# A Geometric Approach for Learning Latent Mixed Membership Models

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

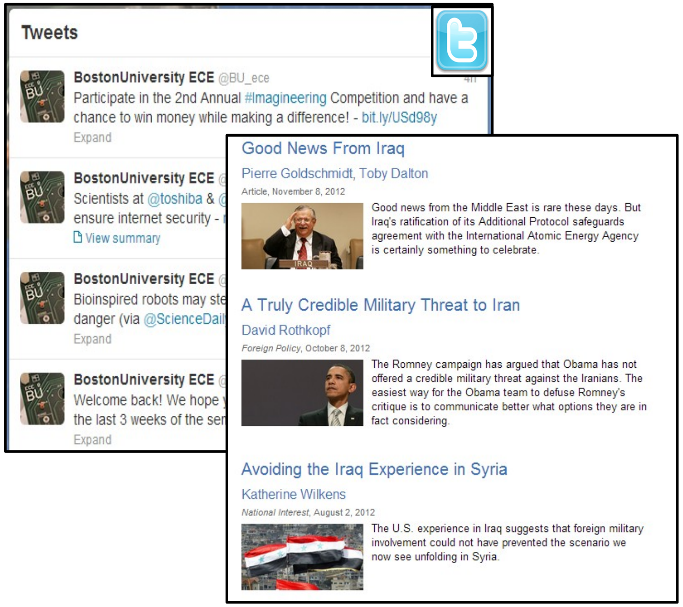
# Outline

- Latent Mixture Models
  - Text Documents, User Preferences, Community Networks, …
  - Overall Goal/Objective: Algorithm with provable guarantees
- Topic Models & Estimation Problem
- Geometric Structure of Topic Models
- Algorithm & Guarantees: Exploiting Geometry
- Rank Aggregation Problem
- Real-World Expts.

Ding, Ishwar, Saligrama, ICML'13
Ding, Ishwar, Saligrama, NIPS'14
Ding, Ishwar, Saligrama, ITA'15
Ding, Ishwar, Saligrama, AISTATS'14
Ding, Ishwar, Saligrama, AISTATS'15

# 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



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
Scientists at @toshiba &
ensure internet security -
View summary

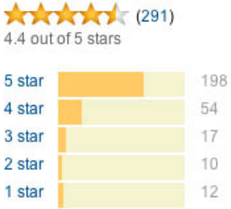
BostonUniversity ECE
Bioinspired robots may ste
danger (via @ScienceDail
Expand

BostonUniversity ECE
Welcome back! We hope
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.

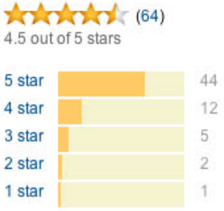
A Truly Credible Military Threat to Iran
David Rothkopf
*Foreign Policy*, October 8, 2012
The Romney campaign has argued that Obama has not offered a credible military threat against the Iranians. The easiest way for the Obama team to defuse Romney's critique is to communicate better what options they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



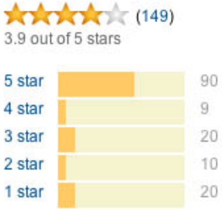
(291)
4.4 out of 5 stars

| | |
|---|---|
| 5 star | 198 |
| 4 star | 54 |
| 3 star | 17 |
| 2 star | 10 |
| 1 star | 12 |



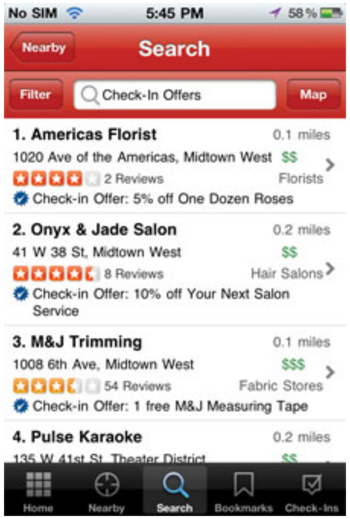
(64)
4.5 out of 5 stars

| | |
|---|---|
| 5 star | 44 |
| 4 star | 12 |
| 3 star | 5 |
| 2 star | 2 |
| 1 star | 1 |



(149)
3.9 out of 5 stars

| | |
|---|---|
| 5 star | 90 |
| 4 star | 9 |
| 3 star | 20 |
| 2 star | 10 |
| 1 star | 20 |

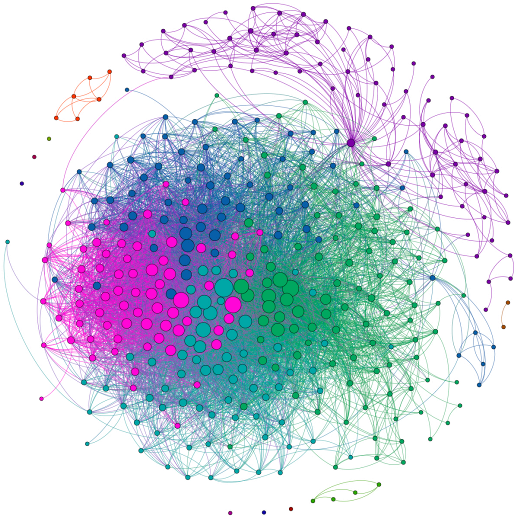


1. Americas Florist 0.1 miles
1020 Ave of the Americas, Midtown West $$
2 Reviews Florists
Check-in Offer: 5% off One Dozen Roses

2. 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

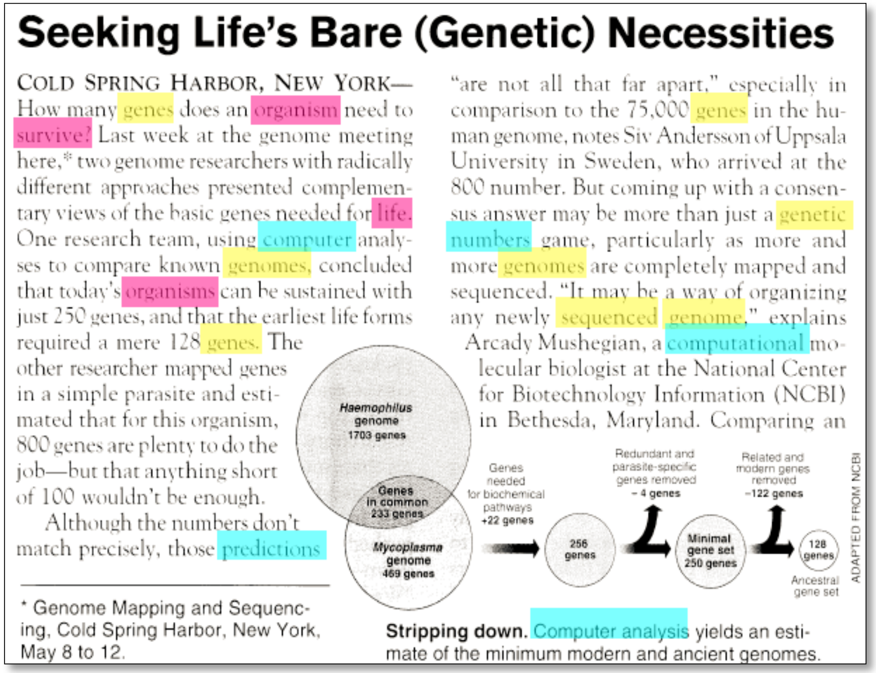
3. M&J Trimming 0.1 miles
1008 6th Ave, Midtown West $$$
54 Reviews Fabric Stores
Check-in Offer: 1 free M&J Measuring Tape

4. Pulse Karaoke 0.2 miles
135 W 41st St, Theater District $$



# Mixed membership latent variable model

Text document:



## Seeking Life's Bare (Genetic) Necessities

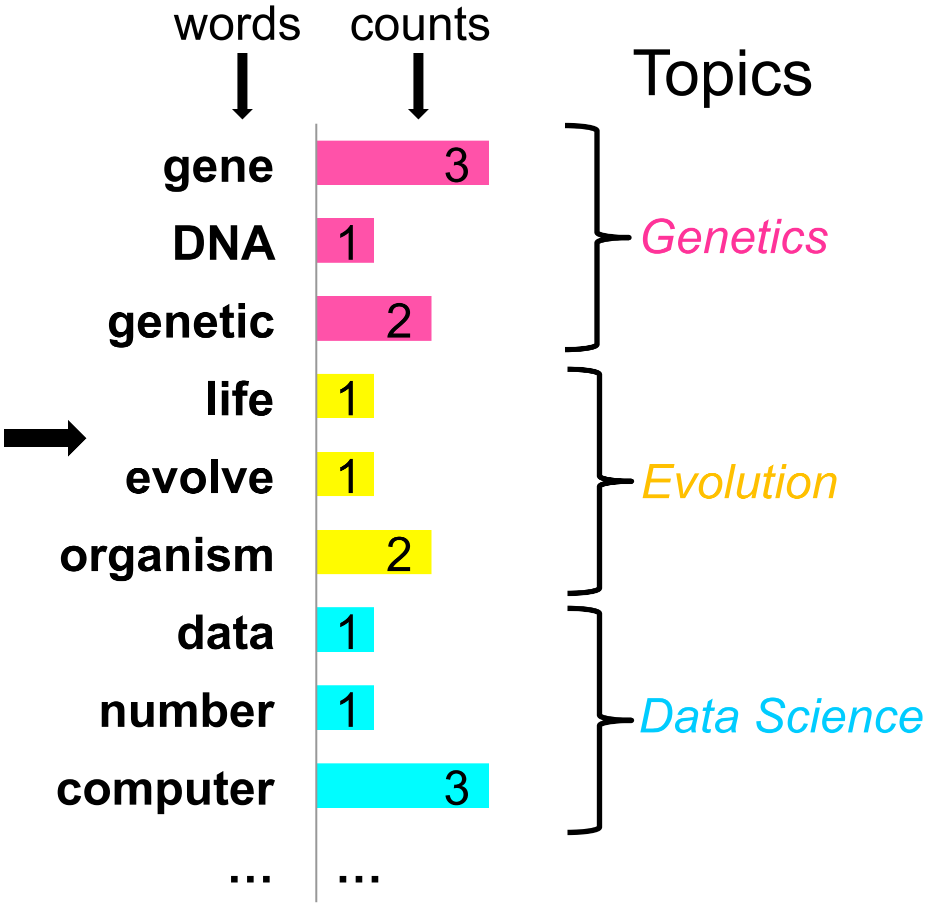
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

**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.

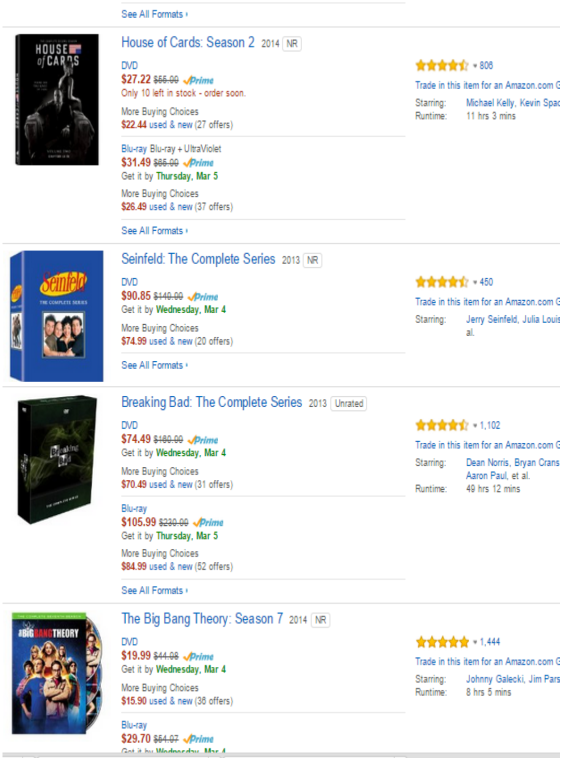


| words | counts | Topics |
|---|---|---|
| gene | 3 | *Genetics* |
| DNA | 1 | |
| genetic | 2 | |
| life | 1 | *Evolution* |
| evolve | 1 | |
| organism | 2 | |
| data | 1 | *Data Science* |
| number | 1 | |
| computer | 3 | |
| ... | ... | |

**document = mixture of latent topics**

# Mixed membership latent variable model

User preferences



See All Formats ›

House of Cards: Season 2 2014 NR

DVD
$27.22 ~~$55.00~~ ✓Prime
Only 10 left in stock - order soon.
More Buying Choices
$22.44 used & new (27 offers)

Blu-ray Blu-ray + UltraViolet
$31.49 ~~$85.00~~ ✓Prime
Get it by Thursday, Mar 5
More Buying Choices
$26.49 used & new (37 offers)

See All Formats ›

★★★★☆ ▾ 806
Trade in this item for an Amazon.com G
Starring: Michael Kelly, Kevin Spac
Runtime: 11 hrs 3 mins

Seinfeld: The Complete Series 2013 NR

DVD
$90.85 ~~$140.00~~ ✓Prime
Get it by Wednesday, Mar 4
More Buying Choices
$74.99 used & new (20 offers)

See All Formats ›

★★★★☆ ▾ 450
Trade in this item for an Amazon.com G
Starring: Jerry Seinfeld, Julia Louis al.

Breaking Bad: The Complete Series 2013 Unrated

DVD
$74.49 ~~$180.00~~ ✓Prime
Get it by Wednesday, Mar 4
More Buying Choices
$70.49 used & new (31 offers)

Blu-ray
$105.99 ~~$230.00~~ ✓Prime
Get it by Thursday, Mar 5
More Buying Choices
$84.99 used & new (52 offers)

See All Formats ›

★★★★☆ ▾ 1,102
Trade in this item for an Amazon.com G
Starring: Dean Norris, Bryan Crans Aaron Paul, et al.
Runtime: 49 hrs 12 mins

The Big Bang Theory: Season 7 2014 NR

DVD
$19.99 ~~$44.98~~ ✓Prime
Get it by Wednesday, Mar 4
More Buying Choices
$15.90 used & new (36 offers)

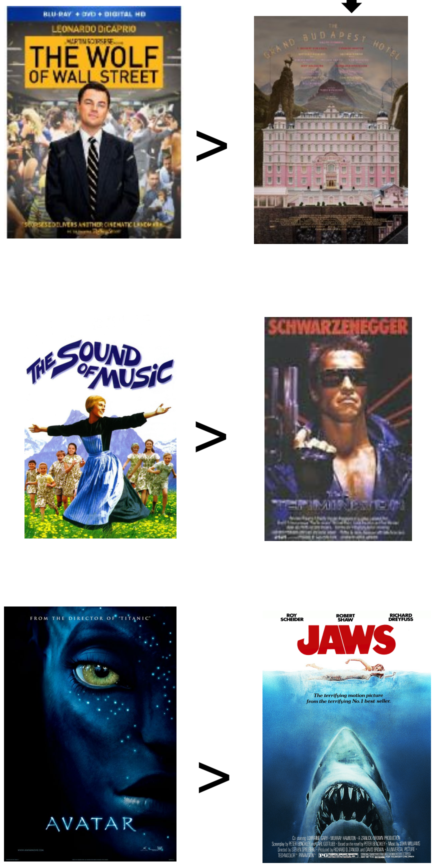
Blu-ray
$29.70 ~~$54.97~~ ✓Prime
Get it by Wednesday, Mar 4

★★★★★ ▾ 1,444
Trade in this item for an Amazon.com G
Starring: Johnny Galecki, Jim Pars
Runtime: 8 hrs 5 mins

→

words counts

Influencing factors



1 "actor"

0 "music"

1 "special effect"

**document = mixture of latent influencing factors**

# Overall Goal

- Learn/Estimate Latent Factors from Observations(docs)

- Goal: develop algorithms with
  - Provable guarantees
    - How many Docs to estimate Latent Factors within a tolerance?
    - Computational Cost: How does Algorithm scale with #params?
  - Good empirical performance
    - Real-world datasets

→ ***Model Fidelity***

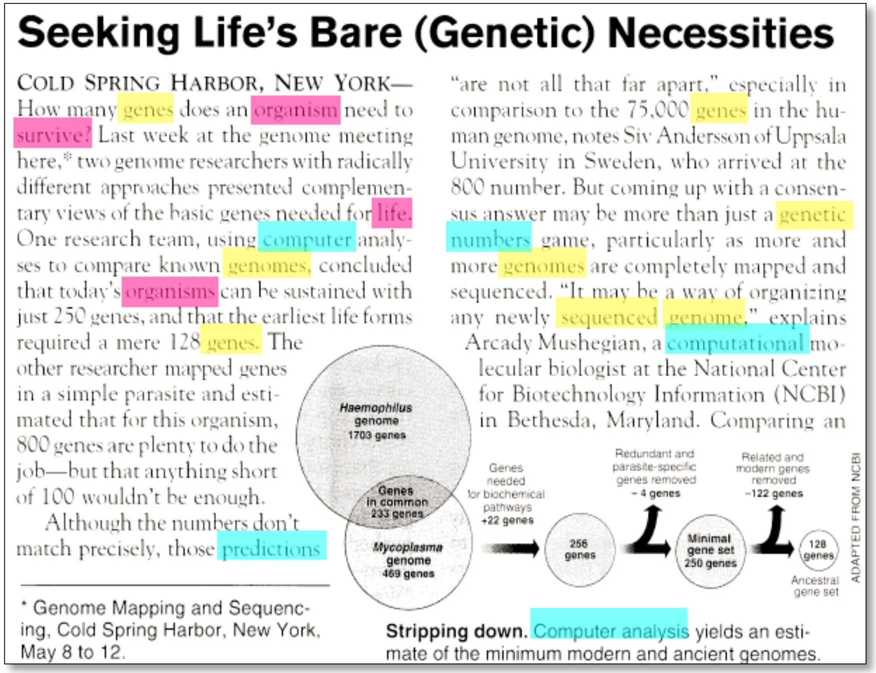
→ ***Web Scale applications***

# Outline

- Latent Mixture Models
  - Text Documents, User Preferences, Community Networks, …
- Topic Models & Estimation Problem
  - Observation Model & Related Work
- Geometric Structure of Topic Models
- Algorithm & Guarantees: Exploiting Geometry
- Extensions to User Preference Model Estimation
- Empirical Results on Real-World Datasets

# "Bag of words" model: a text corpus example

One document in the collection:



## Seeking Life's Bare (Genetic) Necessities

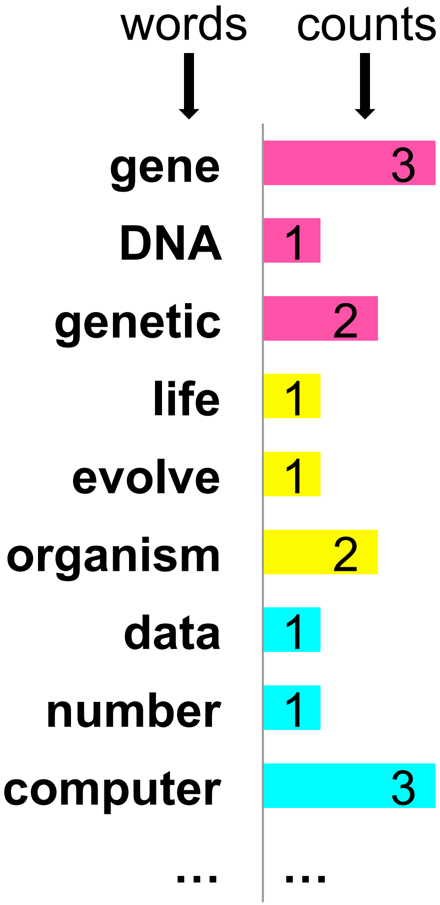
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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

*Haemophilus* genome 1703 genes
Genes in common 233 genes
*Mycoplasma* genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

**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.



$$\begin{array}{c|ccc|}
\text{word } 1 & \beta_{11} & & \beta_{1K} \\
\text{word } 2 & \beta_{21} & & \beta_{2K} \\
 & \vdots & \cdots & \vdots \\
 & \vdots & & \vdots \\
\text{word } W & \beta_{W1} & & \beta_{WK} \\
\hline
 & \text{topic } 1 & & \text{topic } K
\end{array}$$

Topic Matrix - $\boldsymbol{\beta}$

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

$$\begin{array}{r|cccc|}
\text{word } 1 & A_{11} & A_{12} & & A_{1M} \\
\text{word } 2 & A_{21} & A_{22} & & A_{2M} \\
 & \vdots & \vdots & \cdots & \vdots \\
 & \vdots & \vdots & & \vdots \\
\text{word } W & A_{W1} & A_{W2} & & A_{WM} \\
\hline
 & \text{doc. } 1 & \text{doc. } 2 & & \text{doc. } M
\end{array}
=
\begin{array}{r|ccc|}
\text{word } 1 & \beta_{11} & & \beta_{1K} \\
\text{word } 2 & \beta_{21} & & \beta_{2K} \\
 & \vdots & \cdots & \vdots \\
 & \vdots & & \vdots \\
\text{word } W & \beta_{W1} & & \beta_{WK} \\
\hline
 & \text{topic } 1 & & \text{topic } K
\end{array}
\begin{array}{|cccc|}
\theta_{11} & \theta_{12} & & \theta_{1M} \\
\vdots & \vdots & \cdots & \vdots \\
\theta_{K1} & \theta_{K2} & & \theta_{KM} \\
\hline
\text{doc. } 1 & \text{doc. } 2 & & \text{doc. } M
\end{array}$$

Document Distribution matrix - $A$

- column = distb. of a doc.
- $M$ = # docs.

Weight matrix - $\theta$

- column = mixing weights
- $M$ = # docs.

$$\begin{array}{r|cccc|}
\text{word } 1 & A_{11} & A_{12} & & A_{1M} \\
\text{word } 2 & A_{21} & A_{22} & & A_{2M} \\
 & \vdots & \vdots & \cdots & \vdots \\
 & \vdots & \vdots & & \vdots \\
\text{word } W & A_{W1} & A_{W2} & & A_{WM} \\
 & \text{doc. } 1 & \text{doc. } 2 & & \text{doc. } M
\end{array}
=
\begin{array}{r|ccc|}
\text{word } 1 & \beta_{11} & & \beta_{1K} \\
\text{word } 2 & \beta_{21} & & \beta_{2K} \\
 & \vdots & \cdots & \vdots \\
 & \vdots & & \vdots \\
\text{word } W & \beta_{W1} & & \beta_{WK} \\
 & \text{topic } 1 & & \text{topic } K
\end{array}
\begin{array}{|cccc|}
\theta_{11} & \theta_{12} & & \theta_{1M} \\
\vdots & \vdots & \cdots & \vdots \\
\theta_{K1} & \theta_{K2} & & \theta_{KM} \\
\text{doc. } 1 & \text{doc. } 2 & & \text{doc. } M
\end{array}$$

$\downarrow$ $N$ iid samples + empirical words count

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

Observation matrix $\boldsymbol{X}$

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

$$\begin{array}{cc}
\begin{array}{c} \text{word } 1 \\ \text{word } 2 \\ \\ \\ \\ \text{word } W \end{array} &
\left[\begin{array}{cccc}
A_{11} & A_{12} & & A_{1M} \\
A_{21} & A_{22} & & A_{2M} \\
\vdots & \vdots & \cdots & \vdots \\
\vdots & \vdots & & \vdots \\
A_{W1} & A_{W2} & & A_{WM}
\end{array}\right] \\
& \begin{array}{cccc} \text{doc. } 1 & \text{doc. } 2 & & \text{doc. } M \end{array}
\end{array}
=
\begin{array}{cc}
\begin{array}{c} \text{word } 1 \\ \text{word } 2 \\ \\ \\ \\ \text{word } W \end{array} &
\left[\begin{array}{ccc}
\beta_{11} & & \beta_{1K} \\
\beta_{21} & & \beta_{2K} \\
\vdots & \cdots & \vdots \\
\vdots & & \vdots \\
\beta_{W1} & & \beta_{WK}
\end{array}\right] \\
& \begin{array}{ccc} \text{topic } 1 & & \text{topic } K \end{array}
\end{array}
\begin{array}{c}
\left[\begin{array}{cccc}
\theta_{11} & \theta_{12} & & \theta_{1M} \\
\vdots & \vdots & \cdots & \vdots \\
\theta_{K1} & \theta_{K2} & & \theta_{KM}
\end{array}\right] \\
\begin{array}{cccc} \text{doc. } 1 & \text{doc. } 2 & & \text{doc. } M \end{array}
\end{array}$$

$$\downarrow$$

$$\begin{array}{cc}
\begin{array}{c} \text{word } 1 \\ \text{word } 2 \\ \\ \\ \\ \text{word } W \end{array} &
\left[\begin{array}{cccc}
X_{11} & X_{12} & & X_{1M} \\
X_{21} & X_{22} & & X_{2Z} \\
\vdots & \vdots & \cdots & \vdots \\
\vdots & \vdots & & \vdots \\
X_{W1} & X_{W2} & & X_{WM}
\end{array}\right]
\end{array}$$

**Problem**

- Given : $X$ and $K$
- Goal : estimate $\boldsymbol{\beta}$

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

# Related work

Topic matrix Weight matrix

| Method | $\beta$ | $\theta$ | 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*], | Deterministic or Prior | Prior | MAP or ML | Non-convex. Need approximations like MCMC. |
| Method of Moments [*Anandkumar et al.,'12,'13*] | Deterministic and sparse | Prior | Tensor decomposition | No empirical performance reported |
| Topic-separability based [*Several references*] | Approximate Separability | Prior | Geometric | |

# Outline

- 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
- Extensions to User Preference Model Estimation
- Empirical Results on Real-World Datasets

# Approximately Separable Topic Matrix ([Ding-Ishwar-S'14])

$$
\begin{array}{rr|ccc|}
 & & \text{Genetics} & \text{Data} & \\
 & & \text{topic } 1 & \text{topic } 2 & \text{topic } 3 \\
\text{DNA} & \text{w1} & \beta_1 & 0 & 0 \\
 & \text{w2} & \beta_2 & 0 & 0 \\
\text{Computer} & \text{w3} & 0 & \beta_3 & 0 \\
 & \text{w4} & 0 & \beta_4 & 0 \\
 & \text{w5} & 0 & 0 & \beta_5 \\
 & \text{w6} & 0 & 0 & \beta_6 \\
 & & \beta_{71} & \beta_{72} & \beta_{73} \\
 & & & \cdots &
\end{array}
$$

**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]

$$
\begin{array}{r|ccc|}
 & \text{topic } 1 & \text{topic } 2 & \text{topic } 3 \\
\text{w1} & \beta_1 & \le\lambda\beta_1 & \le\lambda\beta_1 \\
\text{w2} & \beta_2 & \beta_{21} & \lambda\beta_2 \\
\text{w3} & \le\lambda\beta_3 & \beta_3 & \le\lambda\beta_3 \\
\text{w4} & \lambda\beta_4 & \beta_4 & \lambda\beta_4 \\
\text{w5} & \lambda\beta_5 & \le\lambda\beta_5 & \beta_5 \\
\text{w6} & \lambda\beta_6 & \lambda\beta_6 & \beta_6 \\
 & \beta_{71} & \beta_{72} & \beta_{73} \\
 & & \cdots &
\end{array}
$$

**Approximately Separable**

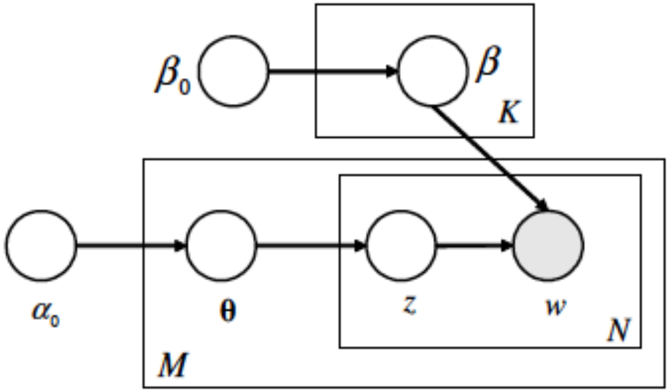
# Is Approximately Separable Fundamental?

- 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



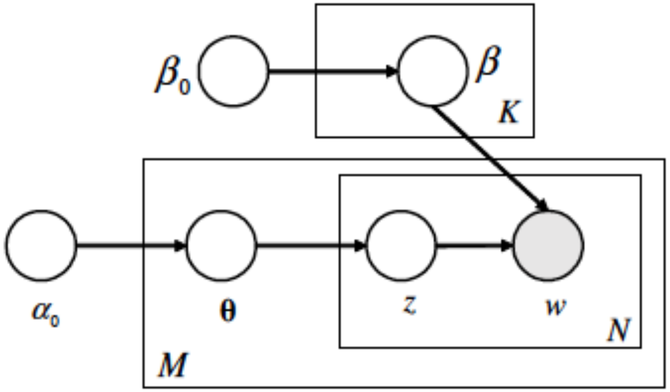
# Why is Separability inevitable for W>> K?

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

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

| 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% |

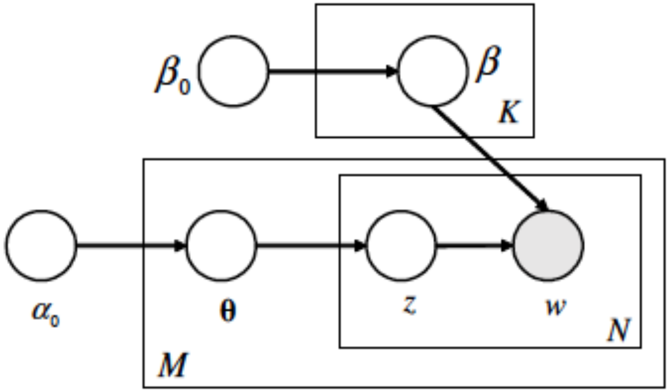
$\beta_0 = 0.01$, 1000 MC runs

- $\beta_0$ 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.

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

- Analysis explains reasoning for this choice!

Generative Model



# Outline

- 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
  - Efficiently Identifying Extreme Points
  - Empirical Results on Real-World Datasets
- Extensions to User Preference Model Estimation

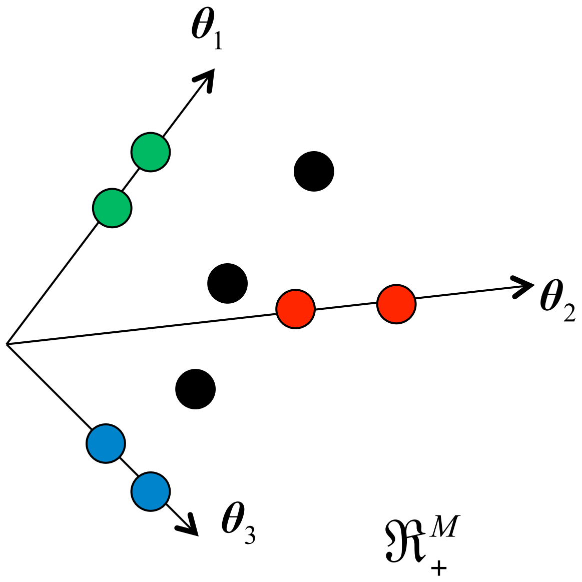
# Key Idea (λ = 0 case)

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

$$\begin{array}{c|ccc|} & \text{topic } 1 & \text{topic } 2 & \text{topic } 3 \\ \hline \text{w1} & \beta_1 & 0 & 0 \\ \text{w2} & \beta_2 & 0 & 0 \\ \text{w3} & 0 & \beta_3 & 0 \\ \text{w4} & 0 & \beta_4 & 0 \\ \text{w5} & 0 & 0 & \beta_5 \\ \text{w6} & 0 & 0 & \beta_6 \\ & \beta_{71} & \beta_{72} & \beta_{73} \\ & & \cdots & \end{array}
\qquad
\begin{array}{|ccc|} \text{doc. } 1 & \cdots & \text{doc. } M \\ \hline \leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_3 & \rightarrow \end{array}$$

**Weight Matrix**

$$\begin{array}{|ccc|} \text{doc. } 1 & \cdots & \text{doc. } M \\ \hline \leftarrow & \beta_1\boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \beta_2\boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \beta_3\boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & \beta_4\boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & \beta_5\boldsymbol{\theta}_3 & \rightarrow \\ \leftarrow & \beta_6\boldsymbol{\theta}_3 & \rightarrow \\ \multicolumn{3}{|c|}{\beta_{71}\boldsymbol{\theta}_1 + \beta_{72}\boldsymbol{\theta}_2 + \beta_{73}\boldsymbol{\theta}_3} \\ & \cdots & \end{array}$$



**Separable Topic Matrix *β***

**Distribution Matrix *A***

# Key Idea (λ = 0 case)

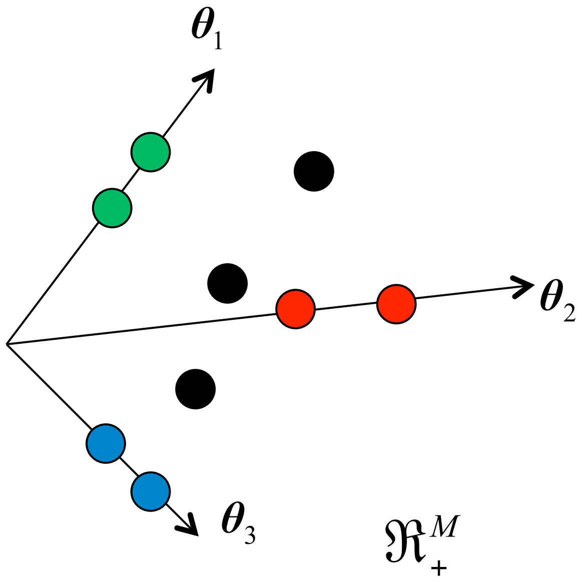
$$\boldsymbol{\beta}_{W \times K} \boldsymbol{\theta}_{K \times M} = \mathbf{A}_{W \times M}$$

$$\begin{array}{c|ccc|}
 & \text{topic } 1 & \text{topic } 2 & \text{topic } 3 \\
\text{w1} & \beta_1 & 0 & 0 \\
\text{w2} & \beta_2 & 0 & 0 \\
\text{w3} & 0 & \beta_3 & 0 \\
\text{w4} & 0 & \beta_4 & 0 \\
\text{w5} & 0 & 0 & \beta_5 \\
\text{w6} & 0 & 0 & \beta_6 \\
 & \beta_{71} & \beta_{72} & \beta_{73} \\
 & & \cdots &
\end{array}$$

$$\begin{array}{ccc}
\text{doc. } 1 & \cdots & \text{doc. } M \\
\leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\
\leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\
\leftarrow & \boldsymbol{\theta}_3 & \rightarrow
\end{array}$$

**Weight Matrix**

$$\begin{array}{ccc}
\text{doc. } 1 & \cdots & \text{doc. } M \\
\leftarrow & \beta_1 \boldsymbol{\theta}_1 & \rightarrow \\
\leftarrow & \beta_2 \boldsymbol{\theta}_1 & \rightarrow \\
\leftarrow & \beta_3 \boldsymbol{\theta}_2 & \rightarrow \\
\leftarrow & \beta_4 \boldsymbol{\theta}_2 & \rightarrow \\
\leftarrow & \beta_5 \boldsymbol{\theta}_3 & \rightarrow \\
\leftarrow & \beta_6 \boldsymbol{\theta}_3 & \rightarrow \\
\multicolumn{3}{c}{\beta_{71} \boldsymbol{\theta}_1 + \beta_{72} \boldsymbol{\theta}_2 + \beta_{73} \boldsymbol{\theta}_3} \\
 & \cdots &
\end{array}$$



**<span style="color:red">Separable</span>**
**Topic Matrix *β***

**Distribution Matrix *A***

# Key Idea

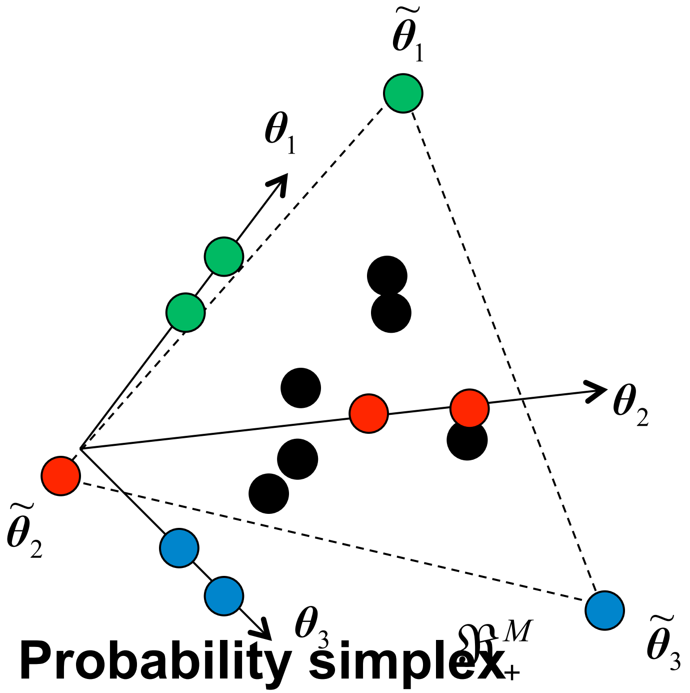
$$\tilde{\beta}_{W\times K}\tilde{\theta}_{K\times M} = \tilde{A}_{W\times M}$$

$$\begin{array}{ccc} \text{doc. } 1 & \cdots & \text{doc. } M \\ \hline \leftarrow & \beta_1\theta_1 & \rightarrow \\ \leftarrow & \beta_2\theta_1 & \rightarrow \\ \leftarrow & & \\ \leftarrow & \beta_4\theta_2 & \rightarrow \\ \leftarrow & \beta_5\theta_3 & \rightarrow \\ \leftarrow & \beta_6\theta_3 & \rightarrow \\ \multicolumn{3}{c}{\beta_{71}\theta_1 + \beta_{72}\theta_2 + \beta_{73}\theta_3} \\ & \cdots & \end{array}$$

**Distribution Matrix $A$**

Novel Word = Extreme Point

Novel word detection + Topic matrix estimation



**Probability simplex**

$$\begin{array}{ccc} \text{doc. } 1 & \cdots & \text{doc. } M \\ \hline \leftarrow & \tilde{\theta}_1 & \rightarrow \\ \leftarrow & \tilde{\theta}_1 & \rightarrow \\ & & \rightarrow \\ \leftarrow & \tilde{\theta}_2 & \rightarrow \\ \leftarrow & \tilde{\theta}_3 & \rightarrow \\ \leftarrow & \tilde{\theta}_3 & \rightarrow \\ \multicolumn{3}{c}{\tilde{\beta}_{71}\tilde{\theta}_1 + \tilde{\beta}_{72}\tilde{\theta}_2 + \tilde{\beta}_{73}\tilde{\theta}_3} \\ & \cdots & \end{array}$$

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

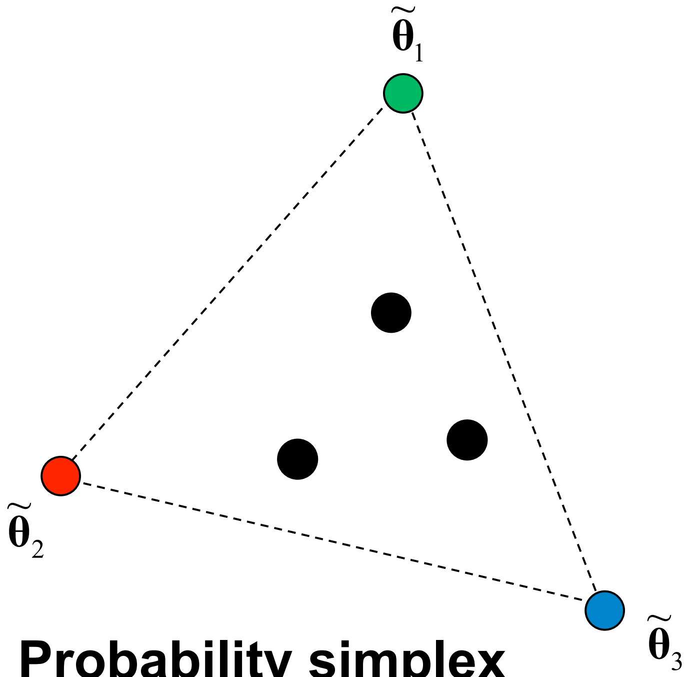
# Finite words/doc.

$$
\begin{array}{ccc}
\text{doc. } 1 & \cdots & \text{doc. } M \\
\end{array}
\left[
\begin{array}{ccc}
\leftarrow & \beta_1\boldsymbol{\theta}_1 & \rightarrow \\
\leftarrow & \beta_2\boldsymbol{\theta}_1 & \rightarrow \\
\leftarrow & \beta_3\boldsymbol{\theta}_2 & \rightarrow \\
\leftarrow & \beta_4\boldsymbol{\theta}_2 & \rightarrow \\
\leftarrow & \beta_5\boldsymbol{\theta}_3 & \rightarrow \\
\leftarrow & \beta_6\boldsymbol{\theta}_3 & \rightarrow \\
\multicolumn{3}{c}{\beta_{71}\boldsymbol{\theta}_1 + \beta_{72}\boldsymbol{\theta}_2 + \beta_{73}\boldsymbol{\theta}_3} \\
 & \cdots & 
\end{array}
\right]
$$

**Distribution Matrix $A$**

$$\xrightarrow{\text{row normalization}}$$

$$
\begin{array}{ccc}
\text{doc. } 1 & \cdots & \text{doc. } M \\
\end{array}
\left[
\begin{array}{ccc}
\leftarrow & \widetilde{\boldsymbol{\theta}}_1 & \rightarrow \\
\leftarrow & \widetilde{\boldsymbol{\theta}}_1 & \rightarrow \\
\leftarrow & \widetilde{\boldsymbol{\theta}}_2 & \rightarrow \\
\leftarrow & \widetilde{\boldsymbol{\theta}}_2 & \rightarrow \\
\leftarrow & \widetilde{\boldsymbol{\theta}}_3 & \rightarrow \\
\leftarrow & \widetilde{\boldsymbol{\theta}}_3 & \rightarrow \\
\multicolumn{3}{c}{\widetilde{\beta}_{71}\widetilde{\boldsymbol{\theta}}_1 + \widetilde{\beta}_{72}\widetilde{\boldsymbol{\theta}}_2 + \widetilde{\beta}_{73}\widetilde{\boldsymbol{\theta}}_3} \\
 & \cdots & 
\end{array}
\right]
$$


$\widetilde{\boldsymbol{\theta}}_1$
$\widetilde{\boldsymbol{\theta}}_2$
$\widetilde{\boldsymbol{\theta}}_3$

**Probability simplex**

**Row Normalized Distribution Matrix $A$**

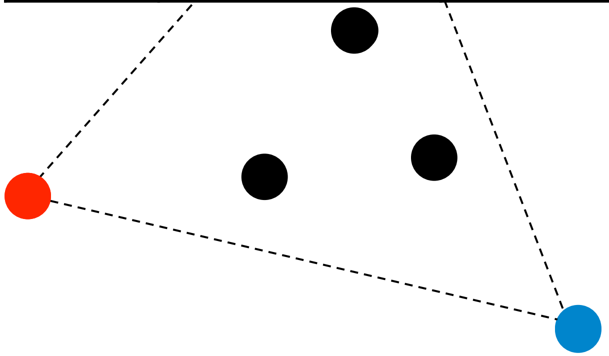
# Finite words/doc. $\tilde{\boldsymbol{\beta}}_{W\times K}\tilde{\boldsymbol{\theta}}_{K\times M} = \tilde{\mathbf{A}}_{W\times M} \approx \tilde{\mathbf{X}}_{W\times M}$

$$\begin{array}{ccc} \text{doc. } 1 & \cdots & \text{doc. } M \\ \left[\begin{array}{ccc} \leftarrow & \tilde{\boldsymbol{\theta}}_1 & \rightarrow \\ \leftarrow & \tilde{\boldsymbol{\theta}}_1 & \rightarrow \\ \leftarrow & \tilde{\boldsymbol{\theta}}_2 & \rightarrow \\ \leftarrow & \tilde{\boldsymbol{\theta}}_2 & \rightarrow \\ \leftarrow & \tilde{\boldsymbol{\theta}}_3 & \rightarrow \\ \leftarrow & \tilde{\boldsymbol{\theta}}_3 & \rightarrow \\ \multicolumn{3}{c}{\tilde{\beta}_{71}\tilde{\boldsymbol{\theta}}_1 + \tilde{\beta}_{72}\tilde{\boldsymbol{\theta}}_2 + \tilde{\beta}_{73}\tilde{\boldsymbol{\theta}}_3} \\ & \cdots & \end{array}\right] \end{array}$$

**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 |



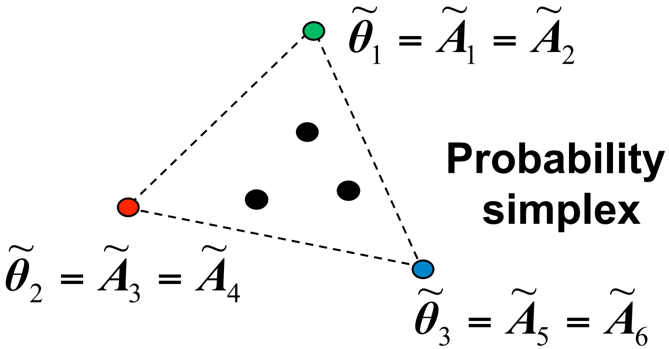
**Probability simplex**

$$\begin{array}{ccc} \text{doc. } 1 & \cdots & \text{doc. } M \\ \left[\begin{array}{ccc} \leftarrow & \tilde{X}_1 & \rightarrow \\ \leftarrow & \tilde{X}_2 & \rightarrow \\ \leftarrow & \tilde{X}_3 & \rightarrow \\ \leftarrow & \tilde{X}_4 & \rightarrow \\ \leftarrow & \tilde{X}_5 & \rightarrow \\ \leftarrow & \tilde{X}_6 & \rightarrow \\ \leftarrow & \tilde{X}_7 & \rightarrow \\ & \cdots & \end{array}\right] \end{array}$$

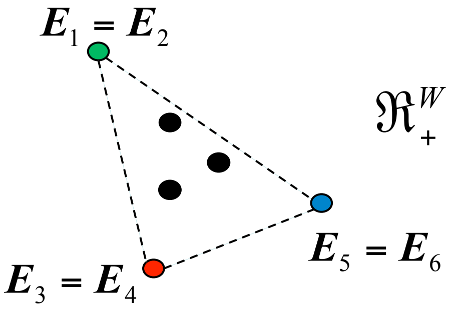
**Row Normalized Observation Matrix: $\tilde{X}$**

$$\underbrace{\widetilde{\beta}_{W\times K}}_{\substack{\textbf{Separable}\\ \textbf{Nonnegative}}}\overbrace{\widetilde{\theta}_{K\times M}}^{\textbf{Nonnegative}} = \widetilde{A}_{W\times M}$$

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


$\widetilde{\theta}_1 = \widetilde{A}_1 = \widetilde{A}_2$
Probability simplex
$\widetilde{\theta}_2 = \widetilde{A}_3 = \widetilde{A}_4$
$\widetilde{\theta}_3 = \widetilde{A}_5 = \widetilde{A}_6$

$$M\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T \xrightarrow[M\to\infty]{a.s} \underbrace{\widetilde{\boldsymbol{\beta}}}_{\substack{\textbf{Separable}\\ \textbf{Nonnegative}}}\overbrace{(M\widetilde{\boldsymbol{\theta}}\ \widetilde{\boldsymbol{\theta}}^T)}^{\textbf{Nonnegative}}\widetilde{\boldsymbol{\beta}}^T = \mathbf{E}_{W\times W}$$

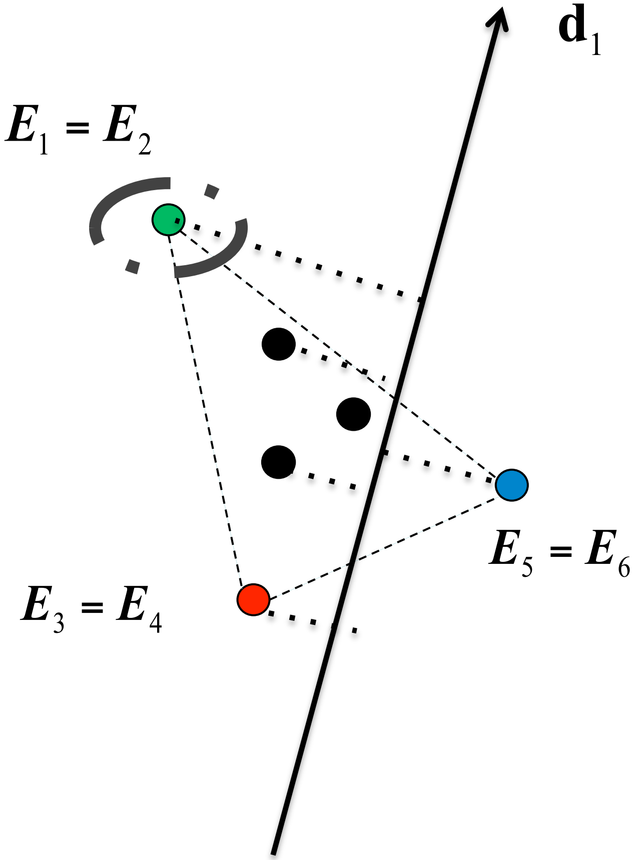


**Novel word detection + Topic matrix estimation**

# Detect Novel Words via Projections

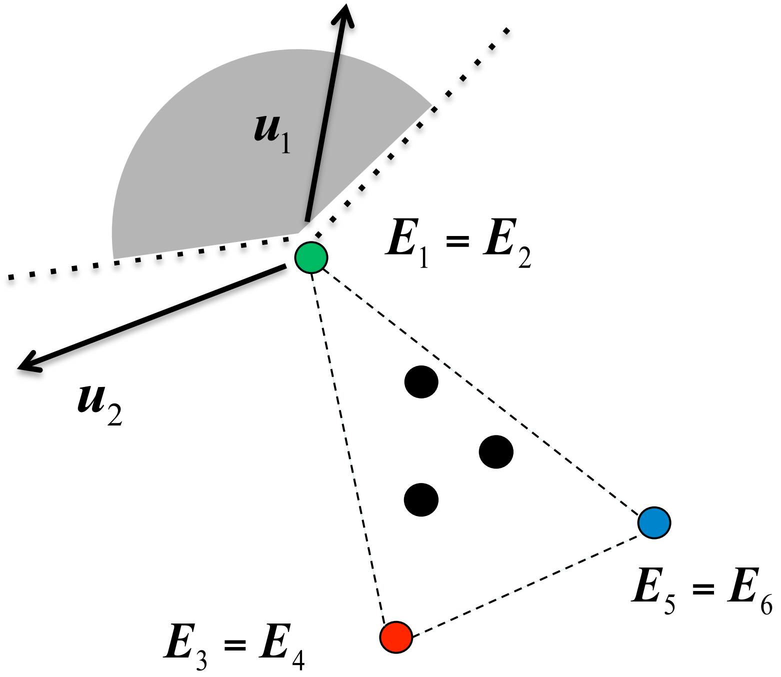
- Max/Min of projection → extreme points of convex hull

- Which directions to use
  →Generate $P$ iid Isotropy directions



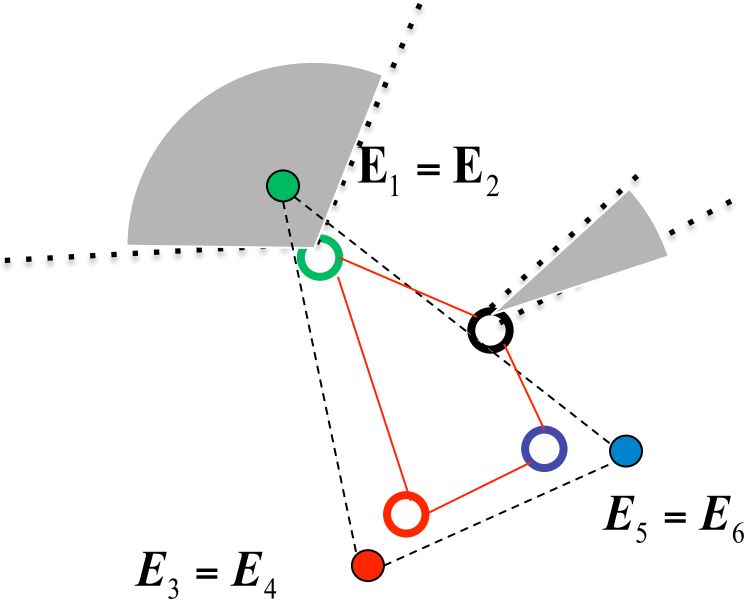
# “Robustness” of extreme points

- Max/Min of projection → extreme points of convex hull
- Which directions to use
  →Generate a few iid Isotropy directions
- Freq. of maximum ≈ Solid Angle of an Extreme Point



# Extension to Approximately Separable (λ > 0)

- Approx. novel words ⇔ larger solid angles
- Solution
  - → Sort solid angles
  - → Take the top – K extreme points



# Main result [Ding et al, ‘13,’14]

- Computational complexity :

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

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

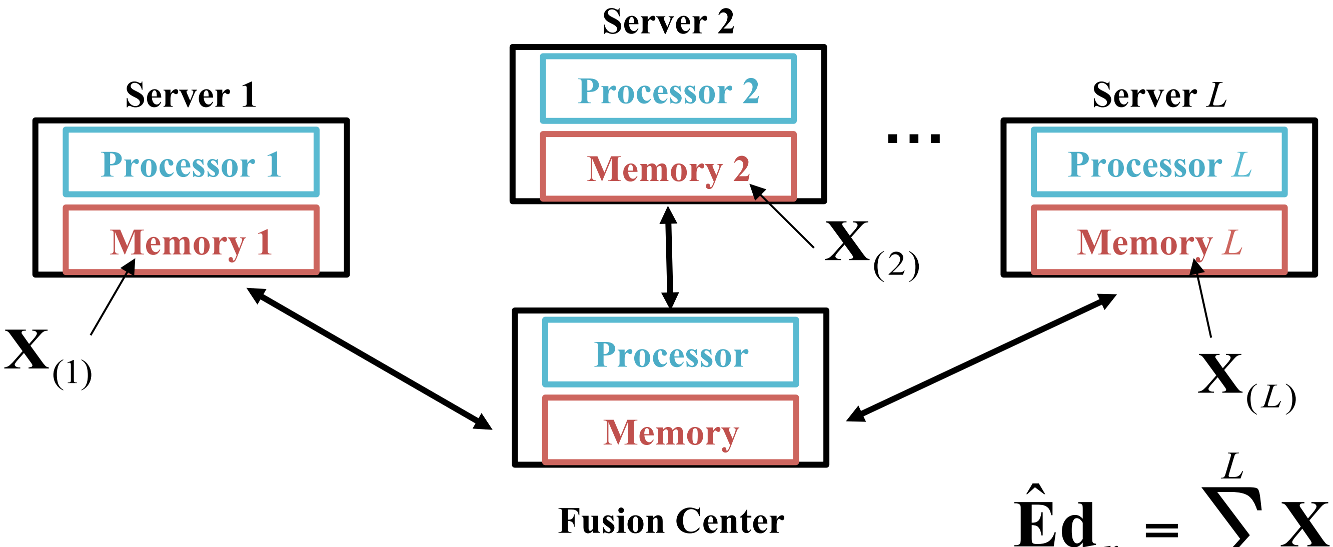
- Sample complexity :

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-$\delta$ if

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

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

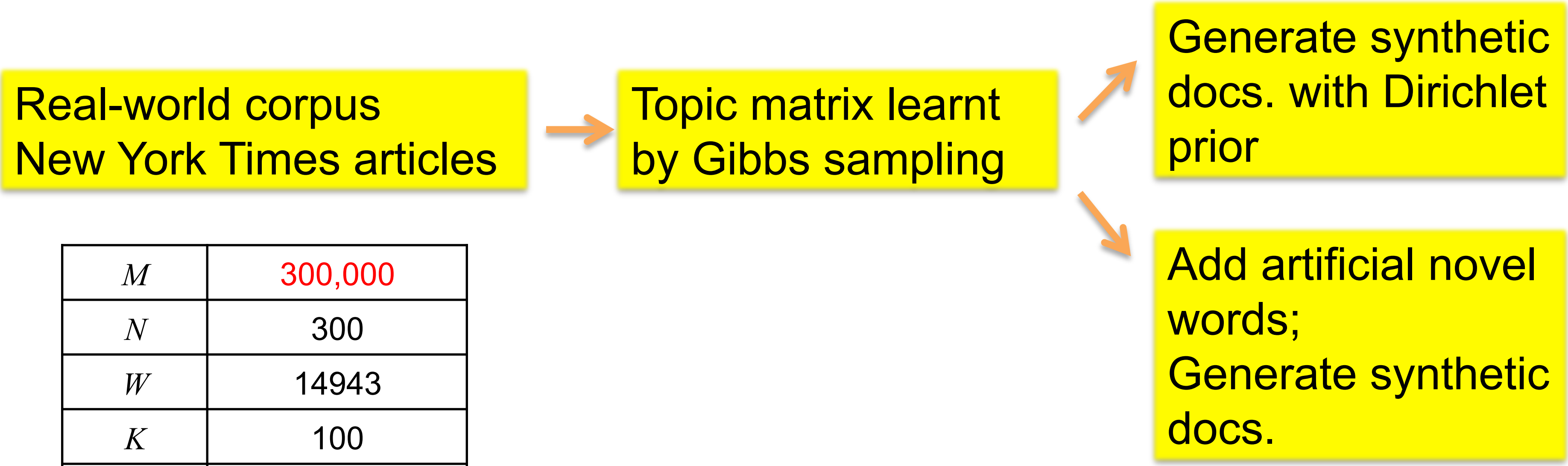
# Distributed Implementation



$$\hat{\mathbf{E}}\mathbf{d}_r = \sum_{l=1}^{L} \mathbf{X}_{(l)}\mathbf{X}_{(l)}^T\mathbf{d}_r$$

- Modern web-scale database are distributed
- $M$ document archived on $L$ servers
- Goal: low communication cost O(WK)

# Experimental Results (semi-synthetic data)



Real-world corpus
New York Times articles

Topic matrix learnt
by Gibbs sampling

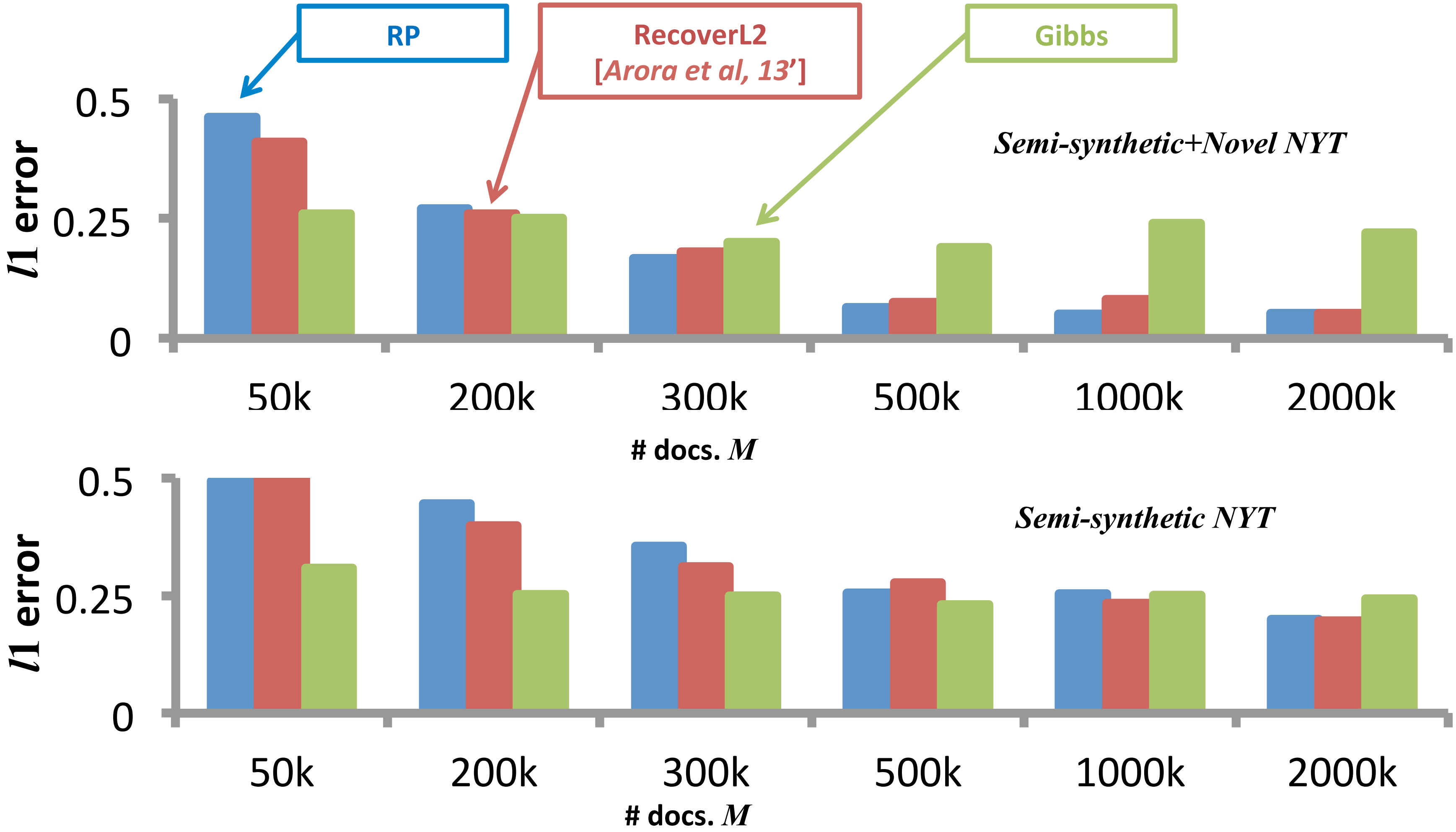
Generate synthetic docs. with Dirichlet prior

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

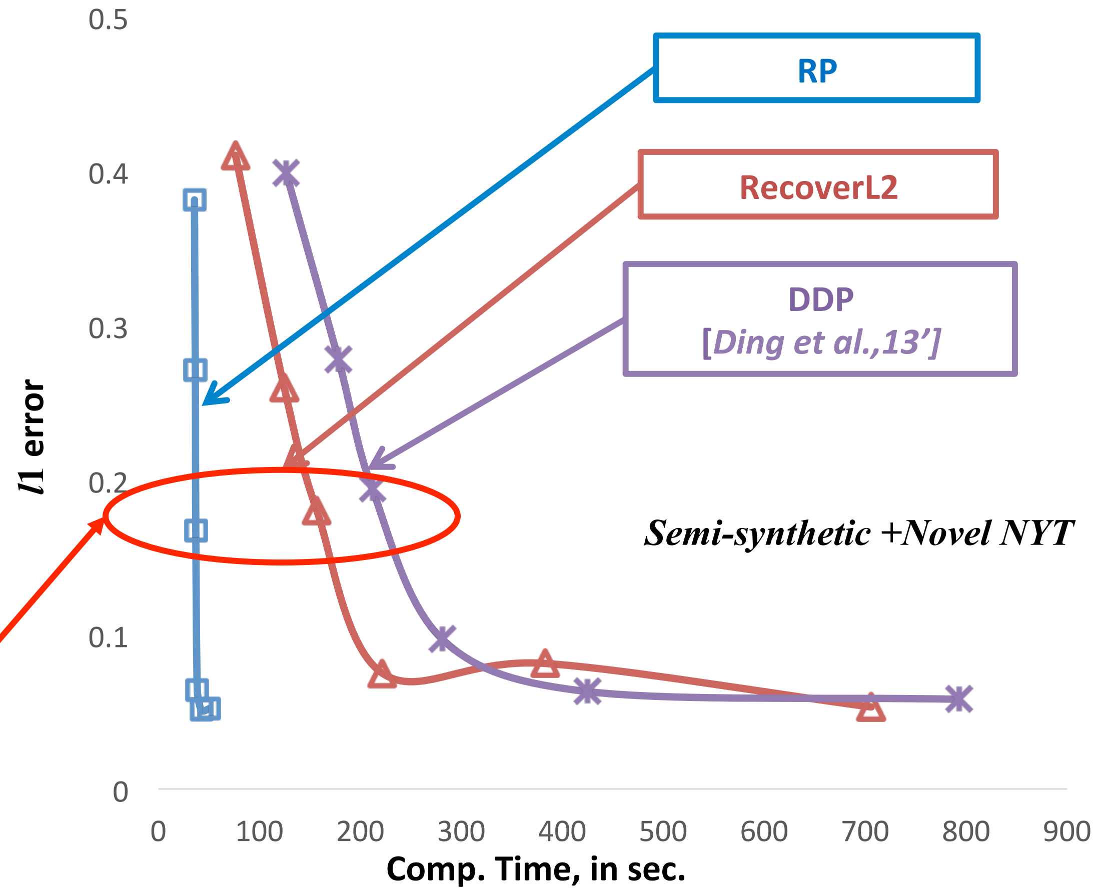
# Experimental Results (semi-synthetic data)



# Experimental Results (semi-synthetic data)

| | *Semi-syn +novel* NYT | *Semi-syn* NYT |
|---|---|---|
| $M$ | Variable | Variable |
| $N$ | 300 | 300 |
| $W$ | 15043 words,<br>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)

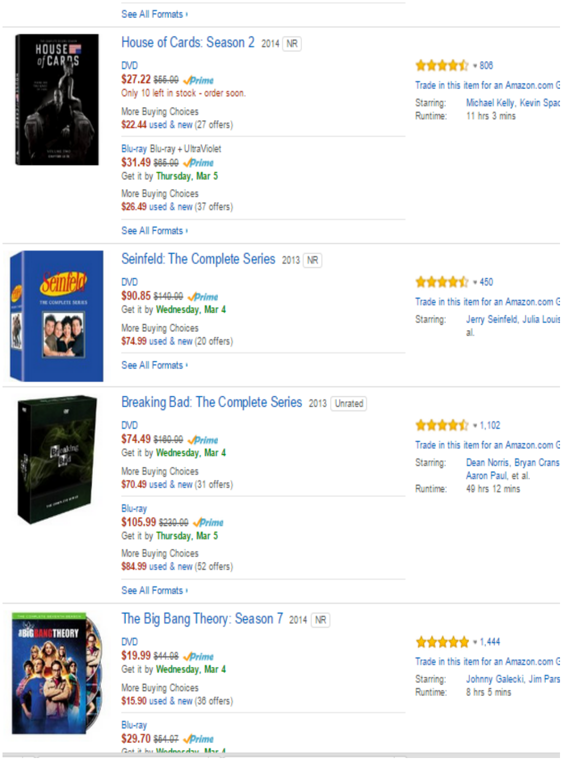
# Outline

- 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
  - Efficiently Identifying Extreme Points
  - Empirical Results on Real-World Datasets

- Rank Aggregation Problem
  - Heterogenous Population ~ Mixture of Mallows Model
  - Reduction to Topic Modeling Problem
  - Empirical Results on Real-World Datasets

# Mixed membership latent variable model

User preferences



See All Formats ›

House of Cards: Season 2 2014 NR

DVD
$27.22 ~~$55.00~~ ✓Prime
Only 10 left in stock - order soon.

More Buying Choices
$22.44 used & new (27 offers)

Blu-ray Blu-ray + UltraViolet
$31.49 ~~$85.00~~ ✓Prime
Get it by Thursday, Mar 5

More Buying Choices
$26.49 used & new (37 offers)

See All Formats ›

★★★★½ ▾ 806
Trade in this item for an Amazon.com G
Starring: Michael Kelly, Kevin Spac
Runtime: 11 hrs 3 mins

Seinfeld: The Complete Series 2013 NR

DVD
$90.85 ~~$140.00~~ ✓Prime
Get it by Wednesday, Mar 4

More Buying Choices
$74.99 used & new (20 offers)

See All Formats ›

★★★★½ ▾ 450
Trade in this item for an Amazon.com G
Starring: Jerry Seinfeld, Julia Louis
al.

Breaking Bad: The Complete Series 2013 Unrated

DVD
$74.49 ~~$180.00~~ ✓Prime
Get it by Wednesday, Mar 4

More Buying Choices
$70.49 used & new (31 offers)

Blu-ray
$105.99 ~~$230.00~~ ✓Prime
Get it by Thursday, Mar 5

More Buying Choices
$84.99 used & new (52 offers)

See All Formats ›

★★★★½ ▾ 1,102
Trade in this item for an Amazon.com G
Starring: Dean Norris, Bryan Crans
Aaron Paul, et al.
Runtime: 49 hrs 12 mins

The Big Bang Theory: Season 7 2014 NR

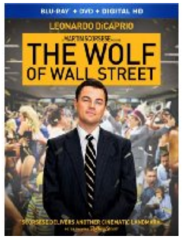
DVD
$19.99 ~~$44.08~~ ✓Prime
Get it by Wednesday, Mar 4

More Buying Choices
$15.90 used & new (36 offers)

Blu-ray
$29.70 ~~$54.07~~ ✓Prime

★★★★★ ▾ 1,444
Trade in this item for an Amazon.com G
Starring: Johnny Galecki, Jim Pars
Runtime: 8 hrs 5 mins

words counts Influencing factors



>

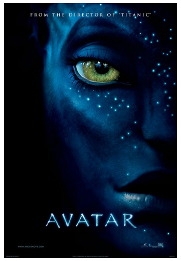
1

*"actor"*

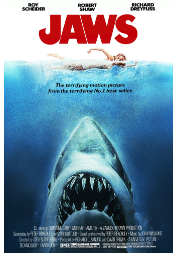
>

0

*"music"*



>



1

*"special effect"*

**document = mixture of latent influence factors**

Prob. Prefer movie 1 over 2 in 1st latent factor

$$\begin{array}{r|ccc|}
 & \text{actor} & & \text{musical} \\
(1,2) \rightarrow & \beta_{11} & & \beta_{1K} \\
(1,3) & \beta_{21} & & \beta_{2K} \\
 & \vdots & \cdots & \vdots \\
 & \vdots & & \vdots \\
(Q\text{-}1.Q) & \beta_{W1} & & \beta_{WK} \\
 & \text{topic } 1 & & \text{topic } K
\end{array}$$

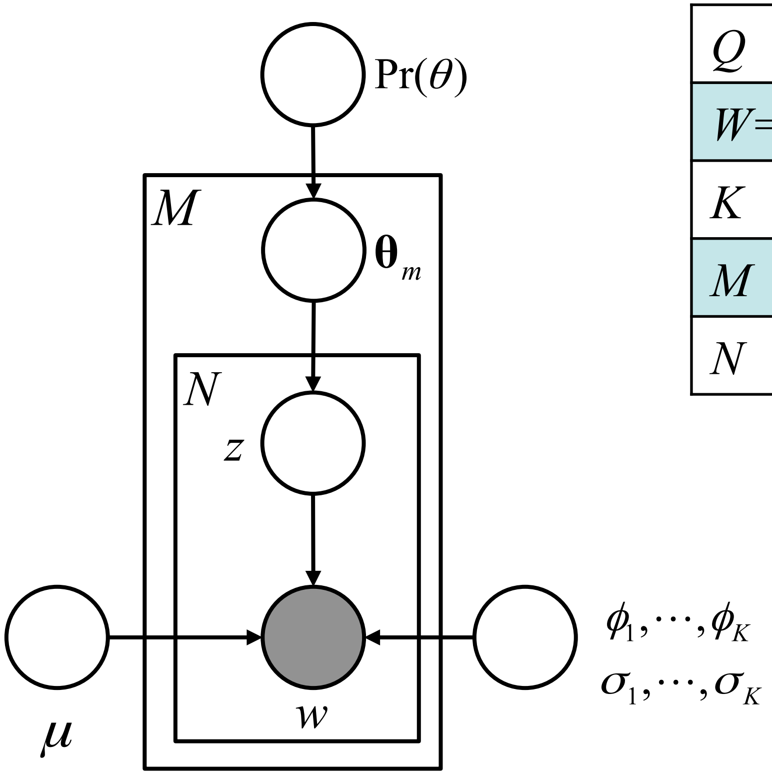
Ranking matrix - $\boldsymbol{\beta}$

- column = “topic”
- $W$ = # ordered pairs = $Q\,(Q - 1)$
- $K$ = # topics

- Generative Model for Latent factor
- Mallows model
  - Baseline permutation $\sigma_0$
  - Prob. of permutation $\sigma$: $Prob\{\sigma \mid \sigma_0\} \propto \phi^{dist(\sigma,\sigma_0)} \Longrightarrow \beta$
  - Heterogeneous population
    - Dispersion factor $\phi$

# Mixed Membership Rank Aggregation Problem

“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

# Related Work

| Category | Models | Issues | Hetero-geneity | User-inconsistency |
|---|---|---|---|---|
| Single ranking models | Mallows[*Mallows,'57*],<br>BTL [*Bradley & Terry,'52*], etc. | Only one global ranking | Only via deviation from base ranking scheme | Via noisy observation model |
| Mixture of ranking models | Mixture of Mallows model [*Lu & Boutilier ICML11, Awasthi et al. NIPS14*],<br>Mixture of single rankings [*Farias et al. NIPS09*],<br>Mixture of BTLs [*Oh & Shah NIPS14*] | Each user is dominated by one type | Multiple mixture components | Via noisy observation model |
| "Topic" ranking models | [*Ding et al. NIPS14*],<br>Mixed membership Mallows, etc. | | Multiple shared rankings | Via probabilistic mixture + noisy observation |

# Approximate Separability

- Most ranking matrix are $\lambda$-Approximate separable, # items $Q$ >> # factors $K$

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

$$\begin{array}{c|ccc|}
 & \text{Rank } 1 & \text{Rank } 2 & \text{Rank } 3 \\
\text{Pair } 1 & 0.98 & 0.01 & 0.01 \\
\text{Pair } 2 & 0.01 & 0.99 & 0.01 \\
\text{Pair } 3 & 0.01 & 0.01 & 0.90 \\
\text{Pair } 4 & 0.98 & 0.90 & 0.10 \\
\text{Pair } 5 & 0.10 & 0.09 & 0.90 \\
 & & \cdots & \\
\end{array}$$

$$\boldsymbol{\beta}$$

Approximately Separable ranking matrix

| $\phi$ | **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

| | |
|---|---|
| $Q$ | # items |
| $W$ | # ordered pairs |
| $K$ | # latent ranking |
| $M$ | # users |
| $N$ | # comp. / user |
| $P$ | # projections |

# Main result

- Computational complexity :

$$O(MNP + Q^2K^3)$$

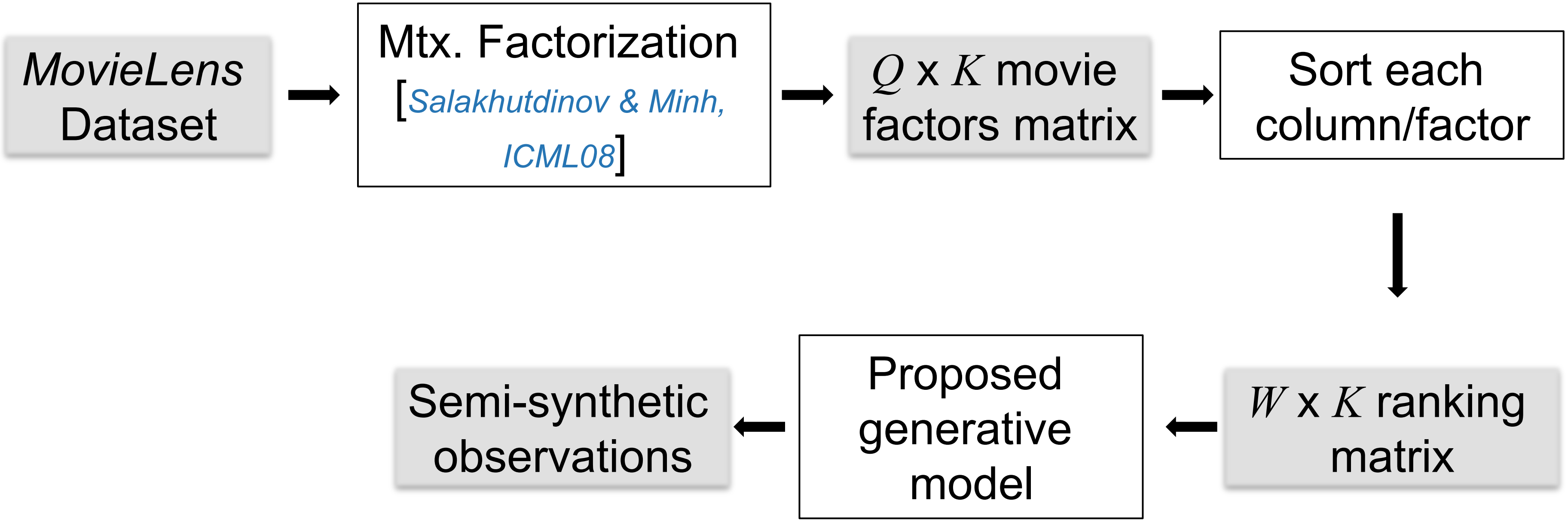
- Sample complexity :
  - If
    - Correlation matrix of weight prior has **full-rank** and
    - The ranking matrix $\boldsymbol{\sigma}$ is $\lambda$**-separable** and
    - $\lambda \leq \mathrm{c}\mathrm{K}^{-2}$
  - Then
    - Proposed Random Projection algorithm can estimate the ranking matrix correctly with probability at least $1\text{-}\delta$ for all

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

# Semi-synthetic data

| | |
|---|---|
| $Q$ | 100 most rated movies |
| $K$ | 10 latent rankings |
| $M$ | 5940 users |
| | ~200K ratings |

- To resemble the dimensionality and characteristics of real-world data



*MovieLens* Dataset

Mtx. Factorization [*Salakhutdinov & Minh, ICML08*]

$Q$ x $K$ movie factors matrix

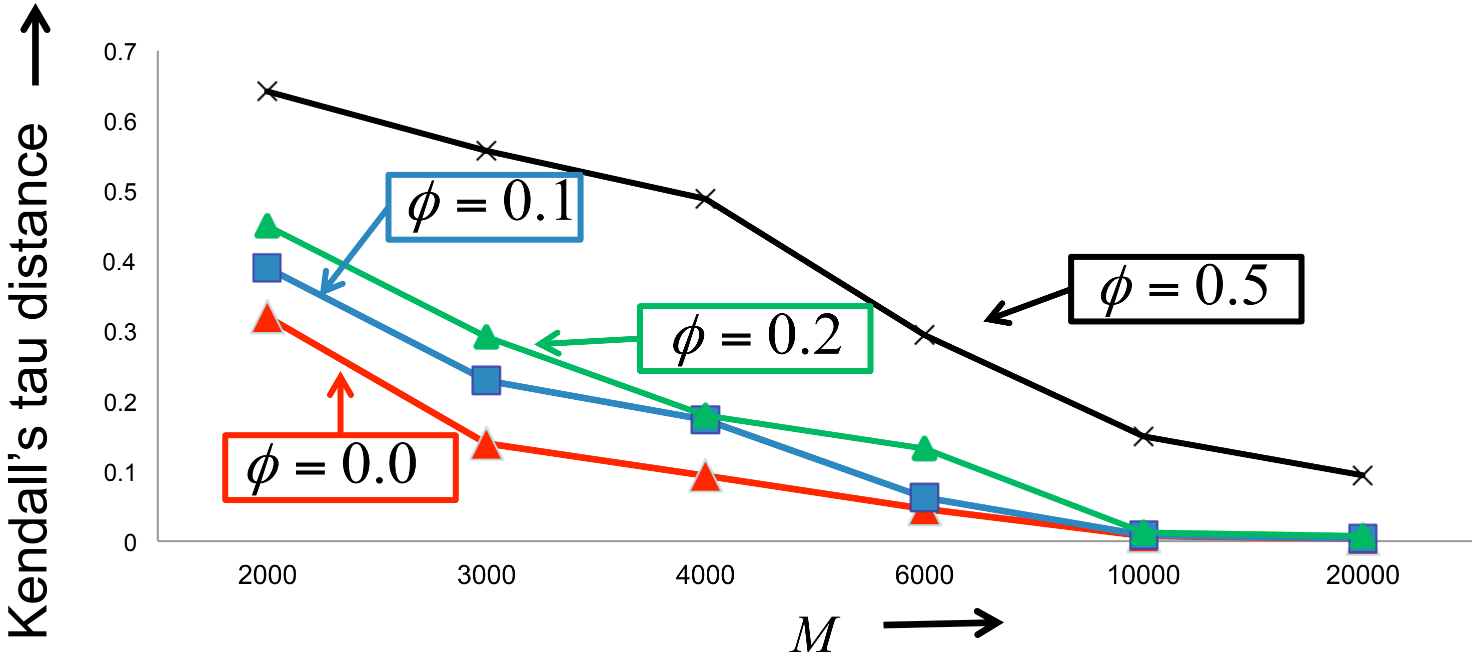
Sort each column/factor

$W$ x $K$ ranking matrix

Proposed generative model

Semi-synthetic observations

# Semi-synthetic data



| | |
|---|---|
| $Q$ | 100 most rated movies |
| $K$ | 10 latent rankings |
| $N$ | 300 comparisons/user |
| $M$ | # user, variable |

- Dirichlet Prior for $\boldsymbol{\theta}$
- Uniform distribution for $\boldsymbol{\mu}$
- $\phi = \phi_1 = \ldots = \phi_K$

# *Movielens* dataset – new comparison

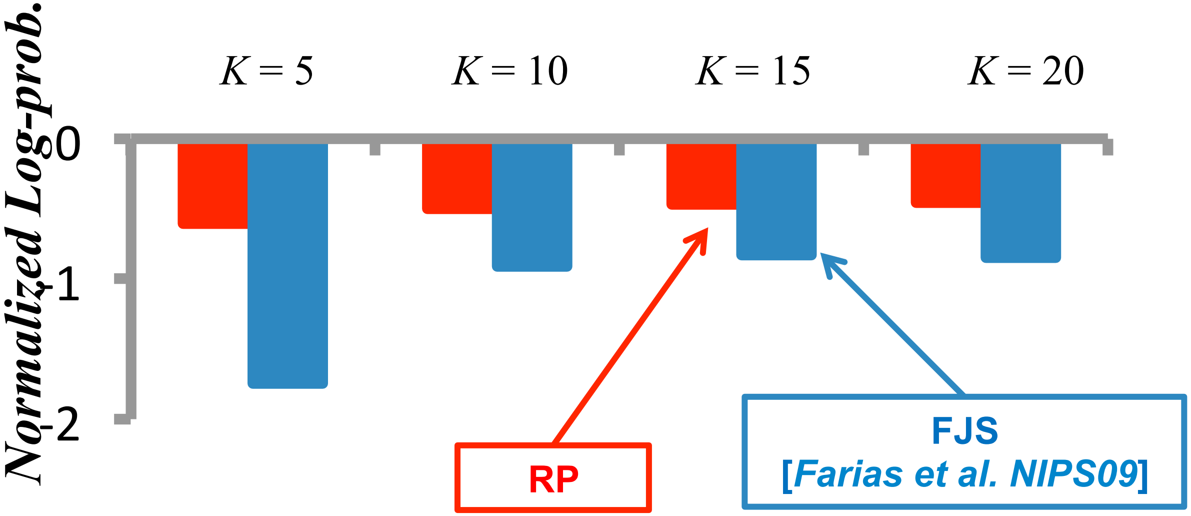
- **Data:** generate comparisons from ratings

  User 1: Movie A, 4 star Movie C, 2 star➔ $w = (\mathrm{A}, \mathrm{C})$ for user 1

- **Task**: predict new comparisons for users in the training set
- **Measure**: predictive log-probability [*Wallach et al. ICML09*]

$$\frac{1}{\text{Total \# test pairs}} \sum_{i,m} \log\left\{p(i \text{ th test pair} \mid \hat{\boldsymbol{\sigma}}, \text{Training pairs of user } m)\right\}$$

| | |
|---|---|
| $Q$ | 100 most rated movies |
| $K$ | variable |
| $M$ | 5940 users |



# *Movielens* dataset – new user

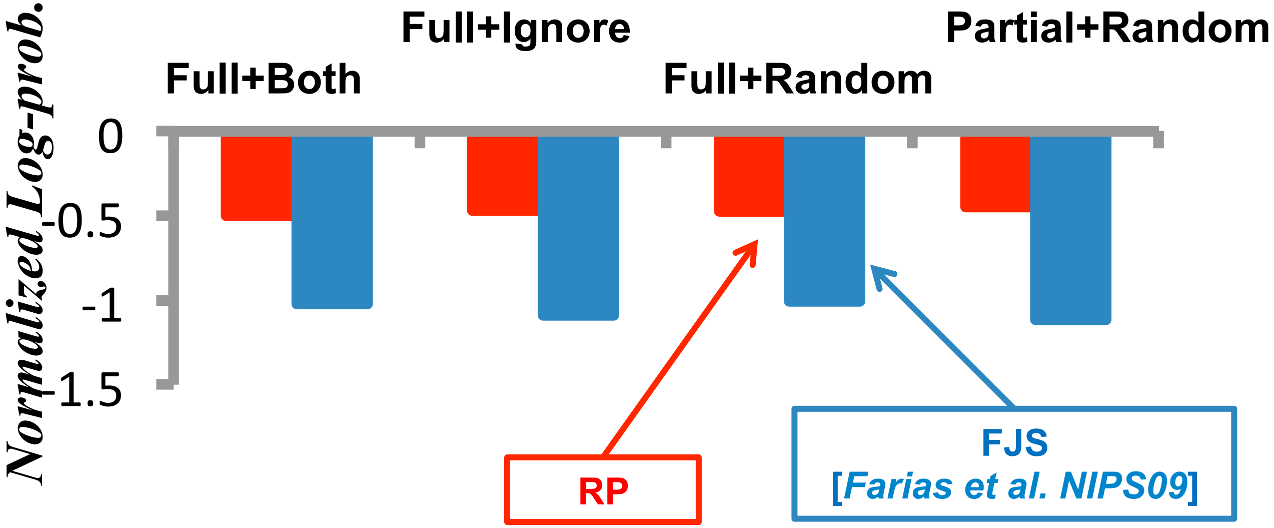
- **Data:** generate comparisons from ratings

  User 1: Movie A, 4 star Movie C, 2 star➔ $w = (\mathrm{A}, \mathrm{C})$ for user 1

- **Task**: predict comparison of new users
- **Measure**: predictive log-probability [*Wallach et al. ICML09*]

$$\frac{1}{\text{Total \# test pairs}} \sum_{i,m} \log\{p(i\text{ th test pair of user }m \mid \hat{\boldsymbol{\sigma}})\}$$

| | |
|---|---|
| $Q$ | 100 most rated movies |
| $K$ | 10 latent rankings |
| $M$ | 4000 *training* user |
| $M$ | 2040 *testing* user |



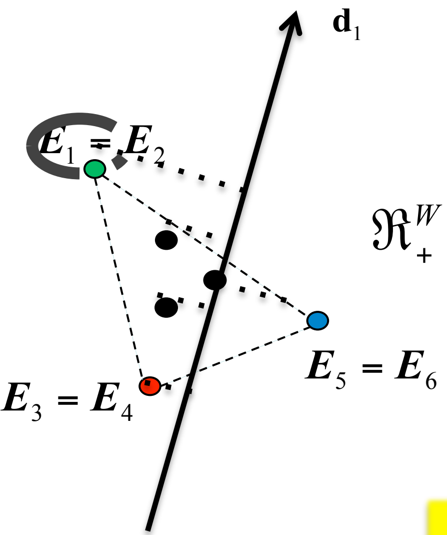
# *Movielens* dataset – predicting stars

- Predict star ratings via ranking models
  - Generate comparisons from training ratings
  - Learn mixed membership Mallows model with Dirichlet prior
  - For each testing movie review:
- 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

# Summary



High-D Latent Factor Models Geometry ~ Approx Sep.

↓

Simple geometric picture

↓

Efficient randomized algorithm

↓

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

$$\begin{array}{c|ccc} & \text{Rank 1} & \text{Rank 2} & \text{Rank 3} \\ \hline \text{Pair 1} & 0.98 & 0.01 & 0.01 \\ \text{Pair 2} & 0.01 & 0.99 & 0.01 \\ \text{Pair 3} & 0.01 & 0.01 & 0.90 \\ \text{Pair 4} & 0.98 & 0.90 & 0.10 \\ \text{Pair 5} & 0.10 & 0.09 & 0.90 \\ & & \cdots & \end{array}$$

$$\boldsymbol{\beta}$$

Approximately Separable ranking matrix