A Geometric Approach for Learning Latent Mixed Membership Models

Venkatesh Saligrama
Boston University, Boston, MA
(joint work with Weicong Ding & Prakash Ishwar)

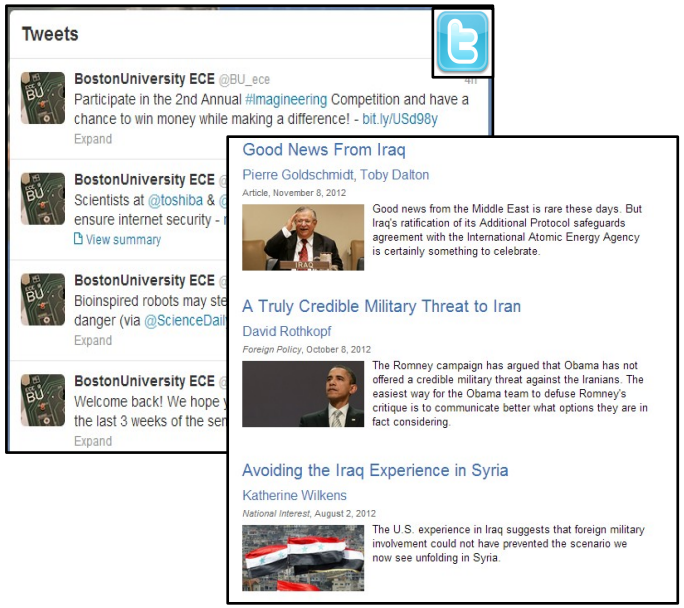
IEEE MLSP 2015

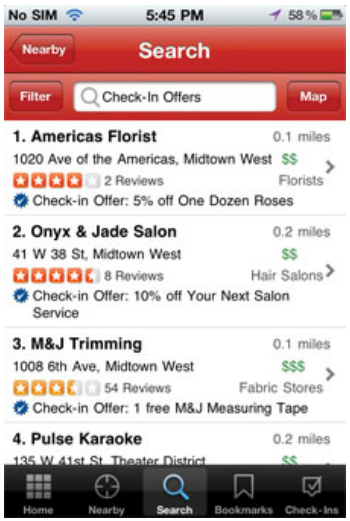
Outline

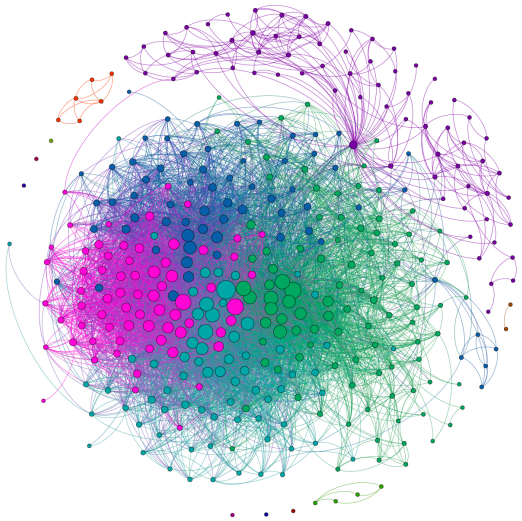
- Examples of Latent Mixture Models
  - Hyperspectral Unmixing, Text Documents, User Preferences, Community Networks, …
- Latent Mixture Model Setup
  - Topic Models
  - Rank Aggregation Models
- Geometric Structure of Latent Mixture Models
  - Inevitability of Approximate Separability & Irreducibility
- Algorithm & Guarantees: Exploiting Geometry
- Real-World Expts.

Mixed membership latent variable model

- Text Docs ← (noisy) mixture of latent topics
- Connections in network ← mixture of latent communities
- User preferences ← mixtures of latent ranking factors







Overall Goal

- Learn/Estimate Latent Factors from Observations(docs)

- Goal: develop algorithms with

  - Provable guarantees
    - How many Docs/Users to estimate Latent Factors within a tolerance?
    - Computational Cost: How does Algorithm scale with #params?

  → ***Model Fidelity***

  - Good empirical performance
    - Real-world datasets

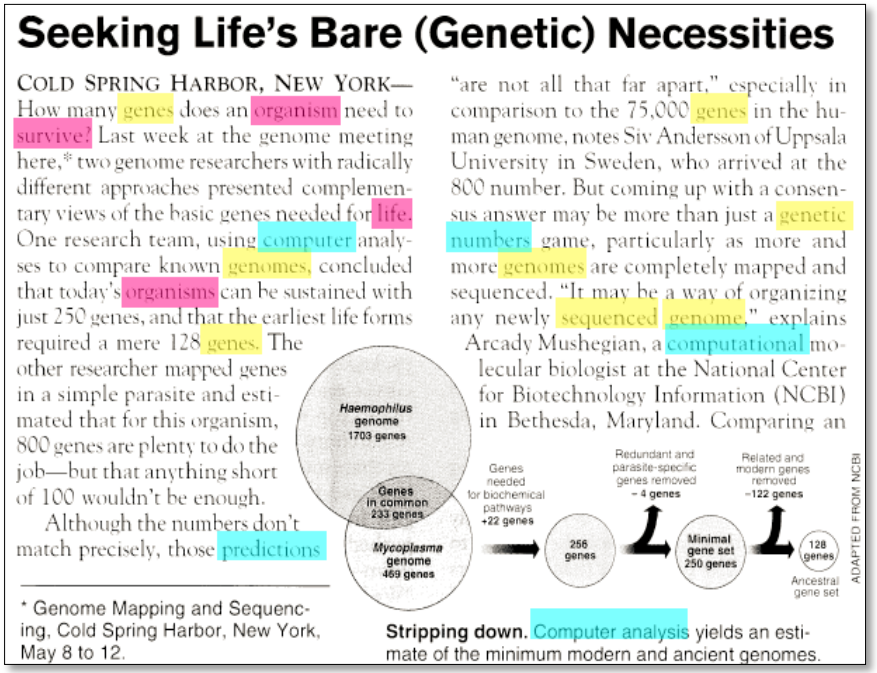
  → ***Web Scale applications***

Outline

- Examples of Latent Mixture Models
  - Text Documents, User Preferences, Community Networks, …
  - Overall Goal/Objective: Algorithm with provable guarantees
- Latent Mixture Model Setup
  - Text Document Topic Models [Blei et. al. 03, …]
  - Rank Aggregation Models
- Geometric Structure of Topic Models
  - Inevitability of Approximate Separability & Irreducibility
- Algorithm & Guarantees: Exploiting Geometry
- Real-World Expts.

"Bag of words" model: a text corpus example

One document in the collection:



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

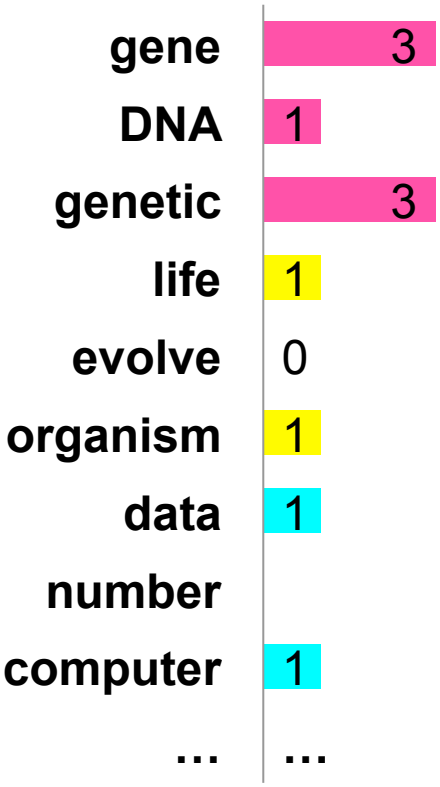
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

words	counts
gene	3
DNA	1
genetic	2
life	1
evolve	1
organism	2
data	1
number	1
computer	3
…	…

Topic Model: Single Document



N IID Words

$$\begin{bmatrix} \text{word 1} \\ \text{word 2} \\ \\ \\ \text{word } W \end{bmatrix} \begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ \vdots \\ A_{W1} \end{bmatrix} = \begin{bmatrix} \text{word 1 DNA} \\ \text{word 2 Data} \\ \\ \\ \text{word } W \end{bmatrix} \begin{bmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \beta_{21} & & \beta_{2K} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \beta_{W1} & & \beta_{WK} \\ \text{topic 1} & & \text{topic } K \end{bmatrix} \begin{bmatrix} \theta_{11} \\ \vdots \\ \theta_{K1} \end{bmatrix}$$

Genetics

Data

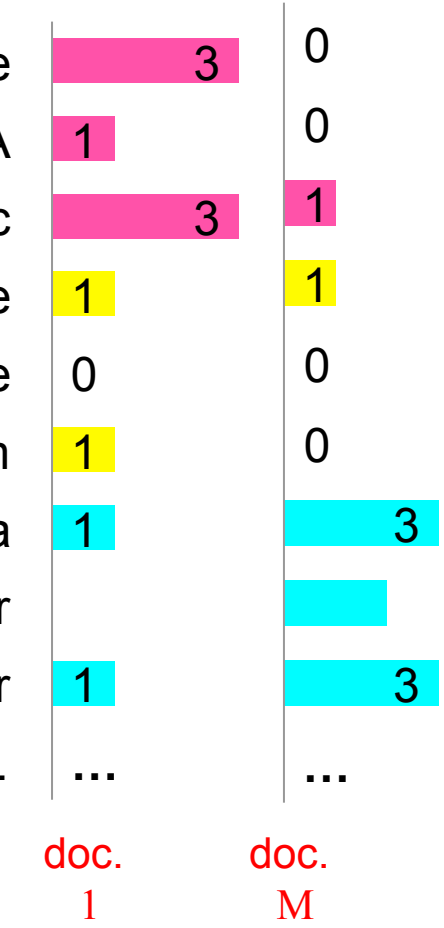
Mixing Wts

Document Distribution

Topic Matrix - $\boldsymbol{\beta}$

- column = topic
- W = vocabulary size
- K = # topics

Topic Model: M-documents



N IID Word/doc

M Documents

$$\begin{bmatrix} A_{11} & A_{1M} \\ A_{21} & A_{2M} \\ \vdots & \vdots \\ A_{W1} & A_{WM} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{1K} \\ \beta_{21} & \beta_{2K} \\ \vdots & \vdots \\ \beta_{W1} & \beta_{WK} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{1M} \\ \vdots & \vdots \\ \theta_{K1} & \theta_{KM} \end{bmatrix}$$

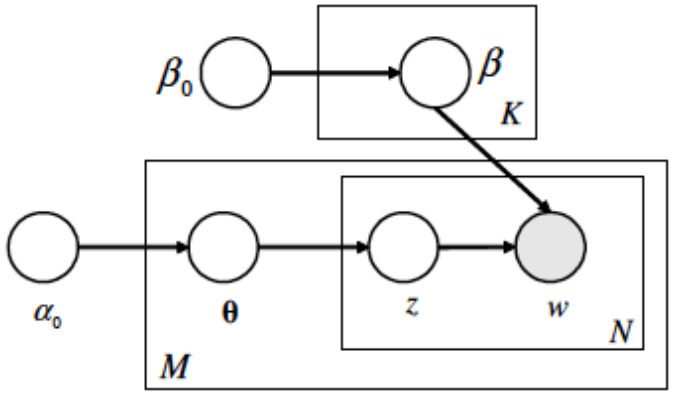
Genetics Data

topic 1 topic K

Observation matrix X

- column = word-freq. of a doc.
- N = # word/doc.

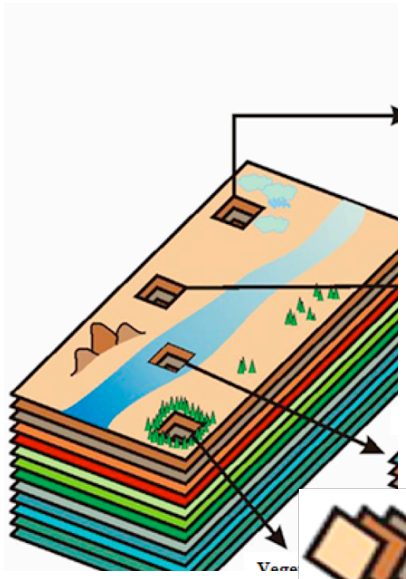
Generative Model [Blei et.al. 2004]

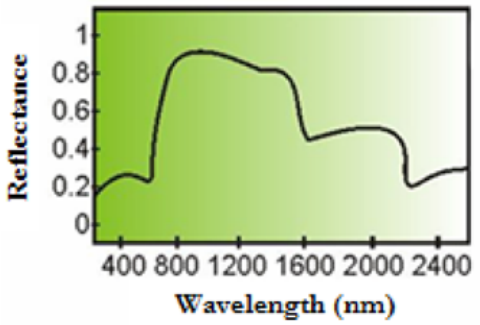


$$\begin{matrix}\text{word } 1 \\ \text{word } 2 \\ \\ \\ \text{word } W\end{matrix}\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1M} \\ A_{21} & A_{22} & & A_{2M} \\ \vdots & \vdots & & \vdots \\ A_{W1} & A_{W2} & & A_{WM} \end{bmatrix} = \begin{matrix}\text{word } 1 \\ \text{word } 2 \\ \\ \\ \text{word } W\end{matrix}\begin{bmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \beta_{21} & & \beta_{2K} \\ \vdots & & \vdots \\ \beta_{W1} & & \beta_{WK} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1M} \\ \vdots & \vdots & & \vdots \\ \theta_{K1} & \theta_{K2} & & \theta_{KM} \end{bmatrix}$$

doc. 1 doc. 2 doc. M | topic 1 topic K | doc. 1 doc. 2 doc. M

$$\begin{matrix}\text{word } 1 \\ \text{word } 2 \\ \\ \\ \text{word } W\end{matrix}\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1M} \\ X_{21} & X_{22} & & X_{2Z} \\ \vdots & \vdots & & \vdots \\ X_{W1} & X_{W2} & & X_{WM} \end{bmatrix}$$



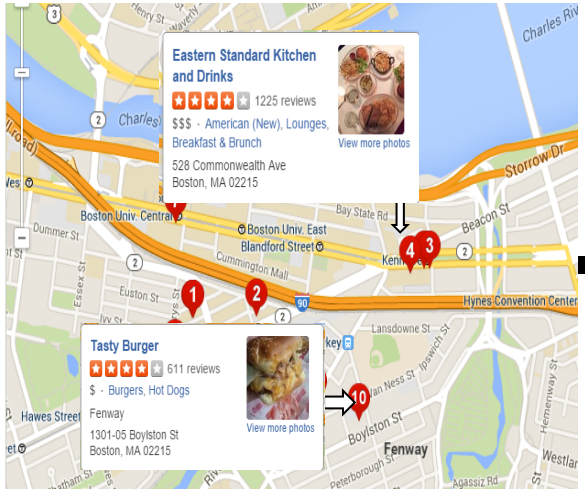


Summary

- Examples
  - Text Documents, User Preferences, Community Networks, …
  - Overall Goal/Objective: Algorithm with provable guarantees
- Latent Mixture Model Setup
  - Text Document Topic Models
  - Rank Aggregation Models [Lu & Boutilier ICML11, Awasthi et al. NIPS14, Oh & Shah NIPS 14, Ding, Ishwar-S 14]
- Geometric Structure of Topic Models
  - Inevitability of Approximate Separability & Irreducibility
- Algorithm & Guarantees: Exploiting Geometry
- Real-World Expts.

Shared Rankings & Pairwise Comparisons

User Observations



check-in/GPS/browsing records

comparisons | counts | Influencing factors

1 — *"$$"*

0 — *"Cuisine"*

1 — *"Popularity"*

User = mixture of latent influencing factors

Ranking Model: M-Users (Noiseless Infl. Factor)

Counts

2

0

4

M Users

N Comparisons

$$\begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ \vdots \\ A_{W1} \end{bmatrix} = \begin{array}{c} \\ (1,2) \\ (1,3) \\ (2,1) \\ (2,3) \\ (3,1) \\ (3,2) \end{array} \begin{array}{c} \text{Rank 2} \quad \text{Rank 3} \\ \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{bmatrix} \theta_{11} \\ \vdots \\ \theta_{K1} \end{bmatrix}$$

Restaurant 1 > 2

$\sigma^3 : 3 > 1 > 2$

Mixing Wts

Price

Popularity

Multiple Rankings

Noisy Influencing Factors

- Sampling: Space of Permutations
  - Binary Matrix ~ Pairwise Comp.
  - Sampled μ_{ij}

$\sigma^1 : 1 > 2 > 3$
$\sigma^2 : 2 > 3 > 1$
$\sigma^3 : 3 > 1 > 2$

Price

Popularity

	Rank 1	Rank 2	Rank 3
(1,2)	1	0	1
(1,3)	1	0	0
(2,1)	0	1	0
(2,3)	1	1	0
(3,1)	0	1	1
(3,2)	0	0	1

ϕ=0

- Mallows Model (1957) ~ K baselines

σ^1 2 3 6 ..... Q

$6 > 1$

σ 6 3 1 ..... Q

$$Prob\{\sigma \mid \sigma_j\} \propto \phi^{dist(\sigma,\sigma_j)}$$

$$Prob\{1 > 2 \mid \sigma^1, \phi^1\} \rightarrow \begin{bmatrix} \beta_{11} & \beta_{1K} \\ \beta_{21} & \beta_{2K} \\ \vdots & \vdots \\ \beta_{W1} & \beta_{WK} \end{bmatrix}$$

Inconsistent & Heterogeneous User Model

$$\begin{matrix}(1,2)\\(1,3)\\\\\\(Q\text{-}1.Q)\end{matrix}\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1M} \\ A_{21} & A_{22} & \cdots & A_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ A_{W1} & A_{W2} & \cdots & A_{WM} \end{bmatrix} = \begin{matrix}(1,2)\\(1,3)\\\\\\(Q\text{-}1.Q)\end{matrix}\begin{bmatrix} \beta_{11} & \beta_{1K} \\ \beta_{21} & \beta_{2K} \\ \vdots & \vdots \\ \vdots & \vdots \\ \beta_{W1} & \beta_{WK} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1M} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{K1} & \theta_{K2} & \cdots & \theta_{KM} \end{bmatrix}$$

Price (Rank 1) — Popularity (Rank K)

user 1 — user 2 — user M

user — user 2 — user M

$$\begin{matrix}(1,2)\\(1,3)\\\\\\(Q\text{-}1.Q)\end{matrix}\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1M} \\ X_{21} & X_{22} & \cdots & X_{2Z} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ X_{W1} & X_{W2} & \cdots & X_{WM} \end{bmatrix}$$

Observation matrix X

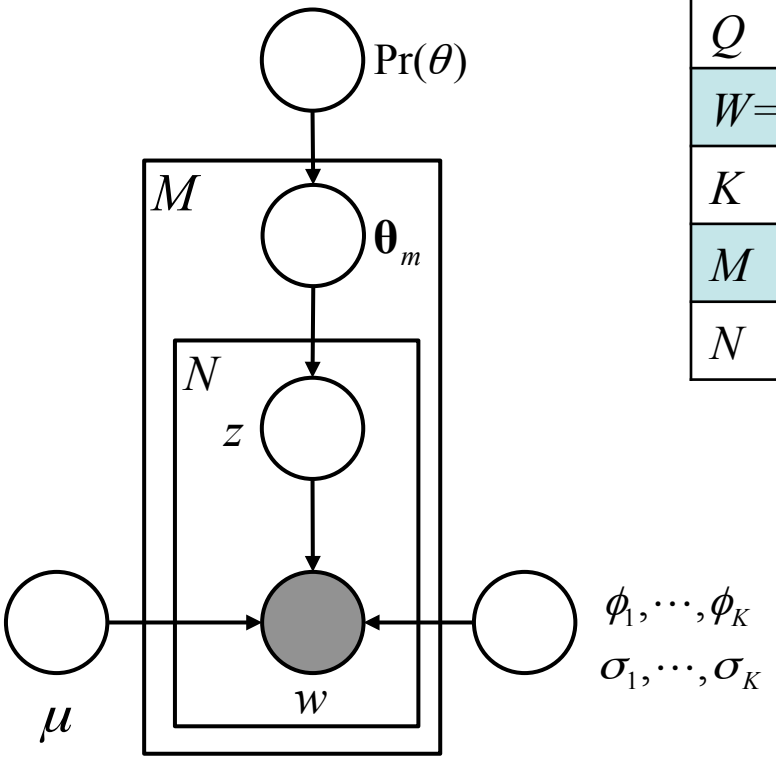
- column = pairwise counts of a user
- # times user m prefers i over j
- N = # comparisons/user.

Deterministic Case: NP Hard

Recover β, θ from A

Rank Aggregation Problem is a Topic Model

"Plate" representation



w: comparisons (i, j)

Key parameters

Q	# items
$W=Q(Q-1)$	# ordered pairs
K	# latent rankings
M	# users
N	# comps./user

➔Words
➔Topics
➔Documents

[Ding, Ishwar, S'14]

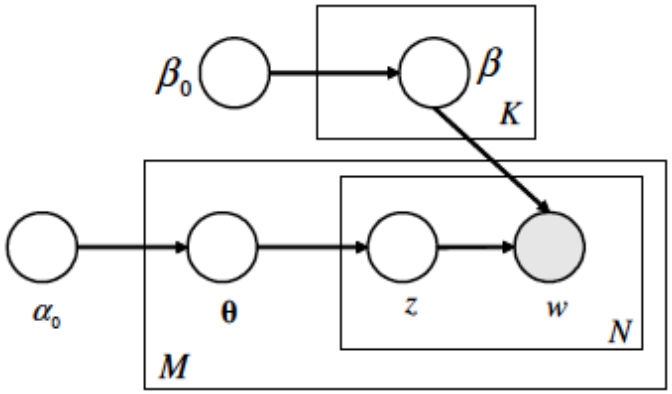
Summary

- Examples
  - Text Documents, User Preferences, Community Networks, …
  - Overall Goal/Objective: Algorithm with provable guarantees
- Latent Mixture Model Setup
  - Text Document Topic Models
  - Rank Aggregation Models [Lu & Boutilier ICML11, Awasthi et al. NIPS14, Oh & Shah NIPS 14, Ding, Ishwar-S 14]
- Geometric Structure of Topic Models
  - Inevitability of Approximate Separability & Irreducibility
- Algorithm & Guarantees: Exploiting Geometry
- Real-World Expts.

Basic Topic Modeling Approaches

Topic matrix Weight matrix

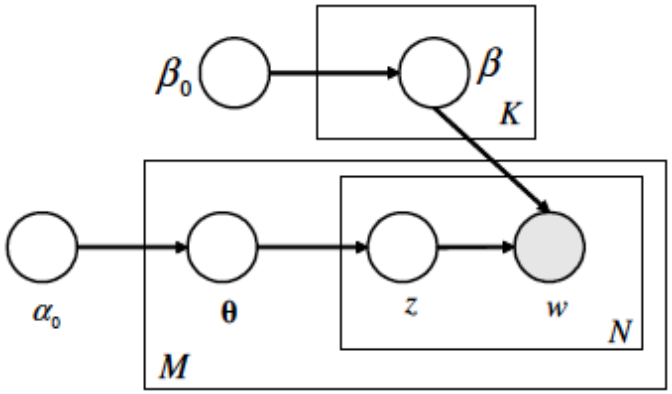
Method	β	θ	Approach	Issues
Nonnegative Matrix Factorization (NMF), e.g., [*Cichocki et al.,'09*]	Deterministic	Deterministic	Regularized joint optimization	NP Hard (Arora'12) Non-convex. Need approximations and heuristics.
"Bayesian Methods" e.g., LDA, CTM [*Blei et al.,'03*],	Prior	Prior	MAP or ML	Non-convex. Need approximations like MCMC. No Guarantees



$$\begin{bmatrix} X_{11} & X_{12} & & X_{1M} \\ X_{21} & X_{22} & & X_{2Z} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ X_{W1} & X_{W2} & & X_{WM} \end{bmatrix} \sim \begin{bmatrix} \beta_{11} & \beta_{1K} \\ \beta_{21} & \beta_{2K} \\ \vdots & \cdots \\ \vdots & \vdots \\ \beta_{W1} & \beta_{WK} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & & \theta_{1M} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{K1} & \theta_{K2} & & \theta_{KM} \end{bmatrix}$$

Bayesian Topic Models are Approximately Separable

[Ding-Ishwar-S'14]



$$\begin{bmatrix} X_{11} & X_{12} & & X_{1M} \\ X_{21} & X_{22} & & X_{2Z} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ X_{W1} & X_{W2} & & X_{WM} \end{bmatrix} \sim \begin{bmatrix} \beta_{11} & & \beta_{1K} \\ \beta_{21} & & \beta_{2K} \\ \vdots & \cdots & \vdots \\ \vdots & & \vdots \\ \beta_{W1} & & \beta_{WK} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & & \theta_{1M} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{K1} & \theta_{K2} & & \theta_{KM} \end{bmatrix}$$

$$\begin{matrix} w1 \\ w2 \\ w3 \\ w4 \\ w5 \end{matrix} \begin{bmatrix} \beta_1 & \lambda\beta_1 & \lambda\beta_1 \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \lambda\beta_3 & \beta_3 & \lambda\beta_3 \\ \beta_{41} & \beta_{42} & \beta_{43} \\ \lambda\beta_6 & \lambda\beta_6 & \beta_6 \end{bmatrix}$$

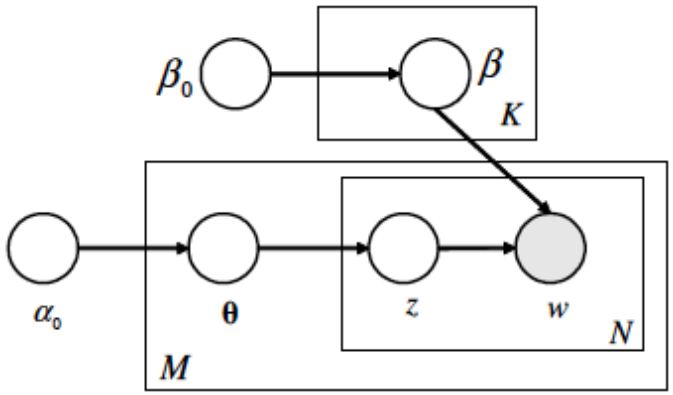
Approximately Separable & High-Dimensionality

- In real-world problems
  - Size of vocab. W >> #. Topics K

Dataset	W	K
Wikipedia	109,611	50
Twitter	122,035	50
New York Times	102,660	100
PubMed	141,043	150

- Main result: Separability is an inevitable consequence of high-dimensionality!
  - Satisfied in estimates produced by NMF, LDA, and other algorithms (Bayesian Models)

Generative Model



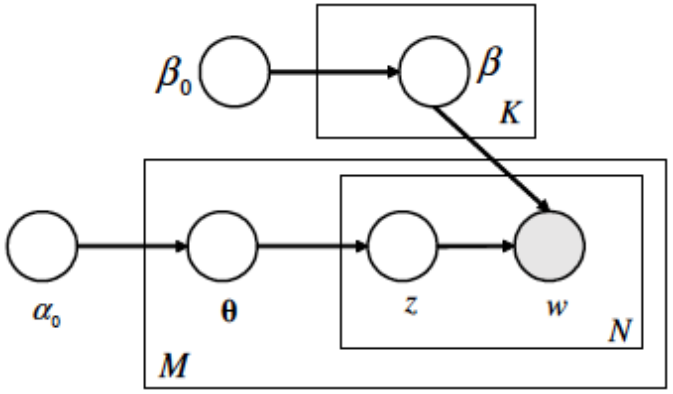
Why is Separability inevitable for W>> K?

- Theorem: Suppose $W \geq tKe^{\beta_0 K \log(K+1/\lambda)}$

$$\text{Prob}\{\beta \text{ not } \lambda - sep\} = O(W^{-t})$$

Dataset	W	K
Wikipedia	109,611	50
Twitter	122,035	50
New York Times	102,660	100
PubMed	141,043	150

Generative Model



Separability in Practice

Dataset	Vocab. size W	# Topics K	Prob. 0.01-separable
NIPS	12,419	50	100±0%
Wikipedia	109,611	50	99.9±0.3%
Twitter	122,035	50	100±0%
New York Times	102,660	100	99.6±0.6%
PubMed	141,043	150	99.9±0.3%

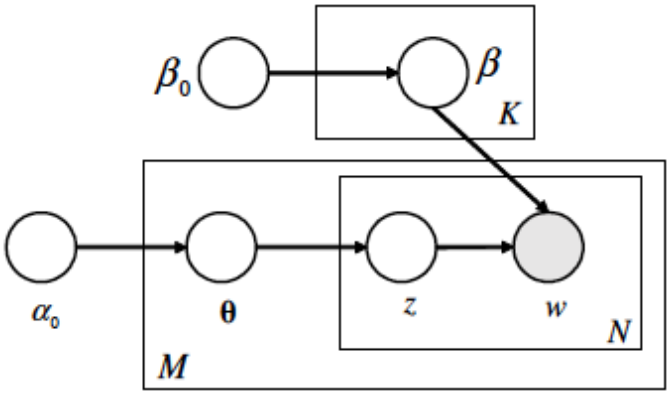
$\beta_0 = 0.01$,
1000 MC runs

- β_0 moderately small positive value in practice.
  - β_0 = 0.01 for K ∈ [50, 200]
- Some packages suggest β_0 = c/W to get satisfactory empirical results.

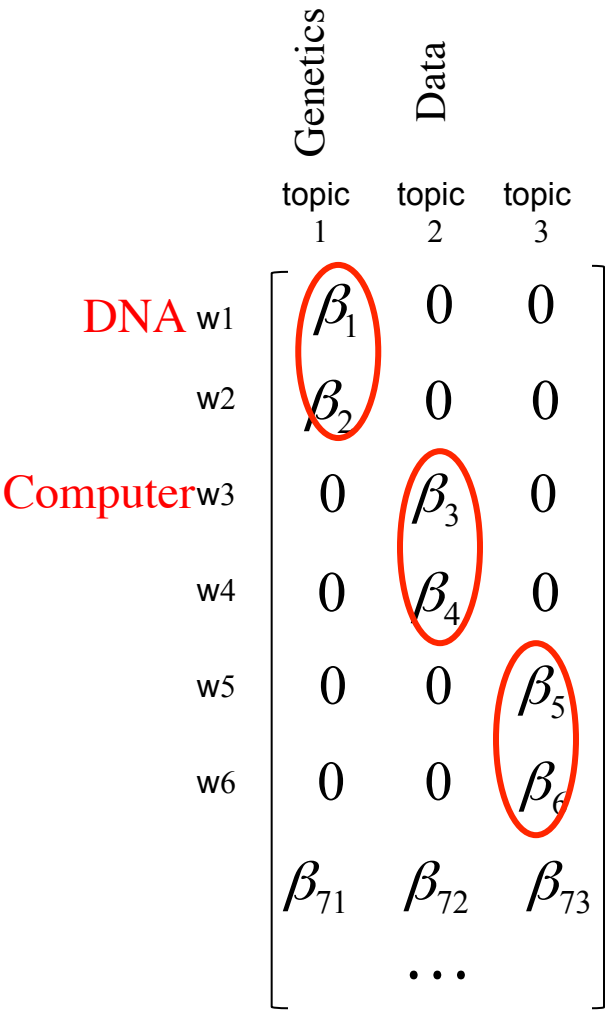
$$W \geq tKe^{\beta_0 K \log(K+1/\lambda)}$$

- Analysis explains reasoning for this choice!

Generative Model



What does Approximately Separable Topic Matrix mean?

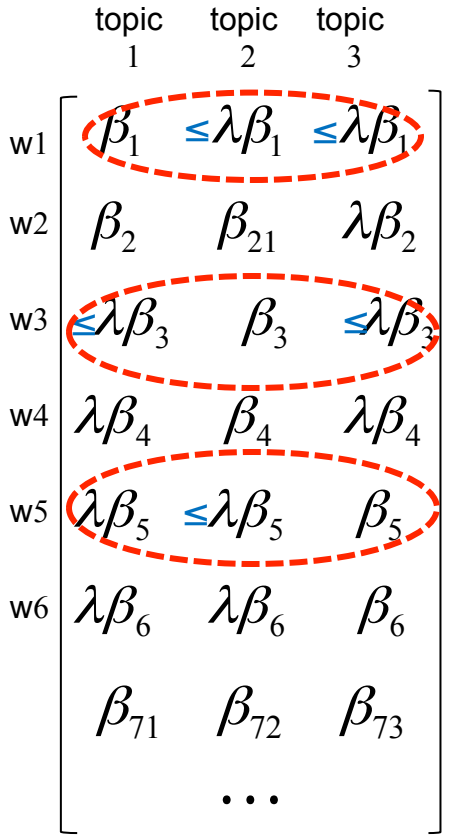


Separable Topic Matrix

λ-approximately separable if one word for each topic is predominantly unique

λ = 0 Case: Novel Word(s) **unique** to each topic

[Boardman'93, Donoho'04], [Arora'13, Ding et. al.'13]



Approximately Separable

[Ding, Ishwar-S 15]

HOW ABOUT RANK AGGREGATION MODELS ?

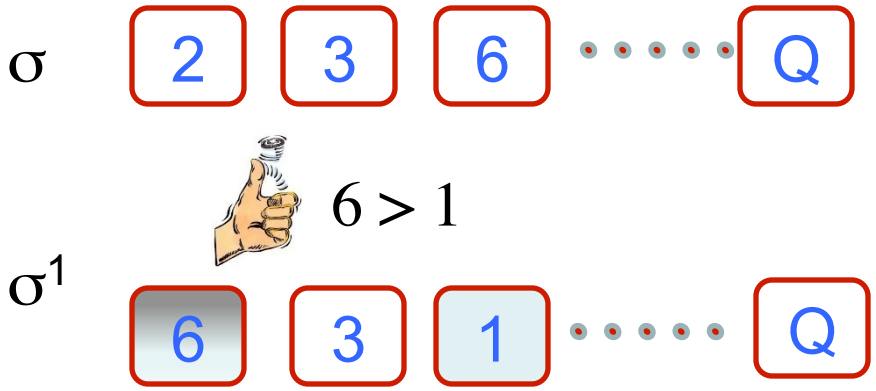
Mallows Rank Aggregation Model is Separable

Obs. Comparisons

$$\begin{bmatrix} 1 & 0.25 \\ 0 & 0.75 \\ \vdots & \vdots \\ 1 & 0.90 \end{bmatrix} \sim \begin{matrix} (1,2) \\ (1,3) \\ \\ \\ (Q\text{-}1.Q) \end{matrix} \begin{bmatrix} \beta_{11} & \beta_{1K} \\ \beta_{21} & \beta_{2K} \\ \vdots & \vdots \\ \beta_{W1} & \beta_{WK} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & & \theta_{1M} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{K1} & \theta_{K2} & & \theta_{KM} \end{bmatrix}$$

Price (Rank 1) — Popularity (Rank K)

user 1 — user 2 — user M

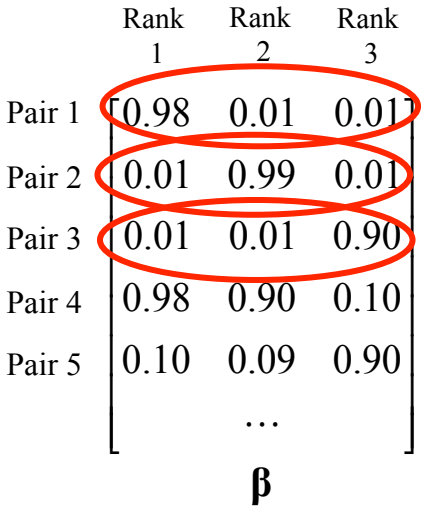


Mallow Rank Aggregation Model Is Approx Separable

Mallows (Noisy Factors) are Approx Separable

- Most ranking matrix are λ-Approximate separable, # items Q >> # factors K

$$\Pr(\boldsymbol{\sigma} \text{ is } \lambda\text{-separable}) \geq 1 - K\exp(-QL(\lambda;\phi)^{-2K+1})$$



Approximately Separable ranking matrix

ϕ	Prob. of 0.05-separable
0.0	93.3%
0.1	87.0%
0.2	79.3%
0.5	42.6%

$Q = 100$

$K = 10$

1000 Monte Carlo runs

Summary

- Latent Mixture Models
  - Many Scenarios: Document Models, Rank Aggregation Models
- Bayesian Setup
  - Latent Factors are approximately Separable
- How to exploit approximate separability?

Key Idea ($\lambda = 0$, No-Noise)

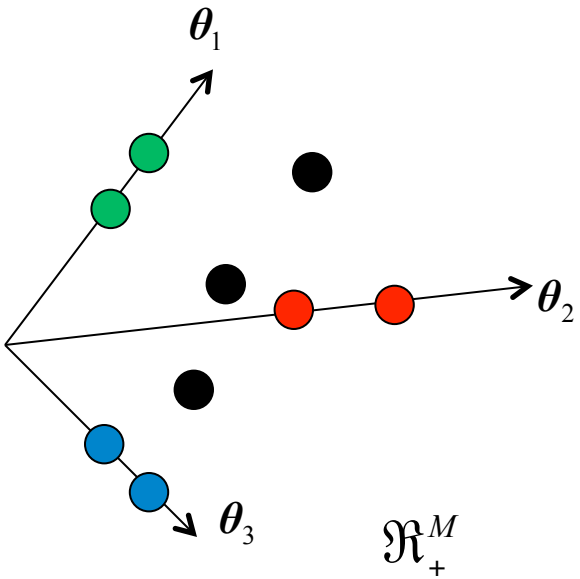
$$\boldsymbol{\beta}_{W\times K}\boldsymbol{\theta}_{K\times M} = \mathbf{A}_{W\times M}$$

	topic 1	topic 2	topic 3
w1	β_1	0	0
w2	β_2	0	0
w3	0	β_3	0
w4	0	β_4	0
w5	0	0	β_5
w6	0	0	β_6
	β_{71}	β_{72}	β_{73}
	...		

Separable
Topic Matrix $\boldsymbol{\beta}$

doc. 1	...	doc. M
$\leftarrow$	$\boldsymbol{\theta}_1$	$\rightarrow$
$\leftarrow$	$\boldsymbol{\theta}_2$	$\rightarrow$
$\leftarrow$	$\boldsymbol{\theta}_3$	$\rightarrow$

Weight Matrix



doc. 1	...	doc. M
$\leftarrow$	$\beta_1\boldsymbol{\theta}_1$	$\rightarrow$
$\leftarrow$	$\beta_2\boldsymbol{\theta}_1$	$\rightarrow$
$\leftarrow$	$\beta_3\boldsymbol{\theta}_2$	$\rightarrow$
$\leftarrow$	$\beta_4\boldsymbol{\theta}_2$	$\rightarrow$
$\leftarrow$	$\beta_5\boldsymbol{\theta}_3$	$\rightarrow$
$\leftarrow$	$\beta_6\boldsymbol{\theta}_3$	$\rightarrow$
$\beta_{71}\boldsymbol{\theta}_1 + \beta_{72}\boldsymbol{\theta}_2 + \beta_{73}\boldsymbol{\theta}_3$		
...		

Distribution Matrix A

Key Idea ($\lambda = 0$ case)

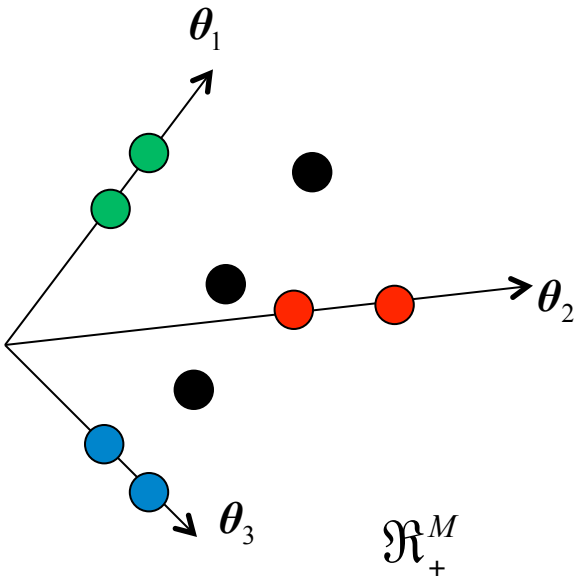
$$\boldsymbol{\beta}_{W\times K}\boldsymbol{\theta}_{K\times M} = \mathbf{A}_{W\times M}$$

topic 1 | topic 2 | topic 3 | doc. 1 | ... | doc. M

$$\begin{array}{c} \\ w1 \\ w2 \\ w3 \\ w4 \\ w5 \\ w6 \\ \\ \\ \end{array}\begin{bmatrix} \beta_1 & 0 & 0 \\ \beta_2 & 0 & 0 \\ 0 & \beta_3 & 0 \\ 0 & \beta_4 & 0 \\ 0 & 0 & \beta_5 \\ 0 & 0 & \beta_6 \\ \beta_{71} & \beta_{72} & \beta_{73} \\ & \cdots & \end{bmatrix} \quad \begin{bmatrix} \leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_3 & \rightarrow \end{bmatrix}$$

Weight Matrix

Separable
Topic Matrix $\boldsymbol{\beta}$



doc. 1 | ... | doc. M

$$\begin{bmatrix} \leftarrow & \boldsymbol{\beta}_1\boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\beta}_2\boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\beta}_3\boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & \boldsymbol{\beta}_4\boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & \boldsymbol{\beta}_5\boldsymbol{\theta}_3 & \rightarrow \\ \leftarrow & \boldsymbol{\beta}_6\boldsymbol{\theta}_3 & \rightarrow \\ \beta_{71}\boldsymbol{\theta}_1 + \beta_{72}\boldsymbol{\theta}_2 + \beta_{73}\boldsymbol{\theta}_3 & & \\ \cdots & & \end{bmatrix}$$

Distribution Matrix A

Key Idea

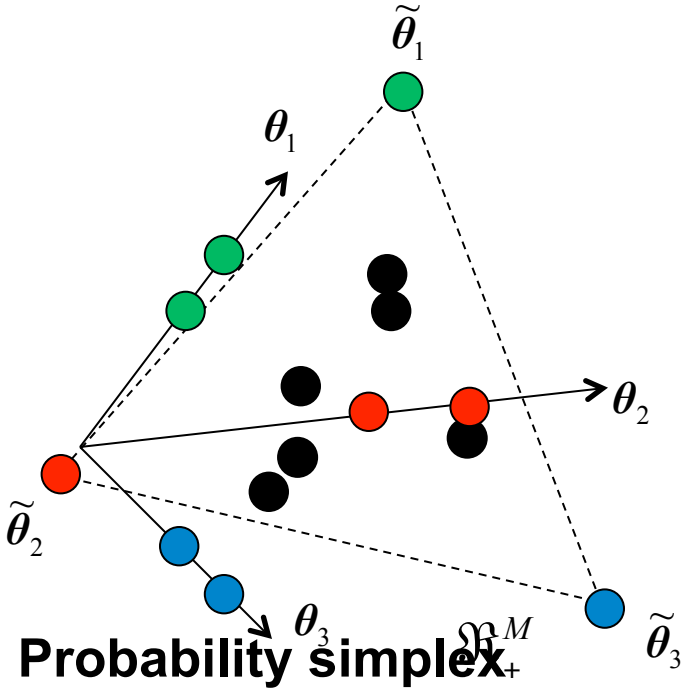
$$\widetilde{\beta}_{W\times K}\widetilde{\theta}_{K\times M} = \widetilde{A}_{W\times M}$$

Novel Word = Extreme Point

Novel word detection + Topic matrix estimation

$$\begin{bmatrix}
\text{doc. } 1 & \cdots & \text{doc. } M \\
\leftarrow & \beta_1\theta_1 & \rightarrow \\
\leftarrow & \beta_2\theta_1 & \rightarrow \\
\leftarrow & & \rightarrow \\
\leftarrow & \beta_4\theta_2 & \rightarrow \\
\leftarrow & \beta_5\theta_3 & \rightarrow \\
\leftarrow & \beta_6\theta_3 & \rightarrow \\
\beta_{71}\theta_1 + \beta_{72}\theta_2 + \beta_{73}\theta_3 & & \\
\cdots & &
\end{bmatrix}$$

Distribution Matrix A



Probability simplex

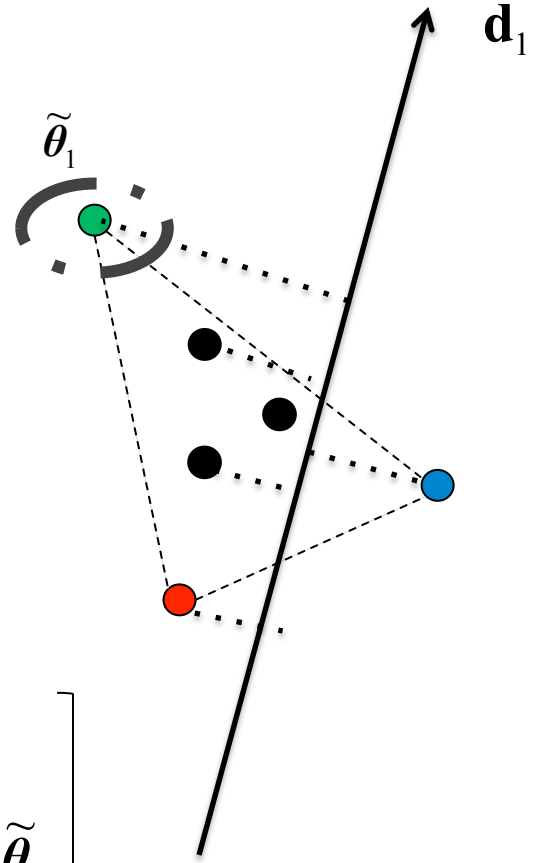
$$\begin{bmatrix}
\text{doc. } 1 & \cdots & \text{doc. } M \\
\leftarrow & \widetilde{\theta}_1 & \rightarrow \\
\leftarrow & \widetilde{\theta}_1 & \rightarrow \\
\leftarrow & & \rightarrow \\
\leftarrow & \widetilde{\theta}_2 & \rightarrow \\
\leftarrow & \widetilde{\theta}_3 & \rightarrow \\
\leftarrow & \widetilde{\theta}_3 & \rightarrow \\
\widetilde{\beta}_{71}\widetilde{\theta}_1 + \widetilde{\beta}_{72}\widetilde{\theta}_2 + \widetilde{\beta}_{73}\widetilde{\theta}_3 & & \\
\cdots & &
\end{bmatrix}$$

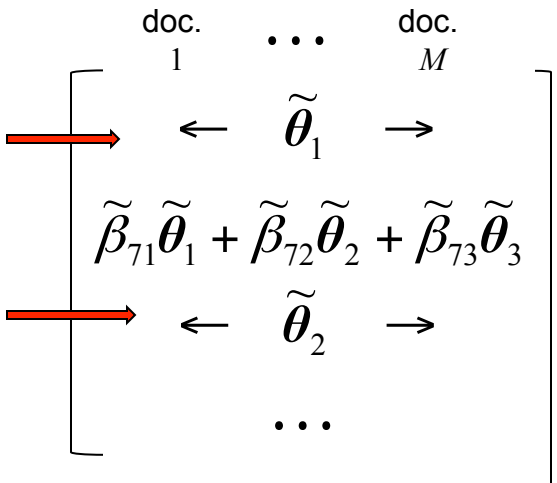
Row Normalized Distribution Matrix $\widetilde{A}$

Detect Novel Words via Projections

- Max/Min of projection → extreme points of convex hull
- Which directions to use

  →Generate P iid Isotropy directions
- Pick words associated with maximum → Recover extreme points from $\widetilde{A}$

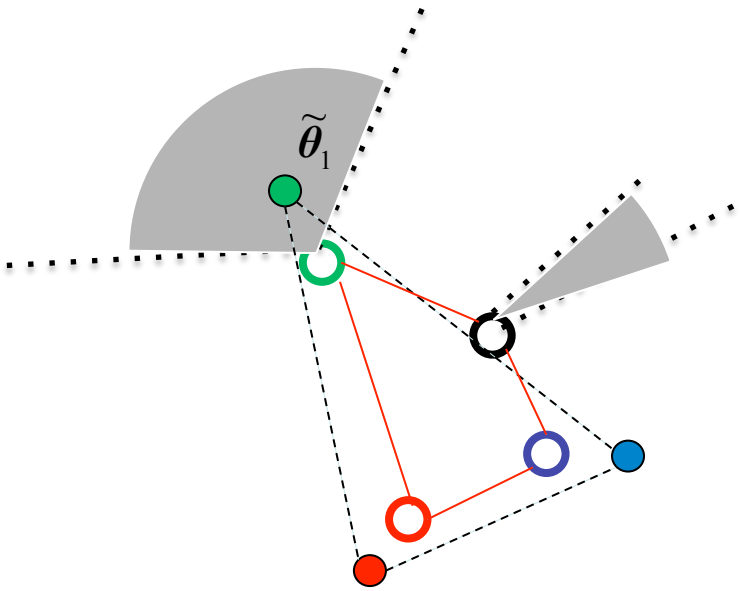




Extension to Approximately Separable ($\lambda > 0$)

- Approx. novel words ⇔ larger solid angles

- Solution
  - → Sort solid angles
  - → Take the top – K extreme points
  - → Freq. of maximum ≈ Solid Angle of an Extreme Point



Finite words/doc.

$$\tilde{\beta}_{W\times K}\tilde{\theta}_{K\times M} = \tilde{A}_{W\times M} \approx \tilde{X}_{W\times M}$$

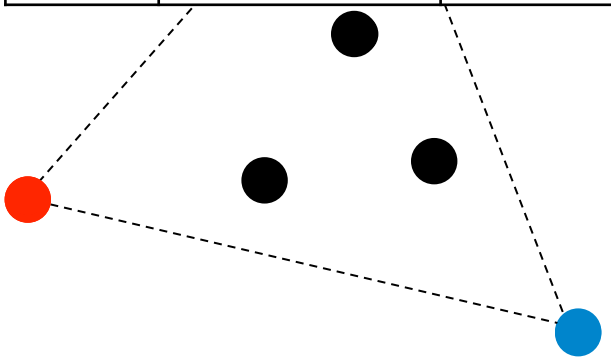
doc. 1 ... doc. M

$$\begin{bmatrix}
\leftarrow \tilde{\theta}_1 \rightarrow \\
\leftarrow \tilde{\theta}_1 \rightarrow \\
\leftarrow \tilde{\theta}_2 \rightarrow \\
\leftarrow \tilde{\theta}_2 \rightarrow \\
\leftarrow \tilde{\theta}_3 \rightarrow \\
\leftarrow \tilde{\theta}_3 \rightarrow \\
\tilde{\beta}_{71}\tilde{\theta}_1 + \tilde{\beta}_{72}\tilde{\theta}_2 + \tilde{\beta}_{73}\tilde{\theta}_3 \\
\cdots
\end{bmatrix}$$

Row Normalized Distribution Matrix $\tilde{A}$

Key issue:
N fixed → perturbation does not vanish

K	# topics	~100
W	vocab. size	~10k
N	#word/doc.	~100
M	# doc.	~100k



Probability simplex

doc. 1 ... doc. M

$$\begin{bmatrix}
\leftarrow \tilde{X}_1 \rightarrow \\
\leftarrow \tilde{X}_2 \rightarrow \\
\leftarrow \tilde{X}_3 \rightarrow \\
\leftarrow \tilde{X}_4 \rightarrow \\
\leftarrow \tilde{X}_5 \rightarrow \\
\leftarrow \tilde{X}_6 \rightarrow \\
\leftarrow \tilde{X}_7 \rightarrow \\
\cdots
\end{bmatrix}$$

Row Normalized Observation Matrix: $\tilde{X}$

$$M\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T \xrightarrow[M\to\infty]{a.s} \mathbf{E}_{W\times W}$$

Main result: Latent Mixture Models

- Computational complexity :

$$O(MNK + WK + WK^3)$$

K	# topics	~100
W	vocab. size	~10k
N	#word/doc.	~100
M	# doc.	~100k

- Sample complexity :

Under the **Simplicial Condition** on $\boldsymbol{R}$', with $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{I}_W)$, the proposed Random Projection algorithm can detect all novel words of K topics with probability $1\text{-}\delta$ if

$$M \geq \text{Poly}\left(W, \log\left(\frac{1}{\delta}\right), K, \frac{1}{N}\right), P \geq \text{Poly}\left(\log W, \log\left(\frac{1}{\delta}\right), K\right)$$

Moreover, if R is **full-rank**, can recover $\boldsymbol{\beta}$ with ε element-wise error with probability $1\text{-}\delta$.

Summary

- Latent Mixture Models
  - Text Documents, User Preferences, Community Networks, …
- Topic Models & Estimation Problem
  - Related Work
- Geometric Structure of Topic Models
  - Inevitability of Separability in high-dimensions
- Algorithm & Guarantees: Exploiting Geometry
  - Efficient Extreme Points Identification
- Empirical Results on Real-World Datasets
  - Text-Document Models
  - Rank Aggregation

Experimental Results (semi-synthetic data)

Real-world corpus
New York Times articles

→ Topic matrix learnt by Gibbs sampling

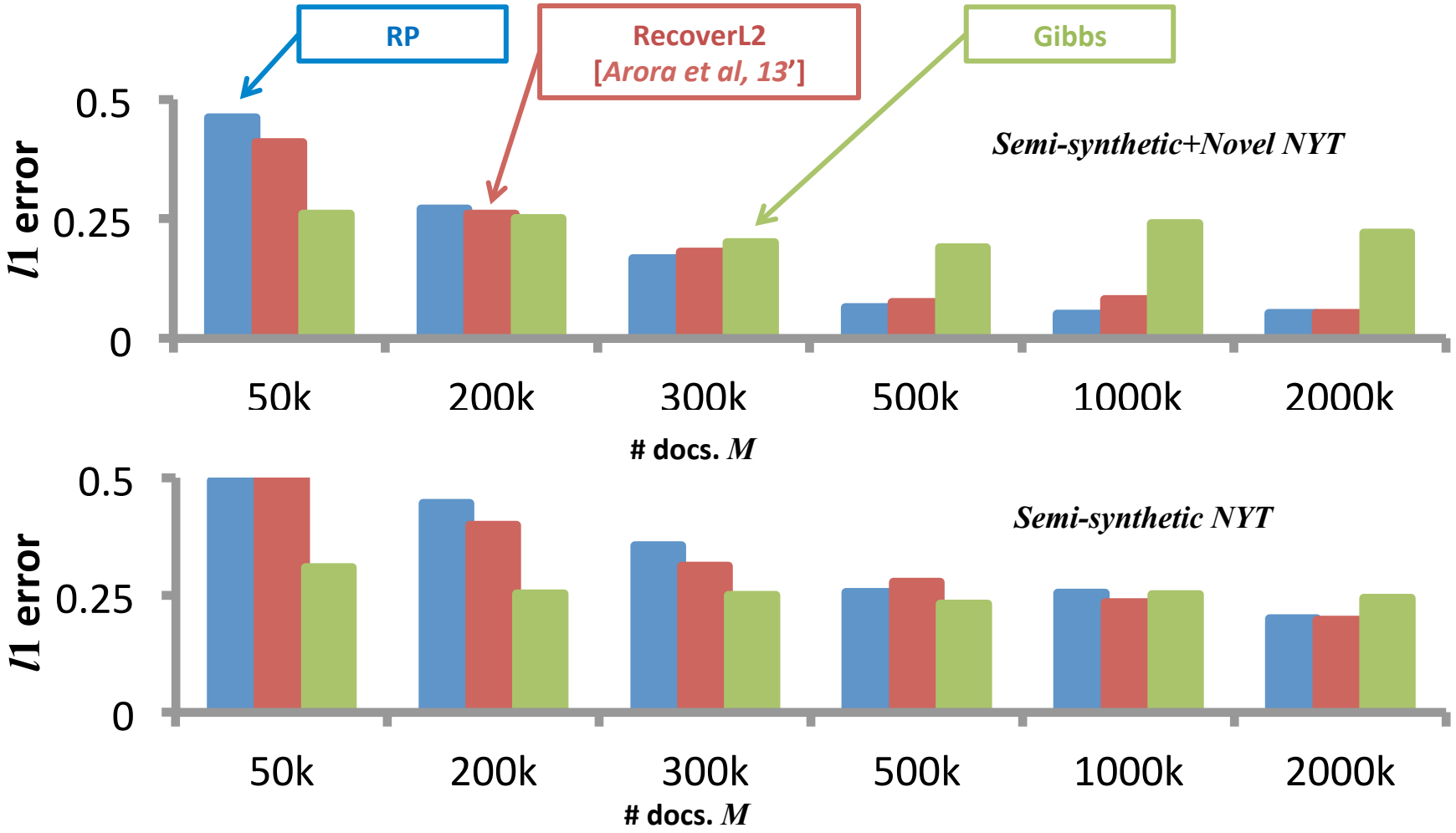
→ Generate synthetic docs. with Dirichlet prior

→ Add artificial novel words; Generate synthetic docs.

M	300,000
N	300
W	14943
K	100
L	200

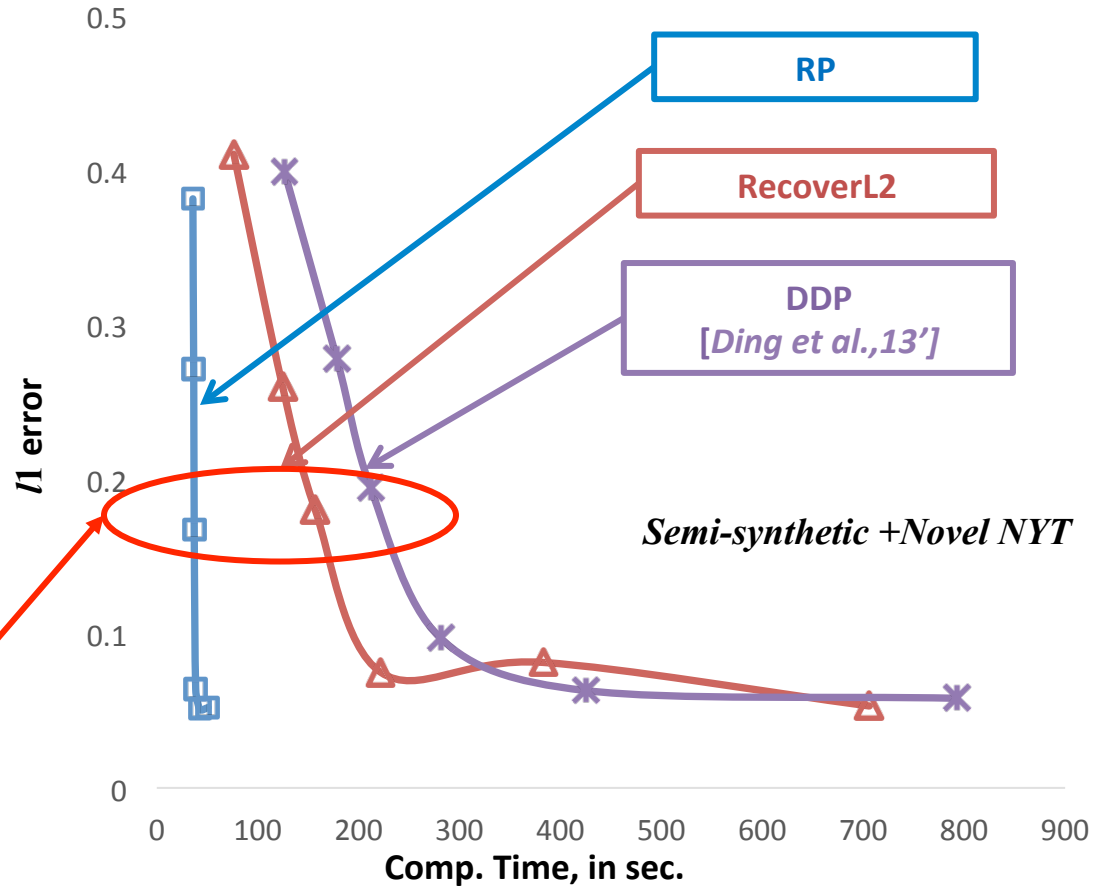
- Semi-synthetic data can resemble the dimensionality and sparsity of real-world data

Experimental Results (semi-synthetic data)



Experimental Results (semi-synthetic data)

	Semi-syn +novel NYT	*Semi-syn* NYT
M	Variable	Variable
N	300	300
W	15043 words, 100 novel	14943
K	100	100
L	200	200



* T (Gibbs) ~ 6918 sec

Experimental Results (Real-World Text Corpus)

New York Times dataset

M	300,000 = 240k train + 60k test
N	300 words/doc. (avg.)
W	14,943
K	50/100/150 topics

Decreasing Prob. ↓

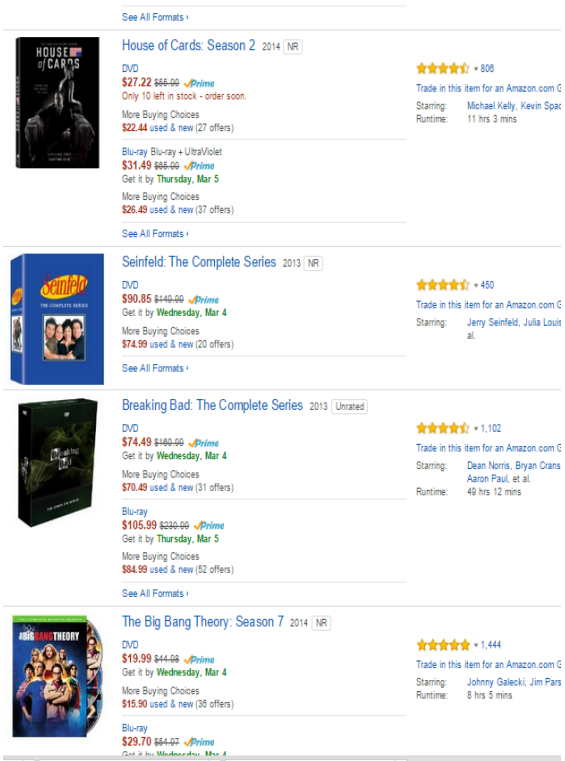
"Weather"	"Emotion"	"Politics"	"Football"
Weather	Feeling	Election	Yard
Wind	Sense	*Florida*	Game
Air	Love	Ballot	Team
Storm	Character	Vote	Season
Rain	Heart	*Al_gore*	Play
Cold	Emotion	Recount	*NFL*

(See Ding et al., '13 for more example topics)

MovieLens Dataset: Ratings → Comparisons

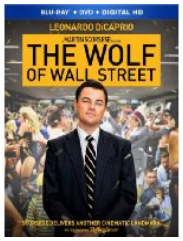
6000 Users, 3000 Rated Movies

User preferences



words counts

Influencing factors



>

1

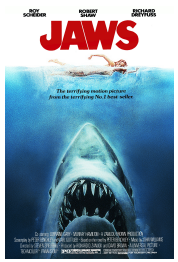
"actor"

>

0

"music"

>



1

"special effect"

MovieLens dataset – predicting star ratings

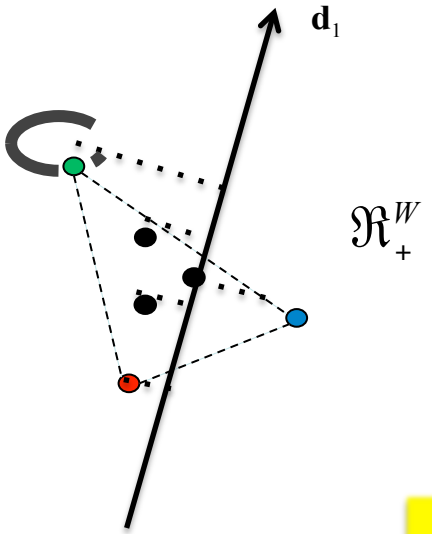
- Predict star ratings via ranking models
  - Generate comparisons from training ratings
  - Learn mixed membership Mallows model with Dirichlet prior
  - For each test movie predict most-likely star rating.

- Measure: RMSE of estimated star ratings

K	PMF	BPMF	BPMF-int	TM(Ding et al.,14)	MMMM
10	1.0491	0.8254	0.8723	0.8840	0.8509
15	0.9127	0.8236	0.8734	0.8780	0.8296
20	0.9250	0.8213	0.8678	0.8721	0.8241

Rating based models

Conclusions



High-D Latent Factor Models
Geometry ~ Approx Sep.

Simple geometric picture

Efficient randomized algorithm

Consistency, efficiency, state-of-the-art performance

	Rank 1	Rank 2	Rank 3
Pair 1	0.98	0.01	0.01
Pair 2	0.01	0.99	0.01
Pair 3	0.01	0.01	0.90
Pair 4	0.98	0.90	0.10
Pair 5	0.10	0.09	0.90
		…	

$\boldsymbol{\beta}$

Approximately Separable ranking matrix

We are looking for interested students & post-docs
Contact: srv@bu.edu (sites.bu.edu/data)