Emprircal Results: Datasets

- **NIPS**: 1,500 NIPS full papers
- **NYT**: Random subset of 30,000 documents from the New York Times dataset
- **Pubmed**: Random subset of 30,000 documents from the Pubmed abstracts dataset
- **20NG**: 13,389 documents from 20NewsGroup dataset

Empirical Results: Assumptions

Corpus	Documents	K	Fraction of Documents		
			$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.9$
NIPS	1,500	50	56.6%	10.7%	4.8%
NYT	30,000	50	63.7%	20.9%	12.7%
Pubmed	30,000	50	62.2%	20.3%	10.7%
20NG	13,389	20	74.1%	54.4%	44.3%

Table: Fraction of documents satisfying dominant topic assumption.

Corpus	K	Mean per topic frequency of CW	% Topics with CW
NIPS	50	0.05	95%
NYT	50	0.11	100%
Pubmed	50	0.05	90%
20NG	20	0.06	100%

Table: CatchWords (CW) assumption with $\rho = 1.1$, $\varepsilon = 0.25$

Empirical Results: Semi-synthetic Data

- Generate semi-synthetic corpora from LDA model trained by MCMC, to ensure that the synthetic corpora retain the characteristics of real data
- Gibbs sampling is run for 1000 iterations on all the four datasets and the final word-topic distribution is used to generate varying number (s) of synthetic documents with document-topic distribution drawn from a symmetric Dirichlet with hyper-parameter 0.01
- Note that the synthetic data is *not* guaranteed to satisfy dominant topic assumption for every document, on average about 80% documents satisfy the assumption

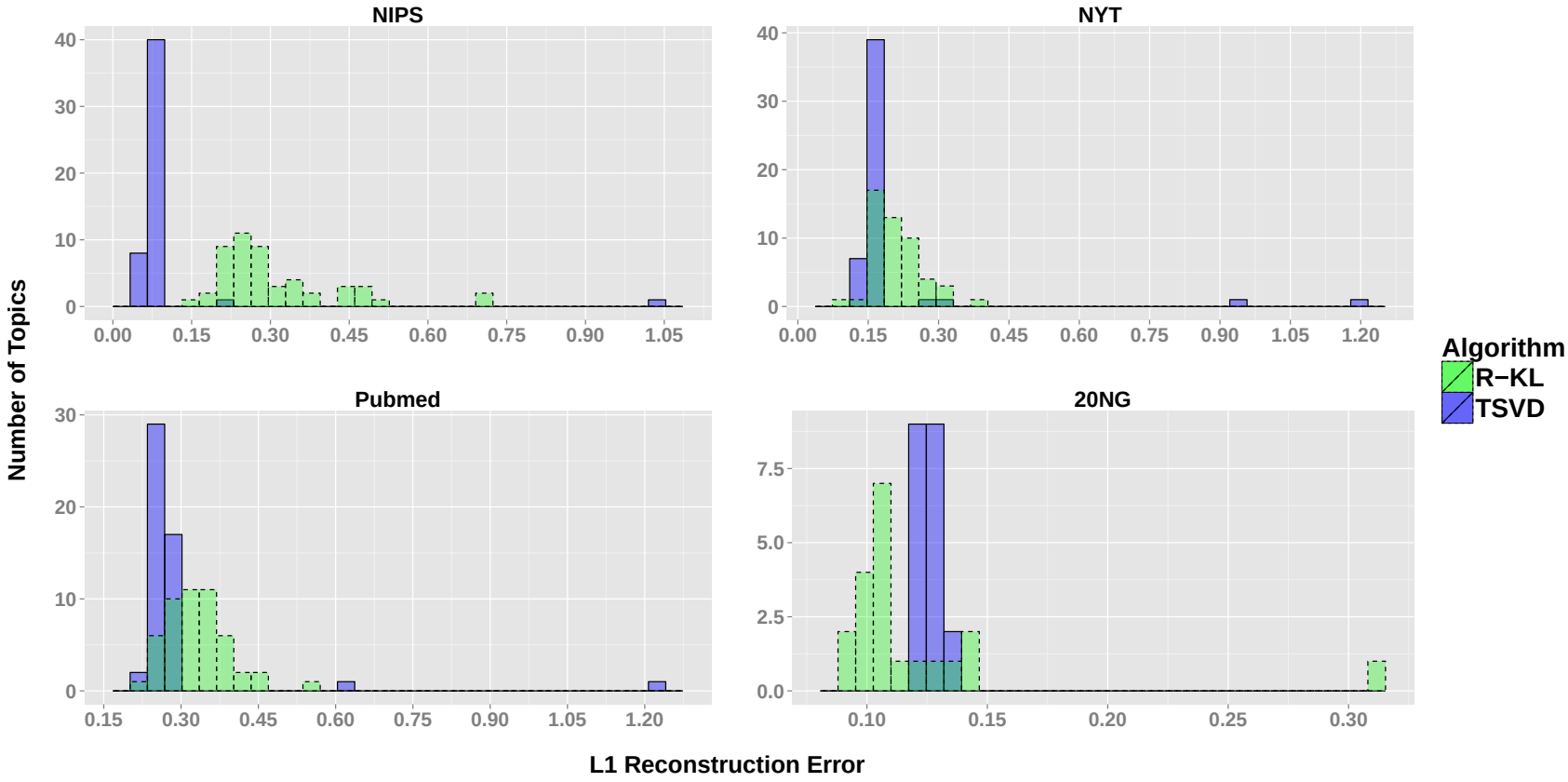
Empirical Results: L1 Recnstruction Eror

L1 reconstruction error on various semi-synthetic datasets. Last column is percent improvement over Recover-KL. Total average improvement over R-KL is **20%**

Corpus	s	R-L2	R-KL	TSVD	% Improvement
NIPS	40k	0.342	0.308	**0.094**	**69.6%**
	50k	0.346	0.308	**0.135**	**56.3%**
	60k	0.346	0.311	**0.124**	**60.2%**
Pubmed	40k	0.388	0.332	**0.291**	**12.5%**
	50k	0.378	0.326	**0.290**	**11.2%**
	60k	0.372	0.328	**0.297**	**9.6%**
20NG	40k	0.126	**0.120**	0.125	-4.1%
	50k	0.118	**0.114**	0.115	-0.7%
	60k	0.114	0.110	**0.108**	**2.5%**
NYT	40k	0.214	0.208	**0.198**	**4.8%**
	50k	0.211	0.206	**0.186**	**9.4%**
	60k	0.205	0.200	**0.192**	**4.1%**

Empirical Results: L1 Recnstruction Eror

Histogram of L1 error across topics for 40k synthetic documents. On majority of the topic (> 90%) the recovery error for TSVD is significantly smaller than Recover-KL.



Empirical Results: Perplexity & Topic Coherence