Understanding the Limiting Factors of Topic Models via Posterior Contraction Analysis

Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei,
Ming Zhang

Presented by Changyou Chen
July 31, 2015

Duke
UNIVERSITY

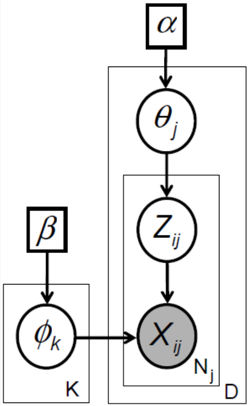
Outline

1 Limiting Factors of the LDA

2 Experiments

Latent Dirichlet allocation (LDA)

ϕ_k: word-topic distributions; θ_j: topic proportion;
Z_{ij}: topic indicators; X_{ij}: observed words



$$\phi_k|\beta \sim \text{Dirichlet}(\beta)$$
$$\theta_j|\alpha \sim \text{Dirichlet}(\alpha)$$
$$Z_{ij}|\theta_j \sim \text{Categorical}(\theta_j)$$
$$X_{ij}|\{\phi_k\}, Z_{ij} \sim \text{Categorical}(\phi_{Z_{ij}})$$

Motivations & Contributions

- Common questions from non-experts in LDA:
  - is my data topic-model friendly?
  - why did LDA fail on my data?
  - how many documents do I need to learn 100 topics?
- This paper provides theory to describe how the following limiting factors affect convergence of the LDA:
  - # documents
  - lengths of documents
  - # topics
  - Dirichlet hyper-parameters

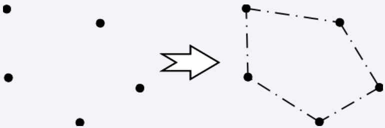
Problem setting

- In LDA, docs are generated from K topics $\phi = (\phi_1, \cdots, \phi_K)$.
- Each doc is associated with a topic proportion vector $\theta_d \in \triangle^{K-1}$
  - equivalently, each doc uniquely corresponds to a word probability vector $\eta_d = \sum_{k=1}^{K} \theta_{dk}\phi_k$
  - observed words are generated from these $\{\eta_d\}$'s, represented with a $D \times N$ matrix
- Problem:
  - how fast (rate) does the posterior distribution of $\{\phi_k\}$'s converge to the true value as D and N approach infinity?

Latent topic polytope in LDA

- Study convergence of individual topic-word distribtion?
  - identifiability problems in LDA: *e.g.*, the label-switching issue
- To avoid such problems, instead of studying individual topics, the *topic polytope* is used as a representation of topic structures in LDA:
  - given topics $\{\phi_k\}_{k=1}^K$, the topic polytope is defined as the convex hull of $\{\phi_k\}$:

$$G(\Phi) \triangleq \text{conv}(\phi_1, \cdots, \phi_K) \tag{1}$$



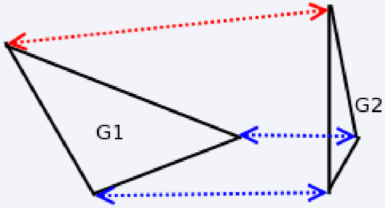
Distance between topic polytopes

- To compare 2 different models, distance between topics need to be defined.
- Define distance between two topic polytopes G_1 and G_2:

$$d_{\mathscr{M}}(G_1, G_2) \triangleq \max\{d(G_1, G_2), d(G_2, G_1)\}, \text{ ,where} \quad (2)$$

$$d(G_1, G_2) = \max_{\phi_1 \in \mathrm{extr}(G_1)} \min_{\phi_2 \in \mathrm{extr}(G_2)} \|\phi_1 - \phi_2\|_2 \quad (3)$$

where 'extr' means the *extreme points* (topics in LDA).
- Equivalent to the well-known Hausdorff metric in convex geometry under mild assumptions.



Posterior contraction analysis

- Posterior contraction analysis describes how fast the posterior of a given subset of data convergence to the true posterior distribution.
- This paper uses posterior contraction analysis to analyze the impact of limiting factors in LDA, *e.g.*, #docs, #topics, lengths of docs.
- In the following:
  - K^*: true #topics
  - K: #topics in a model
  - D: # docs
  - N: document length (assume same document length)

Contraction of the posterior of topic polytope

- Assume mild regularity conditions such that (formal descriptions omitted):
  - topic polytopes are not degenerated or collapsing
  - the prior is dense enough in the space of parameters

Theorem

Let the Dirichlet parameters for topic proportions $\alpha_k \in (0,1]$, and assume either one of the following holds:

(A1) $K = K^*$, *i.e., the true #topics is known;*

(A2) *the Euclidean distance between every pair of topics is bounded from below by a known positive constant* r_0.

then as $D \to \infty$ *and* $N \to \infty$ *such that* $N \geq \log D$, *for some* $C > 0$ *independent of* N *and* D:

$$\Pi(d_{\mathscr{M}}(G, G^*) \leq C\delta_{D,N}) \to 1 \ , \tag{4}$$

where $\delta_{D,N} = (\frac{\log D}{D} + \frac{\log N}{N} + \frac{\log N}{D})^{1/2}$, $\Pi(\cdot)$ *means under the posterior distribution.*

Some observations on the convergence rate

$$\Pi(d_{\mathscr{M}}(G,G^*) \leq C\delta_{D,N}) \to 1,\ \delta_{D,N} = (\frac{\log D}{D} + \frac{\log N}{N} + \frac{\log N}{D})^{1/2}$$

- The proof of the theorem requires $N \geq \log D$.
- Convergence rate: $\max\{(\frac{\log N}{N})^{1/2}, (\frac{\log D}{D})^{1/2}, (\frac{\log N}{D})^{1/2}\}$, $(\frac{\log N}{D})^{1/2}$ does not play a noticeable role empirically (might be an artifact due to the proof techniques).
- The actually rate might be faster since this is an upper bound (there is an lower bound $\Omega(\frac{1}{DN})$ not given here).
- The rate does not depend on #topics K, meaning if K is known or the topics are well-seperated, the inference is statistically efficient.
- In practice, the overfitted setting is preferred, *e.g.*, $K \gg K^*$, which is considered in the following.

Contraction of the posterior of topic polytope

- When neither of (A1) and (A2) hold, the rate is much worse:
  - (A1) $K = K^*$, *i.e.*, the true #topics is known;
  - (A2) the Euclidean distance between every pair of topics is bounded from below by a known positive constant r_0.

Theorem

Under the same conditions as the previous theorem, except that none of the conditions (A1) and (A2) holds, then for $K^* < K \leq |V|$*, we have*

$$\Pi(d_{\mathscr{M}}(G, G^*) \leq C\delta_{D,N}) \to 1 \ , \tag{5}$$

where $\delta_{D,N} = \left(\frac{\log D}{D} + \frac{\log N}{N} + \frac{\log N}{D}\right)^{\frac{1}{2(K-1)}}$.

- This means the convergence is very slow, depending on K.
- It is said underfitting ($K < K^*$) will result in a persistent error even with infinite data, thus not considered.

Outline

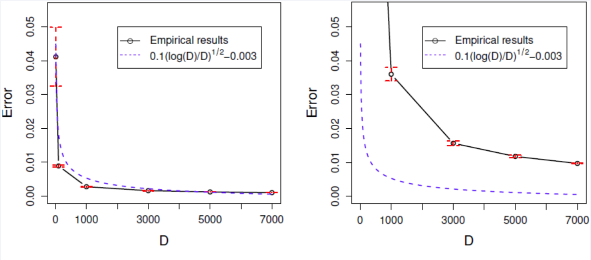
1 Limiting Factors of the LDA

2 Experiments

On synthetic data

- Generate data from an LDA with $K^* = 3$, $V = 5000$, symmetric Dirichlet prior for topic proportions and word-topic distributions to being 1 and 0.01, respectively.
- Variation of parameters: #docs D, length of docs N, Dirichlet hyperparameter for topic-word distributions β, #topics K.
- Use collapsed Gibbs sampler.

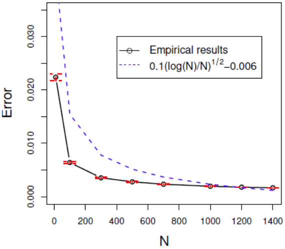
Fixing N: theoretical upper bound: $\propto \left(\frac{\log D}{D}\right)^{\frac{1}{2}}$



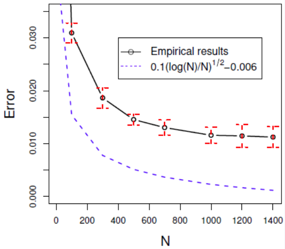
(a) $K = K^*, \beta = 0.01$ (b) $K > K^*, \beta = 0.01$

(c) $K = K^*, \beta = 1$ (d) $K > K^*, \beta = 1$

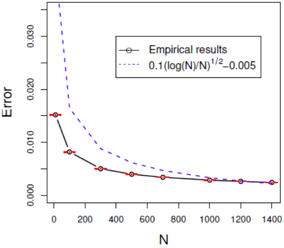
Fixing D: theoretical upper bound: $\propto \left(\frac{\log N}{N}\right)^{\frac{1}{2}}$



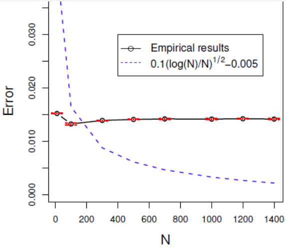
(a) $K = K^*, \beta = 0.01$



(b) $K > K^*, \beta = 0.01$

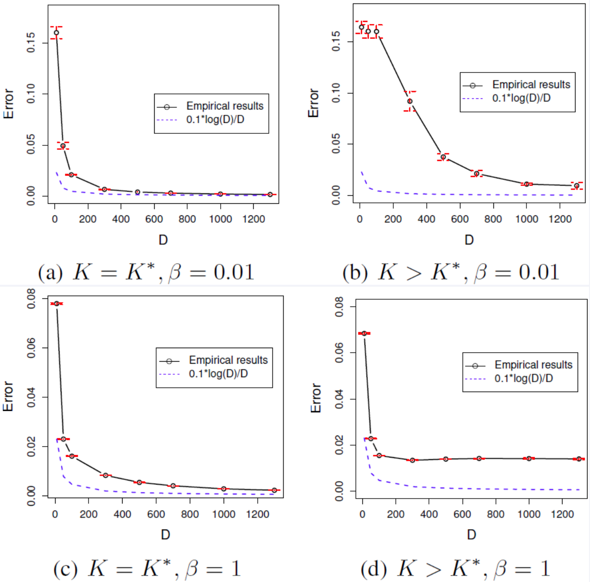


(c) $K = K^*, \beta = 1$



(d) $K > K^*, \beta = 1$

Increasing $N = D$: theoretical upper bound: $\propto \left(\frac{\log N}{N}\right)^{\frac{1}{2}}$



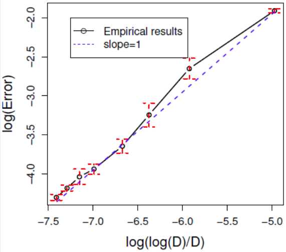
(a) $K = K^*, \beta = 0.01$

(b) $K > K^*, \beta = 0.01$

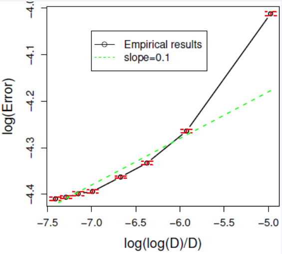
(c) $K = K^*, \beta = 1$

(d) $K > K^*, \beta = 1$

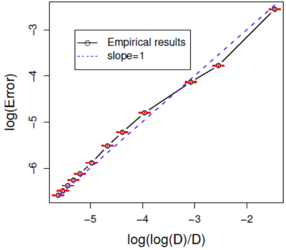
Compared with theoretically asymptotic error rates



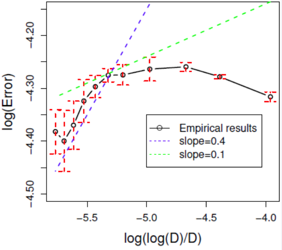
(a) Fixed N, $K = K^*$



(b) Fixed N, $K > K^*$

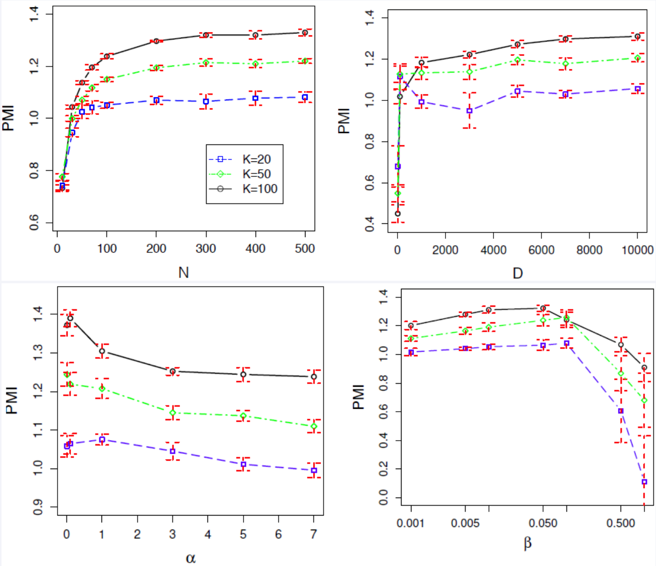


(c) $D = N$, $K = K^*$

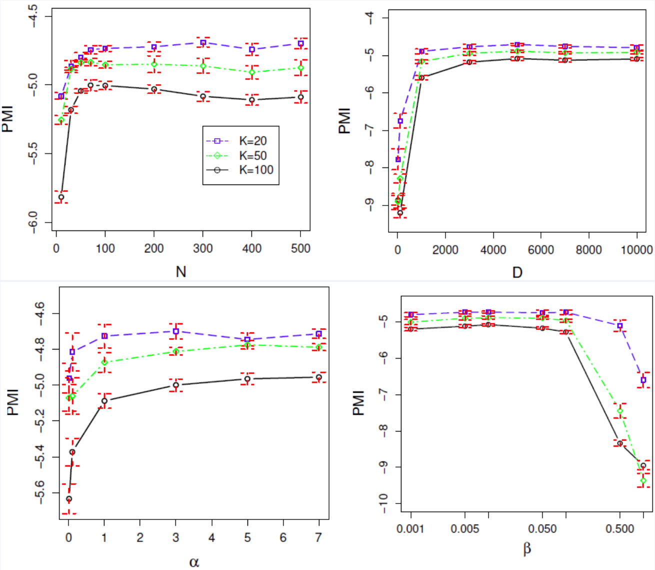


(d) $D = N$, $K > K^*$

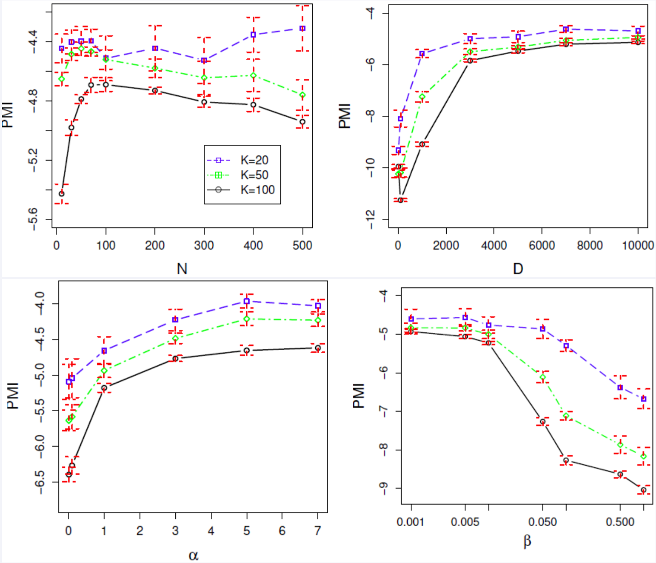
Real data: Wikipedia



Real data: New York Times



Real data: Twitter



Implications and guidelines for LDA

- #docs plays the most important role:
  - it is theoretically impossible to guarantee identification of topics from a small docs
  - once sufficient docs are provided, further increasing the number might not help significantly, unless document lengths are also increased
- poor performance when lengths of docs are too short, even if there are a lot of docs.
- when over fitting ($K \gg K^*$), convergence rates might deteriorate quickly.
- the LDA performs well when the underlying topics are well-seperated.
- if each doc is associated with few topics, the Dirichlet hyperparameter should be set to small

Thanks for your attention!!!