}\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}.$$

Based on the definition of $\mathbf{B}$ and the minimal volume constraint, we can show that the columns of $\mathbf{B}$

1. are unit vectors (i.e., on $bd\mathcal{K}$);
2. belong to $\mathcal{C}^*$.

Due to the second condition of SS, columns of $\mathbf{B}$ must be $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_K)$, i.e., $\mathbf{B}$ is a permutation matrix.

Prior Work on SS

- The SS condition was first proposed in [Huang et al., 2016], and used to ensure identifiability along with determinant minimization on $\mathbf{W}\mathbf{W}^T$. Although the volume of Conv$(\mathbf{C})$ is not discussed in [Huang et al., 2016], the conditions there in fact lead to a topic matrix $\mathbf{C}$ with the maximal volume.

- [Jang and Hero, 2019] proved a result similar to ours for topic models with $V = K$ (vocabulary size being the same as the topic size). Their analysis is built on a formula of the volume of Conv$(\mathbf{C})$ that holds true only when $V = K$.

5. Estimation

How to Compute Volume?

Volume minimization has been widely used in many applications [Craig, 1994, Nascimento and Dias, 2005, Miao and Qi, 2007, Fu et al., 2015, Jang and Hero, 2019].

However, one challenge is that $|\text{Conv}(\mathbf{C})|$ does not take a simple form. It is often approximated by $\sqrt{\det(\mathbf{C}^T\mathbf{C})}$.

$$\sqrt{\det(\mathbf{C}^T\mathbf{C})} = (K-1)!h_{\mathbf{C}}|\text{Conv}(\mathbf{C})|$$

where $h_{\mathbf{C}}$ is the perpendicular distance between the origin and $\text{Conv}(\mathbf{C})$. $h_{\mathbf{C}}$ varies for different $\mathbf{C}$ if $V > K$.

The Proposed Estimator

- We can integrate out the nuisance parameters $\mathbf{w}_i$ with respect to some distribution and estimate $\mathbf{C}$ by maximizing the integrated likelihood.

- We propose to integrate out $\mathbf{w}_i$'s with respect to the uniform distribution over simplex Δ^{k-1}, which induces a uniform distribution on $\mathbf{u}_i = \mathbf{C}\mathbf{w}_i$ over Conv$(\mathbf{C})$.

- The integrated likelihood is given as follows:

$$F_{n\times d}(\mathbf{C};\mathbf{X}) = \prod_{i=1}^{d} \int_{\text{Conv}(\mathbf{C})} \frac{f_n(\mathbf{x}^{(i)}|\mathbf{u})}{|\text{Conv}(\mathbf{C})|} d\mathbf{u}, \tag{1}$$

and the MLE is defined to be $\hat{\mathbf{C}}_n = \arg\max_{\mathbf{C}} F_{n\times d}(\mathbf{C};\mathbf{X})$.

Volume Minimization and Uniform Prior

Consider the noiseless case (or the limiting case as $n \to \infty$),

indicator function

$$\lim_{n\to\infty} \prod_{i=1}^{d} \int_{\text{Conv}(\mathbf{C})} \frac{f_n(\mathbf{x}^{(i)}|\mathbf{u})}{|\text{Conv}(\mathbf{C})|} d\mathbf{u} = \frac{\overset{\downarrow}{\mathbb{1}}(\mathbf{u}_1, \ldots, \mathbf{u}_n \in \text{Conv}(\mathbf{C}))}{|\text{Conv}(\mathbf{C})|}.$$

Asymptotically, maximizing $F_{n\times d}(\mathbf{X}|\mathbf{C}) \Leftrightarrow$ minimizing $|\text{Conv}(\mathbf{C})|$.

- The uniform prior is only used to integrate out $\mathbf{w}$.
- In our theoretical analysis, we do NOT assume $\mathbf{w} \sim$ unif dist'n.

Finite Sample Accuracy

- Condition 1: $\mathbf{u}_1, \cdots, \mathbf{u}_d$ are strict inner points of Δ^{V-1}.
- Condition 2: exists a subset of s ($\geq 2K$) columns in $\mathbf{W}$ that is $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-SS with $\boldsymbol{\alpha} = C_1\sqrt{\frac{s\log d}{n}}$ and $\boldsymbol{\beta} \leq C_2\sqrt{\frac{\log d}{n}}$

$\downarrow$
perturbed version of SS

Theorem 3 (Finite sample error bound)

With high probability,

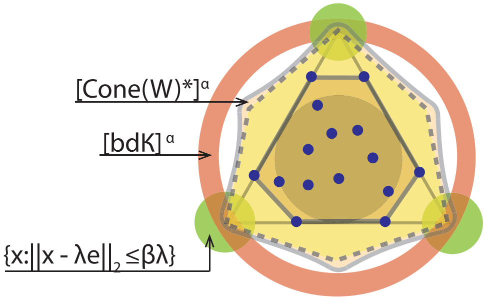
$$\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}) \leq D_2 C_\alpha \sqrt{\frac{s\log d}{n}},$$

where $\mathcal{D}(\hat{\mathbf{C}}_n, \mathbf{C}) = \arg\min_{\mathbf{\Pi}} \|\hat{\mathbf{C}}_n - \mathbf{C}\mathbf{\Pi}\|_2$.

$[\text{Conv}(\mathbf{W})^*]^\alpha$: enlarged cone

$[bd\mathcal{K}]^\alpha$: thickened boundary of $\mathcal{K}$

green balls : small balls around the corners



Comparison with Existing Theoretical Results

$$\sqrt{1/nd} \quad : \quad \text{[Arora et al., 2012]; [Ke and Wang, 2017]}$$

$$\sqrt{(\log d)/n} \quad : \quad \text{Ours; [Javadi and Montanari, 2019]}$$

Two-Stage interpretation of those algorithms:

- (1) Learn a projection onto a $(K-1)$-dim space containing $\mathbf{C}$.
  - $1/\sqrt{nd}$ error rates, since projection algorithms have an effective sample size nd via information pooling across all d documents.
- (2) Estimate $\mathbf{C}$.
  - With separability assumption, (2) becomes a searching procedure incurring less errors than (1).
  - Under volume minimization, (2) becomes a boundary detection procedure, a difficult problem known to have slower convergence rates [Goldenshluger and Tsybakov, 2004, Brunel et al., 2021].

6. Computation

Computation

Our proposed estimator is essentially the MLE from the LDA model [Blei et al., 2003] with a particular choice of prior on $\mathbf{W}$.

Therefore many algorithms developed for the LDA model can be used, such as MCMC, MCMC-EM, stochastic EM, and variational Bayes.

7. Summary

Summary

For a topic model parameterized by $(\mathbf{C}, \mathbf{W})$, we aim to address

- Question 1: under what conditions, the model is identifiable?

- Question 2: For an identifiable topic model, can we provide an estimator for $\mathbf{C}$ with the desired error rate?

For Question 1, we propose to resolve the non-identifiability issue by focusing on convex hulls of the smallest volume, and then provide a set of conditions to ensure identifiability. Our conditions are weaker than the ones from prior studies.

For Question 2, we propose an estimator of $\mathbf{C}$ based on an integrated likelihood and establish the error rate of the proposed estimator, which consequently implies asymptotic consistency of the proposed estimator.