# Lecture 4:
# Topic Modeling in Practice

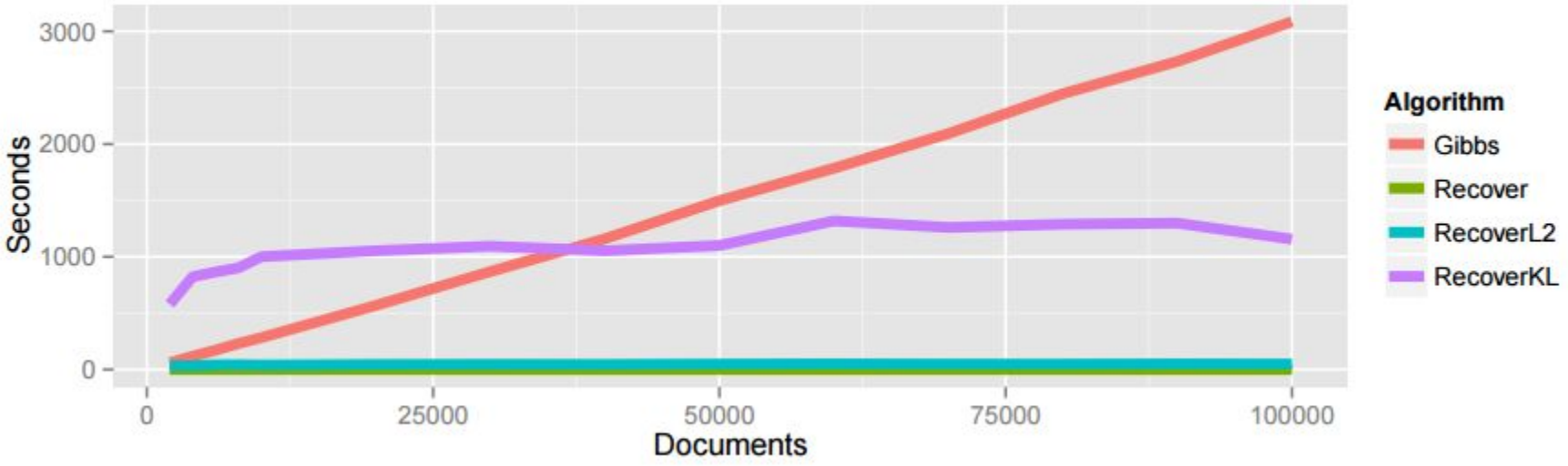
# Evaluation


- Toy Examples: correctness.
- Synthetic Examples: running time, sample complexity, robustness
- Real Data
  - Qualitative evaluation: look at the topics found
  - Quantitative evaluation: held-out likelihood, …
- Real Application: Apply topic models to recommend articles, social science, ...

# Evaluating Topic Modeling Algorithm

- Compare to MALLAT
  (package based on Gibbs sampling)
- Variants of algorithms
  - Recover: Basic algorithm
  - Recover-L2: Try to minimize $||Q-AW||_F$
  - Recover-KL: Try to minimize KL-divergence between rows of Q and AW.
- Data Set: UCI New York Times
  - 295k articles, 15k vocabulary, average length~300

# Running Time

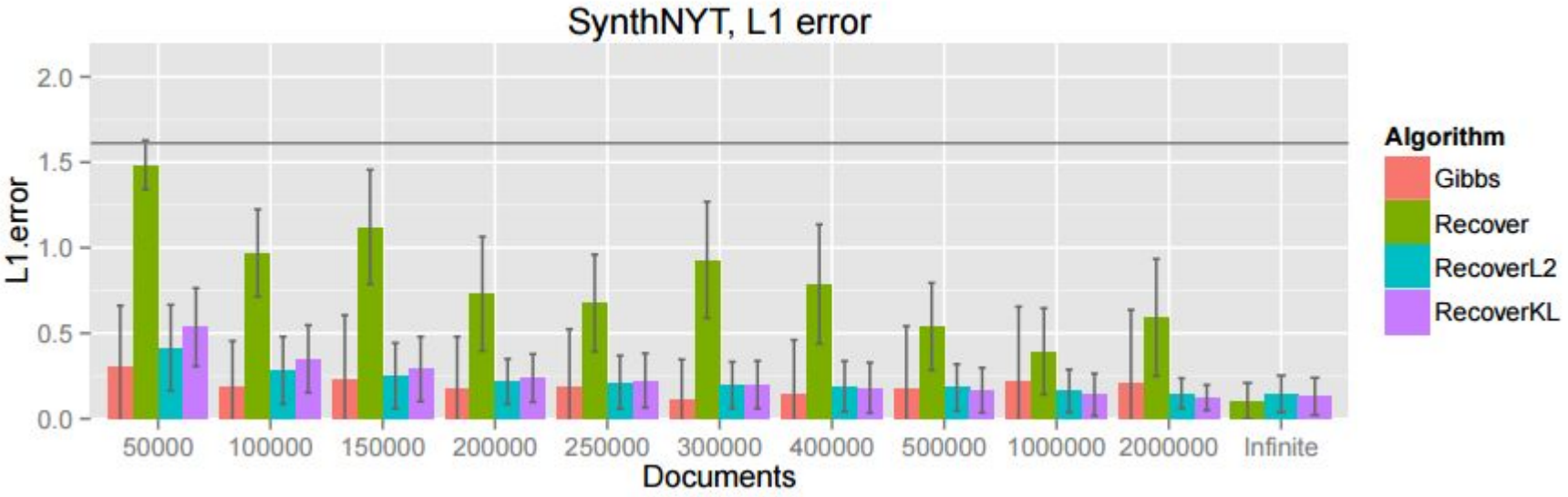



- Algorithms are faster than MALLAT, because most of the work is done on the word-word correlation matrix

# Semi-synthetic Example

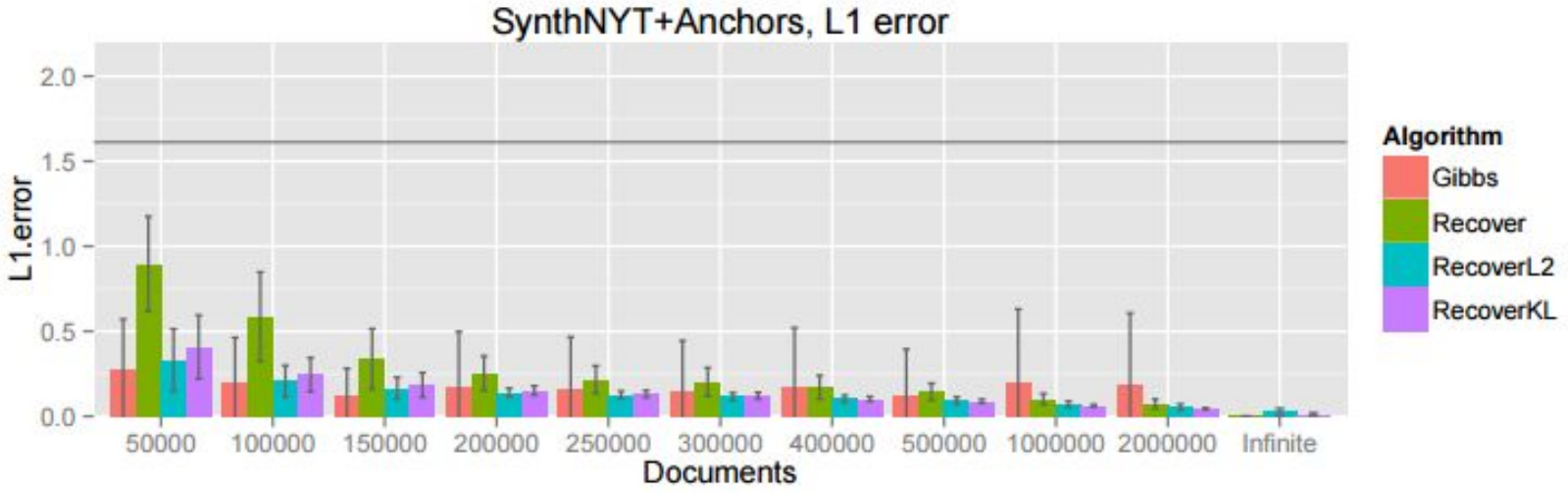
- Idea: Compute topic matrix by running MALLET on NYT data set, then generate synthetic documents.
- Benefit:
  - Has ground truth, measure error in parameter space
  - Easy to tweak parameters (different topic models, topic matrix, # documents, #words, ...)
  - Topic matrix is “natural”
- Data is still generated from the model, hard to evaluate the robustness of algorithm.

# Semi-synthetic Experiments




- Performance is comparable to MALLAT, especially with more documents.
- Does not achieve 0 error with infinite data (not separable)

# Anchor Words?




- Most topics have anchor words.
- Algorithms works OK even when some topics do not have anchor words.

# Real Data (sample topics)

| | |
|---|---|
| RecoverL2 | president zzz_clinton zzz_white_house zzz_bush official zzz_bill_clinton |
| Gibbs | zzz_bush zzz_george_bush president administration zzz_white_house zzz_dick_cheney |
| RecoverL2 | father family **zzz_elian** boy court zzz_miami |
| Gibbs | zzz_cuba zzz_miami cuban zzz_elian boy protest |
| RecoverL2 | oil prices percent million market zzz_united_states |
| Gibbs | oil power energy gas prices plant |
| RecoverL2 | **zzz_microsoft** company computer system window software |
| Gibbs | zzz_microsoft company companies cable zzz_at zzz_internet |
| RecoverL2 | government election zzz_mexico political zzz_vicente_fox president |
| Gibbs | election political campaign zzz_party democratic voter |
| RecoverL2 | fight **zzz_mike_tyson** round right million champion |
| Gibbs | fight zzz_mike_tyson ring fighter champion round |

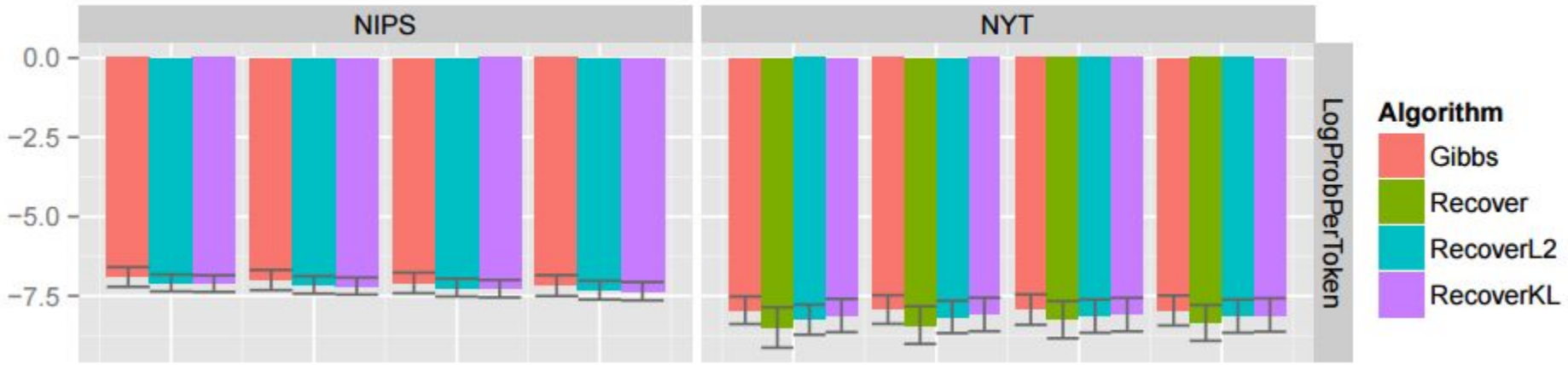
# Real Data (Held-out likelihood)

- Idea: For each document, show a fraction of words, use the learned topic matrix to predict the distribution Pr[z = i|doc]

- For the rest of the words $i_1, i_2, \ldots$
  Score = $\sum_j \log \Pr[z = i_j|\text{doc}]$

- Details matter (how to predict Pr[z=i|doc], fraction of held-out, smoothing…)

# Real Data (Held-out likelihood)




- MALLAT is better, but RecoverKL is close.
- Recover algorithms followed by MALLAT improves held-out likelihood.

# Challenges and New Algorithms

- What if anchor-word assumption is not true?
  - For LDA, can use tensor decomposition [AFHKL'12]
  - Only appear in 1 topic ⇒ Only appear in few topics (subset separable [GZ'15])
  - "Catch Words": words that appear more frequently in one topic than all others [BBGKP'15]
- How to guess the number of topics?
  - Use low dimensional embeddings? [LeeMimno'14]
- Variants of topic models?
  - multilingual, temporal, …
- Make the algorithm more robust?
  - "Fix" the correlation matrix Q [LeeBindelMimno'15]

# Homework

- Homework 1 is out, due in 2 weeks (9/21/2015 in class)
- Latex strongly encouraged.
- Discussions are allowed, but must acknowledge.
- Start early.
- Questions: post on Piazza.

# References

Codes for Recover Algorithm:

http://www.cs.nyu.edu/~halpern/code.html

MALLAT package

http://mallet.cs.umass.edu/

Papers

[LeeMimno'14]

[GZ'15]

[BBGKP'15]

[LeeBindelMimno'15]