

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 \leftarrow (noisy) mixture of latent topics
- Connections in network \leftarrow mixture of latent communities
- User preferences \leftarrow mixtures of latent ranking factors

Tweets

BostonUniversity ECE @BU_ece
Participate in the 2nd Annual #imagineering Competition and have a chance to win money while making a difference! - bit.ly/USd98y
Expand

BostonUniversity ECE @BU_ece
Scientists at @toshiba & @ ensure internet security -
View summary

BostonUniversity ECE @BU_ece
Biospired robots may sta danger (via @ScienceDai
Expand

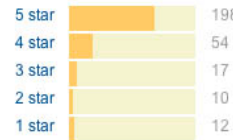
BostonUniversity ECE @BU_ece
Welcome back! We hope y the last 3 weeks of the ser
Expand

Good News From Iraq
Pierre Goldschmidt, Toby Dalton
Article, November 8, 2012
Good news from the Middle East is rare these days. But Iraq's ratification of its Additional Protocol safeguards agreement with the International Atomic Energy Agency is certainly something to celebrate.

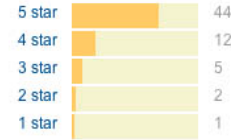
A Truly Credible Military Threat to Iran
David Rothkopf
Foreign Policy, October 8, 2012
The Romney campaign has argued that Obama has not offered a credible military threat against the Iranians. The easiest way for the Obama team to defuse Romney's critique is to communicate better what options they are in fact considering.

Avoiding the Iraq Experience in Syria
Katherine Wilkens
National Interest, August 2, 2012
The U.S. experience in Iraq suggests that foreign military involvement could not have prevented the scenario we now see unfolding in Syria.

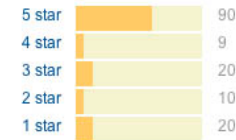
★★★★★ (291)
4.4 out of 5 stars



★★★★★ (64)
4.5 out of 5 stars



★★★★★ (149)
3.9 out of 5 stars



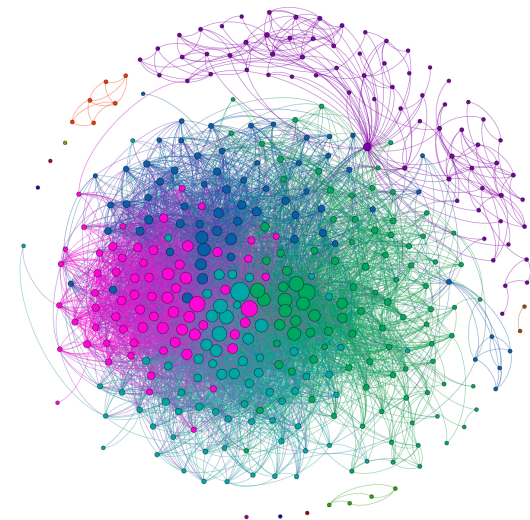
No SIM 5:45 PM 58%

Nearby Search

Filter Check-In Offers Map

- Americas Florist** 0.1 miles
1020 Ave of the Americas, Midtown West
2 Reviews Florists
Check-in Offer: 5% off One Dozen Roses
- Onyx & Jade Salon** 0.2 miles
41 W 38 St, Midtown West
8 Reviews Hair Salons
Check-in Offer: 10% off Your Next Salon Service
- M&J Trimming** 0.1 miles
1008 6th Ave, Midtown West
54 Reviews Fabric Stores
Check-in Offer: 1 free M&J Measuring Tape
- Pulse Karaoke** 0.2 miles
135 W 41st St, Theater District

Home Nearby Search Bookmarks Check-Ins



Overall Goal

- Learn/Estimate Latent Factors from Observations(docs)
- Goal: develop algorithms with
 - **Provable** guarantees → **Model Fidelity**
 - How many Docs/Users to estimate Latent Factors within a tolerance?
 - Computational Cost: How does Algorithm scale with #params?
 - Good **empirical** performance → **Web Scale applications**
 - Real-world datasets

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

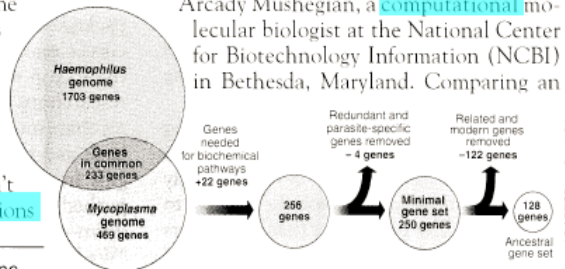
One document in the collection:

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

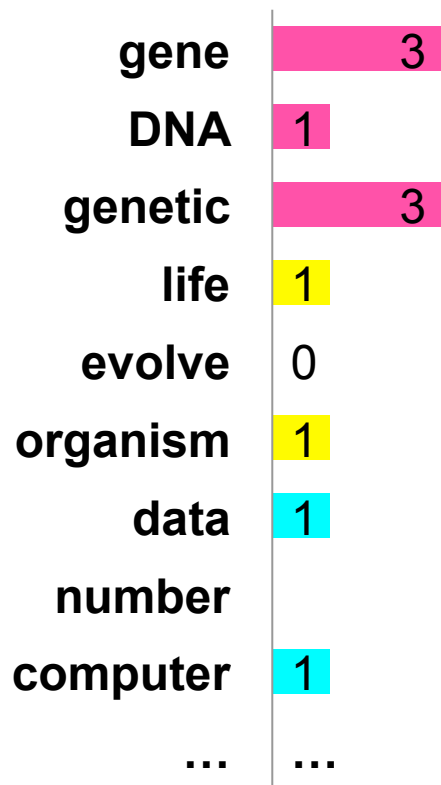


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

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

Topic Model: Single Document



document

N IID Words



$$\begin{matrix} \text{word 1} \\ \text{word 2} \\ \vdots \\ \vdots \\ \text{word } W \end{matrix} \begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ \vdots \\ A_{W1} \end{bmatrix}$$

=

$$\begin{matrix} \text{word 1} \\ \text{word 2} \\ \vdots \\ \vdots \\ \text{word } W \end{matrix} \begin{matrix} \text{Genetics} & \text{Data} \\ \begin{bmatrix} \beta_{11} & \beta_{1K} \\ \beta_{21} & \beta_{2K} \\ \vdots & \vdots \\ \vdots & \vdots \\ \beta_{W1} & \beta_{WK} \end{bmatrix} \end{matrix} \begin{bmatrix} \theta_{11} \\ \vdots \\ \theta_{K1} \end{bmatrix}$$

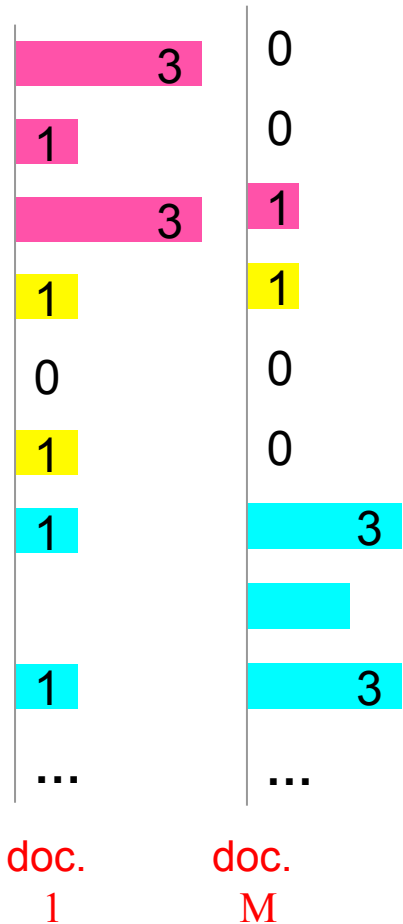
Mixing Wts

Document Distribution

Topic Matrix - β

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

Topic Model: M-documents

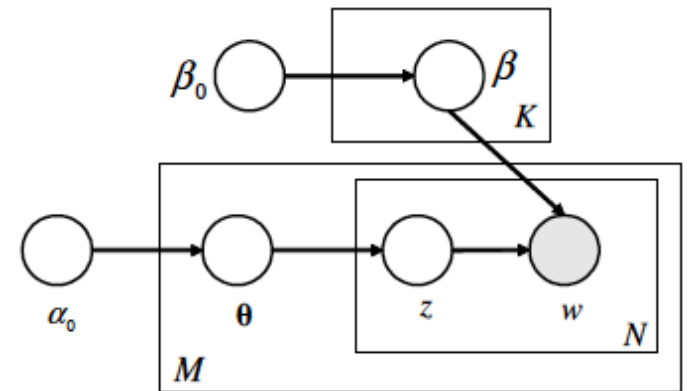


N IID Word/doc
M Documents

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

topic 1 topic K

Generative Model [Blei et.al. 2004]



Observation matrix X

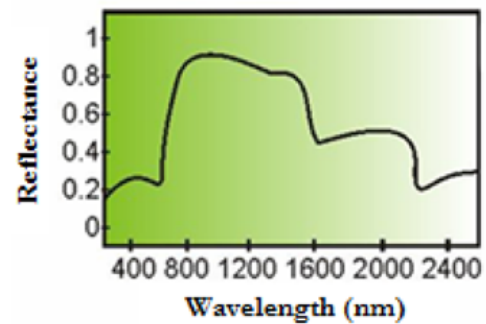
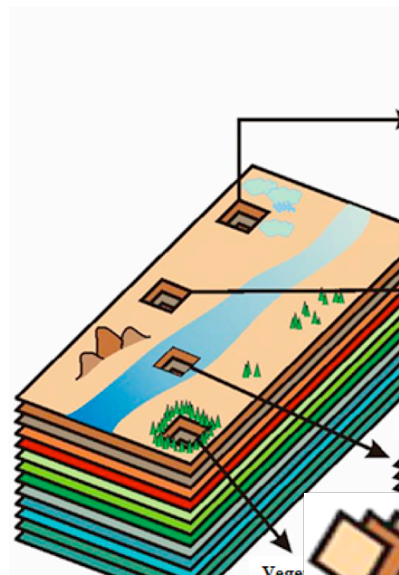
- column = word-freq. of a doc.
- $N = \#$ word/doc.

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

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

↓

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

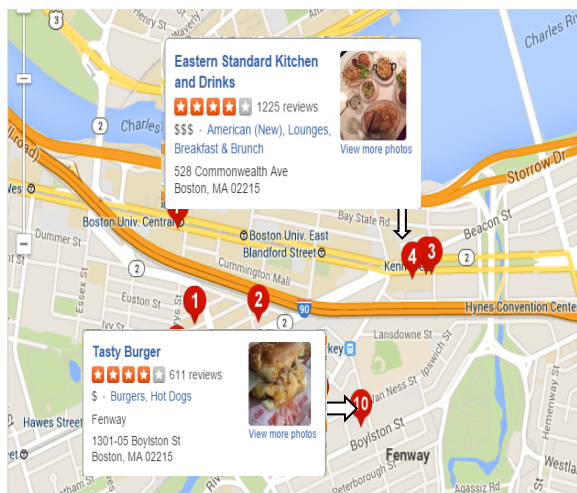


Summary

- 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



comparisons

counts

Influencing factors



Audubon Boston > Island Creek Oyster Bar

Audubon Boston: 44 reviews, \$\$\$ - American (Traditional), 838 Beacon St, Boston, MA 02215.

Island Creek Oyster Bar: 1342 reviews, \$\$\$ - Seafood, 500 Commonwealth Ave, Boston, MA 02215.

1

“\$\$”

Island Creek Oyster Bar > Eastern Standard Kitchen and Drinks

Island Creek Oyster Bar: 1342 reviews, \$\$\$ - Seafood, 500 Commonwealth Ave, Boston, MA 02215.

Eastern Standard Kitchen and Drinks: 1225 reviews, \$\$\$ - American (New), Lounges, Breakfast & Brunch, 528 Commonwealth Ave, Boston, MA 02215.

0

“Cuisine”

check-in/GPS/browsing records

Tasty Burger > Audubon Boston

Tasty Burger: 611 reviews, \$ - Burgers, Hot Dogs, Fenway, 1301-05 Boylston St, Boston, MA 02215.

Audubon Boston: 44 reviews, \$\$\$ - American (Traditional), 838 Beacon St, Boston, MA 02215.

1

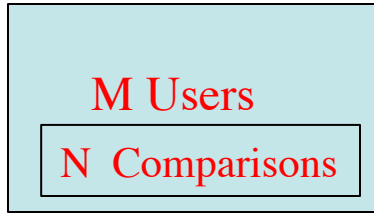
“Popularity”

User = mixture of latent influencing factors

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

Counts

Audubon Boston (44 reviews) > Island Creek Oyster Bar (1342 reviews) **2**
 Island Creek Oyster Bar (1342 reviews) > Eastern Standard Kitchen and Drinks (1225 reviews) **0**
 Tasty Burger (611 reviews) > Audubon Boston (44 reviews) **4**



Restaurant 1 > 2

$$\begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ \vdots \\ A_{W1} \end{bmatrix}$$

=

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

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

$$\begin{bmatrix} \theta_{11} \\ \vdots \\ \theta_{K1} \end{bmatrix}$$

Mixing Wts

Price

Popularity

Multiple Rankings

Noisy Influencing Factors

- Sampling: Space of Permutations

- Binary Matrix ~ Pairwise Comp.
- Sampled μ_{ij}

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



	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

Price

Popularity

$\uparrow \phi=0$

- Mallows Model (1957) ~ K baselines



$Prob\{1 > 2 \mid \sigma^1, \phi^1\}$



6 > 1



β_{11}	β_{1K}
β_{21}	β_{2K}
\vdots	\vdots
\dots	\dots
\vdots	\vdots
β_{W1}	β_{WK}

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

Inconsistent & Heterogeneous User Model

$$\begin{array}{l}
 (1,2) \\
 (1,3) \\
 \vdots \\
 \vdots \\
 (Q-1,Q)
 \end{array}
 \begin{bmatrix}
 A_{11} & A_{12} & \dots & A_{1M} \\
 A_{21} & A_{22} & \dots & A_{2M} \\
 \vdots & \vdots & \dots & \vdots \\
 \vdots & \vdots & \dots & \vdots \\
 A_{W1} & A_{W2} & \dots & A_{WM}
 \end{bmatrix}
 =
 \begin{array}{l}
 (1,2) \\
 (1,3) \\
 \vdots \\
 \vdots \\
 (Q-1,Q)
 \end{array}
 \begin{array}{c}
 \text{Price} \\
 \text{Popularity}
 \end{array}
 \begin{bmatrix}
 \beta_{11} & \beta_{1K} \\
 \beta_{21} & \beta_{2K} \\
 \vdots & \vdots \\
 \vdots & \vdots \\
 \beta_{W1} & \beta_{WK}
 \end{bmatrix}
 \begin{array}{c}
 \left[\begin{array}{ccc}
 \theta_{11} & \theta_{12} & \dots & \theta_{1M} \\
 \vdots & \vdots & \dots & \vdots \\
 \theta_{K1} & \theta_{K2} & \dots & \theta_{KM}
 \end{array} \right] \\
 \text{user } 1 \quad \text{user } 2 \quad \dots \quad \text{user } M
 \end{array}$$

$$\begin{array}{l}
 (1,2) \\
 (1,3) \\
 \vdots \\
 \vdots \\
 (Q-1,Q)
 \end{array}
 \begin{array}{c}
 \text{user } 1 \\
 \text{user } 2 \\
 \vdots \\
 \text{user } M
 \end{array}
 \begin{bmatrix}
 X_{11} & X_{12} & \dots & X_{1M} \\
 X_{21} & X_{22} & \dots & X_{2M} \\
 \vdots & \vdots & \dots & \vdots \\
 \vdots & \vdots & \dots & \vdots \\
 X_{W1} & X_{W2} & \dots & X_{WM}
 \end{bmatrix}$$

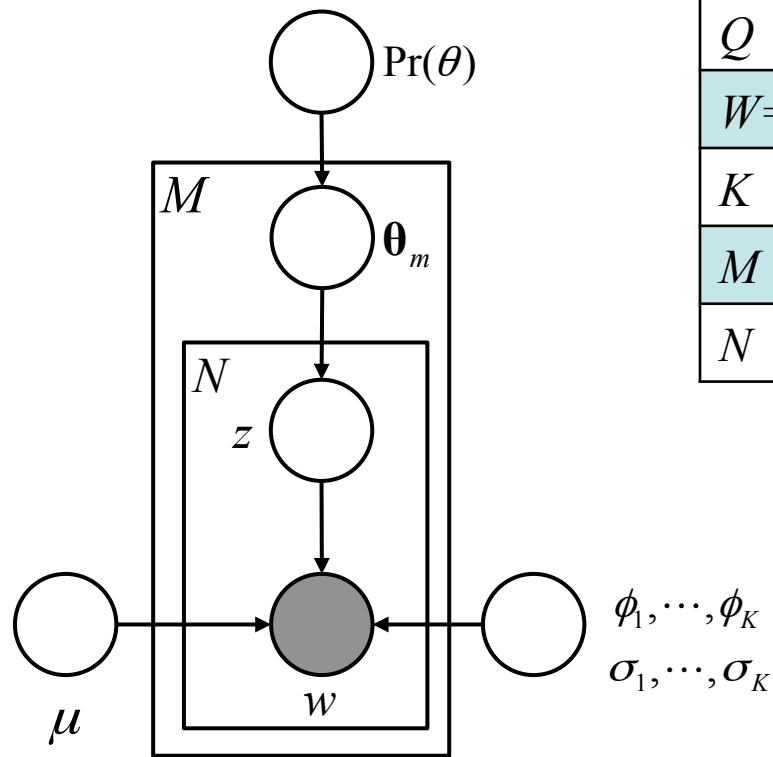
Observation matrix X

- 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]

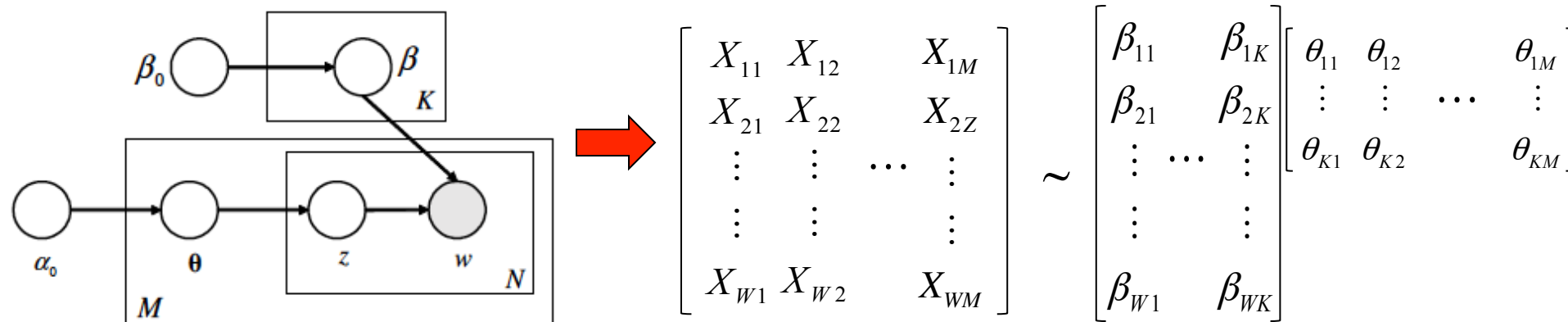
Summary

- 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

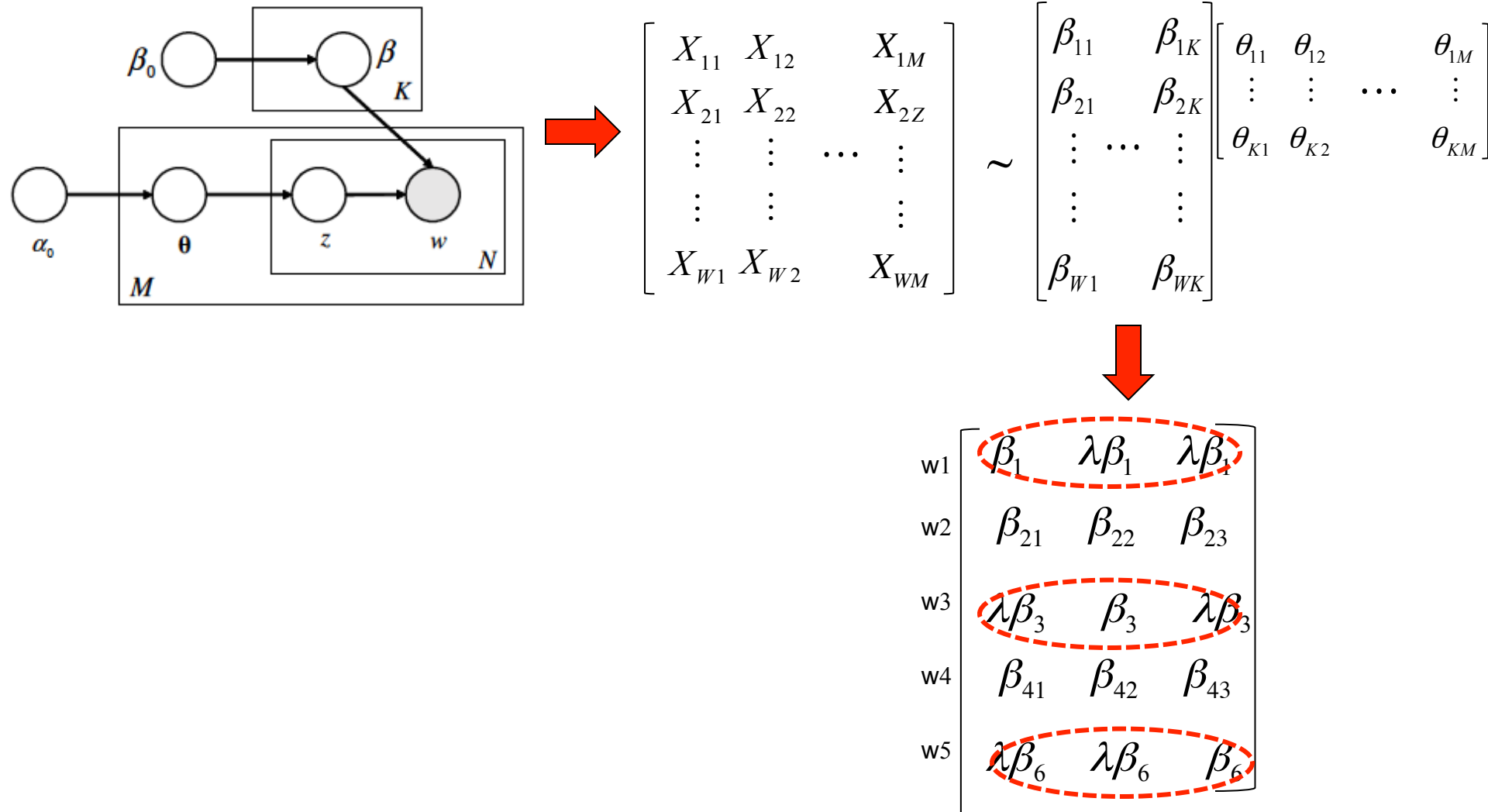
Topic matrix Weight matrix

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



Bayesian Topic Models are Approximately Separable

[Ding-Ishwar-S'14]

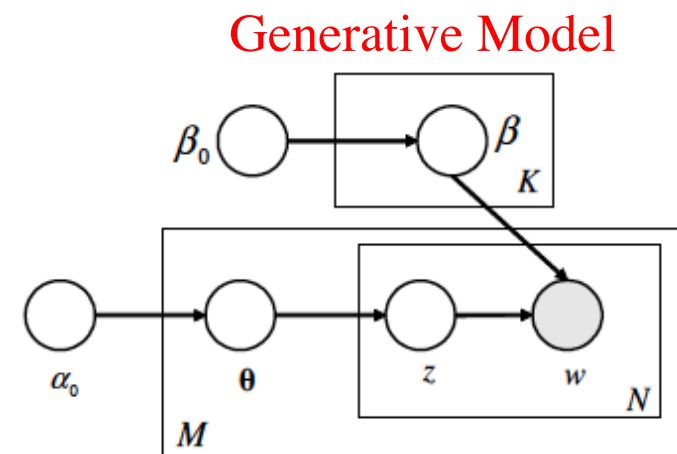


Approximately Separable & High-Dimensionality

- In real-world problems
 - Size of vocab. $W \gg \#$. 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)

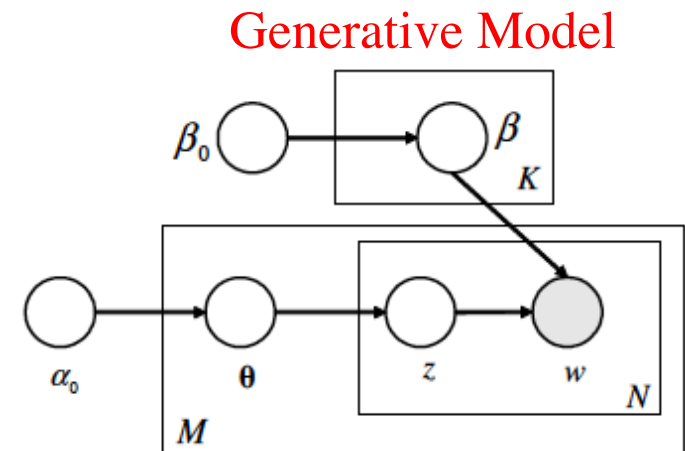


Why is Separability inevitable for $W \gg K$?

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

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

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

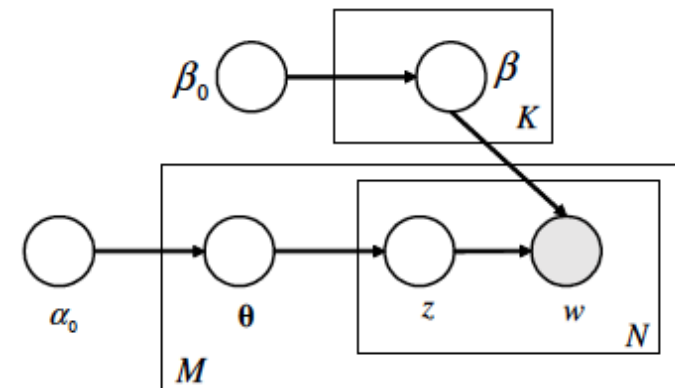


Separability in Practice

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

Generative Model



- β_0 moderately small positive value in practice.
 - $\beta_0 = 0.01$ for $K \in [50, 200]$
- Some packages suggest $\beta_0 = c/W$ to get satisfactory empirical results.

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

- Analysis explains reasoning for this choice!

What does Approximately Separable Topic Matrix mean?

		Genetics			Data		
		topic 1	topic 2	topic 3	topic 1	topic 2	topic 3
DNA	w1	β_1	0	0			
	w2	β_2	0	0			
Computer	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

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

$\lambda = 0$ Case: Novel Word(s)

unique to each topic

[Boardman'93, Donoho'04],

[Arora'13, Ding et. al.'13]

		topic 1	topic 2	topic 3
w1	β_1	$\leq \lambda \beta_1$	$\leq \lambda \beta_1$	
w2	β_2		β_{21}	$\lambda \beta_2$
w3	$\leq \lambda \beta_3$		β_3	$\leq \lambda \beta_3$
w4	$\lambda \beta_4$		β_4	$\lambda \beta_4$
w5	$\lambda \beta_5$	$\leq \lambda \beta_5$		β_5
w6	$\lambda \beta_6$	$\lambda \beta_6$		β_6
	β_{71}	β_{72}	β_{73}	
		...		

Approximately Separable

[Ding, Ishwar-S 15]

HOW ABOUT RANK AGGREGATION MODELS ?

Mallows Rank Aggregation Model is Separable

Obs. Comparisons

$$\begin{bmatrix} 1 & 0.25 \\ 0 & 0.75 \\ \vdots & \dots \\ \vdots & \vdots \\ 1 & 0.90 \end{bmatrix}$$

\sim

	Price	Popularity									
(1,2)	β_{11}	β_{1K}	$\begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1M} \\ \vdots & \vdots & \dots & \vdots \\ \theta_{K1} & \theta_{K2} & \dots & \theta_{KM} \end{bmatrix}$								
(1,3)	β_{21}	β_{2K}									
	\vdots	\dots									
	\vdots	\vdots									
	\vdots	\vdots									
(Q-1,Q)	β_{W1}	β_{WK}	<table border="0" style="width: 100%; text-align: center;"> <tr> <td>user</td> <td>user</td> <td></td> <td>user</td> </tr> <tr> <td>1</td> <td>2</td> <td></td> <td>M</td> </tr> </table>	user	user		user	1	2		M
user	user		user								
1	2		M								
	Rank 1	Rank K									

σ



$6 > 1$

σ^1



Mallow Rank Aggregation Model
Is Approx Separable

Mallows (Noisy Factors) are Approx Separable

- Most ranking matrix are λ -Approximate separable, # items $Q \gg$ # factors K

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

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

	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
		...	
		β	

Approximately
Separable
ranking matrix

$$Q = 100$$

$$K = 10$$

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)

$$\beta_{W \times K} \theta_{K \times M} = A_{W \times M}$$

	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}
		...	

doc. 1	...	doc. M
$\leftarrow \theta_1 \rightarrow$		
$\leftarrow \theta_2 \rightarrow$		
$\leftarrow \theta_3 \rightarrow$		

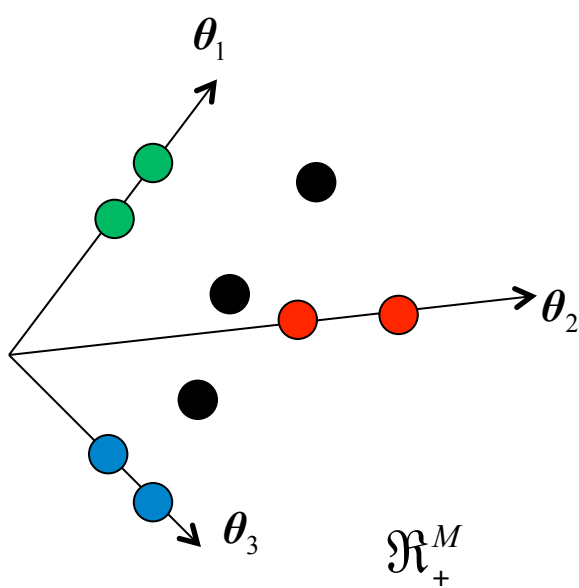
Weight Matrix



doc. 1	...	doc. M
$\leftarrow \beta_1 \theta_1 \rightarrow$		
$\leftarrow \beta_2 \theta_1 \rightarrow$		
$\leftarrow \beta_3 \theta_2 \rightarrow$		
$\leftarrow \beta_4 \theta_2 \rightarrow$		
$\leftarrow \beta_5 \theta_3 \rightarrow$		
$\leftarrow \beta_6 \theta_3 \rightarrow$		
$\beta_{71} \theta_1 + \beta_{72} \theta_2 + \beta_{73} \theta_3$		
	...	

Distribution Matrix A

Separable Topic Matrix β



\mathcal{R}_+^M

Key Idea ($\lambda = 0$ case)

$$\beta_{W \times K} \theta_{K \times M} = A_{W \times M}$$

	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}
		...	

doc. 1	...	doc. M
$\leftarrow \theta_1 \rightarrow$		
$\leftarrow \theta_2 \rightarrow$		
$\leftarrow \theta_3 \rightarrow$		

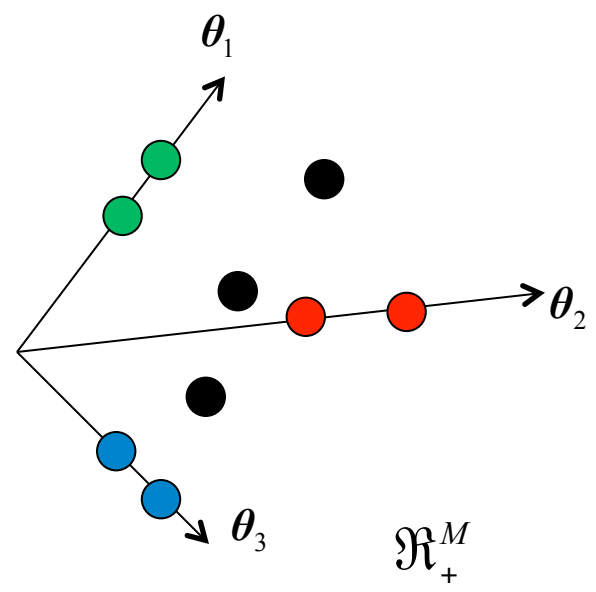
Weight Matrix



doc. 1	...	doc. M
$\leftarrow \beta_1 \theta_1 \rightarrow$		
$\leftarrow \beta_2 \theta_1 \rightarrow$		
$\leftarrow \beta_3 \theta_2 \rightarrow$		
$\leftarrow \beta_4 \theta_2 \rightarrow$		
$\leftarrow \beta_5 \theta_3 \rightarrow$		
$\leftarrow \beta_6 \theta_3 \rightarrow$		
$\beta_{71} \theta_1 + \beta_{72} \theta_2 + \beta_{73} \theta_3$		
...		

Distribution Matrix A

Separable Topic Matrix β



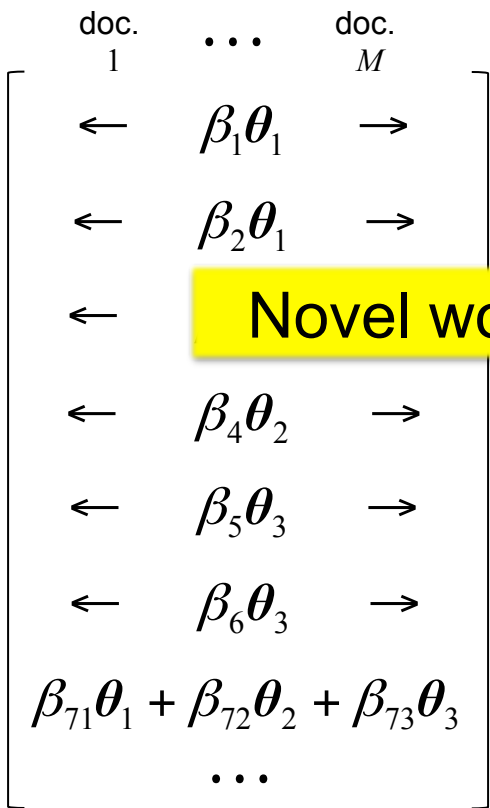
\mathcal{R}_+^M

Key Idea

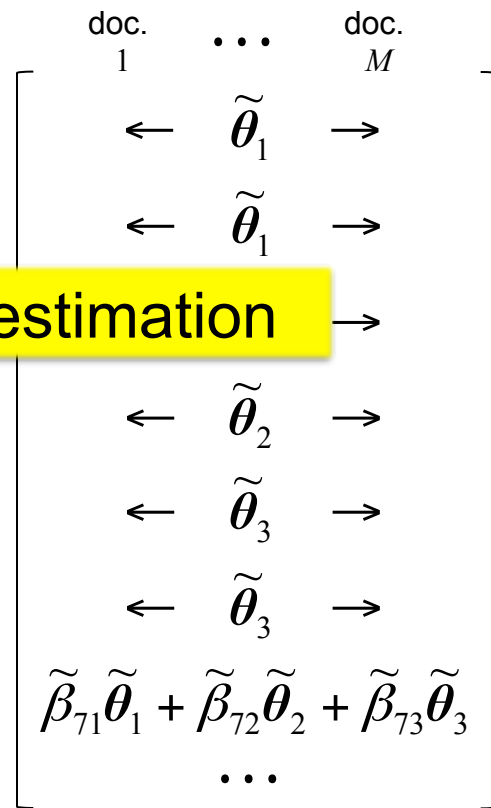
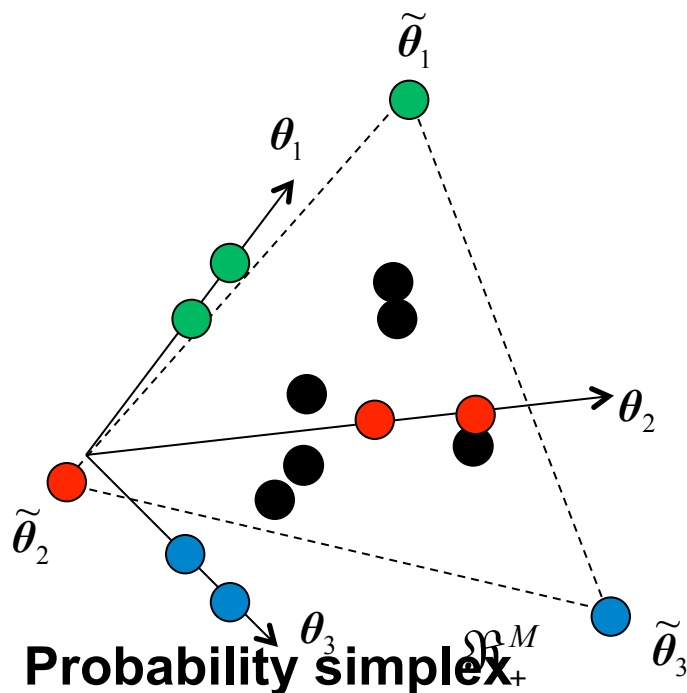
$$\tilde{\beta}_{W \times K} \tilde{\theta}_{K \times M} = \tilde{A}_{W \times M}$$

Novel Word = Extreme Point

Novel word detection + Topic matrix estimation



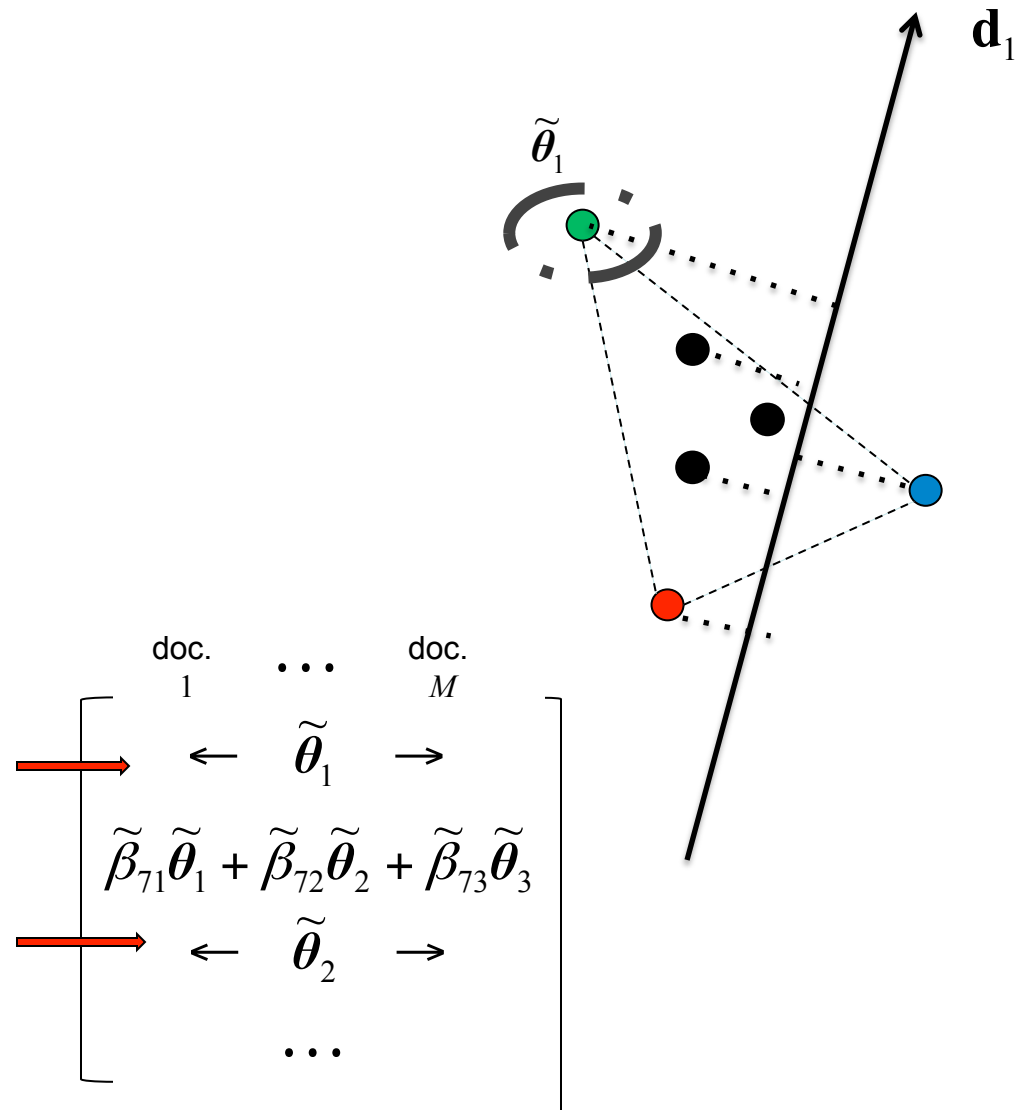
Distribution Matrix A



Row Normalized Distribution Matrix \tilde{A}

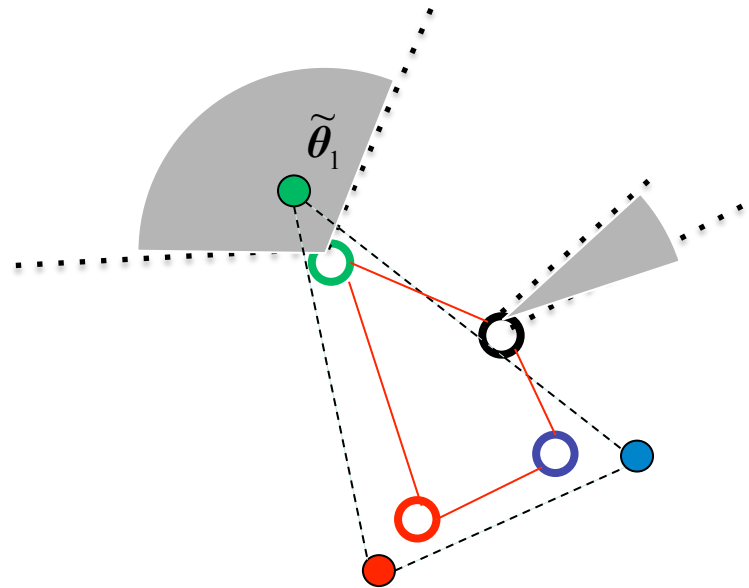
Detect Novel Words via Projections

- 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 \tilde{A}



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

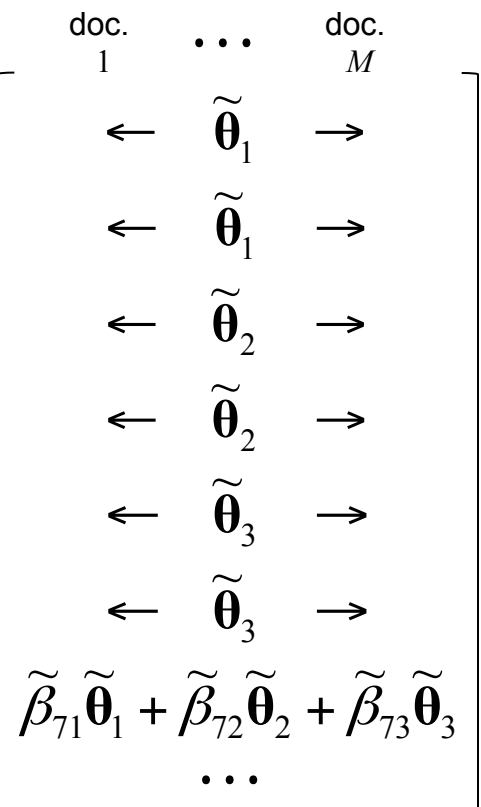
- Approx. novel words \Leftrightarrow larger solid angles
- Solution
 - Sort solid angles
 - Take the top – K extreme points
 - Freq. of maximum \approx Solid Angle of an Extreme Point



Finite words/doc.

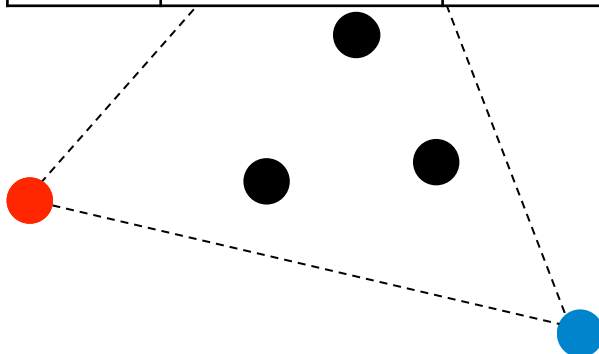
$$\tilde{\beta}_{W \times K} \tilde{\theta}_{K \times M} = \tilde{A}_{W \times M} \approx \tilde{X}_{W \times M}$$

Key issue:
 N fixed \rightarrow perturbation
 does not vanish

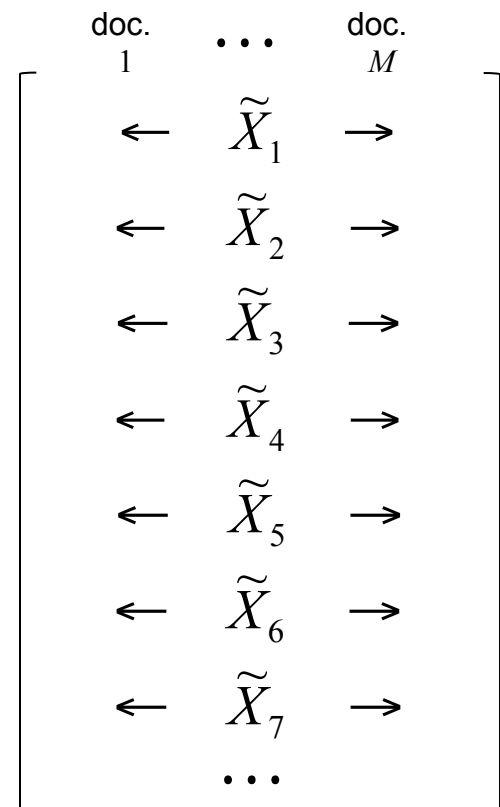


**Row Normalized
 Distribution
 Matrix \tilde{A}**

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



Probability simplex



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

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

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 R' , with $u \sim N(\mathbf{0}, I_W)$, the proposed Random Projection algorithm can detect all novel words of K topics with probability $1-\delta$ 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 β with ε element-wise error with probability $1-\delta$.

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

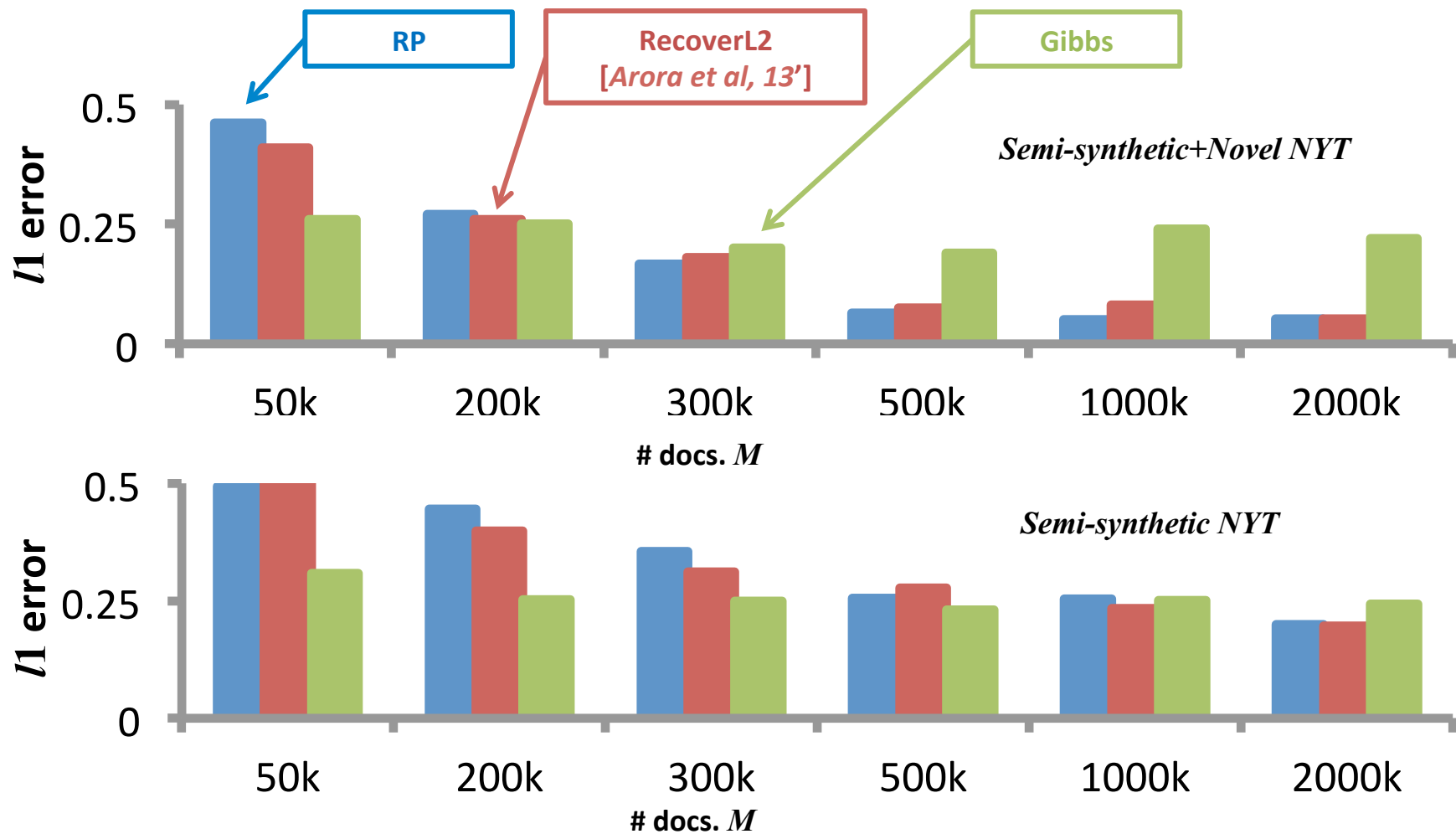
Generate synthetic
docs. with Dirichlet
prior

Add artificial novel
words;
Generate synthetic
docs.

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

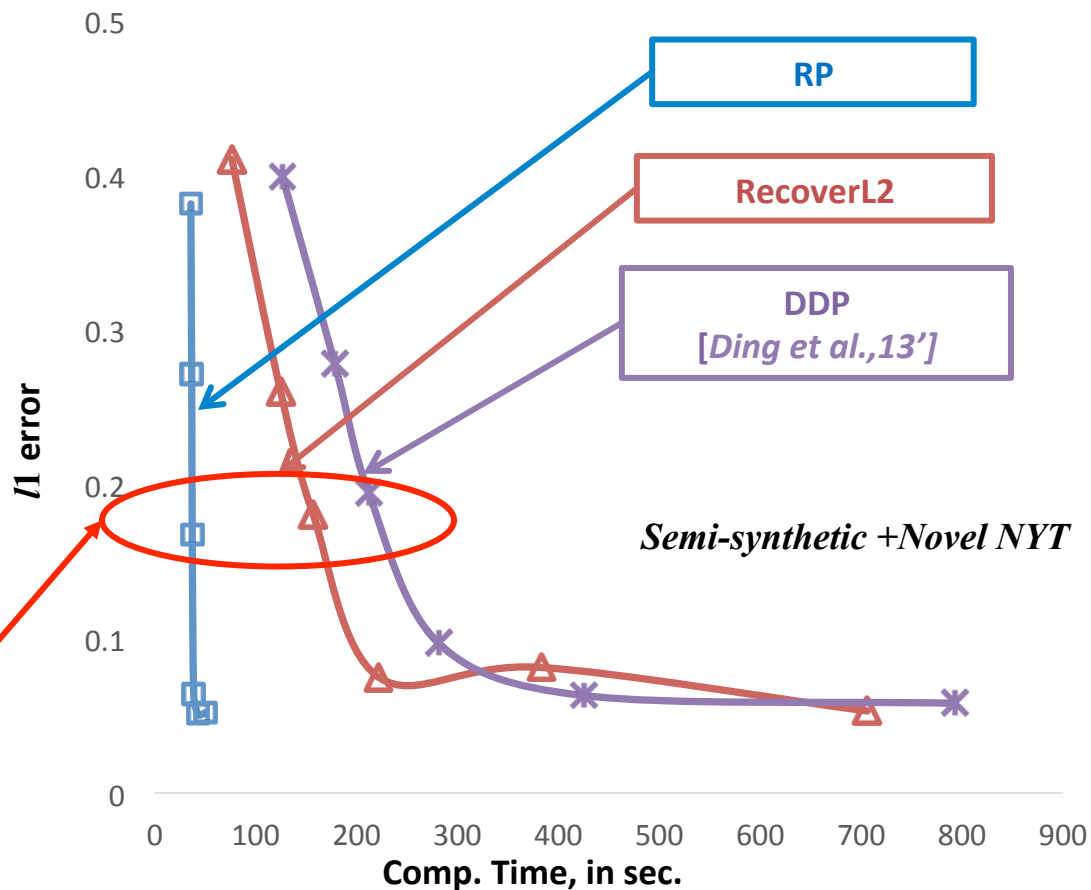
- Semi-synthetic data can resemble the dimensionality and sparsity of real-world data

Experimental Results (semi-synthetic data)



Experimental Results (semi-synthetic data)

	<i>Semi-syn +novel NYT</i>	<i>Semi-syn NYT</i>
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	<i>Florida</i>	Game
Air	Love	Ballot	Team
Storm	Character	Vote	Season
Rain	Heart	<i>Al_gore</i>	Play
Cold	Emotion	Recount	<i>NFL</i>

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

MovieLens Dataset: Ratings → Comparisons

6000 Users, 3000 Rated Movies

User preferences

House of Cards: Season 2 2014 NR
 DVD \$27.22 \$66.00 Prime
 Only 10 left in stock - order soon.
 More Buying Choices \$22.44 used & new (27 offers)
 Blu-ray Blu-ray + UltraViolet \$31.49 \$66.00 Prime
 Get it by Thursday, Mar 5
 More Buying Choices \$26.49 used & new (27 offers)

Seinfeld: The Complete Series 2013 NR
 DVD \$90.85 \$449.00 Prime
 Get it by Wednesday, Mar 4
 More Buying Choices \$74.99 used & new (20 offers)

Breaking Bad: The Complete Series 2013 Unrated
 DVD \$74.49 \$400.00 Prime
 Get it by Wednesday, Mar 4
 More Buying Choices \$70.49 used & new (31 offers)
 Blu-ray \$105.99 \$220.00 Prime
 Get it by Thursday, Mar 5
 More Buying Choices \$84.99 used & new (52 offers)

The Big Bang Theory: Season 7 2014 NR
 DVD \$19.99 \$44.00 Prime
 Get it by Wednesday, Mar 4
 More Buying Choices \$15.90 used & new (38 offers)
 Blu-ray \$29.70 \$64.00 Prime
 Get it by Wednesday, Mar 4



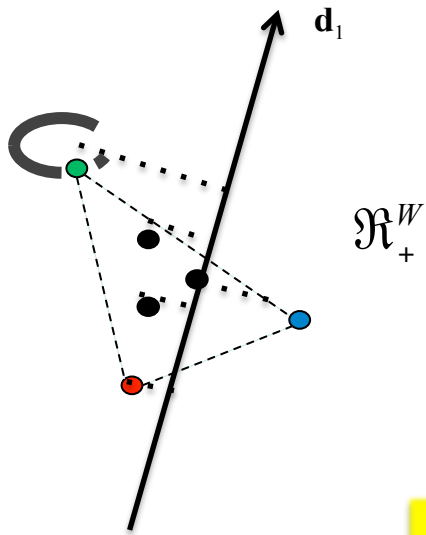
MovieLens dataset – predicting star ratings

- 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 \sim 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
		...	
		β	

Approximately
Separable
ranking matrix

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