Topic modeling with more confidence: a theory and some algorithms

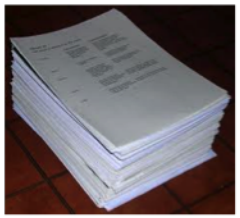
Long Nguyen

Department of Statistics
Department of EECS
University of Michigan, Ann Arbor

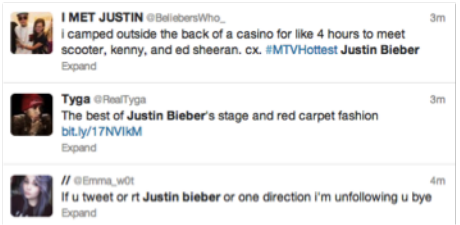
Pacific-Asia Knowledge Discovery and Data Mining,
Ho Chi Minh city, May 2015

Topic models – such as Latent Dirichlet allocation and its variants – are a popular tool for modeling and mining patterns from texts in news articles, scientific papers, blogs, but also tweets, query logs, digital books, metadata records...

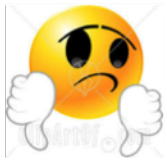
News articles



Scientific papers



Tweets



Applied with varying degrees of success, to diverse domains in computer science and beyond, e.g., biomedical informatics, scientometrics, social and political science, and digital humanities.

An example document from the AP corpus (Blei, Ng, Jordan, 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An example document from the AP corpus (Blei, Ng, Jordan, 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

After feeding such documents to Latent Dirichlet Allocation (LDA) model:

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

An example document from *Science* corpus (1880–2002) (Blei & Lafferty, 2009)

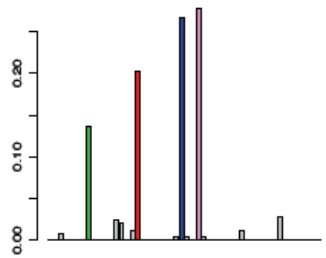
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Eager non-expert consumers of topic modeling often ask:

- is my data LDA-friendly?
- why did LDA fail on my data set?
- shall I fit LDA with 100 topics?
- how many documents do I need?
- and how long should these documents be?
- how do I set tuning parameters?

How do we know that the topics we have learned are "right"?

How do we know that the topics we have learned are "right"?

Why should we care?

Latent Dirichlet Allocation = Finite admixture models

introduced independently as extensions of the mixture of the multinomials

- Blei, Ng and Jordan (BNJ). Latent Dirichlet Allocation. NIPS conference, 2001. 11K citations.

- Pritchard, Stephens and Donnelly (PSD). Inference of population structure using multilocus genotype data. Genetics, June 2000. 14K citations.

Latent Dirichlet Allocation = Finite admixture models

introduced independently as extensions of the mixture of the multinomials

- Blei, Ng and Jordan (BNJ). Latent Dirichlet Allocation. NIPS conference, 2001. 11K citations.

- Pritchard, Stephens and Donnelly (PSD). Inference of population structure using multilocus genotype data. Genetics, June 2000. 14K citations.

models introduced by two papers are *exactly* the same, except that

- "topics" in BNJ = "population structures" by PSD

Latent Dirichlet Allocation = Finite admixture models

introduced independently as extensions of the mixture of the multinomials

- Blei, Ng and Jordan (BNJ). Latent Dirichlet Allocation. NIPS conference, 2001. 11K citations.

- Pritchard, Stephens and Donnelly (PSD). Inference of population structure using multilocus genotype data. Genetics, June 2000. 14K citations.

models introduced by two papers are *exactly* the same, except that

- "topics" in BNJ = "population structures" by PSD
- PSD used Gibbs sampling, BNJ developed variational inference algorithm for MLE

Latent Dirichlet Allocation = Finite admixture models

introduced independently as extensions of the mixture of the multinomials

- Blei, Ng and Jordan (BNJ). Latent Dirichlet Allocation. NIPS conference, 2001. 11K citations.

- Pritchard, Stephens and Donnelly (PSD). Inference of population structure using multilocus genotype data. Genetics, June 2000. 14K citations.

models introduced by two papers are *exactly* the same, except that

- "topics" in BNJ = "population structures" by PSD
- PSD used Gibbs sampling, BNJ developed variational inference algorithm for MLE
- BNJ mostly cited by CS community; PSD by the population genetics community

Latent Dirichlet Allocation = Finite admixture models

introduced independently as extensions of the mixture of the multinomials

- Blei, Ng and Jordan (BNJ). Latent Dirichlet Allocation. NIPS conference, 2001. 11K citations.

- Pritchard, Stephens and Donnelly (PSD). Inference of population structure using multilocus genotype data. Genetics, June 2000. 14K citations.

models introduced by two papers are *exactly* the same, except that

- "topics" in BNJ = "population structures" by PSD
- PSD used Gibbs sampling, BNJ developed variational inference algorithm for MLE
- BNJ mostly cited by CS community; PSD by the population genetics community
- the two communities rarely cite each other

Despite this popularity, there has been almost no rigorous theoretical guarantees or systemmatic analysis of these methods in either literatures

Eager non-expert consumers of topic modeling often ask
(replace "LDA" by "admixture" if you want)

- is my data LDA-friendly?
- why did LDA fail on my data set?
- shall I fit LDA with 100 topics?
- how many documents do I need?
- and how long should these documents be?
- how do I set tuning parameters?

How do we *guarantee* that the topics we have learned are "correct"?

How efficient is the learning procedure with LDA?

Outline of talk

- theory: contraction behavior of the posterior distribution of latent topic structures
  - also, maximum likelihood estimation
  - upper and lower bounds for rate of convergence

Outline of talk

- theory: contraction behavior of the posterior distribution of latent topic structures
  - also, maximum likelihood estimation
  - upper and lower bounds for rate of convergence

- practice: confirmatory systematic study on artificial and real data on the limiting factors of a topic model
  - roles of number of documents, document length, number of topics, sparsity and distance of topics

Outline of talk

- theory: contraction behavior of the posterior distribution of latent topic structures
  - also, maximum likelihood estimation
  - upper and lower bounds for rate of convergence
- practice: confirmatory systematic study on artificial and real data on the limiting factors of a topic model
  - roles of number of documents, document length, number of topics, sparsity and distance of topics
- algorithms: for discovering new topics which encode emerging events

Understanding the limiting factors of topic modeling via posterior contraction analysis. J. Tang, Z. Meng, X. Nguyen, Q. Mei and M. Zhang. *ICML-2014* ("Best paper award").

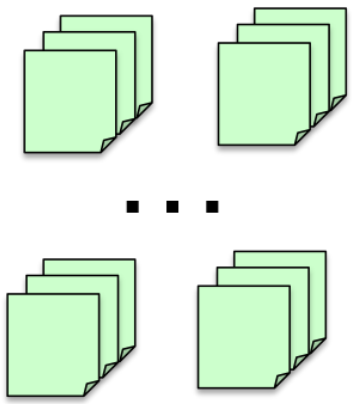
Posterior contraction behavior of the population polytope in finite admixture models. X. Nguyen. *Bernoulli*, 21(1), 618–646, 2015.

Detecting emerging topic models with confidence. Z. Meng, X. Nguyen, A. Hero and Q. Mei. In preparation.

Outline

Topic modeling for documents

Documents:
Bags of words



PLSA (Hofmann et al. 1999)
LDA (Blei, et al. 2003)
HDP (Teh et al. 2005)

....

Topics:
Multinomial distribution over words

information	*0.16*
retrieval	*0.08*
search	*0.07*
...	....

machine	*0.16*
learning	*0.08*
classifier	*0.07*
...	....

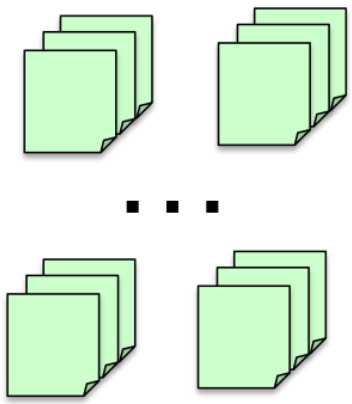
data	*0.16*
mining	*0.08*
knowledge	*0.07*
...	....

web	*0.16*
semantic	*0.08*
content	*0.07*
...	....

Topic modeling for documents

Documents:
Bags of words

Topics:
Multinomial distribution over words



PLSA (Hofmann et al. 1999)
LDA (Blei, et al. 2003)
HDP (Teh et al. 2005)
....

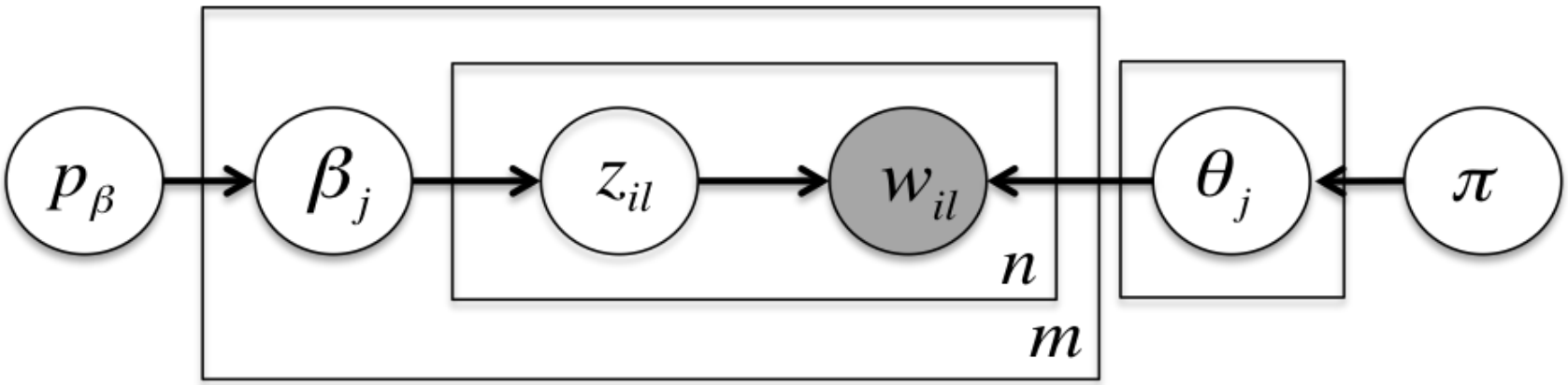
information	0.16
retrieval	0.08
search	0.07
...	....

machine	0.16
learning	0.08
classifier	0.07
...	....

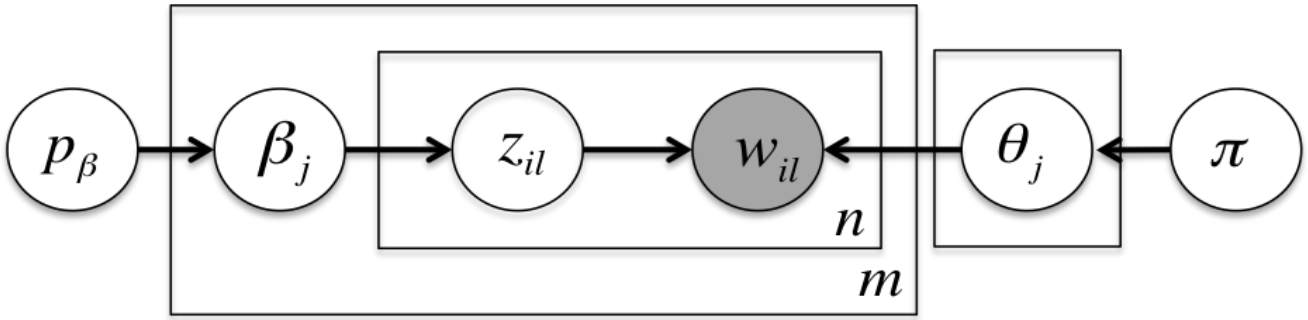
data	0.16
mining	0.08
knowledge	0.07
...	....

web	0.16
semantic	0.08
content	0.07
...	....

Latent Dirichlet allocation (LDA)



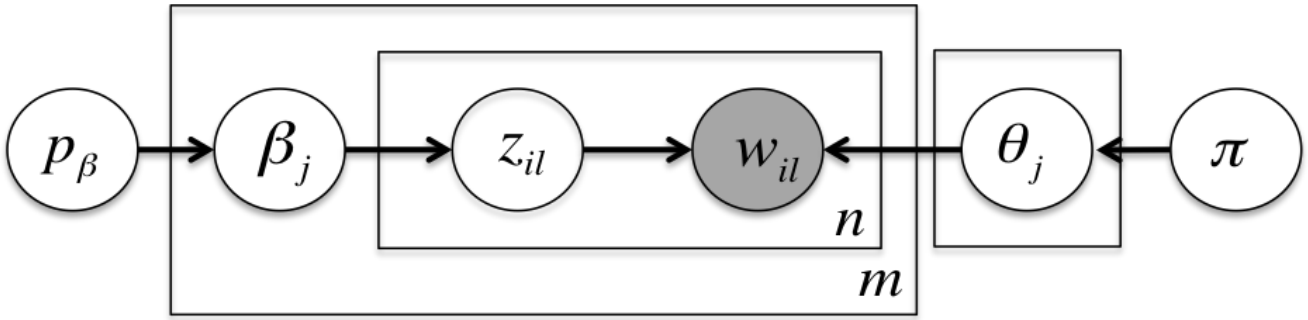
Latent Dirichlet allocation model



Generative process:

- For each $j = 1, \ldots, k$, sample a vector of frequencies $\theta_j \in \Delta^{d-1}$
  - these are called "topics", distributed by a Dirichlet
  - d = vocabulary size

Latent Dirichlet allocation model



Generative process:

- For each $j = 1, \ldots, k$, sample a vector of frequencies $\theta_j \in \Delta^{d-1}$
  - these are called "topics", distributed by a Dirichlet
  - d = vocabulary size
- For each document $i = 1, \ldots, m$,
  - sample a topic proportion $\beta \in \Delta^{k-1}$ (e.g., another Dirichlet)
  - for each word position in document i
    - ★ sample a topic label $z \sim \text{Multinomial}(\beta)$;
    - ★ given z, sample a word $w \sim \text{Multinomial}(\theta_z)$.

Inferential goal: given data of size $m \times n$, estimate the topic vectors θ_j's

Geometric reformulation of LDA

LDA posits that

- there are k topics $\boldsymbol{\theta}_j \in \Delta^{d-1}$, for $j = 1, \ldots, k$
  - $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ may be endowed with a prior Π

Geometric reformulation of LDA

LDA posits that

- there are k topics $\boldsymbol{\theta}_j \in \Delta^{d-1}$, for $j = 1, \ldots, k$
  - $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ may be endowed with a prior Π
- m documents, each with a random vector of proportions $\boldsymbol{\beta} \in \Delta^{k-1}$

Geometric reformulation of LDA

LDA posits that

- there are k topics $\boldsymbol{\theta}_j \in \Delta^{d-1}$, for $j = 1, \ldots, k$
  - $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ may be endowed with a prior Π
- m documents, each with a random vector of proportions $\boldsymbol{\beta} \in \Delta^{k-1}$
- each document $i = 1, \ldots, m$ contains n words generated by frequency

$$\boldsymbol{\eta} = \sum_{j=1}^{k} \beta_j \boldsymbol{\theta}_j \in \Delta^{d-1}$$

Geometric reformulation of LDA

LDA posits that

- there are k topics $\boldsymbol{\theta}_j \in \Delta^{d-1}$, for $j = 1, \ldots, k$
  - $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ may be endowed with a prior Π
- m documents, each with a random vector of proportions $\boldsymbol{\beta} \in \Delta^{k-1}$
- each document $i = 1, \ldots, m$ contains n words generated by frequency

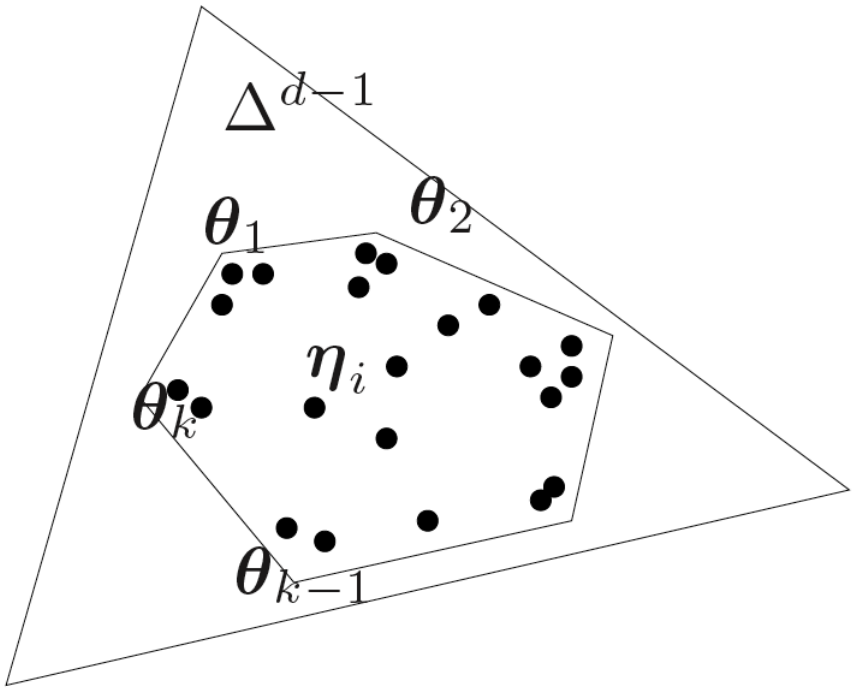
$$\boldsymbol{\eta} = \sum_{j=1}^{k} \beta_j \boldsymbol{\theta}_j \in \Delta^{d-1}$$

geometry: $\boldsymbol{\eta}$ lies in convex hull of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$. Call this topic polytope

Inference problem

Given $m \times n$ data set $\mathcal{D}$, estimate the topic polytope

LDA = Convex geometry problem



- let $G_0 = \text{conv}(\theta_1^*, \ldots, \theta_k^*)$, the convex hull of the θ_j^*, be the "true" topic polytope
- Data are documents, each of which correspond to a random point η_i drawn from inside of polytope G_0

Question

How to estimate convex polytope G_0 based on noisy observation sampled from the polytope?

We ask how fast does the MLE estimate of the convex polytope $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ converges:

$$G \longrightarrow G_0 = \mathrm{conv}(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*)?$$

We ask how fast does the MLE estimate of the convex polytope $G = \text{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ converges:

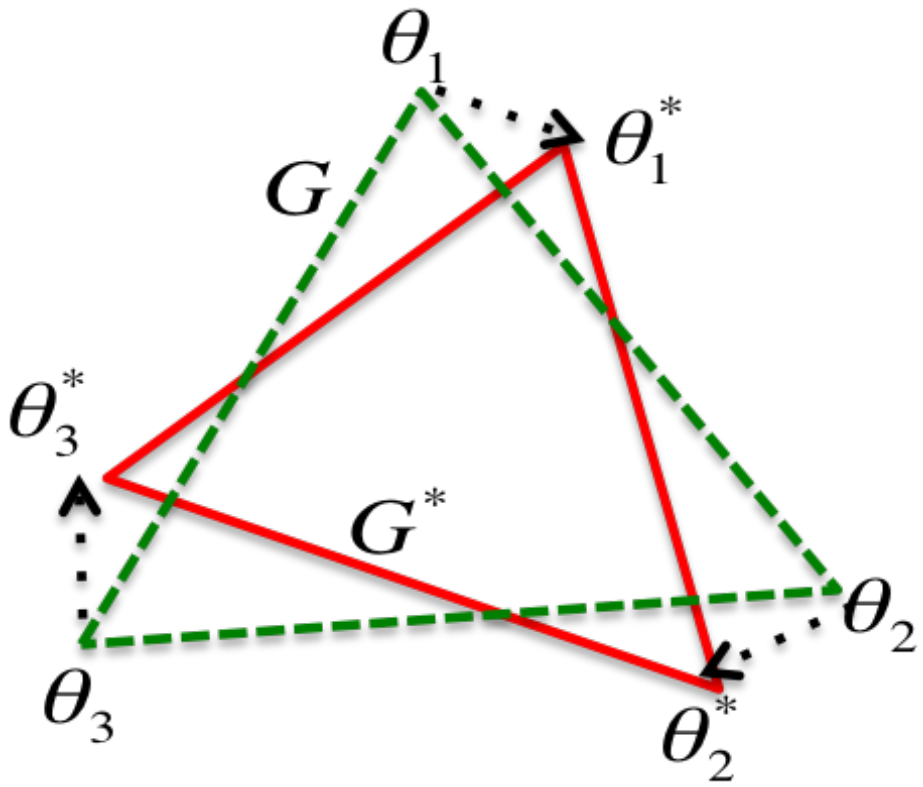
$$G \longrightarrow G_0 = \text{conv}(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*)?$$

In a Bayesian setting, we calculate the posterior distribution

$$\Pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k | \text{Data set } \mathcal{D})$$

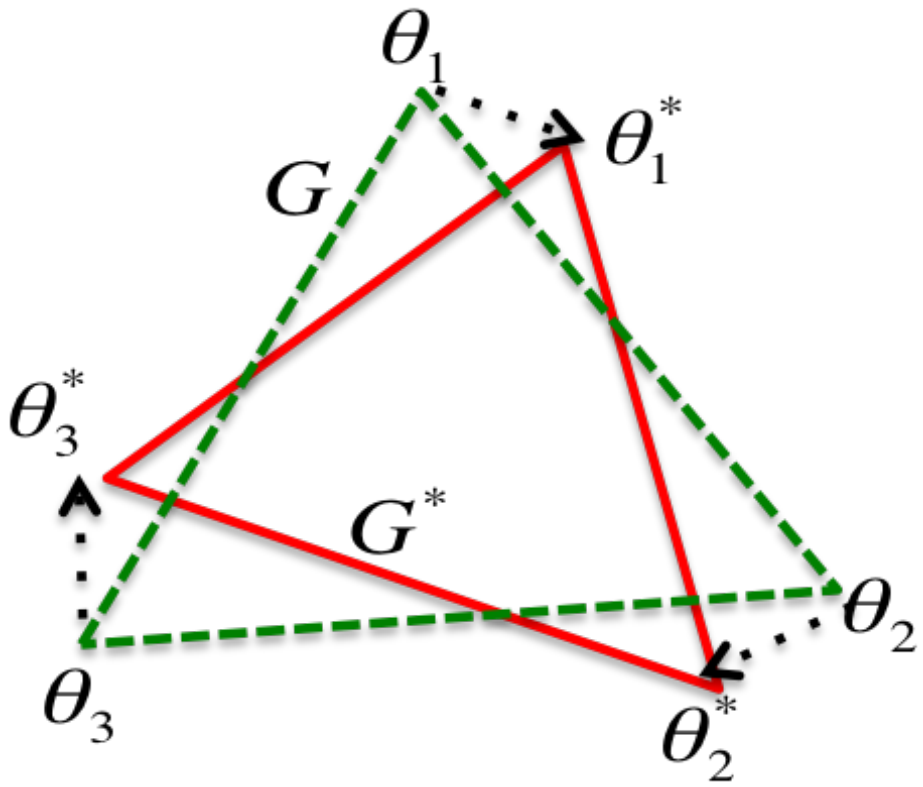
and asks how fast does this posterior concentrates most its mass around the truth $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$, as data size $m \times n$ tends to infinity

Metrics on topic polytopes



If we try to match each vertex (i.e., extreme point) of G by the closest vertex of G^* and vice versa, the minimum-matching distance is the largest distance among the matching pairs.

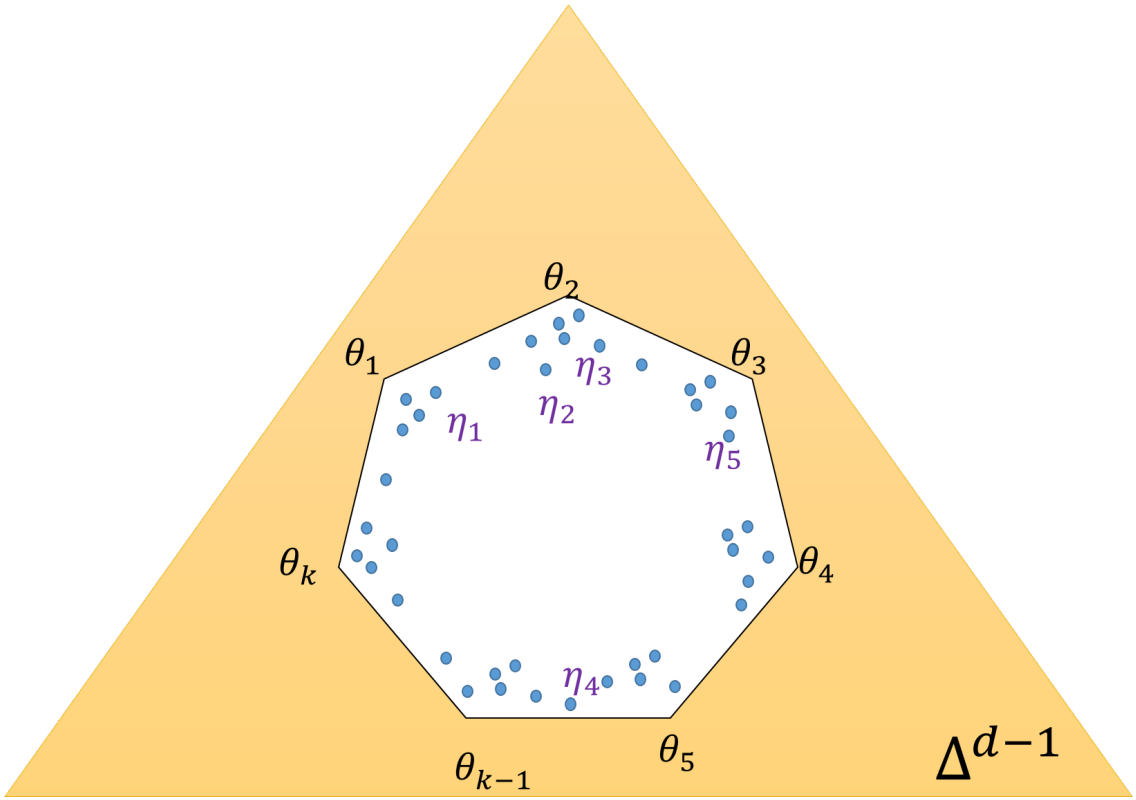
Metrics on topic polytopes



If we try to match each vertex (i.e., extreme point) of G by the closest vertex of G^* and vice versa, the minimum-matching distance is the largest distance among the matching pairs.

Hausdorff distance is the smallest amount of enlargement of G so as to cover G^* and vice versa.

Regularity of mass concentration near polytope's boundary



Intuition:

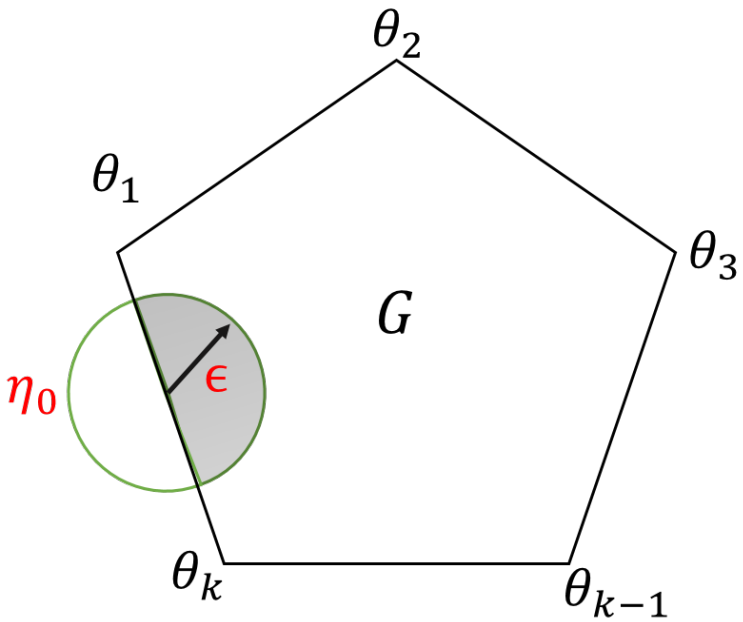
If most documents lie near the boundary of the polytope, then it is easier to recover the extreme points (vertices)

Regularity of mass concentration near boundary

We say the probability distribution $P_{\eta|G}$ on polytope G α-regular if for any η_0 in the boundary of G,

$$\frac{P_{\eta|G}(\|\eta - \eta_0\| \leq \epsilon)}{\mathrm{vol}_p(G \cap B_d(\eta_0, \epsilon))} \gtrsim \epsilon^{\alpha}.$$

where p is the number of dimensions of the affine space aff G that spans G

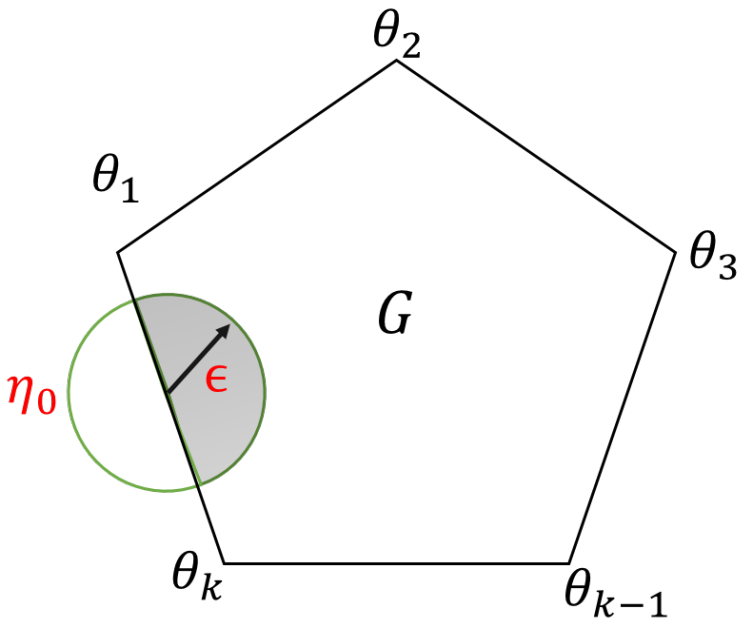


Regularity of mass concentration near boundary

We say the probability distribution $P_{\eta|G}$ on polytope G α-regular if for any η_0 in the boundary of G,

$$\frac{P_{\eta|G}(\|\eta - \eta_0\| \leq \epsilon)}{\mathrm{vol}_p(G \cap B_d(\eta_0, \epsilon))} \gtrsim \epsilon^\alpha.$$

where p is the number of dimensions of the affine space aff G that spans G



Example: uniform distribution on G corresponds to $\alpha = 0$

Regularity Lemma

If $\boldsymbol{\beta} \sim \mathrm{Dir}(\gamma_1, \ldots, \gamma_k)$ where $\|\boldsymbol{\gamma}\|_\infty \leq 1$, then α-regularity holds with

- $\alpha = 0$ if $k - 1 \leq d$
- $\alpha = \sum_{j=1}^{k} \gamma_j$ if $k - 1 > d$

Examples

- Topic modeling: k = number of topics, d = vocabulary size, so $k \ll d$
- Population genetics: k = number of ethnic origins, d = number of DNA alphabets, so $k \gg d = 4$

Technical assumptions

Π is a prior distribution on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ such that the following hold for the relevant parameters that reside in the support of Π:

(S0) Mild geometric properties are satisfied uniformly for all G to disallow degenerate polytopes.

(S1) Each of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ is bounded away from the boundary of Δ^d. That is, if $\boldsymbol{\theta}_j = (\theta_{j,0}, \ldots, \theta_{j,d})$ then $\min_{l=0,\ldots,d} \theta_{j,l} > c_0$ for all $j = 1, \ldots, k$.

(S2) For any small ϵ, $\Pi(\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\| \leq \epsilon \; \forall j = 1, \ldots, k) \geq c_0' \epsilon^{kd}$, for some $c_0' > 0$.

(S3) $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is distributed (a priori) according to a symmetric probability distribution $P_{\boldsymbol{\beta}}$ on Δ^{k-1}. That is, the random variables $\beta_1, \ldots, \beta_k$ are exchangeable.

(S4) $P_{\boldsymbol{\beta}}$ induces a family of distributions $\{P_{\boldsymbol{\eta}|G} | G \in \mathcal{G}^k\}$ that is α-regular.

Theorem 1: guarantee for Bayesian posterior contraction

Suppose the true topic polytope G_0 has at most k vertices, and the above assumptions hold. Let $p = (k-1) \wedge d$.

As $m \to \infty$ and $n \to \infty$ such as $\log\log m \leq \log n = o(m)$, the posterior distribution of the topic polytope G contracts toward truth G_0 at rate $\delta_{m,n}$:

$$\Pi\left(d_{\mathcal{M}}(G_0, G) \leq \delta_{m,n} \middle| m \times n \text{ Data} \right) \longrightarrow 1$$

in probability, where

$$\delta_{m,n} \asymp \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2(p+\alpha)}}.$$

Theorem 2: guarantee for maximum likelihood estimation

If we use maximum likelihood estimation of obtain $\widehat{G}_{mn}$. The convergence rate to the truth G_0 is the same:

$$d_{\mathcal{M}}(G_0, \widehat{G}_{mn}) = O_p\left(\left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}\right]^{\frac{1}{2(p+\alpha)}}\right).$$

Remarks on rate

$$\delta_{m,n} \asymp \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2(p+\alpha)}}.$$

- Presence of m^{-1} and n^{-1} in the contraction rate suggests that if either m or n is small, the rate would suffer even if data size $m \times n$ increases
  - Moral: LDA won't work for many short tweets or very few long articles

Remarks on rate

$$\delta_{m,n} \asymp \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2(p+\alpha)}}.$$

- Presence of m^{-1} and n^{-1} in the contraction rate suggests that if either m or n is small, the rate would suffer even if data size $m \times n$ increases
  - Moral: LDA won't work for many short tweets or very few long articles
- The exponent $\frac{1}{2(p+\alpha)}$ appears intrinsic (there are comparable lower bounds).
  - suppose, $k = 100$ topics, $d = 10K$ words, so $p = \min(k-1, d) = 99$ then the rate becomes

$$\delta_{m,n} \asymp \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2(99+\alpha)}}.$$

  - this is worrisome! Slow rate is due to overfitting the model with more topics than needed.

What if we know exact number of topics k^*?

or if the topic vectors are well separated by a known constant, i.e., $\|\theta_j - \theta_{j'}\| \geq c_0$

Theorem 2

Under such additional assumptions, the rate of contraction becomes

$$\delta_{m,n} = \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2(1+\alpha)}},$$

the exponent is independent of the number of topics

What if we know exact number of topics k^*?

or if the topic vectors are well separated by a known constant, i.e., $\|\theta_j - \theta_{j'}\| \geq c_0$

Theorem 2

Under such additional assumptions, the rate of contraction becomes

$$\delta_{m,n} = \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2(1+\alpha)}},$$

the exponent is independent of the number of topics

Moral: We should not liberally over-fit the LDA with too many redundant topics, for doing so will simply earn us a long list of junks!

Minimax lower bounds

We wish establish a lower bound for the minimax term, say

$$\min_{\widehat{G}} \max_{G_0} P_{G_0}^m d_{\mathcal{H}}(G_0, \widehat{G}) \geq \epsilon_{mn}$$

Why does this bound matter?

given any algorithm for obtaining $\widehat{G}$ there exists an instance of G_0 such that the algorithm cannot estimate G_0 at a rate better than lower bound ϵ_{mn}

Technical assumptions

either

(S5) For any pair of p-dimensional polytopes $G' \subset G$ that satisfy certain geometric properties,

$$V(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G'}) \lesssim d_{\mathcal{H}}(G, G')^{\alpha} \operatorname{vol}_p G \setminus G'.$$

or

(S5') For any p-dimensional polytope G, $P_{\boldsymbol{\eta}|G}$ is the uniform distribution on G. (This actually entails (S5) for $\alpha = 0$.)

Theorem 3 (Minimax lower bounds)

Given mild geometric assumptions

(a) Let $q = \lfloor k/2 \rfloor \wedge d$. Under Assumption (S5), we have

$$\inf_{\hat{G}} \sup_{G_0} P_{G_0}^m d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{mn}\right)^{\frac{1}{q+\alpha}}.$$

Bound is improved to $(\frac{1}{mn})^{\frac{1}{1+\alpha}}$ under exact-fit or well-separated cases.

Theorem 3 (Minimax lower bounds)

Given mild geometric assumptions

(a) Let $q = \lfloor k/2 \rfloor \wedge d$. Under Assumption (S5), we have

$$\inf_{\hat{G}} \sup_{G_0} P_{G_0}^m d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{mn}\right)^{\frac{1}{q+\alpha}}.$$

Bound is improved to $(\frac{1}{mn})^{\frac{1}{1+\alpha}}$ under exact-fit or well-separated cases.

(b) Let $q = \lfloor k/2 \rfloor \wedge d$. Under Assumption (S5') [i.e., $P_{\eta|G}$ is uniform], we have

$$\inf_{\hat{G}} \sup_{G_0} P_{G_0}^m d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{m}\right)^{\frac{1}{q}}.$$

Bound is improved to $1/m$ under exact-fit or well-separated cases.

Remarks on lower/upper bounds

(i) Although there remain some gap between upper and lower bounds, they are qualitatively comparable and both notably dependent on d and k.

If $k \geq 2d$, and letting $m \asymp n$, the rate exponents between two bounds differ by only a factor of 4:

$$m^{-1/2(d+\alpha)} \quad \text{vs} \quad m^{-2/(d+\alpha)}$$

(ii) When $P_{\eta|G}$ is uniform, there is a lower bound which does *not* depend on n

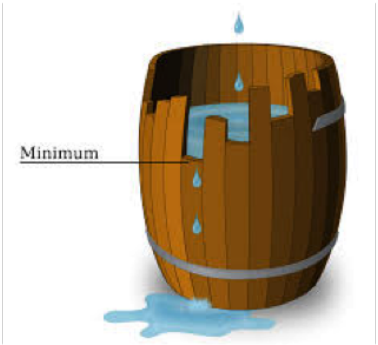
Outline

Simulation and Real Data Illustration

Simulations to confirm roles of limiting factors
Experiments with Wikipedia, New York Times articles, and Twitter messages

Simulation and Real Data Illustration

Simulations to confirm roles of limiting factors
Experiments with Wikipedia, New York Times articles, and Twitter messages

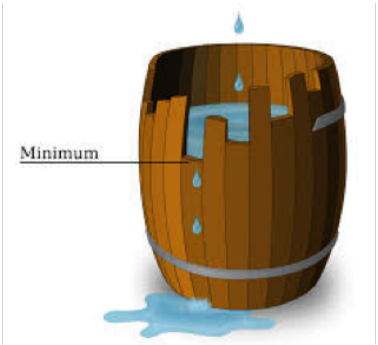


Limiting factors

- $D = m$, number of documents
- $N = n$, document length
- num of redundant topics $k - k_0$
- Dirichlet prior hyperparameters

Simulation and Real Data Illustration

Simulations to confirm roles of limiting factors
Experiments with Wikipedia, New York Times articles, and Twitter messages



Limiting factors

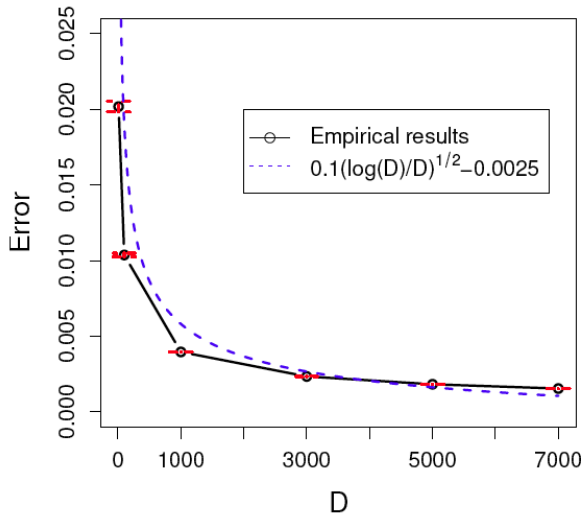
- $D = m$, number of documents
- $N = n$, document length
- num of redundant topics $k - k_0$
- Dirichlet prior hyperparameters

Liebig's law: capacity of a barrel with staves of unequal length is limited by the shortest staves

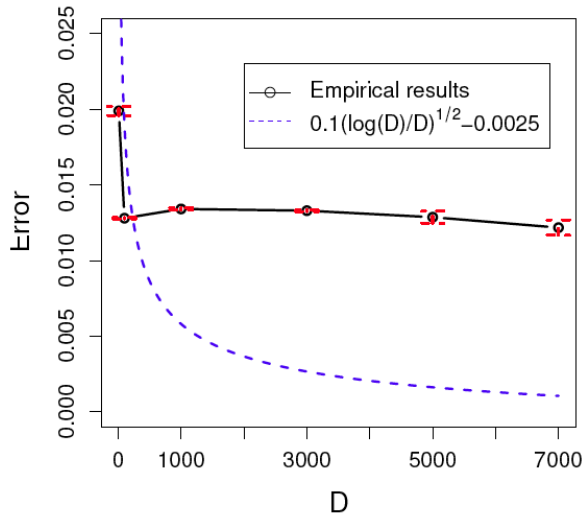
Simulations

We simulate text data with D documents and N words per document

Setting: Fix $N = 500$, let D increase; true $k^* = 3$, overfit $k = 10$



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



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

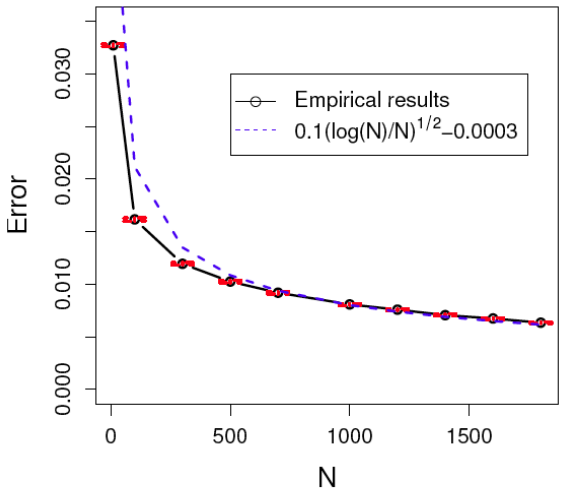
Left: Exact-fitted

$$\left[\frac{\log N}{N} + \frac{\log D}{D} + \frac{\log N}{D}\right]^{1/2}$$

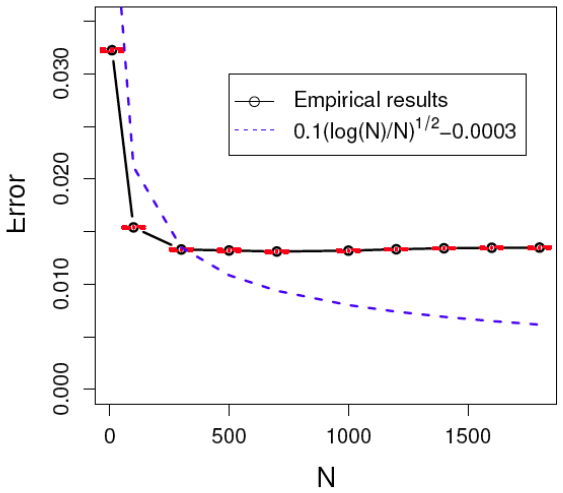
Right: Over-fitted

$$\left[\frac{\log N}{N} + \frac{\log D}{D} + \frac{\log N}{D}\right]^{1/2(k-1)}$$

Setting: Fix $D = 100$, let N increase; true $k^* = 3$, overfit $k = 5$



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



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

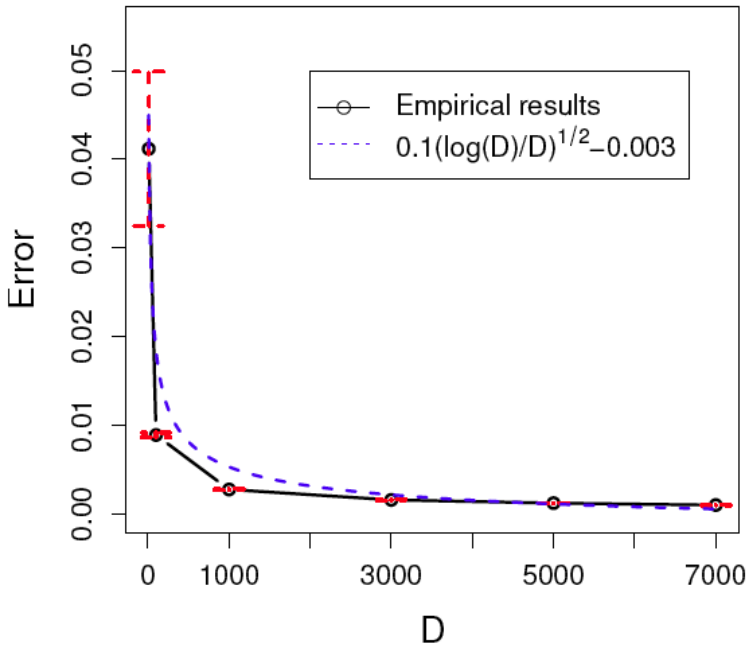
Left: Exact-fitted

$$\left[\frac{\log N}{N} + \frac{\log D}{D} + \frac{\log N}{D}\right]^{1/2}$$

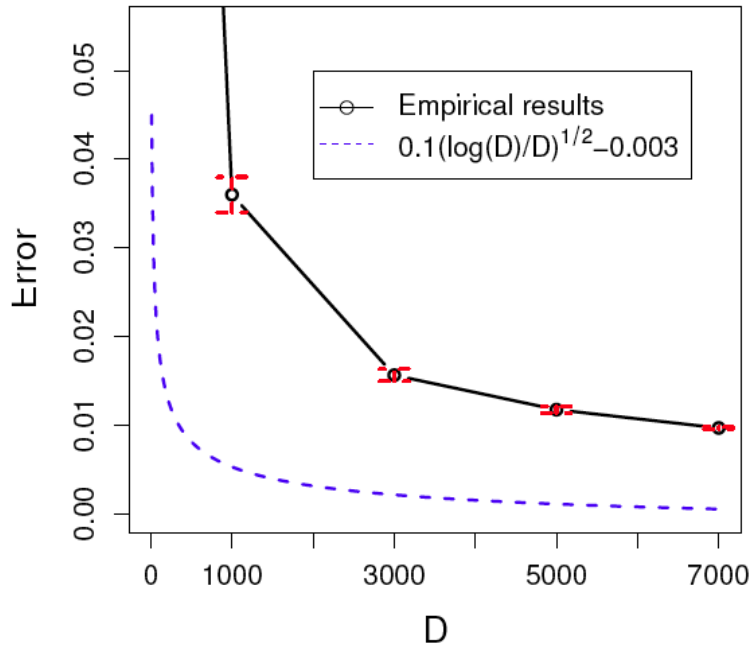
Right: Over-fitted

$$\left[\frac{\log N}{N} + \frac{\log D}{D} + \frac{\log N}{D}\right]^{1/2(k-1)}$$

Better result if the topic is concentrated at a small number of words

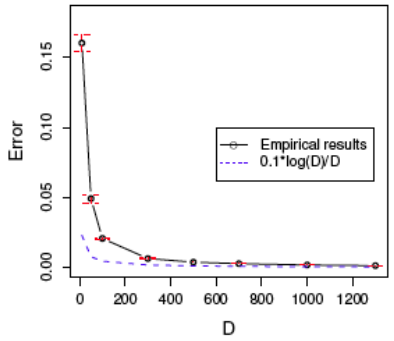


(a) $K = K^*, \beta = 0.01$

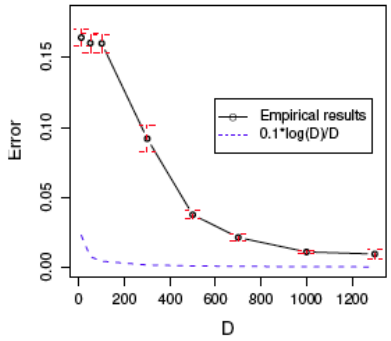


(b) $K > K^*, \beta = 0.01$

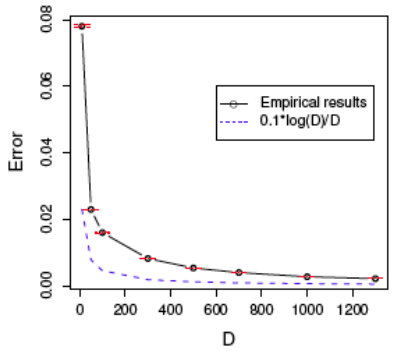
Setting: Let $N = D$ increase; true $k^* = 3$, overfit $k = 5$;



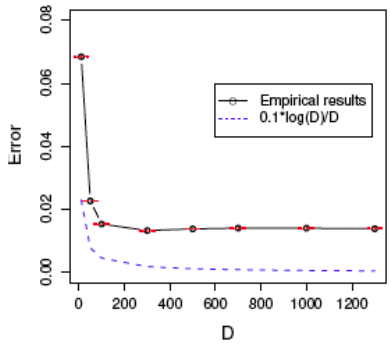
(a) $K = K^*, \beta = 0.01$



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



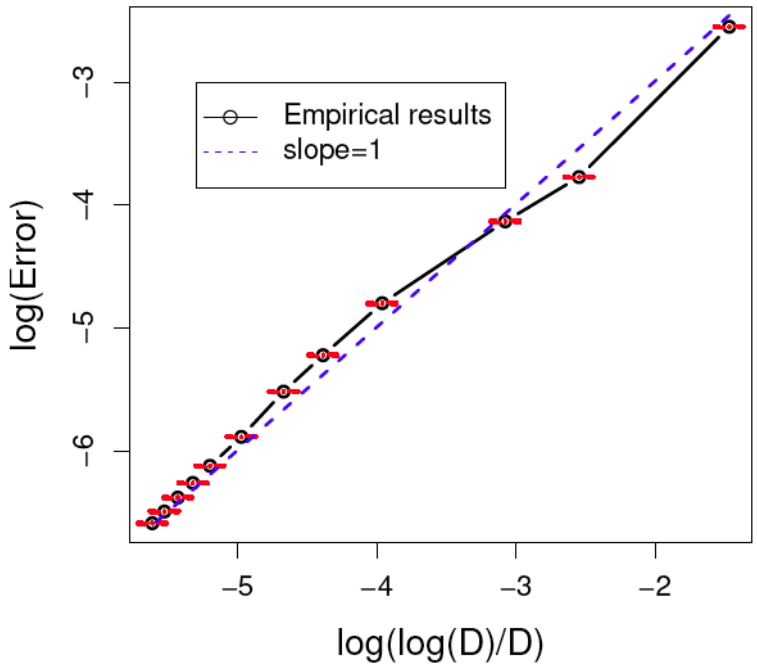
(c) $K = K^*, \beta = 1$



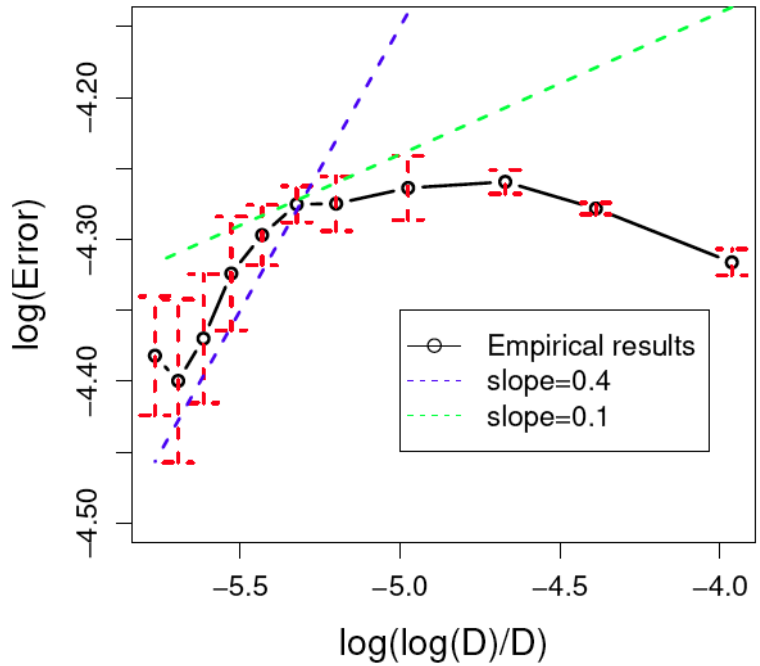
(d) $K > K^*, \beta = 1$

Left Exact-fitted: upper bound $= [\log D/D]^{1/2}$, lower bound $= (1/D)$
Right Over-fitted: upper bound $= [\log D/D]^{1/2(k-1)}$, lower bound $= (1/D)^{2/k}$

Verifying the exponential rate



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

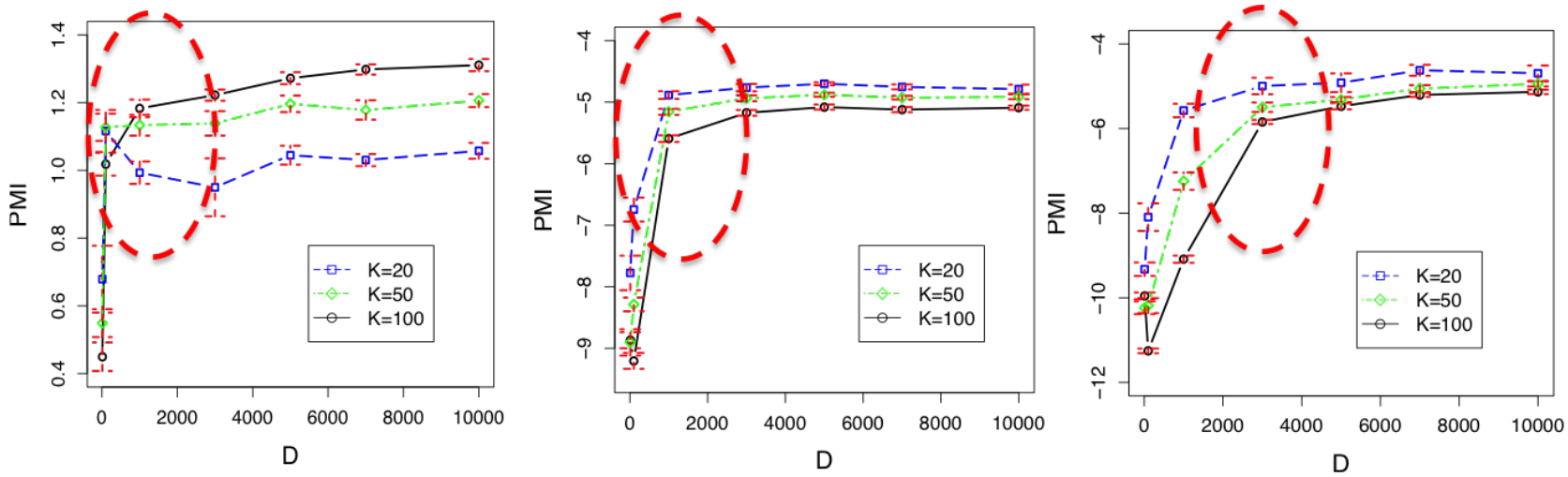


(d) $D = N$, $K > K^*$

Exactfitted: the slope of the logarithm of error is bounded between $1/2$ and 1
Overfitted: the slope is bounded between $1/2(K-1) = 0.125$ and $2/K = 0.4$

Effects of varying number of documents

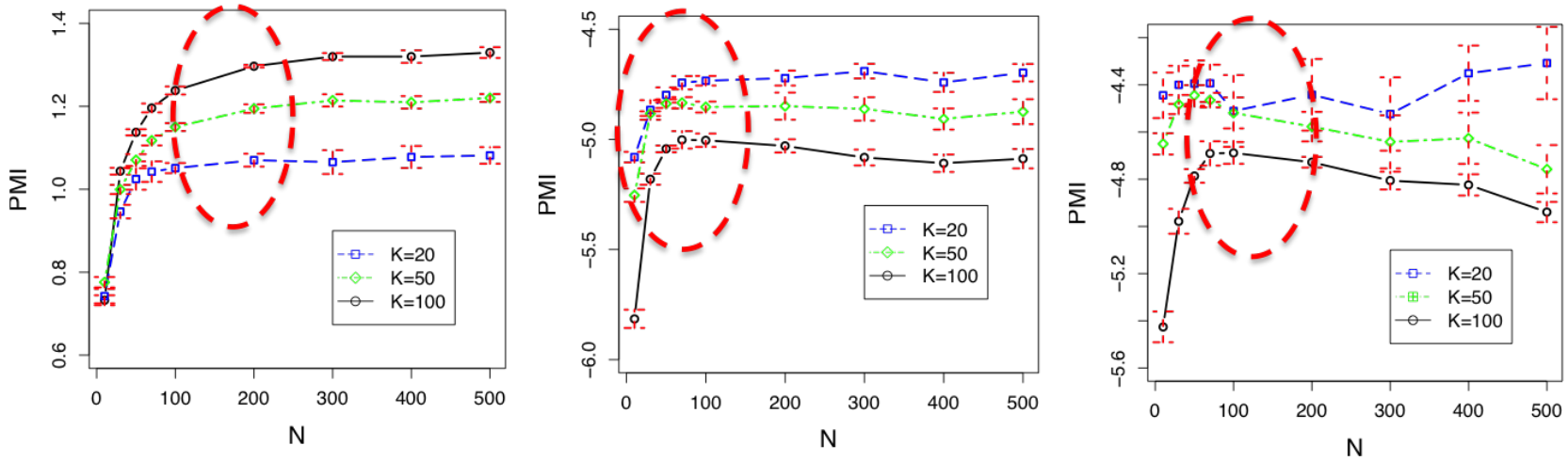
Experiments on Wikipedia pages, New York Times articles and Tweeter messages



- A few (eg., tens of) documents won't work even if they are long
- Performance stablizes after some large D (1000 documents for 100 topics)
- Sample a fraction of documents if too many

Effects of varying document length

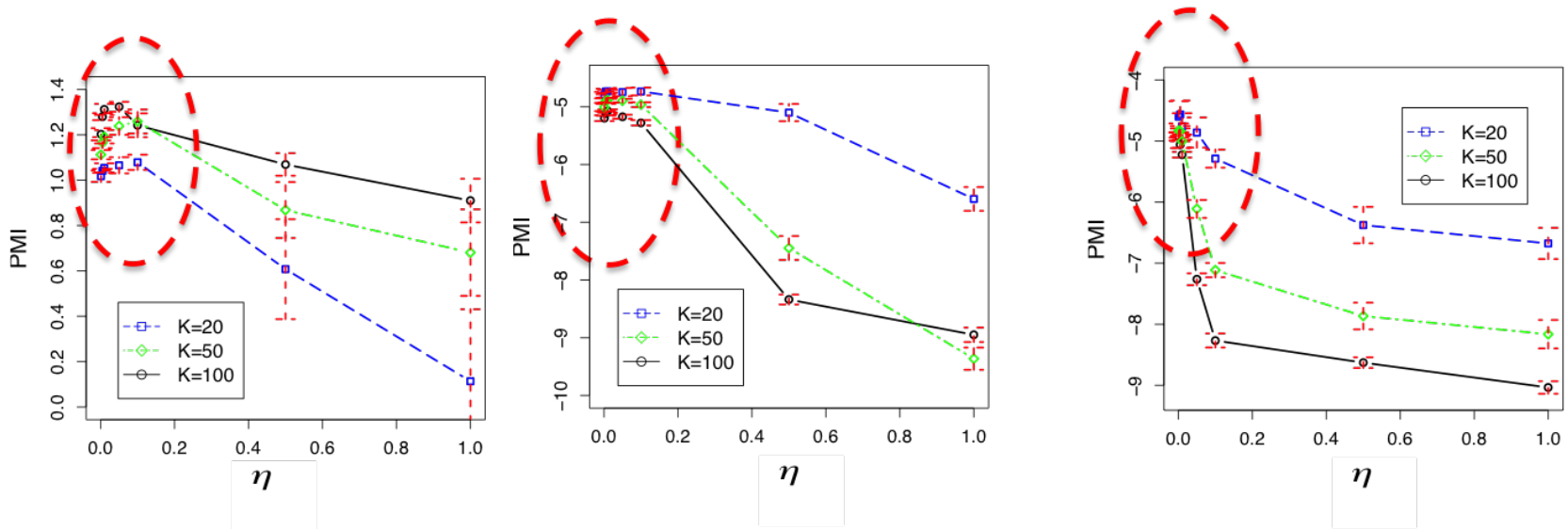
Experiments on Wikipedia pages, New York Times articles and Tweeter messages



- Short documents (eg., less than 10 words) won't work even if there are many of them
- Performance stablizes after some large N (100 words for 100 topics)
- Sample a fraction of words per document if too long

Small tuning Dirichlet parameter for topic vector θ_j's implies that each topic concentrates on few words so they are well-separated, i.e., $\|\theta_j - \theta_{j'}\|$ bounded away from 0

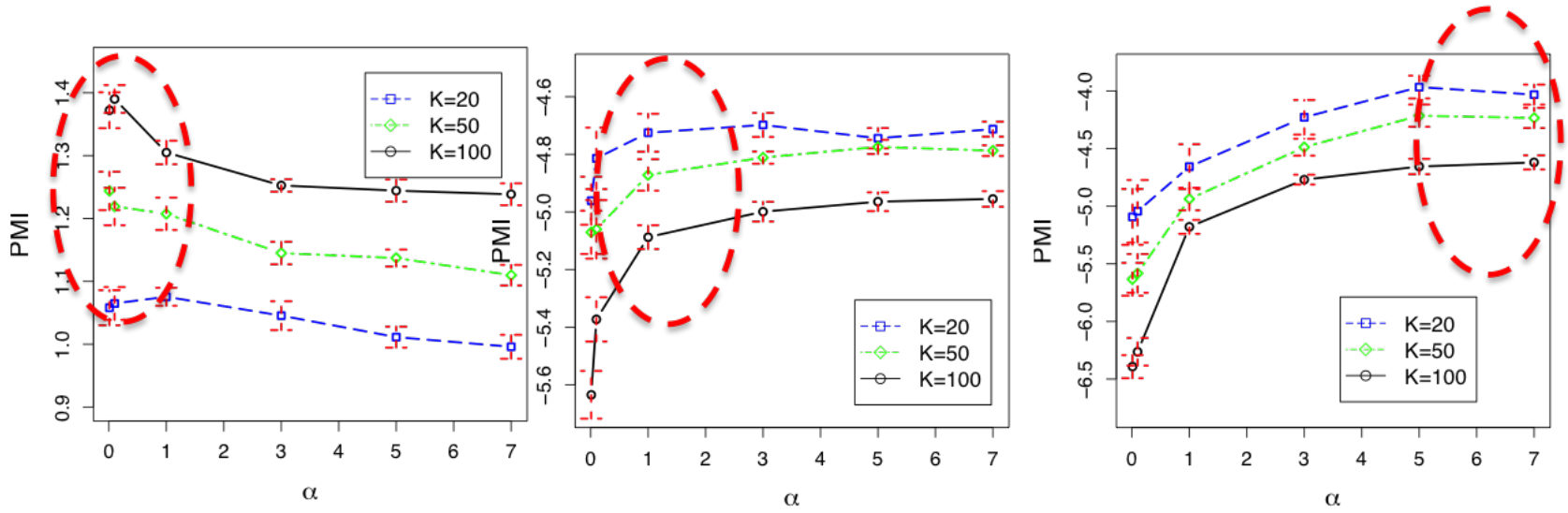
Left: Wikipedia; Middle: NYT; Right: Twitter



Small Dirichlet parameter helps, especially if we overfit.

Small tuning Dirichlet parameter for topic proportion β implies that most documents concentrate being near boundary of topic polytope

Left: Wikipedia; Middle: NYT; Right: Twitter



Different data sets appear to favor different regimes: Twitter messages seem to be more diffuse (and so more similar) in terms of topics than Wikipedia or NYT articles!

Take-away lessons

- number of documents the most important (a few won't work)
- document length plays an useful role too (short documents won't do)

Take-away lessons

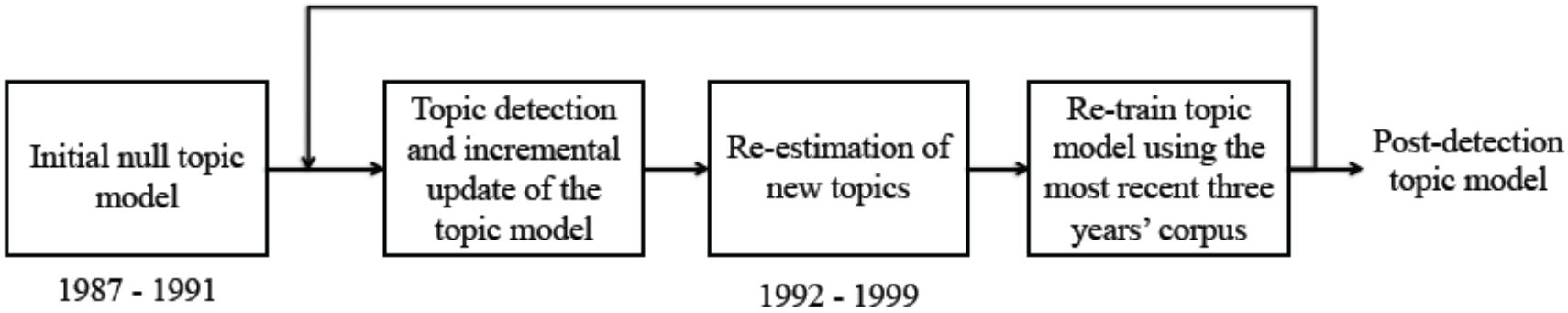
- number of documents the most important (a few won't work)
- document length plays an useful role too (short documents won't do)
- overfitting too many redundant topics dramatrically worsen the learning rate
  - remedies by being cautious and by well-separated topics
  - topics are well-separated by insisting that Dirichlet parameter for random topic vector be small

Outline

- LDA is great for topic modeling with the vast text data from the web
- but data is highly dynamic, and topics subject to change
- can we use LDA to detect emerging events?
- do not want to overfit with too many redundant topics

Detecting emerging topics from NIPS abstracts

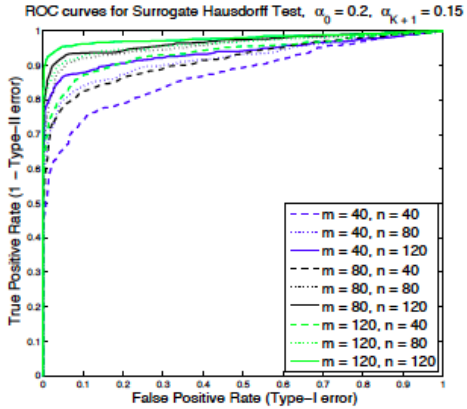
- fitting a null topic model using the first five years of NIPS abstract
- continually testing for presence of new topics in the following years
- deliberately underfit for detection purpose (i.e., alternative hypothesis may be a misspecified model)



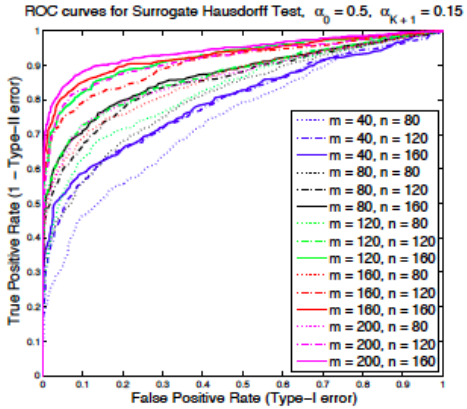
Avoid overfitting with redundant topics

- to avoid overfitting, for alternative hypothesis we add three (3) topics to the polytope and perform a generalized likelihood ratio test
- instead of evaluating the likelihood function, which is computationally intractable, we propose a so-called Hausdorff surrogate test, by making use of the Hausdorff geometry of convex polytopes
- our approach comes with theoretical guarantees of the detection (type-1 and type-2) errors

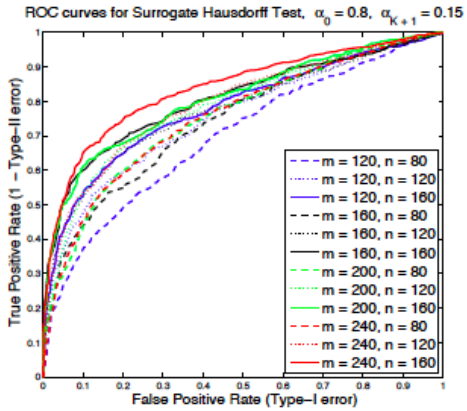
ROC curves of Hausdorff surrogate test (simulations)



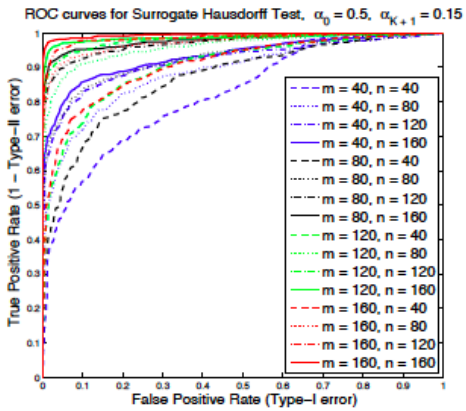
(a) ROC curves for HT-1 ($\alpha_0 = 0.2$, $\alpha_{K+1} = 0.15$).



(b) ROC curves for HT-1 ($\alpha_0 = 0.5$, $\alpha_{K+1} = 0.15$).



(c) ROC curves for HT-1 ($\alpha_0 = 0.8$, $\alpha_{K+1} = 0.15$).



(d) ROC curves for HT-q ($q = 2$, $\alpha_0 = 0.5$, $\alpha_{K+1} = \alpha_{K+2} = 0.15$).

Example of Findings for NIPS abstract corpus

- Earlier years: domination of neural-biological subjects (*rat, hyppocampal, and visual*)
- 1996: new topics of *independence component analysis* in 1996 and 1997 — this anticipates the CFP for ICA in the year 2000 (not included in this data set)
- 1998: Emergence of *support vector machines (SVM)* topic — this concides with the CFP of 1998, where the key words of SVM first appears
- 1999: SVM-related topic appears again, with co-appearance of new words such as *theorem, proof, conditions, bound*

Concluding remarks

Unsupervised learning methods based on complex models are exciting, but how do we interpret what we have found?

- some theory can help us understand our tools better
- give us confidence in how to use such modeling tools effectively
- more confidence in the inferential findings we obtain
- much work remains to be done in these regards